NATIONAL QUALITY FORUM

+ + + + +

EPISODE GROUPER EVALUATION CRITERIA
EXPERT PANEL

+ + + + +

WEDNESDAY
FEBRUARY 5, 2014

+ + + + +

The Expert Panel met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Kristine Anderson and Joseph Cacchione, Co-Chairs, presiding.

PRESENT:
KRISTINE MARTIN ANDERSON, MBA, Panel Co-Chair
JOSEPH CACCHIONE, MD, Panel Co-Chair
STEPHEN BANDEIAN, MD, JD, John Hopkins University Bloomberg School of Public Health
DAVID BODYCOMBE, MSc, ScD, Johns Hopkins University Bloomberg School of Public Health
FRANCOIS DE BRANTES, MS, MBA, Health Care Incentives Improvement Institute*
NANCY GARRETT, PhD, Hennepin County Medical Center
JENNIFER HOBART, MBA, MSc, Blue Shield of California
DAVID HOPKINS, PhD, Pacific Business Group on Health
JIM JONES, MBA, AmeriHealth Caritas
MARJORIE KING, MD, American Association of Cardiovascular and Pulmonary Rehabilitation
MARK LEVINE, MD, Centers for Medicare & Medicaid Services

JIM LOISELLE, McKesson Corporation*
THOMAS MACURDY, PhD, Stanford University
JELANI McLEAN, PhD, MPA, Health Intelligence Company, LLC
DAVE MIRKIN, MD, Milliman MedInsight
JAMES NAESSENS, ScD, MPH, Mayo Clinic
DAVID REDFEARN, PhD
TAMARA SIMON, MD, MSPH, Seattle Children's Hospital
CHRIS TOMPKINS, PhD, Brandeis University

NQF STAFF:
TAROON AMIN, Senior Director, Performance Measurement
HELEN BURSTIN, MD, MPH, Senior Vice President, Performance Measurement
NEAL COMSTOCK, JD, Vice President, Member Relations
ANN HAMMERSMITH, JD General Counsel
ANN PHILLIPS, Project Analyst
ASHLIE WILBON, RN, MPH, Managing Director, Performance Measurement
EVAN WILLIAMSON, MPH, MS, Project Manager, Performance Measurement

* present by teleconference

A-G-E-N-D-A

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

8:33 a.m.

MR. WILLIAMSON: Good morning and welcome to the expert panel meeting. We will now begin. We will start with Neal Comstock who will give us an update on the annual conference.

MR. COMSTOCK: Well, good morning and welcome to NQF. Thank you for coming here and joining us in person as well as on the phone.

I wanted to just introduce myself and tell you a little bit about our annual conference next week which I very much hope you can join us for.

We will have a terrific conference to focus specifically on the type of information, information about healthcare quality that can be useful to actual patients and consumers, and also of course to providers and other healthcare professionals.

There's a great wealth of information that's made available as a result of measures approved by NQF and as a result of many other sources, some reporting by professionals and institutions as well.

And there's a very wide array of sources that actual patients when they need to make healthcare decisions go to to get this information.

There is a wide variety of information that is better -- some which is better than others and that is more accessible than others. And we want to have basically a day and a half conversation with everyone in the quality enterprise about how we can collectively make that information more useful to actual patients and consumers.

And so that's what we're going to do. We will start off with of course remarks from our CEO Chris Cassel. This will be her first annual conference as CEO

of the organization.

We'll have a series of discussions on variations of this topic. But among those who will be speaking are James Guest, the president of the Consumers Union. And of course Marilyn Tavenner, the CMS Administrator.

And on Friday afternoon to close out our conference we'll have remarks from Senator Tom Daschle, the former Senate Majority leader and from Mike Leavitt, the former Governor of Utah and HHS Secretary.

We'll have an engaging and interesting and I believe you'll find informative conference. I very much hope that you can join us for it.

It's at the Marriott Wardman Park Hotel next Thursday and Friday. And you can find more information about it at our website. But please also don't hesitate to ask me any questions or follow up with me afterwards.

Again, my name's Neal Comstock. I'm fairly new to NQF here as vice president for member relations. Welcome to NQF. Welcome to our meeting. Thank you for your time and for your work here today. We very much hope we will see you next week. Thank you.

MR. WILLIAMSON: Great, thank you very much, Neal.

At this point we'll go over some of the logistics for the meeting. And my name's Evan Williamson. I am the project manager for this project. We want to go over and make sure everybody knows the lay of the land here, what we'll be doing the next 2 days, where everything is and how we can functionally operate through this meeting.

So if you want to leave the room here, head out past the elevators and take a right. We have restrooms. We'll be taking three main breaks today, 10:15, 12:45 for

lunch and then again at 3:15 after the breakout session.

We do have wi-fi here. The network is NQF Guests. The login is "guest" lowercase and the password is "nqfguest."

We want to make sure everybody mutes their cell phone during the meeting. We don't want to have any unwelcome interruptions.

We will be using again microphones. Please be sure when you are speaking that you speak directly into the microphones. You see that these microphones, red means that it's on. We can only have three microphones on at once so if you are not speaking and you have finished speaking please be sure to press the speak button again to turn it off.

And you just need to press -- to turn it on. Don't press and hold. We've had issues with that. So these microphones pick up everything for the court reporter.

We want to make his job as easy as possible.

We have provided a discussion guide along with an agenda. The discussion guide will be the main document we'll be using for this meeting. We have printed out a copy for you. It's also available on the SharePoint site. We'll be making realtime changes to it as we go through. You see it's displayed on our two auxiliary monitors on the side. So we'll be going through that as well. Again, that contains all of the key questions for today's meeting that we'll be hoping to answer as we go through. And that's our main document.

I'll introduce the rest of the project staff now. Or I'll let them introduce themselves. I'll start with Ashlie.

MS. WILBON: Good morning, everyone. I think I got a chance to greet everyone individually this morning so I just want to thank everyone for coming.

I am one of the managing directors in the performance measurement department and I've been working on most of our cost and resource use work over the last couple of years. So I'm really excited about this work and I'm excited about the group that we've gathered today. So I'm looking forward to a good meeting. Thanks.

MR. WILLIAMSON: And Taroon?

MR. AMIN: Good morning, everyone. I'm very excited to get started on this work.

My name's Taroon Amin. I'm a senior director here at NQF supporting our cost of care efforts both on the performance measurement side and on the Measure Applications Partnership side of NQF.

In terms of disclosures I just wanted to note that I maintain an academic affiliation with Brandeis University and was part of version 1 of the Medicare Episode Grouper Development Team. But I have since

not been part of that team for almost 3 years now.

MR. WILLIAMSON: Thank you, Taroon. We also have Ann Phillips.

MS. PHILLIPS: I'm Ann Phillips. I'm a project analyst here at NQF.

MR. WILLIAMSON: Additionally next to me we have Helen Burstin.

DR. BURSTIN: Good morning, everybody. I'm Helen Burstin, the senior vice president for performance measurement here at the National Quality Forum.

I'm delighted to see so many familiar faces and we're also -- and some new ones. I'm really thrilled to get to consider this really very important new line of inquiry for us.

MR. WILLIAMSON: We also have Ann Hammersmith, our general counsel. She'll be talking to you all in a minute.

At this point I want to go over the time-line. We went over this in

orientation just to remind everybody of what we are doing today and then we will be doing in the future.

So again, we've gone through our orientation. We've had our information review. We'll be doing our in-person meeting today and tomorrow.

We do have two post meeting calls scheduled on March 12 and March 19. We will have a draft report posted on March 24 for public review and comment. We'll be meeting again on May 14 to review those comments by phone.

Then we have a CSAC review and approval through June. Hope to get endorsement by the board at the end of June. And then our final report will be complete by the 1st of July.

So again, this is a quick trip through this. We appreciate all your effort on this. We know it's going to be a lot of work here going forward. We'll hope to make

it as easy as possible.

At this time I'll turn it over to Ashlie to go over the project scope.

MS. WILBON: So I'll just talk a little bit -- some of these slides you may remember from orientation, but since we're reconvened here in person we just want to make sure that everyone is on the same page as we start the day.

So, our work today is primarily going to be focused around understanding exactly what we mean when we say episode grouper. So how do we define that, how do we differentiate that between other types of measurement systems.

And in doing that we're going to take a lot of time kind of defining what those key elements of a grouper would be. What are some of the key principles when you're defining those and evaluating them. And then what criteria might we use to evaluate them.

And then the last part of the day tomorrow will be used to kind of think through some of the implications that we need to think through before we begin to actually endorse grouper. So that will be a key discussion we'll have tomorrow.

In terms of the scope of this project I just want to reiterate that we are going to be focusing on principles and considerations. We're not going to be actually evaluating groupers today. That's not the purpose and the charge of this group.

We're going to really try to stay away from talking specifically about the merits of specific tools or products, and that the criteria that we would develop potentially could be applied to any grouper.

This criteria will not be developed to evaluate a specific grouper or tool that we may be familiar with. So just want to make sure that everyone understands

this is kind of, just like our measurement evaluation criteria for resource use measures and quality measures, the criteria are broad such that any type of measure could be -- or the criteria could be applied to evaluate any type of measure within those domains. So, just want to make sure that we reiterate that as we go forward.

MR. AMIN: Yes, and I'll just sort of emphasize I think one point that Ashlie is bringing up here. We can't stress how critical this is.

We recognize that when we're talking about principles and considerations many of you in the room have spent a considerable number of hours and days and years working on either developing groupers or using groupers in various different applications.

So the thought of making considerations or principles that are divorced of your own product seems a bit

challenging.

So one of the critical things that we're going to ask you to do in a minute is to make sure that at least everyone -- I mean, it may be challenging, but we're going to ask you to do that anyway.

And at least we can try to understand various different people who have worked with different products. Maybe we can generalize some different principles and considerations across different products.

But it is extremely important that all members of the panel disclose to each other if they have any experience using any of these groupers in any application.

That way we have any potential considerations or your own personal interest in terms of how you developed groupers out on the table so we can all have a pretty open conversation about how these principles may transcend an individual product but

could be applied across different grouper types. So again, I just want to reiterate that.

It's extremely important for the success of this project and the credibility of the outcomes of this group that we make sure that we're at that level of making sure that this is across different products. So, I think that's all I needed to say.

MS. WILBON: And pardon me if you were going to say this, but just to kind of piggyback on Taroon's statement. Considering that all of you guys come from various backgrounds and we've actually convened you because of the expertise that you have.

And you are actually sitting on this committee as individuals because of your expertise that you bring as an individual professional and so forth. And so you're not representing your organization or affiliations that you have with

particular products and so forth.

So that said, I'll hand it over to Ann to carry us through the disclosures process. Thank you.

MS. HAMMERSMITH: Thanks, Ashlie and Taroon, both of whom did a very good job of summarizing key considerations when doing disclosures of interest.

As Ashlie mentioned you sit as individuals so you're not representing an organization that nominated you or that you work for or that you're somehow associated with.

And as Taroon pointed out you're not looking at measures today, you're not looking at individual groupers. You're looking at developing criteria.

And because of that as Ashlie and Taroon noted most, if not all of you are going to have involvement with groupers. You've worked on them, you've consulted on them, and so on.

So what we want to do today is we want to go around the table, tell us who you're with and have you disclose any involvement you've had with groupers.

On the slide you will see the specific disclosures that we are looking for, involvement in the development of an episode grouper system, a personal financial arrangement or affiliation with a specific product or service based on a product. That may be stock ownership. It may be that you consulted with a company on a grouper.

Investment in a specific product by your organization. And employment by or other affiliations with organizations, companies, or other entities that own, develop, or use episode groupers.

We are not looking for you to summarize your résumé. We're looking for you to make specific disclosures regarding the work of the committee today.

So, with that I'm going to start

with the co-chairs. We can go around the table. To the extent people are on the phone I will call on them.

MS. MARTIN ANDERSON: Thank you, Ann. I'm Kristine Martin Anderson. I'm currently employed by Booz Allen Hamilton.

I've had two interactions with grouper development, one in the late nineties where CareScience, my former employer, developed a grouper that ultimately was not taken to market because we weren't satisfied with its overall reliability of performance.

And then now I work for Booz Allen and Booz Allen is a subcontractor to Brandeis University and their CMS contract on the open source grouper.

DR. CACCHIONE: I'm Joe Cacchione. I'm with the Cleveland Clinic.

I'm on the scientific advisory board for United Healthcare which has an ownership of Optum Insights. And we have

customized some grouper tools for local use only with Optum. Otherwise I have no other affiliations that are material.

MR. HOPKINS: I'm David Hopkins from Pacific Business Group on Health. And I don't have anything to disclose that fits any of those bullets. But I do have some experience with groupers, so a couple of things.

Back in the early two thousands we got an AHRQ grant and worked with Doug Cave as he was actually developing his grouper system. And we were looking at the variation in costs among physician groups with Blue Shield of California.

More recently I was chair of the technical efficiency committee for IHA's pay-for-performance program and we did some work with -- what was the name at the time? It was Medstat, or Thomson Reuters, or something, one of those. At any rate the owner of MEGS. So I got some familiarity

with MEGS but never had any involvement of that type.

MR. BODYCOMBE: Hi, I'm Dave Bodycombe. I'm with Johns Hopkins Bloomberg School of Public Health. I direct research and development for the ACG System. It's a commercially available case mix adjustment predictive modeling tool. We do not have an episode grouper component to that so I don't think any of these particular conflicts would apply.

DR. LEVINE: HI, I'm Mark Levine. I'm a physician employed by Centers for Medicare and Medicaid Services and I'm the clinical lead for the development of the Medicare episode grouper.

DR. BANDEIAN: I'm Steve Bandeian. I'm an internist at AHRQ. I developed at AHRQ an analytic system that includes groupers as -- episode groupers as a component. And I have participated in the CMS project.

MR. TOMPKINS: HI, I'm Chris Tompkins. I'm on the faculty at Brandeis University. I'm the project director for the CMS support contract to develop the public source grouper.

MR. JONES: I'm Jim Jones. I'm vice president at AmeriHealth Caritas. It's an IBC-owned Medicaid plan.

I don't have any specific disclosures that fit the descriptions there. But like everyone else I've used various tools for network tiering and performance contracting, MedStat, BPS, tools like that.

MS. HOBART: I'm Jennifer Hobart at Blue Shield of California. I'm also on the technical committee of IHA California PFP, the technical committee of CHPI which is a California collaborative that among other things is working towards an all-claims database, and worked with PBGH and other entities.

I think in terms of particulars

Blue Shield has had a grouper in the past. They had Optum ETGs which we suspended for awhile but now we're re-initiating, putting in a grouper, and the various collaboratives use groupers.

MR. MCLEAN: Hi, I'm Jelani McLean. I'm the head of cost analytics for BCBSA Blue Distinction Center Program, so developing the entire methodology from administrative claims all the way to evaluating the facilities.

And in that component obviously would be some sort of grouper. So I've had experience with a lot of customer groupers when they come for transplants, so forth and so on.

MS. SIMON: I'm Tamara Simon. I'm a pediatric hospitalist at Seattle Children's.

I've been involved in -- I haven't been involved in the development of an episode grouper system, but I have been

developing through the Centers for Excellence for Children with Medical Complexity a pediatric medical complexity algorithm that is through funding from AHRQ and CMS.

It's essentially designed to be a publicly available algorithm to identify children with medical complexity. Our center has done comparisons with the 3M CRG system.

MR. REDFEARN: I'm David Redfearn. On Monday of this week I retired from WellPoint after 31 years with the company.

While I was at WellPoint -- WellPoint licenses the Optum ETG product and the Truven MEG product. I've had fairly extensive experience with both of them.

I've also spent some time looking at the Optum procedure episode grouper. And most recently I've been trying to take a look at the 3M patient-focused episodes

model.

So I've had wide experience, in fact, hands on experience with trying to run these suckers which is sometimes a challenge. But no conflicts at all.

MS. GARRETT: Good morning, I'm Nancy Garrett. I'm the chief analytics officer for Hennepin County Medical Center which is a safety net care provider in Minneapolis.

And I don't have any conflicts to disclose of that nature. I have worked with various episode groupers from a payer perspective in past positions in my career.

And at NQF I'm involved in a couple of other committees, the cost and resource use as well as the risk adjustment and socioeconomic status group.

DR. MIRKIN: Hello, I'm Dave Mirkin. I'm the chief medical officer for Milliman MedInsight analytic platform.

And as far as I know I have

nothing to disclose, but I do need to disclose that Milliman on the consulting side works with a number of organizations. And I wouldn't know if they actually contributed to development of an episode grouper.

MR. NAESSENS: Good morning. I'm Jim Naessens, a health services researcher at Mayo Clinic.

I've been involved with evaluating groupers including MEG and ETGs and PROMETHEUS, but haven't been involved with any development and as far as I know Mayo Clinic has not been involved in developing an episode grouper.

MR. MACURDY: Hi, I'm Tom Macurdy. I'm a professor of economics at Stanford University but I also serve as the senior research associate at Acumen LLC.

Acumen has held the evaluation contract for evaluating episode groupers for CMS since 2008. And we've had extensive

experience with ETG grouper, the MEG grouper and the 3M grouper.

Acumen also is a payment support contractor for the hospital payment system with Medicare and in that role we have developed groupers for pay-for-performance sort of schemes.

DR. KING: Hello, I'm Marjorie King. I'm a clinical cardiologist working at an acute rehab hospital in the metropolitan New York area, Helen Hayes Hospital affiliated with Columbia University.

For disclosures I was involved in the Brandeis PCPI et cetera initial product that was developed for CMS and am now in the clinical work group of the CMS episode grouper project.

MS. HAMMERSMITH: Okay, thank you. I'm going to call on some people who may be on the phone so that they can do their disclosures. Is Francois de Brantes

on the phone?

MR. DE BRANTES: Yes, I am. Hi, Ann. Can you hear me?

MS. HAMMERSMITH: Yes.

MR. DE BRANTES: All right. So I'm Francois de Brantes. I'm the executive director of the Healthcare Incentives Improvement Institute. And I led the development of the PROMETHEUS payment model which created as a part of that payment model a grouper tool called the Evidence-informed Case Rate Analytics.

I then worked with Brandeis on the development of a prototype for what we refer to as version 1 of the Medicare episode grouper. And HCI3 is also a subcontractor on the development of the current versions of the Medicare episode grouper.

In addition to that HCI3 has a relationship with the SAS Institute in which the SAS Institute has developed a new

episode system that's based on our ECR analytics called the SAS episode analytics.

MS. HAMMERSMITH: Okay, thank you. Is Dan Dunn on the phone? Is Jim Loiselle on the phone?

MR. LOISELLE: Yes, good morning, everyone. Jim Loiselle from McKesson Corporation.

Going down the bullets no direct involvement with developing episodic groupers. But at my work through McKesson we have in various business units we have deployed and/or implemented as OEM partners ETGs and PEGs from Optum, MEGs from Truven, PROMETHEUS tools in our payer solutions as well as I've evaluated internally for McKesson the 3M grouper as well.

MS. HAMMERSMITH: Okay, thank you. Thanks, everyone, for those disclosures. I just want to give you a few additional reminders.

The most important one is that we

expect you to participate in the committee's work in an open way. We expect you to listen to each other, keep an open mind.

We realize that you all have ideas and opinions which is part of the reason you're on the committee, but this is a group process.

If you are ever in a committee meeting doing work with the committee and you believe that a fellow member is biased, is unable to participate in an open and fair manner you should bring that to our attention immediately.

If you want to bring it up openly in the meeting you are entirely welcome to do that. You can go to your co-chairs who should then go to NQF staff, or you should go to NQF staff directly.

What we don't want is any committee member sitting thinking that there is bias or something improper is going on and not speaking up.

So, in that spirit do you have any questions of each other, or of me, or anything you'd like to discuss based on the disclosures this morning?

Okay, thank you.

MR. WILLIAMSON: Thanks a lot, Ann. And we're running exactly on time. Let's see how long that lasts.

Now we'll move into setting the stage. We have Ashlie Wilbon and Taroon Amin.

First, we'll just quickly go over the agenda of the meeting. We just went over the welcome, the project purpose and the time-line. We just did the disclosures of interest. We will now set the stage.

After that we will review key definitions. We will then review the existing NQF resource use measure evaluation criteria.

After that we will define the key modules for episode groupers followed by a

public and member comment period.

After that we'll have lunch. We then have breakout sessions where we will use those defined modules to really talk about how we construct and evaluate an episode grouper.

We'll then convene back as a full group where we'll review the work of the breakout groups, going over the principles for constructing and evaluating. After that we will adjourn.

We do have a dinner planned tonight that is optional but we will get a final headcount at lunch. It's located just a block away from the hotel so we hope that most of you will be able to join us.

MR. AMIN: Evan, before you move on if we can just go back to the slide right before lunch.

I just wanted to point out to the committee one of the critical things that we're going to be doing today is there's a

degree of flexibility in the agenda here in terms of how we define the key modules.

NQF staff has sort of developed a straw person for the committee to react to in terms of what the key modules are for episode groupers. And we can discuss that at further length.

In general terms the clinical logic, construction logic and adjustments for comparability.

Our goal is to ensure that those are appropriate modules and the components within those modules are appropriate. So, once we have that structure in place we'll use the breakouts to then do deep dives in each of those modules.

So, by no means is this setup set in stone. The purpose of the morning session is to go through those modules and ensure that we're all comfortable with that construction or at least can live with that construction and then do a deep dive later

on in the day.

So there is a high degree of variability here but we wanted to at least start with a structure and ensure that we had something to start with and then we can make some adjustments as we move forward.

So you may have noticed that as you reviewed your discussion guide that there's a lot -- much of the structure is already set up for you to react to. But you shouldn't feel constrained by that structure as it's set up.

MR. WILLIAMSON: Thank you very much, Taroon.

MR. DE BRANTES: Evan, this is Francois. Just a question on the breakout sessions. How is that going to work for us on the phone?

MR. WILLIAMSON: Yes, we will have a dial-in available. We have a speaker sub-conference that we'll pull the groups into. We have one group set for this main

conference room and so we'll make sure that you guys are able to participate via phone.

MR. DE BRANTES: Okay, thank you.

DR. BANDEIAN: This is Steve Bandeian. I'm sorry to raise this but I'm a little -- I mean, while I understand that there are modules I'm -- and I've read through the document so I kind of know what's there I sort of think that actually higher-level discussion prior to consideration of modules is, you know, may be worth considering.

And while it may well be true that almost any grouper would have these modules that somehow seems a bit more detailed than sort of the very high-level concepts of what is required for the grouper to be acceptable.

And so to me I appreciate all the work that's been done here, but it does seem to me that some higher-level concepts may be worth considering first.

MR. AMIN: So, let me propose this in terms of how we were thinking about it. And if the committee feels strongly about that we can have some fuller discussion.

So, the goal of this isn't -- of my statement wasn't to jump right into the conversation around those modules.

The agenda is set up to first have some overarching considerations of how we're thinking about this space. And then that would certainly be the opportunity to have general higher-level conversation.

And more importantly, there is a section right afterward which is to define the critical components of what a grouper entails. And in that period we can also talk about general constructs that seem appropriate for the group that may need to be discussed in broader detail, or from a broader context. See if that's sufficient to the group and the chairs.

MR. REDFEARN: First, a procedural. Should we use a little rule of turning our signs sideways if we want to talk so that the chair can recognize us?

DR. CACCHIONE: Yes, I think that would be a good idea.

MR. REDFEARN: But, my comment was that I think we're going to start talking about definitions. And I already have comments about the definitions. I don't think the definition is broad enough to encompass all the variety that's out there. So I suspect we're going to get into some of these issues before we drill down just inevitably, just based on what we're seeing.

MS. WILBON: So, this is one of our template slides that I'm sure most of you have seen if not on the orientation a few weeks ago. Just to kind of give a little bit of context on NQF and the work we've been doing and how we conduct our work

in terms of using multi-stakeholder groups to build consensus around different measurement, quality measurement and cost measurement topic areas.

And for this particular process we have named you guys an expert panel. Generally the expertise for our steering committees tend to be more multi-stakeholder and representative of our eight membership councils which include consumers, providers, health professionals and so forth.

Because the task of this group is much more specific and technical we have convened a group that as you can hear from the introductions around the table that there are multiple stakeholders represented, but that the people we've actually asked to participate on the committee have that very specific technical expertise.

But I did want to add that because we are a membership organization and we do represent a multi-stakeholder group

that the work of this group will be shared with our multi-stakeholder group.

And we would like to work with you guys in terms of the report that we put out to make sure that it is a product that can be shared with the multi-stakeholder group and that is understandable and digestible for a broader audience than just a very technical group.

Although we understand obviously that you guys, we're asking you to do a very technical task and the context of this meeting will be very technical. So I just wanted to kind of bring that context in as we embark on this journey.

MR. AMIN: Just quickly, sorry. I will also note that the work of this committee clearly impacts both our endorsement process. And we will have a discussion as well about potential implications to the Measure Applications Partnership and potential considerations for

applications of measures that might be coming out of episode groupers. And that will be part of our day two path forward discussion.

MS. WILBON: Thanks. So, a lot of these things we've talked about before. And some of these things will be covered in upcoming slides, and particularly the "why now?" so I'll kind of skip over that.

But historically the purpose of NQF endorsement has been to adopt standards that can be used to be compared, to make national comparisons around different quality measurement topics. Particularly in the last few years we've moved into the cost measurement space.

So generally endorsed measures are deemed to be kind of national standards for measuring these topics. So that theme is going to kind of carry through as we get into the criteria for episode groupers as we kind of think about whether or not that

concept of having a national standard or particularly having one particular method for measuring episodes in a particular way is applicable in terms of the endorsement context that we have used in the past. So we'll kind of refer back to that as we go.

And that also kind of encompasses the balancing the flexibility in some of the grouper methods in that many of the tools have user options that allow users to choose different methods depending on what the intended use of their analysis is. So we'll talk a little bit about that as well as we go forward.

Clearly there is a cost imperative that there is a need for more tools and measures to measure costs in the healthcare system. And a lot of policy implications some of which are listed here in terms of legislation around physician feedback programs, value-based payment modifier and so forth. I won't read them

all off.

MR. AMIN: Yes. And clearly this discussion broadly has implications for both commercial and public applications.

So that, again, we want to keep this conversation broad. It obviously has a lot of implications for various programs.

Chris, do you have a question?

MR. TOMPKINS: Yes, it's more of a state-setting question I guess as we sort of feel our place here.

When I think of NQF I think sometimes in terms of your mission is to uphold I'll just call them minimum standards. You probably don't call them minimum standards, but standards of acceptability.

And it's possible that many measures that are purporting to do the same thing or similar things can be acceptable in their own way.

We say we're going to do this,

we're going to say we're going to do it that way. And then somebody else comes along, does a measure development activity. We say we're going to do something similar but we're going to choose a different pathway.

In general are you trying to look for standards that are minimum which could accommodate a lot of flexibility discretion among the building of episode systems in such a way that NQF could see that several of them meet those standards because they say what they're going to do and they do what they're going to say.

Or, part two is sometimes I think of NQF as this best in class kind of thing. In other words, discriminating criteria that say yes, two of them are reasonably good but we are going to choose the winner.

So anyway, you can comment or not comment on that. Are we attempting the latter?

MS. WILBON: I will just say real

quickly and I'll just have Kristine add on. I think we're still trying to figure some of that out. And I think some of your questions may be -- the discussion on day two will help us flesh some of that out. But I'll --

MS. MARTIN ANDERSON: I think we're not -- definitely not at the point of trying to do a best in class here in this particular area.

But one thing I would offer is that there's a lot to be learned from past efforts to look at and proper ways to approach sort of methodologies. And I think the most recent one in my mind for NQF was around risk adjustment. Where in the end of the day it really turned out that it's not so much exactly how you do it, it's that how do you know if when it's done it's good.

So I think from that perspective I always think about us looking from the endpoint backward.

And it may be that there are many different methods that can produce a good result and this committee needs to talk about that.

But the question is what's a good result. And in context of how it's being used, or intended to be used.

So I think if we keep ourselves at that level it will be easier to try to figure out what kind of criteria should there be that you could then say this is a good episode grouper without getting into this is how you create an episode, you know, being so prescriptive about the how.

DR. BURSTIN: Yes, that was great actually, Kristine.

I think the only thing I'd add to Chris' question is I think this may not be the same space as the measure space we have lived in traditionally of individual measure by measure.

And I think we're open to

whatever emerges out of this, what is the right approach.

We do hear a lot from the field of people wanting to at least have some confidence that if they're using different systems the results are somehow comparable.

I think that's going to be something -- from an end user perspective people will want to feel comfortable that the end results of the use will not disadvantage one group or another. But again, I think that's to be told as you go through your process.

So I think you should assume this is a very open-ended assignment and we're really in a space we've not been in before. So we really look to your guidance.

MS. WILBON: This is a very busy slide, but I will just highlight a few things.

The purple boxes are highlighting some of the other work that we have going on

in the cost measurement space. We've definitely grown in terms of the type of work we've been taking on in this space over the last few years.

And those purple boxes are superimposed upon another kind of framework in the blue and the green boxes that kind of show how we think about kind of cost measurement in the context of efficiency and value.

And that cost measurement really along with quality is how you come up with your efficiency signal. And that the efficiency signal potentially with the cost and quality in combination with stakeholder preference is how you get your value, how you better understand value.

So, the purple boxes within the different blue boxes is kind of explaining the different parts of work and which parts of the model they're addressing.

So, in the value box, the big

value box you can see we're doing a project around measuring affordability for consumers.

That's a piece of work that is sponsored by the Robert Wood Johnson Foundation and it's really focused on kind of understanding what types of measures and measurement concepts are important to consumers in understanding affordability and how they can make decisions about purchasing and engaging with the healthcare system.

There's another effort just under that in the linking cost and quality project also sponsored by Robert Wood Johnson Foundation in which we're producing a white paper that will discuss some of the methodological challenges around combining costs and quality signals to get an efficiency signal, what that looks like, the different approaches that there may be to get to an efficiency signal. And we're convening an expert panel to discuss those

issues as well.

And then the episode grouper work we have in the resource use space as well as a parallel effort we have with the standing committee for cost and resource use measurement in which they are using the consensus development process to evaluate cost and resource use measures.

So, our current effort, we're reviewing three cardiovascular measures for cost and resource use and that is ongoing.

We also have an ongoing effort through our Measure Applications Partnership in which there is a subset of one of the MAP committees that has been developed to address, to discuss affordability and develop an affordability family of measures, and kind of think about some of the high-leverage opportunities there are to identify measures and measure costs at the system level.

So, that's kind of some of the

work we're discussing. You can kind of see how it's somewhat connected in the context of what we're doing here today.

MR. AMIN: I'll just reiterate on that, our conceptual framework that we've been working with at NQF is that in order to really understand efficiency you need to look at costs in relationship to quality.

And really what differentiates value is taking into account preferences of various different stakeholders. So our goal is to try to move toward measures and measurement of efficiency which really includes both signals, to be able to really understand the efficiency of providers and the health system broadly.

So again, this work fits in the context of broader work that some members of the committee are very familiar with as being part of the Cost and Resource Use Standing Committee that is essentially overseeing the body of this work.

So, the current landscape for groupers. Again, we have -- many of you in the room are obviously very familiar with them. But for those of you that are not episode groupers have, you know, this is an established space in some ways and new in some ways.

There's been established players in the episode grouper market from many established people including Optum Insight, former or still part of I think United Healthcare, and various different other products that are in the market for the commercial population that have been used for commercial, potentially profiling, for provider profiling and potentially for pay-for-performance applications.

There has been increasingly new work that has been in play for an episode grouper for the Medicare population which Tom and Chris obviously are very familiar with and others in the room as well clearly.

And so there's various different tools. And a lot of what we heard is that it clearly varies by region. There is not one national standard.

And there is a concern -- and there are various concerns about episode groupers that we've heard from our stakeholders. And we will explore these challenges during the course of these two days.

The first which is not an insignificant challenge is the complexity of the groupers makes it very difficult to do an evaluation of them and to understand what the cost implication on the other end when you're being profiled, what you're actually being profiled for.

The transparency of these groupers varies. Understanding how decision logic or how individual claims are being assigned to various different episodes, how various episodes relate to one another.

There is a varying degree of transparency in the market.

Part of what this initiative is intending to do is to create increased transparency or expectations for increased transparency for products that are in the market both for consumers and purchasers, clearly, and also for providers who are being profiled using these products going forward.

There's obviously a lot of challenges for providers who are being profiled using multiple different grouper systems and are being given different information, different results. And different methodologies causes a lot of challenges in terms of being able to understand how to improve.

We also recognize and we're obviously not walking into this blindly that this effort has clear market implications where the efforts of many different

commercial products that are in the market.

And also there are various different proprietary components of these groupers that should not be underestimated.

Again, the goal of NQF's effort in this space is to move toward national standards of how to measure cost and resource use using episode groupers as one potential approach, and to keep transparency at the forefront of that effort.

And so NQF's role in the evaluation of groupers is very new. We are in, in a lot of ways, uncharted territory for NQF. And so we will be asking a number of series of path-forward questions around can episode groupers be evaluated in isolation of their -- can they be evaluated just in terms of their output, meaning the episode grouper measures as in the way that they're slated to be used for the Physician Feedback Reporting Program and potentially other value-based purchasing applications.

Or can they be looked at in -- or should they be looked at in totality, meaning the episode grouper, all of its components and its output.

So, this is a very new space for NQF. Again, we're looking for some guidance. And the guidance here will be translated to other governing bodies of NQF that will evaluate the recommendations of this expert panel. And those will mainly be our Consensus Standards Approval Committee and the board who will both be looking at the recommendations of this committee in terms of what NQF's future role will be in the actual evaluation of groupers going forward.

That seemed to have initiated a lot of comments so I'll turn it to the chairs to manage that.

DR. CACCHIONE: Steven, you had a question or a comment?

DR. BANDEIAN: On the previous

slide where you have the different boxes and the different colors the point that I'd make is that if one is talking about sort of cost of care or efficiency one might want to think about going back and looking at the existing types of things that you've been looking at and approving.

Because all cost and efficiency measures are really part of a more comprehensive picture of care. So one could look at, you know, an emergency room visit, what the cost of that is. But if you didn't consider what happened to the patient after they leave the emergency room you may not have a very good understanding of what the implications were of that care in the emergency room.

So what I'm trying to say is I suspect that this episode discussion may well ultimately move to looking at the whole range of cost measures because there is a lot of interconnectedness between the

different ways of looking at costs.

And so it may be that there will be implications of this episode work that feed back to how you've been thinking about cost of care measures in other contexts.

Even for the consumer, by the way. Because when the consumer has a knee problem and is thinking about going to Dr. Jones or Dr. Smith ultimately what's important is what the total cost to the consumer will be, likely, from beginning to end which is almost an episode type of concept.

So, all I'm trying to say is if one sort of goes up to an abstract level and thinks about how do we measure efficiency broadly, cost and quality, I think you'll find that all of these areas have a lot of interconnectedness. And it may be worthwhile to try to puzzle that out. At some point in the longer term.

DR. CACCHIONE: The episodes tend

to be somewhat arbitrary in terms of their time constraints or the constraints that are put on them by how we do that.

Mark, you were next I think.

DR. LEVINE: Yes, just the observation that on the following slide when you talk about the current uses of groupers, that really is just a population and geographic look at uses.

But I wonder whether or not we would be wise to consider the use case for groupers in general.

What are the use cases that we have? And what are their purposes? A grouper that is good at one use case might not be applicable in another use case.

And I suspect that we're going to need to evolve different standards and different approaches for looking at different use cases of groupers. So I welcome discussion about use case.

MS. MARTIN ANDERSON: Yes, that

is on the agenda.

MR. TOMPKINS: I think I just had a quick clarifying question. Under "Challenges" the first two bullets are complexity and transparency. So if you understand that complexity is a challenge then you move on to the next bullet, transparency.

Does that mean that the methods that are used in the grouper are disclosed? Or does it mean that, for example, that they are proprietary and undisclosed? In other words, are these two separate bullets?

You can have a complex system that is fully disclosed in which maybe some people understand it and some people don't, versus you could have a system of any complexity that isn't disclosed and it becomes literally kind of a black box.

MS. MARTIN ANDERSON: One of the elements that's come up often in NQF review of measures is could somebody repeat the

results on their own. So there's a level of typically that would be required.

Now, if they could handle the complexity is a whole nother issue. Could they do it, right? But is it transparent enough that someone could recreate the results for themselves.

And assuming they have the capability with the complexity which oftentimes people do not, and/or the access to the data that would allow them to do it which oftentimes they do not. So I think they're two very separate things, but important.

MR. HOPKINS: So, just to extend that issue a little bit more.

So, episode groupers are by nature very complex. They are difficult to understand. By busy physicians, certainly by lay consumers.

So often when I have heard people raise issues around transparency what

they're really saying is, you know, I don't have time to look at it, it's too complicated. It's really about complexity.

I haven't met an episode grouper yet that one can't delve down into the deepest part of it and look at codes if you want to do that.

So, I haven't seen that transparency is an issue. But maybe I'm missing something.

MR. LOISELLE: This is Jim Loiselle. May I interject?

Yes, I think the distinction, and I think the previous commenter made that point, is that the greater variation comes in how you apply the episode, whether it's for a payment purpose, an initiative purpose, or an analytical purpose, or an efficiency purpose. The outcomes are very variable even on the same grouped information.

So I think that looking at the

use case question about how you use this information is really a much wider and broader discussion than actually the evaluation or the creation of the episodes themselves.

MS. MARTIN ANDERSON: Thank you.

DR. LEVINE: I was just going to suggest that perhaps the issue is really not complexity or transparency but understandability, that it must be committed to the user in a way that they understand what's going on and can therefore interpret the results.

MS. MARTIN ANDERSON: Let's have this conversation -- that's a great point and let's have that conversation when we talk about the applicability of the NQF endorsement criteria because I think that comes up, that usability element comes up very clearly, as do some of these issues around transparency.

DR. BANDEIAN: Just briefly. So,

why are complexity and transparency important? That also then relates to the use case.

But part of the use case is also what do we want people to do with this. And so if we want doctors to trust and feel comfortable to be able to use the information to improve the care that they're providing that I think has -- if one says that, that statement has a whole series of logical consequences that I think are -- what I would argue would be the principal things that this committee should define as criteria.

What needs to be in a grouper so that the medical community looks at this and says we trust it, we're comfortable with it, we think it's fair, but even more than that it's providing us information that we can use to do a better job.

And if we can then say what is logically required so that physicians across

the country feel that this is something that they can trust and use to transform the care that they provide, those logical requirements I think would go a long way to what you want to specify.

MR. DE BRANTES: Evan, this is Francois. Unfortunately I don't have a table tent that I can raise so I don't know if I'm -- so I'm raising my hand, but if I'm not in order --

MR. WILLIAMSON: That's fine. Whenever you want to talk just go ahead and speak up.

MR. DE BRANTES: All right, thank you. So, my concern about these comments is that it seems to me that we're veering from a task which to me seems to be pretty clear, and that is establishing some criteria that others can use, i.e., other committees in NQF ultimately will use to evaluate a grouper as opposed to establishing criteria that prejudge groupers.

And I think this is an important and dangerous line that we shouldn't cross.

So in other words, criteria to evaluate a measure should include things such as, and it's on the list, reliability and validity testing and so on and so forth.

Then the burden is on the developer of the grouper to demonstrate that they have and they can meet those criteria of validity and usability and so on and so forth.

Some of these issues such as transparency and understandability by physician, I mean that's fine. But you know, to a large extent it's irrelevant.

And I'll tell you why it's irrelevant from my perspective which is if someone wants to develop a grouper that they feel is valid and that a committee might feel is valid but is completely not understandable by the field, it will essentially fail.

Now, then it's NQF's decision as to whether or not it wants to spend time reviewing those types of submissions. But I'm not sure that we should stand in prejudgment of the submission of potential developers of groupers.

MS. MARTIN ANDERSON: Thank you, Francois. And I think we are going to get deep into this conversation when we talk about what the endorsement criteria should be. And so thanks for offering that and to Steve, and to Mark, and to David, and I think we'll be getting deep into that.

I think we want to dive into now the definitions of a grouper because I think we have plenty of debate to have around that too. That's where we'll attack the use case issue, for what purpose are you developing the grouper.

So, Taroon?

MR. AMIN: Yes, I think actually Evan is going to lead that section. Evan,

take it away.

MR. WILLIAMSON: Great. At this point we'll be starting with the key definitions of episode groupers to make sure that we are all speaking the same language and have a general agreement on the definition.

So these are a straw man. These are provided for talking points. So we will pull them up here. We have these five discussions. The full discussions are listed in the discussion guide on page 3. So, we are asking key questions about these definitions.

So the first question we want to go through is describing the purpose and function of an episode grouper. And so we have two definitions here laid out that will help us get to that and where we have a definition of an episode and then definition of an episode grouper.

So, as far as the episode we have

an episode of care is defined as a series of temporally contiguous healthcare services related to the treatment of a given spell of illness or that is provided in response to a specific request by the patient or other relevant entity.

Do we have comments on the episode definition?

DR. CACCHIONE: One thing that I would say here is that the episode is not just related to the treatment but also to the condition itself.

I mean, I think that comorbidities that confound an illness are very important in terms of the providers. So it's not just related to the treatment arm. David?

MR. REDFEARN: There is sort of an implication here that an episode is sort of a clinically homogenous set of complaints. It's sort of driven off of the diagnosis. And a patient that has multiple

diseases is going to have multiple episodes.

First, there are groupers that are driven by procedures. For example, the Optum PEG Procedure Episode Grouper. So what triggers the episode is a procedure being performed.

And that may be fairly homogenous with regard to the underlying condition but not necessarily. You can do the same procedure for multiple underlying conditions. So you have to expand it a little bit to take into consideration when procedures drive the groupers because they're out there, they're being used.

The other is something I ran into in terms of looking at the new 3M patient-focused episode model in which there is only one episode active at a time for a member.

So what happens when you have a member that has multiple comorbidities? They're all in the same episode. They're all lumped together.

And 3M argues that that avoids the difficulty of parsing the utilization out by disease when a patient has multiple diseases. And we all know there's lots of work that's been done to show, for example, ETGs and MEGs carve things up differently when you do that. So basically we're just going to avoid it, we think that's too hard to do, and lump it together.

So it's not necessarily driven by diagnosis and it's not separate by underlying diagnosis codes. So I think we need to expand the definition a little bit.

MS. SIMON: As the pediatrician in the room I just want to point out that I agree with your statements. And with children, particularly with children we have acute episodes of illness for the vast majority of healthy children that are out there.

And then we have these incredibly complex children who are born with chronic

conditions and continue to have chronic conditions. So when I read about episode of care I really struggled with acute versus chronic conditions, and really understanding how long an episode of care might last for a chronically ill child.

MR. JONES: That's the exact point I was going to make, building on what David said. In that one challenge that I've always found and quite frankly the reason that we stopped using some of these tools was how do you really put bookends around something that has no clean period.

So, I almost would argue that it should not be considered an episode.

MS. GARRETT: I was going to talk about chronic conditions as well. And they just don't fit very well here and I don't think we can leave them out. So I'm not sure what the answer is, but I don't feel that the treatment of a given spell of illness is broad enough for what we're doing

here.

MR. MACURDY: I guess I think the definition is broad enough. Because in the cases of multiple comorbidities which come up a lot in Medicare you can, I mean the illness is kind of the comorbidities themselves and the combination.

I understand that you get a lot of combinations of illness as a consequence but that's actually the way a lot of the risk adjustment models work. And then the issue becomes, well how do you have not too many kind of conditions or episode kind of constructions. But I think the definition is broad enough.

And the issue on chronic care, that's true it's not a well-defined period. But this doesn't necessarily have a definition of a well-defined period. It can be over an extended period of time. You have an illness and there's a particular kind of sequence of care you're going to

have. So, I don't think you have to modify the definition accordingly.

DR. CACCHIONE: I think there's some -- or there's anchoring in all of our heads when we think about episodes around time constraint.

And you know, I think that to somebody's point earlier we may have to think about this thing more broadly and think about things differently. Because we all have this preconceived notion about an episode being anchored in time. And I think that that might not be the case when we walk out of this room.

MS. MARTIN ANDERSON: One thing to ask is do you all -- I'm hearing various levels of support for even including the words "of care" right? So, an episode of care. And is it just treatment, or is it also natural progression of disease, or a period of time that a disease exists.

Because this and the next

definition both are anchored in how the care is provided which I think is the signal we see through the data. But are you all in support that we're looking at episodes of care? Or episodes of illness? What is the feel of the group there? Dave?

DR. LEVINE: I think it goes back again to use case in the sense that if you're looking at an episode of care for hypertension or diabetes your use case may require looking at it over 20, 30, 40 years in order to -- if what you're looking at, what your endpoint is trying to get to is how does care influence the outcomes of the disease.

So, it needs to be very flexible I think at this level of definition of what is an episode. And then as you get into a particular use case each use case may be defining what it means by an episode in a much more flexible way.

So I think we're going to need to

wind up eventually looking at a whole set of criteria for different use cases. But in the overall way I think the definition as presented is appropriate.

MR. BODYCOMBE: When I looked at this I thought what happens between episodes. And you know, there are episodes of management, there are episodes of prevention.

And in fact, I would argue that an episode of care is in a sense a performance measure of a poor job at management or prevention. It's a failure.

So you could actually use it as an outcome measure. You shouldn't be having an episode of care if you have a well-managed patient.

DR. BANDEIAN: Episode of care, I read this definition and I was kind of okay with it.

But to have a complete picture of the consequences of the care that is

provided for a particular condition one may need to look at other conditions so to speak.

So, for example, a person has a hip fracture and it's treated and the patient is discharged. And a few weeks later the person develops a pulmonary embolism or deep venous thrombosis.

Well, that's sort of itself an episode, you know, the treatment of the deep venous thrombosis or the pulmonary embolism.

But if it were the case that that was caused in effect by the care or lack thereof during the hip fracture care somehow one needs to take those two things into account.

Because if one only looks at the care or the surgery and not some of the consequences of the care or the surgery that may create new condition episodes one may be having an incomplete picture.

So, I'm okay with the concept of

episode of care more or less with how it's defined and also including treatment episodes, surgical episodes, et cetera.

But that's not necessarily the unit of analysis on which one makes judgments as to whether what was being provided is the most efficient or the best possible. One needs to look at the interconnectedness of these things.

DR. CACCHIONE: Do you think that when you use the term related to the treatment -- so, a pulmonary embolism that occurs after a hip replacement is thought to be causal and related because of some -- whether it be some comorbidity.

So, is it covered in the definition by saying health services related to the treatment of a given -- does that suffice for the definition?

DR. BANDEIAN: It may well be that one can do something of that sort to tweak it. Yes, it may well be.

But I guess what I'm trying to say is there are, you know, a clinical entity may give rise to complications which are also clinical entities. And to have a complete picture one may need to make connections between episodes.

So one might say what are the costs of the condition plus the cost of complications which are fairly attributed to the base episode.

MS. MARTIN ANDERSON: We're going to take these -- did you have another question? These questions that are here, or these comments that are here.

But then I think I'm being persuaded to Mark's argument that we better talk about purposes and function. Because I think it's going to be hard to agree on a definition if we don't understand the breadth of purposes and functions in this room. So let me just --

MR. DE BRANTES: This is

Francois. I just, you know, sometimes it's good to go back to the origin of the concepts because the concept of an episode of medical care was developed -- was at least written about in March 1967 by Dr. Jerry Solon.

And his definition of an episode of medical care is as follows. An episode of medical care is a block of one or more medical services received by an individual during a period of relatively continuous contact with one or more providers of service in relation to a particular medical problem or situation.

And since then pretty much everyone has built groupers around that definition.

Now, the relationship of one episode to another episode, and how someone might construct it, and link a complication to a core episode and so on and so forth, those are design definitions for those who

will submit the episode groupers. And at some point they'll have to justify why they made those decisions.

Here we're talking about a base definition of what is an episode. And I would submit that we go back and use Jerry Solon's.

MS. MARTIN ANDERSON: Thank you, Francois, that's a good suggestion. That sounds like a good definition. Jelani?

MR. MCLEAN: I'm okay with the concept of episode of care, but I do go back to Mark's point about the use case. And you know, and to Jim's point about constraints around time.

When you work with groupers a lot you find out that either the time frame for the standard software is too long or it's too short.

I would argue that the key thing that's just really missing here is an episode is either designed to -- or an

episode is either an objective to get to a certain point of care, so a certain state for the patient, or a certain time of care based on some time constraint. But it's not really just time. It could be the alternative of I'm trying to achieve a certain state for a given patient. I think that's the one pressing thing that's missing in the definition.

MR. REDFEARN: Maybe I'm getting down -- I have a tendency to get down to a practical level real quickly. And I don't mean to disrupt things.

Going back to the acute versus chronic things, when you're actually using these things you have to deal with chronic episodes.

And I think what we need to ask the groupers is that they produce data with enough flexibility so that based on your use case you can do what you need to do with them based on your practical considerations

and what you're trying to do.

Typically at WellPoint we run 2 years of data with 3 months of run-out. That's our production. But we're running it on 35 million members.

Now, we would like to go out to 3 years but there are technical constraints in doing that.

And then from a practical point of view, and we deal with chronic -- what the grouper defines, the ETG grouper defines as a chronic episode, we chop them up, analyze them. That's generally the default way we do it.

But there is flexibility in the grouper that you could say I want to look at 2 years at a time. Or maybe if we could run 3 years of data we want to look at 3 years of data or something. So you have to have that flexibility and it's based I think on the use case.

DR. KING: And I was just going

to comment on hospice care, end of life care. I'm just concerned that this definition we have here is a little bit too narrow.

I really like that historical definition a lot better. I think it covers a lot more.

MS. MARTIN ANDERSON: Let's jump into the -- thank you, that was all great input. And one thing I've learned is that the NQF staff is really good at taking that input and then giving us something else to react to.

So we're not going to try to write the definition here, although Francois, I would invite you to send in the one that you offered on the telephone to the NQF staff so they can also take a look at that and bring that back to us.

MR. DE BRANTES: On its way.

MS. MARTIN ANDERSON: Let's jump into this first question on the purpose and

function of an episode grouper. For those on the phone we're looking at slide 20. Mark?

DR. LEVINE: You do it in the singular. Is there a purpose and a function.

MS. MARTIN ANDERSON: Purposes and functions, yes.

DR. LEVINE: I guess the real question is how many purposes and how many functions.

MS. MARTIN ANDERSON: Let's figure out what this group at least thinks the range is.

MR. REDFEARN: I have one at least to add. One of the things that I have done and we had done for some of the ACO and patient-centered medical home pilots that we've been working on, we have used episode of care models in which you assign a physician to the episode to link patients to the physicians.

Rather than using the Dartmouth method of just sort of counting PCP direct interventions, we actually put them all around and said well, what episodes are being managed for that patient, what physicians are managing those episodes, and then linking the patient back to a physician so that you can assign the patient to the physician and the ACO. So that's one thing that's not mentioned here.

We didn't get very far with the methodology. Everybody's kind of gone back to the Dartmouth methodology because that's the default, but I think that's a very interesting and possibly productive use of the groupers.

DR. CACCHIONE: Attribution. So you say you didn't get very far using it as an attribution tool?

MR. REDFEARN: Well, we actually did it in California and got about a year into it until the company decided that no,

we're going to go with the default methodology of Dartmouth.

Interestingly the medical groups we were dealing with in California which was Healthcare Partners and Monarch, very, very large medical group, so large that they're almost like insurance companies, didn't have a problem with the methodology.

In fact, they kind of liked it because they thought it did a better job of actually identifying what physicians are actually managing the care for the patients. Because they didn't want to have people mis-assigned, thrown into the group that they have to figure out how to deal with. They wanted to know who was actually seeing their physicians already in the medical group. So they were very happy with it.

But our network folks kind of said well, that's not the way the industry is going so we're going to default back to the Dartmouth methodology of just looking at

the PCPs which I had a lot of problem with. I didn't really like that methodology. I think it was over-simplistic.

So it's these operational things that happen that kind of go in a different direction. But I think the method worked fine.

DR. BANDEIAN: Episode is something that we're sort of focusing in on as sort of the starting point.

I think actually episodes are a means to an end. They are not an end of themselves.

What I mean by that is the purpose of all of this is to try to accurately understand the efficiency of care.

So, for example, if one just uses a hospital admission as a unit of analysis and does not take into account what happens after the hospital discharge one might have a misleading impression as to the efficiency

of the care. I mean, if the patient is rehospitalized 3 days later or what have you.

So, the reason why episodes are a useful approach is because care is often provided over a period of time. And to have a complete picture of the care one has to look over a period of time and link things together.

So, to me the -- I mean, maybe not the very top-level principle, but the second to the top level would be to have a valid basis for comparing resource use. And that's what we're trying to accomplish.

It's not necessarily -- and episodes are a means to that end. But the goal I should imagine is not to have a perfect episode grouper, but rather to have a valid measurement of resource use and episodes are a means to that end I would submit.

MR. LOISELLE: And this is Jim

Loiselle. Just to add to that.

From all of these definitions I think the important concept that needs to be built in is the concept of variation, is whether it's warranted or unwarranted. That's really what you're using these for at some level. That term "variation" needs to be built into these definitions.

DR. LEVINE: While variation and efficiency of care are important attributes and outcomes of groupers that doesn't address what is the purpose.

The purpose of the grouper, you know, there are basically two things. One is judging providers and helping to score their efficiency and use it for tiering, or for value-based purchasing, or for improvement purposes, or whatever have you.

Another is these are also applied for populations of patients for purposes of bundling and payment and other things.

And perhaps in David's use case

where you have different people in the same company some of whom thought it was good and others who didn't, maybe they were looking at it from different purposes. Looking at perhaps even the same attributes of variation efficiency, you know, capture, et cetera.

So I wonder whether or not we wouldn't be wise just to continue the discussion about what are the use cases. What are people using groupers for? What is their intent? What are people hoping to get from the output of groupers?

MR. DE BRANTES: This is Francois. To Mark's point, in addition to broadly speaking network management, network design which includes efficiency measurement, and tiering, and so on and so forth, and payment there's a third use that I can think of which is calculating the price of an episode for public transparency purposes. And that's a function that's

being increasingly done in states around the country in response to the lack of transparency on price information.

DR. CACCHIONE: We have used it for standardization of care as well, and for quality purposes. Using -- understanding what's in an episode and understanding that variability to the point earlier, understanding that variability to help reduce that variability and used to prescribe care paths and things like that. So, I think there is a quality purpose to these as well.

DR. BANDEIAN: Following up on Mark's comment and also I think what I heard Francois say.

I think the two basic use cases are measurement on the one hand of efficiency or at least the cost of the care of a particular set of clinical problems, and the other use case would be a bundle payment. Maybe there's a third or fourth,

but to me those are the two principal use cases that I'm aware of. One is a measurement purpose and the other is a pre-payment purpose or definition of a payment.

MS. MARTIN ANDERSON: I think we've heard four use cases so far, at least articulated. And we can see if there are more to add to the list.

One, the most frequently cited so far is a measurement of resource use, or comparison of clinicians around resource use, or something that's around that to that effect.

There's been an example of not just resource use but quality, right? Is this series of treatments effective in the quality of care which, Joe, is I think what you were getting at.

We've also heard that payment, right, a mechanism to bundle payment.

And then I think a fourth one that is always important to keep in mind is

that it's being also used operationally within in this example health plans to try to tackle other operational challenges like attribution.

And I'm sure there are perhaps if we had a broader stakeholder group there would be other examples of operational uses where it's a convenient way to solve a problem that's related to -- and it's not a problem that can be solved with looking at a single incident or a single form of care. David?

MR. HOPKINS: I think you missed Francois' suggestion which is actually a very good one.

More and more we're talking about price transparency these days. And very seldom do I hear people notice that what really matters to the consumer who's got skin in the game now is what's it going to cost me to go through in fact an episode of care.

Rather, we talk about oh, the hospital is going to cost you this, and the pathologist will charge that. That's not the answer the consumer is looking for. It's got to be built around episode.

MS. MARTIN ANDERSON: Right, the consumer-driven one.

MR. MACURDY: A classification that commonly gets used is whether an episode is patient-centric or provider-centric which I actually think is a good categorization to use to kind of organize these uses.

MS. MARTIN ANDERSON: And Tom, does that map to specific use cases as well? I mean, I can map them in my head, but are there other use cases that we're missing in either the patient-centric or the, those of you that are providers, a provider-centric view of what's valuable in getting an episode?

MR. MACURDY: I think between the

notion of a patient-centric and provider-centric almost everything goes under those classifications because in any case you're doing it from the perspective of providers in terms of the care they're providing. And you can do cost, you know, resource use, quality, et cetera.

And then the other one is from the patient perspective which would hit David's point on cost transparency. But it's kind of then reorganized from the patient's perspective irrespective of what provider they're getting, what is the kind of sequence of care, cost of care, quality of care, et cetera. So I think almost everything can be categorized under those two categories.

I mean, everything you mentioned are uses but there was large overlap in what you discussed.

DR. LEVINE: I just wanted to add a nuance to the efficiency use because I

think there's two things that -- there's two different use cases.

One is actual use of resources in which case you did want some kind of standardized pricing methodology for that so you could compare physicians or health systems equitably.

And the other one is use of different priced services. So, if you have -- I think that's how United uses the ETGs to tier their Premier Network. They looked at use of more expensive specialists I think in the episodes. So I think there's those two nuances.

MR. LOISELLE: When you say population or person you're really talking disease management or something less broad than that?

DR. CACCHIONE: Are you addressing that to the last speaker?

MR. LOISELLE: The one before that. I didn't get a chance to interject.

When you talk about population management or person, the person as opposed to the provider, the physician. Is it really the disease management application, or is it a cost management application? Where at the person level does this become relevant?

MR. MACURDY: Actually, I guess I'm still not very clear on your question. So, you're saying --

MR. LOISELLE: The application of the use case. Is the use case evaluating the cost or the disease management which is above and beyond the cost. Obviously it's an outcome question. How far do you take this in evaluating the performance at a person level?

MR. MACURDY: Well, I mean all of the above. There are instances where if you take a look at, say, the quality measures or kind of how they're evolving one is a measure of cost. Another one is a measure of re-hospitalization, you know, various

kinds of healthcare sequences.

So, I mean I think they're pretty broad in terms of the way they're done. And I didn't mean to restrict it to either cost or just purely efficiency. It's just kind of outcomes.

And outcomes I kind of view as all those combined. I mean is just a way of aggregating across a variety of outcomes. But if you --

MR. LOISELLE: Yes, exactly. Cost is just one --

MR. MACURDY: Yes, sure, but that's all I meant.

MR. LOISELLE: Okay, just a clarification.

MR. MACURDY: Yes, the notion is, you know, rehospitalization from a provider's perspective -- I mean, a patient can have a rehospitalization and from one provider's perspective it may be the case that they aren't very accountable for it,

another provider is. But from the patient's perspective they had one, so.

So when I meant provider-centric, I mean that -- it's a different perspective for each set of providers.

MR. MCLEAN: I was going back to David's point about the patient and the value. One of the ways we use, or we're looking at using groupers is understanding for us it's members but patients to understand the value they're going to get.

And then I would argue it's not just cost. We're looking at the cost and the balance between cost and quality. Because what good is buying a service if you don't get your outcome that you desire.

So, trying to figure out groupers, how to use a grouper to look in that holistic view of what is the most effective in that sense outcome for a patient. And also evaluating providers by doing that and saying this provider is good

at providing the holistic view of care for best value for a patient.

So yes, I just -- my key point is just I would argue that value is not just. Efficiency is not just cost. It is some sort of relationship between cost and the outcome or the end result.

MS. HOBART: I think in terms of the use cases there's really two dimensions which weave together a bunch of the things we've been talking about.

So one is what are you trying to find out or accomplish, like efficiency versus quality. And the second is pragmatically, be it WellPoint or whoever, how are you going to operationalize that and do it. And that's where you tend to get into sort of pragmatically being driven in terms of the time period that you can look at.

So to me naturally the episode of care, it's a fuzzy line when it starts or

stops. So you can go from a day to a DRG to the events around the procedure to in the end, you know, life is an event or an episode. And you're getting into population health, and they're all really episodes. So to me you're just going to have to kind of make a somewhat arbitrary decision about how you're defining the episode for a particular use case. It's not going to naturally define itself.

MS. MARTIN ANDERSON: Thank you. I'll give the last two words here and then we want to try to move onto the third question.

DR. LEVINE: I see us coming to sort of a categorization in a way all based upon a series of P's. I think there are two basic categories: payment and performance.

And within each of those there are a series of P's too. For payment purposes it's a population approach and it needs to be a patient-centered mechanism in

order to evolve the episodes.

Whereas for performance it needs a provider-centric approach. If you're really looking at how well does a provider perform you need a whole different approach to the construction of episodes than you would have otherwise. So, is the series of five P's properly arranged?

MR. BODYCOMBE: I wonder if, since we've talked about quality, we've talked about cost, if there's not you might consider a bundle of quality measures and cost measures that might be associated with an episode, and would those be considered something that NQF, for instance, might wish to approve on their own.

And that way you kind of mix and match different episode groupers. Well, I like the way this grouper handles this particular episode, but for this other kind of episode I prefer the other kind. You know, I hate to complicate your work but

that could happen.

MS. MARTIN ANDERSON: I think we're going to get to the how do you go about getting to endorsement from all this. So let's hold that thought.

I think we want to move on now to the third question on how an episode grouper differs from a case mix or risk adjuster.

I don't know that we need to focus too much on that last after the comma "or other measurement systems." I think the real question here is how is an episode grouper different from the other types of constructs that are currently evaluated.

MR. JONES: I think the biggest difference is -- I just view them in two broad categories, a member or total cost of care based versus provider-centric, episodic. So, tools like CRGs, ACGs, population-based. You know, your total cost of care, your illness burden, your risk score. Whereas the episodic is completely

different in that there's a trigger event, there's the clean period, there's all that focus on how that particular disease or episode of care was actually managed as opposed to the total burden of the member. Does that make sense?

MR. TOMPKINS: Part of my -- I'm just pointing out that there are actually two different words here. It might be equivocating or semantics.

In the first part of the question it's an episode grouper. And then the sentence ends with "systems." Because there is such a thing as an episode system. And an episode system I would argue properly configured or fully configured would include a case mix adjuster and maybe some other features to it. But, all right, I'll come back to that.

The grouper itself is a portion of the function of the entire episode system where you're trying to make logical

decisions about which disparate data elements ought to be, quote unquote, "grouped" to become clinically meaningful or otherwise actionable.

And then the episode system takes advantage of that, or uses that as a basic engine, but then does some other things too. For example, in that clinical context what is average performance or, quote unquote, what is "expected." And as soon as you start to say well, what's average performance or what's expected, controlling for what? That's the case mix adjuster.

So I think a fully episode system would incorporate the risk adjustment into it because you want, in my book, all analytical comparisons should be actual to expected.

MS. MARTIN ANDERSON: In that construct that you just laid out, right, you have a grouper which is a sequence of how we're putting together care. And we'll get

into -- or disease over time.

And then you have, I presume, measures that you create around a grouper that are then risk-adjusted. Is there such a thing as an episode that's risk-adjusted, or is it really just the metrics that are, you know, built on top of the grouper that actually get risk-adjusted or in whatever way?

DR. CACCHIONE: I think there is risk adjusting to an episode. I mean, I think that was your question.

MS. MARTIN ANDERSON: Is there a risk adjustment applied actually to the episode, or is it to the elements of the care that you're evaluating within the episode? So, for instance, cost, or quality, or any other type of outcome.

MR. REDFEARN: When I saw the section on risk adjustment I did kind of a double take.

Because my working assumption is

that an episode of care model should produce a clinically homogenous group of patients. So the fundamental episode methodology should have built into it some sort of thing that you could call risk adjustment, or some adjustment. So it should produce a clinically homogenous group.

But the odd thing is from a practical point of view that's not the way a lot of the groupers work.

For example, ETGs now for some episodes will have a layer on top of it and will generate up to four levels of risk that is layered on top of the grouper.

The MEGs which is sort of designed from a disease progression, that's the basic underlying model, they will still tell a lot of people using the models well, go license the DxCG risk model and put patient risk on top of the MEGs that you already have.

So, from a logical point of view

I think all episode modules should produce clinically homogenous groups of patients. From a practical point of view typically you have another process that's layered on top of it because the underlying model I guess doesn't produce a homogenous enough group. So, they sort of say well, we didn't quite get there so here, use this too. So it's kind of an odd dichotomy.

But from a theoretical point of view they should -- the groupers themselves should produce a homogenous group of patients I think.

MS. MARTIN ANDERSON: Do others agree with that?

DR. BANDEIAN: I understand exactly what you're saying. There are lots of ways to skin a cat. And I'm not necessarily sure that one necessarily wants to get into the, at least at this initial stage of the discussion, whether it's okay to do kind of risk adjustment after you've

constructed your episodes. I personally think it is and we could have a longer discussion about that.

I'm sorry, I kind of want to go back to a very high level. To me, in terms of, for example, how is it different from DRGs it's again the episodic nature of things.

But why do we care about episodic nature? The only reason why we're doing this I think in terms of the measurement. And I think the pre-bundled payment is a completely different kettle of fish and I would actually advocate perhaps we don't talk about that and focus on measurement because we already have a huge set of things to talk about.

So, basically from a measurement perspective the systems however they're constructed are intended to lead to a result, a conclusion that says Dr. ABC or Healthcare System XYZ is providing care of

some acceptable standard at a lower cost than other people are. So that's a good thing. So we want to reward them somehow.

So the question really is is that a true statement. Or if we have a whole bunch of methodologists sitting around the table and we poke holes in the methodology and say it didn't take into account this, this, this and this, and basically you can't draw any conclusions from the methodology, do you have confidence that when the system says episode cost is 1.3 times benchmark that that's a reasonable statement to make?

So I think that largely it's a question of what do you need to do to have validity so that when you look at the output you have confidence that it's -- that's true, that you can actually hang your hat on it, that it is therefore reasonable for Medicare to give a bump up in payment under the value modifier, that it's reasonable for a health plan to kick out a doctor because

they have a high episode score. So, that's sort of one use case. It's a measurement use case.

Now, there's a different use case which is still measurement which is now we're a group practice, or an ACO, or what have you. How can we improve what we're doing and what can we do within our ACO or whatever to get a better high-level aggregate score.

So, to me the principal use case that we should be focusing on here are measurement. And there are two types of measurement. The sort of external to the providers saying you're doing a good job or bad job, and then within the provider community what can we do to get a better score.

And in both cases what's -- a critical issue is are the conclusions correct. Or if we had a whole bunch of methodologists sitting around and throwing

rocks at the methodology and looking at statistical outputs it would be obvious of course you can't draw a conclusion on this.

DR. CACCHIONE: In terms of specified level of quality --

DR. BANDEIAN: Yes. And the only reason why I phrase it that way is because we still have got a long ways to go on the quality space. And so kind of short-term we have to a little bit fudge on the quality.

Because obviously if XYZ --

DR. CACCHIONE: I know what you mean.

DR. BANDEIAN: -- it's pretty complicated. And maybe we should try to take that on.

But obviously if XYZ Healthcare System or ACO or practice is lower-cost and also very low quality then that's not a very good situation either.

But these things are being used to draw conclusions to reward under a value

modifier, or alternatively within an ACO or an organized delivery system of some sort as a guide to what they can do to improve their performance.

MS. MARTIN ANDERSON: I want to give just one more piece of clarity here which is that while we're talking about uses because it's important to understand the context of all these high-stake uses as we're coming up with the criteria, NQF has not historically endorsed specific measures or whatever for a specific use.

So I think what we need to keep in mind, not saying that would never happen, but I think what we need to keep in mind is these uses are context for the rest of our work on what are some of the criteria.

And it would be the measure developer would state what their intended use is and under what context they've developed such a thing.

The criteria need to support the

evaluation by the committee of whether or not it's acceptable in terms of whatever criteria we get through next.

And then users will do what users do with things after they're endorsed and out there. And so I don't think we need to -- I'm saying that because I don't think we need to restrict that we're only going to think about one kind of use.

I think it's important for us to understand there are broad and high-stake uses, and keep that in mind in the context of our work. Rather than thinking that we need to come up with criteria for one type of use or another type of use at this stage. So I think it's a good idea to keep it broad and that gives us context.

So we have I think a couple of more tents up.

MR. JONES: I just wanted to echo a couple of things that I heard and underscore a point that Dave made earlier in

that when you're using these tools to, quote, "gauge efficiency" if there's not that flexibility to level for your differences which are oftentimes very large in how you're contracting with providers, you know, you really need to have that flexibility in there where you're solving that price equation and you're teasing out differences due to mix and volume.

I don't know if we want to add that to criteria anywhere, if anybody agrees or disagrees with that.

DR. CACCHIONE: Are you getting at this idea that you assume a standardized cost model?

MR. JONES: Yes.

DR. CACCHIONE: So that we don't get corrupted by contractual relationships between whoever the purchaser is and the provider.

MR. JONES: Yes, I think that's key. And that was a cause of a lot of

challenges that we had.

DR. CACCHIONE: Okay. Jelani? Jelani, move a little closer to the mike because you sort of drop off.

MR. MCLEAN: Sorry, I just don't talk that loud. I'll make sure I stay closer.

One of the things I think we're missing, we're overlooking is groupers don't -- historically groupers don't evaluate providers. The analytics you put around it evaluates the provider. So therefore while I agree with case mixing which is more around the provider and the mix the provider actually has and the -- and the risk adjustment portion I'll get to in a second.

But I don't think you can case mix from a provider standpoint within the grouper because the grouper is focused on the patient, the population that they're using in it.

And so you would essentially if

you're using claims have to match the claims to the provider. And then it's a whole nother algorithm that you're going to apply to it. There would be another requirement and criteria you would have to put in your evaluation of the grouper and its effectiveness, and how good it really is.

And I don't think it's something you want to go down that path, having the experience with trying to match claims to providers is a challenge in itself.

To the risk adjustment portion, risk adjustment from my experience is all about the data and the population that you're trying to evaluate.

Trying to do that within a grouper while I agree is useful, I agree with Steve it's probably more practical to do that after or before you've put the data within the group inside the grouper because of the fact that it's all about the data that you're having.

The population, for example, transplants. Everyone is risky. But there is a large variation of risk within the population receiving the transplant. But how do you do that as opposed to evaluating a cardiac care facility, or cardiac population? It's totally different. So therefore I would argue that that may be a bit of a challenge and a bit extreme with trying to evaluate a grouper.

DR. CACCHIONE: One last comment before the break because we're over the break. So Tom, if you could give us a zinger for the last comment before the break.

MR. MACURDY: I don't know if I want to do that. I just wanted to note I don't think there's a sharp distinction between the risk adjustment case mix adjuster and the grouper. I think the best way to do it is to illustrate it.

I kind of see a continuum between

a bundler and a grouper. If the services you're looking at are provided by one provider you call it a bundle, and if it's across providers you call it a grouper.

And to illustrate the difference here is -- I mean, DRGs are a good example, either MS DRGs or APC DRGs, et cetera. There the risk adjustment is partly involved in the bundling because you essentially are moving the DRG around depending upon the risk characteristics of the patient.

Another way to really get at this issue is if you take a 3M bundler or grouper versus, say, an ETG MEG grouper the main distinction between those two is that the 3M combines its risk adjustment with the grouping. So it'll take all expenses after a certain period of time irrespective of what the circumstances are and try to fix it all with the risk adjustment.

Whereas the ETG grouper and the MEG grouper try to select a particular set

of those services in those categories and then do the risk adjustment separately.

One other point I wanted to make is it is true that it would be nice to have groupers have homogenous patients. And you get one patient per item. And that's the challenge.

I mean, the difficulty is if you take the MEG grouper you can get up to 1,800 categories, and if you take the ETG you get about the same order and you get hardly any individuals per group and then you can't do any benchmarking. So that's always the challenge. You're always going to have a heterogenous set of patients. You're going to have to be able to do some kind of adjustments.

DR. CACCHIONE: For those of us who treat patients we know that it's a -- who we treat is heterogenous.

(Laughter)

MR. MACURDY: That's the reason.

It's because -- yes, you can get them homogenous, it's just then you get really small cells and you can't do very much. So everybody is their own special case.

MS. MARTIN ANDERSON: Thank you. I think we're going to take our break. Just to give you a closing comment to think about during the break is that keep in mind that our objective here isn't to push the science in a certain direction, it's to acknowledge the state of the science and come up with some criteria that can live within where we are.

So I think it's important that we -- as we continue through the criteria we keep that in mind. We're not trying to create criteria to box developers in. We're really trying to figure out how given the state of this business could we assist NQF in evaluating how to do endorsement for where we are today. And it might look different in the future. So, let's take our

break.

(Whereupon, the foregoing matter went off the record at 10:24 a.m. and went back on the record at 10:44 a.m.)

DR. CACCHIONE: Ashlie, do you want to go ahead and get started with the criteria?

MS. WILBON: Sure. There's a few people missing. We'll go ahead and get started without them.

So this next portion of the discussion is designed to give you guys an overview of our existing criteria that we use to evaluate kind of standalone cost and resource use measures in our consensus development process.

And this section is really focused around kind of giving you some broad protocols for how our criteria is applied and then walking through each of the four or five criteria.

And then we'll have a discussion

about which of those criteria we think could be applied to episode groupers and if there are criteria that are missing or ones that don't apply, do apply we'll have that discussion. So I'll just kind of go through all the criteria and then we'll open it up for discussion if that's okay.

So, there are essentially four kind of core criteria that we use to evaluate resource use measures: importance to measure and report, scientific acceptability of measure properties, feasibility and usability and use.

There is a fifth criteria that we'll talk about that is applied only when we have identified when there are measures that have similar specifications and we have identified them as similar or competing. And we'll talk a little bit about that as well.

So, some of the key principles that guide the application of the criteria.

So, within the four kind of major criteria that we described there's two must-pass criteria. And the criteria are applied in a hierarchical manner so they're in a specific order.

And the measures have to meet the importance to measure and report criteria in order to move onto the next criteria which would be scientific acceptability of measure properties. And once they pass that criteria they move onto the other two.

If they don't pass these two criteria the measure doesn't -- the remaining criteria aren't applied and the measure cannot be recommended for endorsement.

Within each of the four overall criteria there's a series of subcriteria which really are used to provide the additional detail. So, how do you know if the scientific acceptability -- if the measure is scientifically acceptable. How

do you know if the measure is important. And so there's a series of subcriteria within each of those major criteria that we'll discuss in some detail.

Also, the criteria that were kind of originated or out of the quality measurement side were really designed to parallel best practice for measurement development.

So, some of the way that we structured this discussion with this group is to kind of think about some of the key principles that should be applied when developing episode groupers and identifying which criteria might be applied to kind of parallel those key principles or considerations so that there's some alignment of those ideas.

And generally the application of the criteria require both evidence and expert judgment. So, not everything is black and white. There's usually a matter

of degree in judging whether or not a criteria has been met.

And generally all the criteria are rated as we go through, walk these through with our committees we ask them to rate the overall criteria and some of the subcriteria on a scale of high, moderate and low and insufficient, and then at the end make an overall recommendation depending on the criteria that have been met throughout the evaluation process.

So, actually we'll just pause here for one second. Some of the questions that we'll be asking you guys to address, and we'll come back to these after we kind of walk through each of the four criteria, is whether or not these criteria can be applied to episode groupers. Of the major criteria that apply how might the subcriteria also apply to groupers.

And then trying to find out whether or not there are other major or

subcriteria that should be considered that aren't listed here that we haven't captured already in some of our existing framework for thinking about other types of measures.

So, the first criteria, and I'm on page 5 of the discussion guide. What we have on the slides is kind of a summary of what's in the discussion guide. So if you want some additional detail you can kind of read along as I go.

So, the importance to measure and report criteria is used to determine if the measure focus or the topic is important in making significant contributions towards understanding healthcare costs for a specific high-impact aspect of healthcare.

So, for example, is it important to measure the cost of hip and knee replacements in an over-65 population. So it's really the topic itself and whether or not it's important to measure in the context that the developer is suggesting.

And then to determine whether or not there's variation or demonstrated high-impact aspect of healthcare or overall poor performance.

So, in the submissions we really are asking the developers is this an area of healthcare that we know there's a lot of variation already that this measure is going to help illuminate or help us better understand that variation or poor performance in that area? So we're really just trying to understand the need for measuring this topic with this particular measure for this population, et cetera.

So the subcriteria are really focused on having the developers identify which major national health goal or priority that this measure would help to address, that there is a demonstrated resource use or cost problem and an opportunity for improvement.

And we're also asking them to

explain the intent of the measure and the types of costs and ensure that the types of costs they're capturing are actually consistent with the intent of the measure and that those costs are important to measure for that particular topic area.

So, for scientific acceptability this criteria is focused on determining the extent to which the measure is reliable and valid, and produces consistent and credible results about the cost of resources used to deliver care.

Again, the two main components of this criterion are reliability and validity. And within the reliability criteria there's two additional kind of micro-criteria if you will that look at the preciseness of the specifications and whether or not they can be used to reproduce or facilitate consistent implementation of the measure.

And then that there are testing results submitted that demonstrate the

results are repeatable. So that's generally some statistical analysis of the measure results or the data elements.

MR. AMIN: So, I just wanted to reiterate here that when we're talking about the preciseness of the specifications that will be the specifications in what we're describing will be the module discussion that will -- which will be the next agenda item.

So that would consist of the clinical logic, construction logic and adjustments for comparability broadly at this point unless we decide that there are other specifications that we would need to evaluate as part of an episode grouper.

MS. WILBON: So, the next major subcriteria within scientific acceptability focused on the validity of the measure.

And there's several bullets here. I guess I'll go ahead and read through them just to make sure we're all on the same

page.

That the measure specifications are consistent with the measure intent. So are they actually measuring what they said that they are intending to measure with the measure results.

That the validity testing demonstrates that the measure data elements are correct and the measure score accurately reflects the cost of care.

That exclusions are supported with clinical evidence or a rationale or analysis of those exclusions. That the exclusions are transparent. That the evidence that exclusions are applied due to patient preference are also disclosed.

That an evidence-based risk adjustment strategy if it is applied that it's based on patient clinical factors that influence the measured outcome.

That there's adequate discrimination and calibration of the risk

model or rationale to support why they have not chosen to use a risk adjustment method.

That the scoring and analysis of the measure produces statistically significant and practically and clinically meaningful differences in performance.

If they have chosen to use multiple data sources we ask them to demonstrate through their analysis that the results are comparable between those two data sources.

Generally we don't run into this issue as much with the cost measures because they tend to all be specified using admin claims data so that tends to be a moot point. But it is part of the kind of framework for the criteria.

MR. AMIN: Ashlie, let me just point out though while it's not necessarily from a data source perspective, there are -- the bar here is that if there are multiple methods meaning other multiple risk

adjustment methodologies or multiple costing approaches there that only one is specified so that it can actually produce comparable results for an individual provider.

If they're using one particular measure we can't have the same measure having both a standardized pricing approach and an actual prices paid approach because those two obviously wouldn't be comparable even though they're using the same NQF measure number.

So it does come up in other ways in terms of episode groupers that we should consider because we are looking for precise specifications and not really looking for potentially additional variation or flexibility that are typically designed in these types of products.

MS. WILBON: Thanks, Taroon, that's a really important point.

The last kind of micro-criteria if you will within the validity subcriteria

are around disparities. And that if there are disparities in care that have been identified for this particular topic area or measure focus that the measure actually allows for the identification of those disparities through some mechanism, stratification or what have you.

MR. AMIN: So just one other point of clarification that I want to just make here is that when we're talking about reliability and validity testing we offer the opportunity to do that at the data element level or the performance measure score level.

But when we talk about validity testing here we're really -- it could be at any one of those levels. And when we're looking at testing that also includes testing of the risk adjustment model.

So, those are two different components that we would be looking at. And again, that would potentially translate

potentially translate to how we're looking at testing of episode groupers. And so we'll explore that in more detail later on in the discussion.

MS. WILBON: Thanks. The third criteria is around feasibility. And the goal of this criteria has been to assess the extent to which the required data elements are readily available and can be captured without undue burden and implemented for performance measurement.

So, the subcriteria for this major criterion focus around whether or not the required data elements are routinely generated through the delivery of care, that the data elements are available in electronic sources and that a data collection strategy can be implemented without undue burden.

And this criterion has also in our resource use work tends to include an assessment of any cost or financial burden

to implement the measure. So any measures that require some type of purchase of risk adjustment software or licensing or anything to be able to run the measure, that that is taken into consideration in terms of the evaluation process as well.

The fourth criterion is around usability and use. And the goal of this criterion is to assess the extent to which potential audiences which encompass kind of our stakeholder and membership councils, so the consumers, purchasers, providers, policymakers and others are using or could use the performance results for both accountability and performance improvement.

And the subcriteria are focused around the developer explaining or demonstrating how the measure is currently used, or how they expect the measure will be used. So how they plan for it to be used. And a public reporting or accountability application.

That the measure -- if it's already in use we're asking them to show data that demonstrates that there is some type of improvement or, you know, understanding of cost and resource performance over time. And that any benefits of the measure outweigh any unintended consequences. So, asking them to think about if there are any unintended consequences of the measure that they've thought those through and that weighing those between the positives and the negatives of the measure, that the benefits outweigh those unintended consequences.

And then the last one which seemed to come up already in some discussion is around whether or not the measure can be deconstructed to facilitate transparency and understanding. So, based on provider or clinician receiving a measure score, can he or she go back into the measure and figure out exactly what that score represents and

what they're actually being measured on.

MR. AMIN: So Ashlie, before you move on on this criteria I just want to reiterate something that Kristine said and I think will translate to multiple different components over the next two days.

So, NQF's current criteria requires essentially reliability and validity for the measure to be used for both accountability and performance improvement applications. It does not draw distinctions between accountability applications, meaning between public reporting or payment applications.

And so the idea here is that the criteria should apply broadly for all applications.

Now, we've had a lengthy discussion around use and we will have additional conversations around use as we go through each of the modules. So if there is a belief that depending on which

accountability application the grouper potentially is intending to be used for that if the criteria do need to be different that needs to be really clearly laid out.

Because as current standard at NQF there is no -- you use the same criteria and think about it broadly for accountability applications and performance improvement. So, I just wanted to kind of reiterate that.

And especially also this last subcriteria around transparency and understanding was also another key subcriteria as we were looking at measures that were a result of episode groupers in our first evaluation of cost and resource use measures. So again, that would be another logical subcriteria that might require more exploration as we move forward.

MS. WILBON: So, the last criteria. Again, and this is one -- criterion. And this is one that again we

generally only apply if the four previous criteria have been met and if the measure or measures that are under review have been identified as being related or competing with other measures under review or currently in the portfolio of endorsed measures.

And for resource use measures the way that we've kind of conceptualized this, taking into consideration that there are different components to cost measures that we may want to consider in this analysis of related and competing, that we take into consideration whether or not it's been a per-episode or per capita measure, whether or not they're applying the same types of costing methodology, so actual prices paid versus standardized prices, the types of costs that are being measured, and the actual population that's being addressed within the measure.

So is it an all-population

measure, total cost. Is it focused around a specific disease condition like diabetes or cardiovascular disease.

And a measure that we would call competing would actually share all of these same characteristics. And we have generally done some analysis with the committee to determine whether or not both measures are needed, or is there some justification for having both measures endorsed at the same time considering they are similar in many aspects.

We can have some discussion. I think this issue has come up already I think with Chris' question in the beginning about whether or not we would want to potentially endorse multiple groupers, or is there a best in class if you will.

So with that we can kind of go back to the questions that we're asking the group to consider and have the co-chairs take it away.

MS. MARTIN ANDERSON: So I think the thing that I'm struggling with in reading through these criteria is that the episode grouping is really, it's unit of analysis. It's a building block for measures.

And so it's a way of taking different kinds of healthcare services and putting them together into often clinically meaningful groups.

But from that then you use that to build measures. You might have a disease-specific measure of provider efficiency, or you might have a cost measure, or a resource use measure.

So, this is all geared toward individual measures and that's why I think we're going to have to really change the language in order to make it work for evaluating what is really a building block towards measurement.

MR. TOMPKINS: I have comment

that's similar to that, namely that in a sense an episode grouper is like a sausage maker, or it's actually maybe a sausage maker in reverse, right, because it takes the scrambled and unscrambles.

And I could deliver you an episode grouper and say, you know, if you run this properly it will produce 600 different measures for you.

And then you could convene panels around -- 600 panels if you want around each measure and probably apply many of these criteria to each one at a time. Is this important? Is it important to measure resource use for transplantation? Is it important to measure resource use for Band-aid placement? Some will be yes, some will be no. So the importance question is, you know, magnitude and so forth, or variation issues.

And then reliability the same way. Have I measured the transplantation

costs reliably so that you can repeat them, and that they have fairly narrow confidence intervals, for example. Again, that applies.

Feasibility as well seems duck soup here, especially with Ashlie's side comment that these often just rely on administrative data.

I think when the trip-up comes in is in the validity and the usability. Because if you were going to -- in my mental thought experiment there of giving you a grouper and saying you can evaluate 600 measures, the validity question comes in. Because sometimes the groupers in that sorting-out process, in that parsing-out process are using presumably consistent logic for doing so.

And therefore there's an efficiency to you examining the logic by which that is done which would cross the 600 measures to a large degree as opposed to

each one at a time.

But my short comment is that most of these criteria apply because you're looking at the end use, the measures themselves, the reliability, the importance and even the validity. But it's nice to have an engine or a grouper that systematically gives you logic that you can review in advance which gives you a head start on the 600 measures that you would otherwise be evaluating.

MS. MARTIN ANDERSON: Since the first two comments took us here I just want to make sure we have just a little bit of discussion around this topic of is there value in evaluating a grouper as well as the outputs from a grouper that might be around measures themselves. Or do you just focus on the measures and by default you're looking at the grouper.

I see this as the reverse of the bundling -- I mean, the composite

discussion, where originally all measures needed to be endorsed and then a composite could be endorsed.

And at first it was all the measures inside the composite had to be also endorsed.

So the question is if you have a measure that's an episode-based measure does the grouper itself also have to be endorsed in order for the measure to be endorsed. So there's a relationship here.

So I just want to hear your thoughts on this concept of the grouper and also of measures, and how it might impact how we do our work this afternoon.

MR. LOISELLE: This is Jim Loiselle. I struggle with the context. I never thought of a grouper in the context of a measure. Obviously it populates outputs and the outputs are variable. So I think the focus should be on the grouper itself.

Because you can create measures

that are valuable that are not otherwise outputs of the grouper. It's whatever analytical process, or payment, or clinical process you add on top of it where measures become specific. That's my struggle with thinking of a grouper as a measure-based.

MS. MARTIN ANDERSON: Well no, I'm not saying that, it's just that there are current measures that are approved that are actually episode-based.

MR. LOISELLE: Yes.

MS. MARTIN ANDERSON: Right? Where you say, you know, there's total cost of care over an episode. So the question is do you agree that NQF should also look at the grouper itself as something to be endorsed, or are we looking at the outputs that are used in specific ways for endorsement.

MR. LOISELLE: That's I think a whole topic of discussion and an afternoon in and of itself, just going on the total

cost element. Cost is variable, and trying to differentiate performance from a payment perspective, or a clinical utilization, or care management perspective, what cost you use might be a different thing. It all depends what you want to use the grouper for.

Because we evaluate and look at cost. A tertiary hospital is less expensive than a primary inner city hospital attached with a medical school. So trying to use cost variations as a topic, that's not really an output of the grouper, that's the intelligence that you apply afterwards. The groupers themselves just might create that number but there's too much variation in even the term I think in just saying total cost.

MS. MARTIN ANDERSON: Even 30-day mortality, that is an episode that has been defined just for the purposes of calling mortality at 30 days. So let's continue.

MR. REDFEARN: My work has mostly been focused on using episodes of care as a foundation for provider cost efficiency profiling.

And there's a certain amount of variability when you do that across time. When you repeat the measures of the physicians. Physicians can change. They can move around a little bit about how efficient they look.

And I was curious about whether -- ETGs has been our default tool. I was curious about if you pull ETGs out, plug something else in and run the rest of the analysis in exactly the same way to see what kind of results you get. And I've done that. I did it using MEGs. I plugged MEGs in.

And the interesting outcome for my work is that essentially the results are very similar. I'm exaggerating but it's almost like I don't care how the groupers

carve things up. They carve them up into groups that make some sense and I use them for my analysis.

Now, it doesn't mean that they're exactly the same. Things move around. Doctors move around. But they don't move around much more than they move around across time when I use the ETGs. So, my argument there is it's the measure, it's not the grouper that you would want to really focus on.

I really am struggling with the idea of how you can evaluate a grouper. How could you look at ETGs and MEGs and say clinically I think one is better than the other. It makes more sense. It's all very specific to how you're using them and whether you can justify them.

Because I know there's a lot of really smart people that develop those two models and they did it -- they ended up doing it differently. And how can you say

that one is better than the other. It's just how you end up using it.

DR. CACCHIONE: I think it's the output that we're really looking here for. I mean, I think that to go that far up the chain, I mean I think people -- there are a lot of different groupers that are out there that work.

But we've had the same experience. We've evaluated two different grouper tools and have come up with very, very similar outcomes with as best we can tell different methodologies. So you can arrive at the same thing that is valid and has valid outputs. I don't know that we need to go that far up the chain to evaluate the individual tool. Jim?

MR. JONES: I agree and I do think they produce similar results. But in terms of looking at a grouper I think that we should consider having certain criteria that they must certify that they've done

properly.

An example of that is that when they're normalizing your data, when they're using their reference data sets that they should not be allowed to market to a Medicaid plan and allow that Medicaid plan, for example, to run the grouper based on a commercial reference set, for example. Those things I think should be disclosed. Because I have run into that problem before and I was quite shocked.

DR. CACCHIONE: But that's not a problem of the tool itself, that's the input that was used to --

MR. JONES: It's not the tool, it's just -- exactly.

DR. CACCHIONE: -- into the tool. So I think that is -- so I'm not sure it's the grouper itself, but it's the inputs in, the data inputs. Tom?

MR. JONES: More like rules of the road for groupers.

MR. MACURDY: So I guess I want to indicate that I disagree with this. And the reason why is I've looked at a lot of grouper output. They don't give the same answers.

And the difficulty is if you're using -- I'll use the specific example of using this for payment purposes. There is going to be the provider, say the physician who's going to ask why did I get stuck with that claim and you're going to have to have an answer to that question.

And the challenge is with a lot of the groupers, especially if you take a Medicare population where they're very complex kinds of cases. There's a lot of comorbidities, there's a lot of competition for where the claim could be assigned. And if you have a rule the claim goes to one and only one spot there's a back-end kind of logic that's kind of going on there which is somewhat of a black box.

Which I've spent a lot of time on groupers and you can't understand. The people who develop the grouper don't really understand how those rules get applied and why the claim went where it did.

So if you have a provider who got stuck with a $10,000 home health claim and you can't explain to them why they got that claim it may be an okay measure but it's going to be a real challenge.

Because at some point the grouper has to be actionable on the part of the provider so they can make that correction. If there's something in the back that nobody really understands why it got assigned, well, you've got your measure and that's fine.

Depends on the level you're looking at. If you're at a really high level maybe it doesn't matter. But I can tell you when you get down to the individual level and you're actually docking somebody's

pay based on this you're going to have to have a pretty clean explanation as to why.

MR. DE BRANTES: This is Francois. I'm going to build exactly on Tom's comments because whether you're using it for pure payment purposes, or whether you're using it for analytic performance evaluation purposes you have to -- the physicians are going to look at not just which claims were assigned but whether or not the condition or the episode that was triggered was in their estimation a valid episode.

So in other words, based on the criteria that the episode grouper is using to determine whether or not a patient has essential hypertension, or hypolipidemia or ischemic heart disease, or some other conditions, or had one procedure versus another procedure, is that even a valid determination.

So in other words, does that

actually match the clinical evidence of that patient based on the physician's medical records.

And if it doesn't, I mean if there's no matching whatsoever then it's a completely invalid output.

So, I think we're caught in a dilemma because -- and in full disclosure I had this conversation with NQF way over a year ago when they started this process of establishing criteria for measuring efficiency.

And at one point we were looking at whether or not we would bother applying for into that process. And the answer at the end was no because the language used to determine these criterion have nothing to do with a grouper.

And so I think it's the language that has been developed and the evaluation criterion used by NQF traditionally are for measures.

And here we're not talking about a measure. We're talking about a grouper which is a process to assemble claims into logical units of inference, or as logical as they can be. And then used for various purposes as Mark and Steve and others have mentioned.

And so if you're going to evaluate the logic of that grouper to determine whether or not it does have any resemblance to the medical reality of the patient it requires a different way of looking and evaluating and establishing criteria than I think NQF has done in the past.

So, if you ignore the grouper itself then you're asking people to submit for potentially 300 measures or 500 measures. Or maybe they decide oh, I'm just going to file for diabetes but how does that even make any sense.

So I think we're -- and again,

this is like an 18-month now, or 24-month discussion. And we need to resolve this today. Because otherwise I don't even know why we're on this call.

DR. CACCHIONE: Tom, I'm going to come back to you. I want to ask a question, then we'll go to Dave.

What do you consider the source or the benchmark -- what do you consider the source of truth when you compare grouper tools to grouper tools?

And what do you consider, as you have done the analytics on this what do you consider the source of truth and what are you establishing as the benchmark?

MR. MACURDY: I said you got different results. I don't know the source of truth.

I mean, ultimately what -- I mean, I don't think I'm the best person to judge on that sort of thing. Ultimately what you need are the clinicians you're

trying to provide information to and incentivize to provide better care as to whether it's something that's understandable to them and actionable and moves them in the direction that whoever happens to be paying them or supervising them wants them to go.

DR. CACCHIONE: The problem is that most of the providers are in an information void. Being on the front line they're in an information void.

MR. MACURDY: But that's part of the goal of the grouper is to give them better information. And that is doable, it's just right now the way the -- you know, it's difficult.

DR. CACCHIONE: Most of them can't spell "episode" right now or understand it.

(Laughter)

MR. MACURDY: Well, if it's not understandable to them you haven't been successful on doing it. It's going to be

very difficult to make a payment modifier based on that.

DR. CACCHIONE: David.

DR. MIRKIN: I may be getting ahead of the discussion but Tom was talking about attribution which is a very important issue obviously.

But I'm just wondering given all these experts around here is an attribution rule, even a rule set, essential to the definition of a -- essential part of an episode grouper? Or is there so many ways to slice and dice attribution is that an entirely different topic?

MR. MACURDY: I don't think it's a different topic. I mean, attribution, first of all it can be sliced and diced different ways.

The easiest way to see that is if you take -- just take concretely, say, a hospital admission. Well, depending on which kind of provider group you're

evaluating, be it the physician, be it the hospitalist, be it the anesthesiologist, what's relevant for them in terms of grouper may be different.

If it's provider-centric attribution is really fundamental. If it's patient-centric it's not so fundamental because patient-centric is the patient is looking and they may not care which providers are giving them services, how that sequence is put together.

But from a provider perspective I think it's absolutely essential. But it's also multiple ways. Is is -- completely. For one set of providers it's one, for another set of providers it's another. So there's no uniqueness there.

MS. GARRETT: So, I'm just building on what Francois said. I really agree that we're talking about evaluating building blocks, not a measure. And so I think we're going to have to really change

the criteria.

And so it would just be helpful to understand a little more of the context of why the group is convened. Is it because CMS is going to be required to bring their publicly available measure through the NQF endorsement process? Is that the reason? And if so we have to figure out how to do that, how to get ready for that. Or is that still in question.

And then are the commercial grouper companies going to -- is Optum going to bring ETGs forward for NQF endorsement. I mean, what's their rationale for doing that and would they even do it. So I have some questions about what are we doing here and why.

MS. MARTIN ANDERSON: And this issue has been around for awhile. Because the day that measures that were based on episodes start coming in for endorsement you're already dealing with, well, what do

you do with the grouper and how do you evaluate the underlying grouper. So I think it's been around for awhile.

I think what we'll do because as this panel was empaneled for, we're going to go forward saying you actually can. For the rest of this afternoon we're going to go forward and say you actually can evaluate a grouper. And we're going to try to figure out how you do that. Even if it means changing this criteria.

And just keep in mind a couple of things. One is how useful is it if when you define it everything passes. So everybody who has a grouper brings it in, they're all endorsed, we're in the same spot. You're still then looking at the measures.

And that's okay, that can be an outcome, but at least there's a vetting process.

And doing that while also not trying to shape the science through what

we're doing. So I think we have a needle to thread that I think will be hard. But I think we should now just jump to saying okay, we don't have to start with the NQF's measure process. It was a place to look at, to think about.

I think we've heard already that there is at least some interest in the components that might not look the same but be similar around the scientific acceptability on reliability and validity. And then I think there's been a focus on validity but both apply.

And then there also has been a number of comments on the feasibility elements and usability.

So, let's try to figure out should we eliminate any of those categories. Should we just say, hey you know what, I can't get my head around importance, I can't get my head around whatever.

And then say what else would we

add. If we're going to do breakout groups we have to decide what we're breaking out to talk about. So, let's start with at least what are these high-level categories that we think are important if you're actually going to endorse a grouper. Steve?

DR. BANDEIAN: Hi. I obviously missed most of the discussion. My daughter is on their way to the emergency room right now with paroxysmal atrial tachycardia but I can't do anything about it so I might as well come back to the conversation.

(Laughter)

DR. BANDEIAN: I think she's fine. The doctor said she was laughing and taking pictures of herself with the cell phone at the student health. And her mom is in South Africa. My wife, her mom, is in South Africa.

(Laughter)

DR. BANDEIAN: Anyway, so I obviously missed most of this.

The importance issue -- let me just -- that to me seems to me to be kind of an issue of use of NQF resources. Like, if an illness affects a -- well, as long as there are enough people to do statistics on the illness and if somebody can do a useful set of measures that are helpful to that illness even if it's pretty rare in the scheme of things I wouldn't know that that means that it shouldn't be done.

So I don't quite understand the purpose of importance other than prioritization of NQF resources. Is that really what the purpose of importance was?

MS. MARTIN ANDERSON: When we're looking at individual measures it's really is this important to measure. Does it have value in being endorsed and worth the resources.

I think the question is how does that apply to one grouper system versus another. Can you look at one and say is

this one important, is that one important. Or are they already fundamentally so similar in their --

DR. BANDEIAN: Yes. I guess, again, I missed almost all the discussion. To me it does seem as though one should look at the system to see how well the system works.

There needs to be some checking to make sure that the logic for individual conditions is okay as well. That may be a different level of scrutiny. But I could easily imagine that in a comprehensive system that addresses all conditions there would be some conditions that wouldn't make it onto anyone's top 100 list. Even there though.

So you might think about some alternative sort of lesser resource-intensive way of being able to pass judgment on -- you know, there's like probably, I don't know, depending upon how you slice and

dice there could easily be one to two thousand conditions in the entire universe.

MS. MARTIN ANDERSON: So let me ask another parallel question here to help get us focused again.

Is it possible or is it desirable to endorse a grouper, or are you talking about endorsing the episode construction for some subset? So, is it -- do you evaluate whether or not a grouper for AMI, or a grouper for diabetes, are you evaluating them all separately or could NQF steering committees actually evaluate the whole grouper itself? Can you create a list of sample conditions or something that would allow you to evaluate the entire grouper all at once?

MR. DE BRANTES: This is Francois. Just a couple of points.

So, on the importance you can boil it down to potentially how much of total spend does the grouper cover.

Because if, you know, you could theoretically have someone go in and submit for endorsement a grouper that only covers two conditions, or one specific set of procedures. That would be a very small percentage of total spend and specify Medicaid, Medicare, commercial.

Versus a grouper to Steve's point which is an entire system that covers everything in which case obviously it has a lot of importance because it's going to cover 80 percent, 90 percent of total cost of care.

So, all the rest of the questions become meaningless because if you're covering 80 percent of care of course you're covering -- it's important. Then you've got lots of variation and so on and so forth.

So that might be a way to cut to the chase at least from a -- is it important? Yes, if it covers a lot of -- a significant percentage of total cost of care

it's important.

And then to your other question about should NQF waste its time evaluating a grouper for one condition or two conditions. I mean, that's NQF's decision.

MS. MARTIN ANDERSON: Okay. Mark.

DR. LEVINE: I again think that we should bifurcate in the sense that if one is looking at a grouper whose intent is to inform upon payment, spending, financial, population performance then, for instance, looking at importance you'd look at percentage of spend and what it is that you're -- are you coming up with valid measures that reflect your intent.

But if the intent of the grouper is to inform upon practice and performance then you might be able to look at importance might be the amount of a physician's practice that you're actually able to address, et cetera. So again, separate

criteria for separate use cases would seem to be in order.

MS. SIMON: I would be concerned about framing it in the context -- or importance in the context of total spend because children will lose out. And I think that's pretty important to incorporate total population. If you're going to improve overall health.

MR. REDFEARN: Just a slightly different take on that.

In the conversation I've had with the 3M folks before they went into this full-blown PFE the argument that they were making is that the real episodes are basically hospital-based episodes. It's an admission plus things that surround that admission and discharge.

And the explanation for that is that's where the money is. That's where the expensive care occurs and that's what you should be focusing on.

So it doesn't have to account for 80 or 90 percent of the total experience to be important I think in that context.

MR. MACURDY: I wanted to address the earlier question about whether you have to evaluate the whole grouper or particular components of the grouper. It depends on the grouper.

And the best example there is if you were to look at groupers like ETG, less so MEG.

The difficulty there is when you look at the particular outcome for one measure there is this competition that's taking place on the back end about where a claim goes. So there you almost have to understand what overall the grouper is doing to be able to figure out why it did the assignment it did.

Another -- and 3M was a good example. If you take 3M, you could take each individual one and they're modular so

they're pretty well self-contained and you wouldn't really need to understand what was happening with the other episodes to be able to do that. So I think it really depends on the grouper constructed.

DR. MIRKIN: I would like to agree with Mark that it really depends on the use case and maybe expand that a little bit that the use case would include population.

So, for pediatrics the use case would be if we're going to evaluate the quality of care in the pediatric population that's one use case. And I think Francois would agree with that in terms of we might want to measure total cost of care for that population in that setting. And I also agree that they're also different.

And then finally, I know it's not directly related to this, but again in my role as a software basically marketer or seller for MedInsight one of the things that

happens is we produce so much information to our clients is that they basically want to winnow it down.

So I do think that there needs to be some way of maybe prioritizing the actual -- what's most important for a particular use case. Which is different than evaluating it for endorsement. I think that's -- it sort of fits in that same bucket.

MR. BODYCOMBE: I think I would have to put myself along the lines of Kristine's argument about I think there is some sense in looking at that.

Episodes are kind of a commodity. And they exist in a context. And when you're evaluating a performance measure I think you need to think about is it important to look at this measure in the context of an episode, or would it be just as valid outside of that episode.

And as I think I was indicating

before, yes, it really depends on the specific instance. You know, like every episode grouper is different. They define episodes somewhat differently.

It gets back to that use case. When you get down at the micro level maybe one's better than another, but that all gets lost when you look at a global evaluation of it.

MR. HOPKINS: So, Kristine, you've asked the right and the very difficult question. Can we mention endorsing a grouper?

I don't know the answer to that. It's difficult to see how one would go about that.

If one did I think you'd reach the conclusion that most of the groupers that are widely used meet these criteria.

And I sure can't imagine anybody trying to answer which is the best in class. Some are probably better at doing some

things and some at others.

By the way, the ones that I'm familiar with have all been put together with very solid clinical expert panels. There were a bunch of clinicians that advised these folks on how to do the grouping and all of that.

And yet they come out somewhat differently. So, it's hard to imagine that one could endorse groupers.

On the other hand, you have to consider the implications of not endorsing a grouper. Which is what we saw in one of the recent Cost and Resource Use Steering Committees where they were grappling with a specific use and very specific to a condition as I remember or procedure.

And are we going to have expert advisory committees for NQF go through this exercise every time? Of deconstructing the episode that's under consideration and rehashing the same issues.

And by the way, while I have the floor I just want to note that a couple of times we've sort of passed through some mention of actual versus standardized pricing. I hope we can come back and have a discussion of that as a property of actually a grouper but also the use of groupers.

Because I don't want to see what went on in that steering committee again which is the whole discussion about what's right, is actual versus standard, standardized. And no, you shouldn't use actual pricing because of X, Y and Z. Let's see if we can at least resolve that one here.

MS. MARTIN ANDERSON: I think two reasons we passed by it. And we'll see if either one of them changes. Because I did hear it.

One is it's getting into a detail. We have to figure out how it's going to fit in the framework. So is that

going to be something that gets into the scientific acceptability? Or is it something where the science differs and you just want transparency? I think we're just going to have to deal with that. There's lots of issues like that. Should a claim go in one episode versus many.

Those are kinds of examples of things that are the detail of constructing a grouper and using a grouper. And we just have to decide if they have any place in how you evaluate groupers given that our objective is not to tell someone what they have to do to get to a grouper.

I know the couple of times that we've kind of gone down that route I think we've regretted that from a point of view outside of episodes.

DR. BANDEIAN: I'm still having trouble with -- and I'm sorry, maybe the discussion has moved on. But I'm still having trouble with the issue of importance.

And to me there are, you know, this is a great big country. There are all sorts of healthcare programs, and there are all sorts of users who may have different priorities than national, you know, the Medicare program, what have you.

So, to me the only issue with importance is is it worthwhile NQF spending its time on the subject given that you have a limited amount of time.

So it's kind of like the Supreme Court deciding not to issue certiorari to consider an appeal.

And so I could see that you might say well, you know, it may well be worthwhile but we can't look at it.

Now, having said that, to me on the other point which is I think you do need to make sure that the methodology in the grouper system as a whole is sound. And so I think the system does need to be evaluated.

And then it would seem to me at least for those conditions which are quote unquote "very important" then one would need to make sure that the logic for, for example, coronary disease or heart failure was sound and that you would look at some outputs from the system relative to coronary disease or heart failure and convince yourself that when this grouper is saying that somebody has a high cost for heart failure or coronary disease that that is a valid conclusion and that the user can act upon that conclusion.

With regard again to these -- and Tom is exactly right that there is competition for the assignment in the background. And so that's again part of the reason why you need to look at the system as a whole. But it also might be for these things that are, quote, "less important" maybe there would be a less resource-intensive way of at least saying if they've

done XYZ in developing their logic for these 500 lesser conditions then it's probably good enough and we'll sort of provisionally or kind of give it a 50 percent endorsement as opposed to the intense review of the critical -- I'm being a little, you know, I don't know the right language here.

But I'm saying for the things that are really, really, really important, dollars or lives, I think you need to actually be convinced that the system is functioning properly.

For the other stuff, well, it may be very important to the patients who have those conditions, et cetera, et cetera. But given that you have limited resources you may need to come up with a sort of less resource-intensive way of saying in general it looks okay and maybe eventually we'll be able to get around to looking at the specific details.

MS. MARTIN ANDERSON: So, we are

a little over time. Thank you, Steve.

We're going to start talking a bit now about the key modules. I just want to ask you guys is it something you really feel like you need to get out now? Because we have a whole -- we'll be in deep on this for the rest of the day. Okay.

Marjorie? You haven't said much so go ahead.

DR. KING: I just wanted to get in the point that as a clinician and as someone who has submitted a measure for endorsement, it's been endorsed and re-endorsed, et cetera, I understand where you're coming from with this.

But you may want to think about that the importance and the scientific applicability are not the pass-go steps for this particular project.

That the pass-go steps really are is it understandable, is it valid, is it reliable. And they may want to go up to the

top.

Because this importance thing is one paragraph and you're done for this.

MS. MARTIN ANDERSON: So yes, we've got that. I think for now we'll leave it in because there was some debate about it, but does not mean it will be -- in any way what's done for measures just be adapted here. So, Taroon?

MR. AMIN: Okay, great, Kristine. So, again, so for context where we're going now is that we wanted to introduce the criteria, what we've been using.

It doesn't -- you know, that it's a starting point and we'll continue to have this discussion as we talk through each of the modules both today and tomorrow. So we'll come back to this to make sure that we're all in alignment.

And so what we're really talking about now is that when we're talking about the scientific acceptability of the grouper

and the grouper measures that come out of the grouper we wanted to start with a straw person of the key elements of the grouper.

And these are by no means intended to be sequential, first and importantly, nor is it intended to be a steady state meaning that these are all adjustable.

So the questions that we want you to consider as we go through this section are do these modules reflect the major elements of episode groupers. Let's just call it episode groupers, not construction but episode groupers.

And again, we want to have the most diverse perspective around that, around every type of episode grouper that's out there, ETG, MEGs, the PROMETHEUS product, the Medicare grouper.

So do these modules reflect the major elements? Are there elements missing?

And secondly, we want to look at

the components within the modules and we want to make sure that that association is appropriate, or at least that we can live with the classification system that we've developed.

Again, for context what we want to be able to do after this is take whatever key modules we come out with. You know, let's assume that we're starting with construction, clinical and adjustments for comparability, and that we'll break out into groups and each group will be responsible for each one of these modules.

They will be evaluating the best practices -- sorry, that's not the right term -- the key principles that one should keep in mind for each of the modules and the components.

Also, think about what use cases may do to the specification. If there is some guidance related to how the use might change the construction of the module.

And the relevant criteria for each of the modules. So those are the key things that we'll be looking at this afternoon.

So, what I want to present to you now is just this straw person for your reaction. And the key takeaway before we break for public comment and lunch and where we want to get to is ensuring that we're in general agreement around the construction of these modules and the components within the modules.

So let's get started with construction logic. Although it probably would have been easier to start with clinical logic now that I think about it. Go ahead.

DR. BANDEIAN: Why -- I'm just again curious why do we want to do this? Why do we want to identify the modules?

MR. AMIN: So, the way that we're -- the reason why I wanted to set it up this

way is because we still think that -- so, the purpose of NQF endorsement, let's just take it all the way back, is that we want to have some standardized specifications for national comparison.

And so if you're using this episode grouper we want -- the straw person we're working from here is that these modules represent the specification of an episode grouper system. That is where we're starting.

Now, if that -- if others disagree with this setup, that this is not really the specification of what an episode grouper contains we can have that discussion. But maybe if I can just start with where we -- start here and then we can disagree if this is not an appropriate approach.

DR. BANDEIAN: I'm sorry to press the point but I'll --

MR. AMIN: Go ahead, please.

DR. BANDEIAN: -- one more time. So, for example, missing data. Why is that there? Well, presumably it's there because it somehow relates to perhaps validity, or maybe to feasibility or some such.

So, and why are complementary services here? Complementary services I think, if I understand what that means and I think I do understand what it means would be like anesthesia. And so here a person gets surgery. Now, how are we going to put the anesthesia with the surgical episode perhaps. Maybe that's what's intended or not.

MR. AMIN: Perhaps, yes.

DR. BANDEIAN: Again though that strikes me as a validity issue. Because if the anesthesiologist puts on his claims or her claims -- my mom is an anesthesiologist -- you know, COPD because that's what he or she is most concerned about as a risk factor for the patient, but the patient is actually

undergoing cardiac surgery it should probably go into the cardiac surgery episode as opposed to the COPD episode.

But that again is sort of an issue of validity of the construct.

So it seems to me that these are all things that are in grouper systems as a means to an end.

And I think there really are two ends. And number one is validity and number two is actionability of the outputs.

And so to me I would first try to flesh out those concepts. Because maybe there are different ways of accomplishing the goals of validity and actionability, and not necessarily these specific components.

MS. WILBON: Right. So let me just try to give you a little context on kind of where our starting point framework is for doing kind of evaluation of measure work.

Generally the way we've

structured our evaluation of measures on the resource use side, we're starting with quality and then kind of moving into resource use, is identifying what information do we need from the developer in order to (a) understand how the measure works.

Which these components essentially represent if a developer was to submit a grouper to us, for example, we would want to know how they've handled these different things, and explain and describe their rationale for why they decided to do it that way to make sure that it aligns with the intent, that it aligns with the intended use that is transparent.

So without having this type of information that you can't evaluate it without kind of understanding how these different components work.

And similarly on our quality side we essentially asked for the developer to

submit specifications on the measure. And the criteria are applied in context of what has been submitted by the developer.

So we may find that these different components, that the validity criteria, whatever we decide those are, need to be applied to these different elements in order to determine whether or not the approach that's been specified is valid.

So, these are really kind of to help us understand (a) what types of information we would need to understand how the construction of the grouper has been proposed in order to give an evaluator an idea of what we're looking at. So that's kind of the context that we're coming from.

And the criteria piece will come as we kind of identify what those are and we figure out how we would actually determine whether or not it's appropriate.

DR. BANDEIAN: Okay. I'll just try one more. See to me, I heard what you

said, but to me what I would suggest is an alternative. But I understand you're pretty far down this road.

Is first to have a discussion of what the threats to validity are. Identify the threats to validity. And so, and then -- as opposed to saying let's have this component, identify what is going to put validity at risk. And have a bulleted list of these things. And then you would ask the developer what are you doing.

So rather than calling it a module I would say to the developer what are you doing to deal with complementary services. Not necessarily a module, but just the question -- or maybe I'm now getting tied into semantics.

DR. CACCHIONE: There's a little semantics here. I think that -- David, did you have a comment?

MR. REDFEARN: I just -- we're going to split up in groups and I guess

mechanically I'm concerned about that process.

I think we have a very wide spread of opinion here and I am concerned about chopping us into pieces so that we are not exposed to that wide range of experience. That's just mechanically an issue that I'm a little concerned about.

Especially since we're chopping it into pieces that the group has widely divergent opinions on whether that's the right way to do it.

MR. AMIN: So, I mean again, we have a tall order in the next two days. So, part of this is to achieve a little bit of efficiency.

However, this discussion right now and the discussion at the end of the breakouts is to bring some of these pieces together and to make sure that we have that diversity of opinion across the various different groups.

We're open to making some changes here, but we're also trying to make sure that we're --

DR. CACCHIONE: I think we ought to just keep with the program the way it is. I think we'll have a rich discussion afterward. Because I think there will be some efficiencies gained by being in breakout groups.

And so as much as, David, I hear you and respect that, I think that there is some efficiency in breaking out and them reassembling to do a debrief.

MS. MARTIN ANDERSON: I think one other thing we want to remind everyone is we're just, we're testing this. So this is one bite at an apple that we have to eat over a couple of months.

So, it may not work. We may find that this isn't the right set of issues to discuss that the NQF staff has laid out for us.

But I think some of the right issues are at least in here. And then we can add others that we need to put together.

I do think there is a lot of diversity of experience in the room. And so we'll -- I guess the one risk that we run is that we'll get back together and we'll have to rehash each section.

But I think in that sense if that's the reality that is going to have to be the reality. We have to hear all viewpoints and see how we can advance.

I actually don't know how we're assigned, so. Are we choosing ourselves or are we assigned?

MR. DE BRANTES: Can I make a couple of comments? This is Francois. Can I just make a couple of comments?

MS. MARTIN ANDERSON: Yes, go ahead.

MR. DE BRANTES: Okay. So, on the construction logic I felt that there

were a few issues that were blended together that might best be kept separate.

So, if you -- because as you look at both -- you almost have to look at construction logic module and the clinical logic module and the components that you would put in there together.

So, the clinical logic has as I read it mostly issues around the rules to trigger episodes and close them and so on and so forth.

But then in the construction logic you also put clinical hierarchies. So the methods used to define the hierarchy of codes and condition groups.

And it seems to me that that's an inherent part of the clinical logic. Because if you're going to figure out how to trigger an episode and you're blending in some kind of clinical hierarchy in your definition of codes that's going to impact inherently how the grouper works to trigger

an episode. So I thought that was kind of putting something there that belonged in the other one.

And the construction logic to an extent as I read through this seems to be dealing mostly with what decisions are made to assign services to different episodes which is a very complex and important issue to discuss.

I'm not sure I would blend that with the attribution. Because the attribution here as I read it is really about attributing how claims are assigned to responsible entities.

Usually claims are not assigned to responsible entities because the claims emanate from an entity. It's the episode that ends up by being assigned to an entity.

So (a) let's be clear about what you're actually asking here.

And second, if it is about the assignment of episodes to providers that's a

completely separate issue from the construction logic and the clinical logic.

And we had a little bit of discussion earlier today. It deserves a separate conversation. And blending it in with the construction logic is going to -- not going to be particularly helpful.

MR. AMIN: So, if it's okay with the chairs maybe I can just walk through these three modules really quickly, just talk about what's in them, and then we can open it up to see.

That is exactly the type of feedback we're looking for, Francois. And we can decide what to do with attribution. We can maybe have a separate group to discuss it, or we can figure that out.

But let me just make sure that we're all on the same page. I don't want to assume anything here.

So, the construction logic essentially is the methods of assigning

claims beyond that which is associated with the clinical logic.

So you can think about the clinical logic as essentially the individual episode, and then the construction logic is essentially how you're dealing with episodes and how they relate to one another. So, hierarchies, concurrence of clinical events, things of that nature.

There is some components here that may not be directly related to that topic around missing data, how missing data is handled in the system.

And then essentially what are the resource use service categories that are built into the episode grouper, meaning -- resource use service category would include sort of durable medical equipment, or pharmacy claims, things of that nature. What are the categories that these services are assigned to.

And attribution would be

essentially -- I think the intent of this was to describe what Francois is describing which is how the episode essentially is attributed to an entity.

The clinical logic on the next slide is essentially the definition of how the individual episode is constructed, meaning the trigger and end mechanisms and potentially interactions of comorbidities and how that is handled in the system.

And then finally, adjustments for comparability includes inclusion and exclusion criteria broadly, meaning the claim line or other data quality exclusions, high-dollar claims, Winsorization, any other approach that's included there outside of the trigger and end mechanisms.

Your risk adjustment methodology, your stratification approach, if any, the costing method which David was referring to before around actual versus standardized pricing approaches, and then the scoring

methodology of how you come up with the -- whether you're using an O/E ratio, observed to expected and how one could interpret that.

DR. CACCHIONE: There seems to be a lot of overlap in these. I mean, these are sort of arbitrary, the buckets.

MR. AMIN: Absolutely.

DR. CACCHIONE: Because I look at clinical severity levels versus stratification versus risk adjustment. And I think that some of the uncomfortable -- I just have sort of -- because it just doesn't -- it's not clean.

MR. AMIN: Okay.

DR. CACCHIONE: And I'm not sure that I have a solution for you. Just it doesn't feel quite right looking at these because there is a lot of overlap.

David, you have some comments?

DR. MIRKIN: I was just going to say what you just said pretty much.

(Laughter)

DR. MIRKIN: Other than maybe -- one thing that helped me was to get rid of the modules and get rid of the titles for the modules and just say there's a bunch of topics that in order to get through them they're going to have to divide the group up to attack those topics.

And maybe as we discuss individual topics we may say this is a total overlap. Or maybe, I don't know if you want to spend the time going through some of that now? Because I do agree there's a lot of overlap.

And when you actually look at those individuals who are actually building episode groupers these are all part of one, you know, one thought process. There's one logical process that they go through. So, anyway.

MS. MARTIN ANDERSON: We'll take proposals. We've got a half hour to figure

out what we're going to do after lunch. So if there are alternate proposals let's get them out.

DR. CACCHIONE: So, do we want to hold to this sort of breakdown of the three modules? Or do we want to think about things differently and break it down differently? Mark?

DR. LEVINE: There's an interface between the grouper and the user of the grouper that is murky in my mind.

For instance, one talks about attribution. That will depend upon what the user is intending to do with the underlying technology that is able to group claims in an appropriate way.

And so if we're really going to be looking at what are the criteria for a grouper we also need to discuss when does the responsibility of the grouping and we pass it over to the user to be able to pick up at this point in time and apply it to a

given use case.

And that's within the context of these broad use cases of a patient-centric approach versus a provider-centric approach which I think is a useful bifurcation.

DR. CACCHIONE: Is that a bifurcation that we ought to think about in terms of the breakouts? Thinking about the use of these as a patient. And I think, Tom, you were the first one to -- but breaking it on that -- that's the breakout? Is it around patient-centric versus provider-centric?

And then addressing all of these issues. David?

MR. HOPKINS: Sort of addressing the same issue. I think there's some logic to having the breakouts. It matters less to me what the labeling is of construction versus clinical logic than that the bullets underneath are meaningful. So we can have discussions of those.

But I've heard suggestions that there's two big bullets here that really ought to be discussed by all of us. So one is attribution and the other is what I guess is referred to as costing method. Those are big topics and I don't think they should be limited to breakouts.

MR. MACURDY: I just wanted to note that I mean, obviously, I have no objections to discussing how to rearrange these. But I assume that when we're all done it's not going to be super satisfactory to everyone already.

And the fact that there's overlap here I would have thought people would be happy with because everybody wanted to have one meeting to start with. The fact that there's overlap is kind of fine.

I mean, no matter what -- if I did mine, people would object. You do yours, it seems like it's fine.

I think maybe underneath if

there's maybe following up with David, if you want to have like attribution across many maybe we can do that, something of that nature so that we can get more discussion and have the breakouts and then people can share their ideas when they get back.

But I would have thought trying to reorganize them is going to be a lot of time for something about where we're going to be.

DR. CACCHIONE: To staff, I mean, other than the fact that these -- you had to draw lines somewhere. That's great. I mean, is there going to be enough -- with all the overlap you guys will feel comfortable with synthesizing this into some sort of logical rule around these different modules?

MR. AMIN: So, these, just for context these modules are reflected in terms of the specifications that we get for cost and resource use measures already.

And we recognize that there is some overlap, again, specifically around how we handle the severity levels and the comorbidity interactions. I think that's what's causing a lot of -- and the clinical hierarchies which is what's causing some concern.

But part of what we're trying to get from this is which of these components can we sort of eliminate or combine. And so if we could just, you know, if we go through the discussion it will become very natural to us which ones to take away.

So in summary, I feel pretty comfortable that we could probably get that information from the work groups.

MS. WILBON: I think from a broader perspective it would be useful to know whether or not there are things that are glaringly missing.

If it's an issue that there's overlap like Taroon I'm less concerned about

that, as opposed to making sure that we captured everything that potentially we might want to evaluate if we were to look at a grouper or individual episodes within a grouper.

So, if we could kind of maybe, like David suggested kind of take a step back, try not to look at the labels that we've given these different buckets and kind of look within the buckets and determine whether or not there's anything here that is -- or anything that's missing.

Again, I think the overlap is something that we can address as we kind of synthesize the information and hear your discussion. It'll give us an idea of how we might actually frame the criteria.

Because the criteria have to be based on something. It has to be based on either the components that we're asking for within the grouper, or things that we think represent what a grouper is.

So, I think I'm still struggling in understanding, you know, getting some consensus from the group on where you are with what we think actually consists of the grouper. So, I don't know if that helps anyone.

MR. MCLEAN: So, one, I start off by saying I'm okay with doing it this way if you want to do it that way.

I think where I'm struggling though, where I think most people is struggling is when I step back for a program for BDC and select a grouper that I want to use for this evaluation I don't look at it this way.

I look at it in a holistic form of what I want and what my use case is, what my purpose is. If I'm looking at it from a provider evaluation, or I'm looking at it from a patient-centric evaluation.

And then I look at for each component in a holistic form. I don't look

at it in parsing it out.

And so it would be easier for this group to take a step back and say well, this is your user case. Now, what components need to go from your perspective into that grouper?

So if we want to make a change that would be my suggestion. But otherwise I would just do it the way it's set up.

MS. HOBART: I had a comment and a question. My comment was basically the same, that I think if you're a health plan or something else you're applying the whole package.

So you want it to have certain characteristics, that it's easy to explain, it's consistent over time and other things, that it's going to be holistic, not so much individual components as how they come together. So at some point I think we need to put that openly.

My question is when we're talking

about the provider-centric versus patient-centric whether that's the organizing principle of the logic as to whether it's tagged to the person versus the provider, or it's the use case as to who's taking the results and using it. I just wasn't clear.

DR. CACCHIONE: So I guess one of the questions -- and I'm going to answer a question with a question and get consensus from the group.

Do you think that within each of these modules that we should -- is it important to think about a construct that would divide this into a patient-centric and a provider-centric within each of the modules? And to see if within each of these modules there is a separation about these issues?

Or is there just enough of an overlap that it shouldn't be -- that we should just sort of go generically and look at these modules?

Because in the absence we are talking here, in the absence of any concrete proposal we're going to stick with what staff has put together for our discussion purposes simply because we'll sit here all day and try to figure out how to break it up.

So, I would just say we'll stick with what staff proposed. And I just offer that up as a potential way to sort of take it a different route within each of those modules.

MR. MCLEAN: You guys put it together. I think at the end of the day, like someone said earlier no one's going to be satisfied with either way. Let's just go with the way you guys kind of have set it up. I think we'll get it done either way.

MR. NAESSENS: It's always good to think about examples, and I think those two examples of the patient perspective or a provider perspective would help us think of

more of the issues that will come out in each one.

DR. CACCHIONE: Whoever the leads are of the staff maybe they can at least direct a discussion in that way.

MR. HOPKINS: Can someone help me understand better what that's -- provider perspective versus patient perspective? I'm not sure.

MS. MARTIN ANDERSON: I think what we just want to do is say would you evaluate a grouper differently in whatever you're discussing if you were taking a perspective of a provider use case, and there were some examples of those, or of a patient-centered use case, and there were some examples of those.

Because I think we're getting at the question of does the actual evaluation need to be divided according to the use. And I think we can only take that up at a detailed level. And so we're saying

consider that in each of your subgroups, whether or not you would be stating what's important to evaluate differently had you had one or the other use case in mind. And that way we can bring that back into the dialogue.

MR. HOPKINS: So what's the patient-focused use case? I'm just not --

MS. MARTIN ANDERSON: Well, I'll give you the -- I take all these notes. Or actually, Mark said payment versus performance. You use those. And Tom used patient-focused episodes versus provider-focused episodes.

So, whereas under performance a provider-centric view of looking at, say, resource use and whatever outcome based on a provider. Or under payment where it might be more focused on, you know, you're paying at a patient level so it's more patient-centered, or population-centered groups of patients. So that was part of your -- Mark,

that was your summary of a proposal?

DR. LEVINE: Yes. Let's look at it in terms of patients with diabetes. A patient-centric approach would be looking at diabetes over a period of time and it would consider nephropathy and retinopathy and all of the different kinds of complications. All of those can apply to a patient.

Whereas a provider, a particular physician might be a nephrologist, or he might be an ophthalmologist, or something like that who would be looking at only one part of that whole continuum.

So a patient-centric is looking at a continuum of care whereas a provider-centric is looking at it within the provider's context.

MS. HOBART: So I think this goes back to my question. I think we've actually thrown two different concepts out there.

Because I think one of the use cases was around transparency to the member

or patient in terms of what their overall cost liability would be. Which that's why I'm saying that's saying it's just from the perspective that you're the user and that's different than a provider wanting to know his or her efficiency.

But then to say if your cut for a provider by a population clinical cut versus maybe their practice cut it's a different thing.

So that's why I was asking are we talking about the logic of the attribution, or are we talking about who's using the information? And if we get some more into the member liability question. And I think they're two different things.

DR. CACCHIONE: So we're going to take -- I think that we've sort of come to a closure on this. At least we have a roadmap for the breakouts.

And I don't know, are there more that you feel like you need to cover?

Because we're sort of -- I think we've at least come to a consensus as much as the group is going to come to consensus on how we're going to discuss this, how we're going to break out the discussions.

Is there more you need to cover from the staff perspective?

MS. WILBON: I think --

DR. CACCHIONE: Because we'd like to spend -- we'd like to spend some time on this cost issue and this attribution issue which are sort of in a lot of people's -- in the forefront of a lot of people's thoughts.

MS. WILBON: I would say let's move on. I think we're still a little bit unsure where the group is going and how -- I think we're just going to --

MS. MARTIN ANDERSON: We're going to go with your design.

MS. WILBON: Okay. Okay.

MS. MARTIN ANDERSON: And see where it takes us. We are organic.

Was there something you needed to add? I didn't want to cut you off.

MR. NAESSENS: No, I was just going to give another example from our practice. It's more or less separating off the referral patients from the community or population-based. It's a very different focus and a very different thinking about the groups.

MS. MARTIN ANDERSON: Thank you. Okay. So you want to go to attribution or pricing?

DR. CACCHIONE: David, do you want to start us off on cost issue? Because this was something that --

MR. HOPKINS: So, I'm stretching this memory for this discussion that took place at a steering committee. And it was a lot of these -- you folks were there. And so, I wasn't actually but I read all this.

What I think I recall was this resistance to a measure using actual pricing

because actual pricing reflects local market conditions. And therefore we could not compare performance, you know, across the country.

And I found that to be a really interesting and somewhat strange argument because for those of us who work in the purchaser and payer domain we care a lot about these market factors that do influence pricing.

And the fact that episode groupers can reveal some of the variation in pricing is part of what makes them important to us.

So we had a very strong reaction to a group that seemed to kind of get dug in on that particular issue.

So maybe we could discuss it here. Maybe I would hope that we could be more open to the idea that actual pricing does have its place in episode groupers and their applications and certain use cases.

MS. MARTIN ANDERSON: So, just a proposal for where this might belong. How you think about, whether you're looking at standardized pricing or actual pricing or cost to whom, whatever, probably doesn't affect much how somebody would form the group through the construction logic or the clinical logic.

But it does come into play when there is this question of risk adjustment and just adjustments of output that you can -- around specific types of measures that are dollar-oriented.

So, maybe the best place to put it for now is in that group that's going to handle talking about adjustments that are made.

I know it's very specific to a certain kind of measure in a way. I mean, I don't know whether or not you could evaluate -- I could imagine the person who creates the groupers may or may not have a

preconceived notion of exactly how that gets done. You know, many of them give it to you both ways and then users choose which way they want to use it.

So in evaluating the grouper I think we have to decide -- I want to hear from this group -- is this really an element of evaluating the grouping methodology, or is it really an element of evaluating the user of the measure.

MR. HOPKINS: But if we punt on that here then I just foresee -- I'm trying to be really practical. I foresee that every time a grouper or a measure, specific measure that uses one of these groupers comes up on the table at future steering committees the same discussion is going to go on. And depending on the makeup of the committee you'll come out one way or another.

I just feel like we ought to weigh in on it. It doesn't feel to me,

Kristine, quite like an adjustment issue. It feels like kind of a standalone issue.

MR. JONES: I think we need to discuss that at detail. Because I think the criteria should be how flexible the grouper is to accommodate those adjustments.

In that you want to see what your actual network looks like, for example, if you're going to tier it and develop select networks and higher copays and all that.

But if you're trying to actually see which providers are more efficient, you want to flatten and just make sure you're solving for mix and volume changes.

So I haven't seen too many good, acceptable responses from the big players in their ability to do that. So I think if they can disclose how they plan to do it, or how they do do it and how easy it is for us to adjust I think that's a key, key criteria.

Because if you can't take the

output of this outside of your building an actual, you know, and face out to the docs in a credible way there's really no value.

MR. DE BRANTES: Well, this is Francois. Again, I think it depends on the use. Because if you're going to use a grouper for payment purposes you're not adjusting for price. I mean, that defeats the entire purpose of using a grouper for bundle payments.

MS. HOBART: I was going to say basically the same thing. I think it's a question of functionality, not a black or white is it standard cost or whatever.

I mean, there are different situations where you're going to want to look at the utilization pattern, you're locked into some crazy contract. It's not going to make sense to look at the cost or whatever. So, for me I would want the functionality to look at both utilization and the cost as a dependent variable. And

it's not a higher-level conversation, it's just functionality.

MR. MACURDY: I just want to go back to Kristine's point. I mean, most of the use of the groupers that I've seen you can either use real prices or use standardized prices. And obviously I totally agree with you that depending on how you're using it if a hospital costs twice as much in one area versus another that actually matters. Because if you're trying to get lower costs then that's what you do.

On the other hand if you're trying to do quality measures you might use standardized price.

But I think it's -- and I'm going to use the term "side issue." And I don't mean that it isn't important, but from a grouping perspective once it's grouped you can use either real prices or standardized prices.

And if somebody kind of says,

well, the only way we'll report the measure is in standardized prices because well, okay, that I agree with you I don't understand. It depends on the purpose.

But as far as the challenges you have with regard to the grouper it can go either direction. I mean, you can even use another price mechanism as well if you need to and that's sometimes done.

It depends on its context. But I think it's an easy sort of thing at the end just to say use these prices, or use those prices, or use yet a third price level.

DR. CACCHIONE: So Mark?

DR. LEVINE: I think this goes back to a point we were discussing earlier, when does the grouper's responsibility end and the user's responsibility take over. And we're going to need to have some discussion about that.

This is one area in which we might want to have a defined handover point.

And there may be others that we should discuss.

DR. CACCHIONE: Do we want to talk about attribution now? I think about the groupers and attribution. Those were two big themes that we wanted to have.

MR. MACURDY: Well, I think attribution is easy too because in some cases groupers don't do attribution. And they explicitly don't do attribution. They leave it up to the user to do attribution.

And if you're talking the mainline groupers like ETG, or MEG, or 3M don't do attribution. So that has to be a feature that's added on.

I mean, if you want to basically say they have to have attribution as well you're not going to get very far in this regard.

But I mean, it depends on its purpose. I mean, if it's for provider payment there's going to have to be some

attribution role, but that can be kind of done separately.

DR. CACCHIONE: So it's almost like a cost issue. David?

MR. REDFEARN: I think attribution is sort of an add-on. But if you want to look at it another way, I mean the ETGs, for example, the current version of ETGs generate what's called an attribution file. And they give you data to let you get yourself in trouble. The user can get into trouble.

There's a whole host of pieces of data that you can make choices in terms of cost, who the first provider was, who the last provider was. There's a whole host of kinds of information.

So, one thing I would look at a grouper if I am doing provider profiling I would say does the grouper provide information for me that helps me make these kind of decisions on the back end. But

that's not a core -- it's not the grouper, that's just what information it provides that you can use.

Because MEGs makes by default I think, just says the most expensive one. And the PFE doesn't do it at all. So, that's just the dimension.

MR. MACURDY: I just want to note both ETG and MEG explicitly say that they'll provide you information but it's up to the user to do attribution. They don't do attribution.

DR. LEVINE: There's a responsibility I think of the grouper to align the technology and the approach to attribution that enables appropriate attribution.

In other words, if the intent of the grouper is to inform upon provider performance and they're using a patient-centric method of evolving the grouper they're going to have trouble in coming up

with an appropriate attribution algorithm.

So, one of the things that we might want to consider for a grouper is that they are explicit in terms of what the attribution opportunities are as a result of their grouping technology.

DR. BANDEIAN: We've actually spent a little bit of time on this topic, the so-called off-ramp discussion that we've had.

And again, if one steps back and says well, why are we talking about episodes. And ultimately the purpose of that I think is to assess the efficiency of the care being delivered and the further discussion about care being provided over a period of time.

The problem with the sort of conventional standard whole episode concept is you're attributing the entire enchilada to one or more or a team of physicians who may or may not have had really an

opportunity to influence what happened during certain aspects of the care.

So, if you think, for example, about say hip fracture and there is the acute care of that and then there might be the rehabilitation portion of the care.

The folks who are involved in the rehabilitation may not necessarily have had much opportunity to influence the acute care side of things.

So again, depending upon what one wants to do, how one wants to implement it one might attribute the entire episode to one or more physicians, or one might split the episode into pieces that reflect what different roles and responsibilities people had at different times.

So again, in turns out that this actually, it can be potentially quite complicated. But certainly the traditional way of implementing this which I guess to use Tom's language would be the sort of

patient-centric I suppose whole episode approach is to try to attribute the entire episode to somehow a small number of physicians or some such.

An alternative implementation tries to figure out what parts of the episode different people were responsible for and/or had control over, and then attribute that to those people.

And there are pros and cons of both approaches and I suspect there could be lots of discussion. But it is itself a fairly complicated topic.

MR. MCLEAN: I'd add to Steve's point about it being very, very complicated.

Provider attribution, and I think I said it earlier, I actually believe it should be a separate discussion from episodes. I don't think -- just like I mentioned about risk adjustment.

Simply for the fact that the purpose of a grouper is to figure out how to

take these claims or this data and group it into some meaningful information.

Now, how you want to use that, do you want to attribute it to five different providers with one main provider and then some other ancillary services and those providers? Or do you want to just do a population-based analysis? That's more about the user and the use case.

Do you want to risk-adjust or do you want standardized cost, the actual cost? All those things are about the use case and what you are trying to attempt to do with this grouper.

The more complex you make things with the grouper with the attribution and the adjustments you're relying heavily on the reliability of data. And anyone who's ever tried to match claims data knows it's very, very complicated.

Because a lot of times it's not there. And what groupers typically do when

data is not there is they drop the episode. Which is also problematic because the reliability and validity discussion we had, you don't have the samples now.

So, the more complex -- although it sounds great at a very high level I think when you get down to thinking about it these discussions I think should take place, and they're very important, but outside of evaluating an episode.

MS. MARTIN ANDERSON: So I think at least what I've heard so far is that these are both issues of user preference, not just they are related to a grouper.

And I've also heard a perspective that says maybe at a minimum we need some transparency on what the grouper allows. And maybe others feel like there ought to be standards for what a grouper needs to be able to allow to occur.

But regardless we can pick those up in breakouts as needed if there is a

specific point of view someone wants to put forward in terms of how you would consider these issues if you're doing endorsement, or whether you would just leave them outside.

So I think it's good that we took this time to hear the perspectives on these two topics. If there are other topics like that I'm sure we'll be taking some time to hear perspectives too.

The good news is we actually don't have to solve everything right here, we just have to keep inching forward in our work.

So, I know that you had a member comment scheduled for now before lunch?

MR. WILLIAMSON: Yes. Any public comments in the room?

OPERATOR: At this time if you would like to ask a question or have any comments please press *1 on your telephone keypad. We'll pause for just a moment to compile the Q&A roster.

MR. WILLIAMSON: Operator, we have one comment in the room.

MS. RUBIN: Hi. First, thank you for the opportunity to comment. Today's discussion was very insightful and helpful.

From a physician's perspective evaluation of episode groupers need to include all components within the episode for physicians to have actual information to know how they are evaluated and to improve upon their care.

Also, the construct of an episode may be different based on what you are evaluating and what is included in the episode.

And that's all I have to say. My comments are just reflective of the discussion that occurred today. Thank you.

OPERATOR: There are no public comments or questions over the phone line.

MR. WILLIAMSON: Thank you very much. At this point we will break for

lunch. And so lunch will be served in the other half of the room over here. We will be convening again at 1:15 for breakout sessions.

Actually, we could -- we'll say 1:10 and we'll give you some instructions. We'll be taking you to other rooms for the breakout session. But we'll reconvene at 1:10.

(Whereupon, the foregoing matter went off the record at 12:29 p.m. and went back on the record at 1:18 p.m.)

MR. WILLIAMSON: So, as we break into our module groups, again construction logic, clinical logic and adjustments for comparability we're going to really discuss the principles for constructing an episode grouper. And we might get into criteria for evaluation. We'll see if there's enough time on that. But we really want to dive into the principles right now.

So, in that regard we're going to

examine the key questions that we've outlined. We're going to further define the module and key elements, and really identify the principles and considerations.

So we've broken the group up into three groups here. We have construction logic will be in this main meeting room, clinical logic will be in the other half of the room and adjustments for comparability will be downstairs with me in the 8th floor conference room.

Taroon Amin will be facilitating group A, Ashlie will be facilitating group B and again I have group C.

So at this point we're going to -- or Ashlie, do you have any other? Okay. So at this point for people on the phone we're going to be going into a speaker subconference in this room. So Operator, if you could please pull in Jim Loiselle and Francois de Brantes. They'll be staying in the main meeting room here on the line.

And for the rest of you on the phone, I'm sorry, you won't be able to listen in for the group discussion. But we will be reconvening again at 3:15 on the line to recap the breakout session.

So Operator, are you able to do that?

OPERATOR: Okay, one moment. Okay, you did say Jim and Tim's line, correct?

MR. WILLIAMSON: Jim and Francois, yes.

OPERATOR: Jim and Francois. Okay, one moment. Okay, I've pulled all three of your lines into the subconference.

MR. WILLIAMSON: Thank you. Jim and Francois, are you on the line?

MR. LOISELLE: Yes, I'm here. This is Jim.

MR. DE BRANTES: Yes.

MR. WILLIAMSON: All right. Excellent, okay. So we're all set here in

the main room.

And so again, if group B wants to meet in the room next door and group C wants to meet me by the front desk here in the conference center I'll take you downstairs and we can go ahead and get started.

MS. WILBON: I would suggest if you're not taking your laptop to bring your paper copies of the discussion guide because we'll be referring to several questions in there as the group goes along.

(Whereupon, the foregoing matter went off the record at 1:21 p.m. and went back on the record at 3:35 p.m.)

MR. WILLIAMSON: At this point we're going to go over the breakout sessions. We're going to do kind of a report-out and overarching discussion of each breakout session.

We'll start with the clinical logic module, or clinical logic group. We may move away from the module moniker for

this.

But I think, again, a lot of great discussion over the last 2 hours so we want to make sure we share that with the whole group, get input, feedback. And so we'll go ahead and kick it off here.

MS. WILBON: We have two spokespersons. Marjorie and Jennifer are going to kind of partner up and present kind of what the group came up with.

We have a document on Word -- we weren't able to get it into slides -- kind of summarizing what we came up with.

And we also have our notes on the notepad paper in the back of the room. I don't know if it's legible but if you were trying to figure out where we were going you can kind of see where we ended up. So thank you, Jennifer and Marjorie.

MS. HOBART: So, I'm tag-teaming this with Marjorie so I'll start.

So our group really came up with

three categories of criteria. So the first criteria was just around basic what I'd call software functionality which was just considered the minimum to even be considered for the episode grouper.

So, you had to have something that would be relatively easy to implement, there's clear documentation, it's reliable just in terms of if you do the same thing twice you're going to get the same answer. If it's hosted at a vendor site that there's a clear security protocol for PHI information. There's a plan for user support and maintenance.

So, those are just examples, but just very standard expectations if you have some sort of software product. And then after that you can start talking about the content.

So then we got into what was the actual clinical logic that was put onto that software and Marjorie's going to talk about

that.

DR. KING: And that's what we spent most of our time talking about. We looked at the subgroups under clinical logic section, the second page in the back, and talked about well, should we have specific criteria within each? Should we require the measure -- it's not really a measure, but the grouper submitters to describe how they're addressing each of those criteria, the evidence to support it, the triggers, end mechanisms, the clinical severity and the comorbidities.

And we basically ended up with we really want them to give us a coherent description of why they are using this clinical logic, the evidence to support the clinical logic, how they are going to deal with triggers and end mechanisms, how they're going to deal with some sort of -- risk adjustment may not be the right term, but how they're going to separate the

different types of patients with different levels of clinical severity into different buckets to compare across providers, how they're going to deal with clinical characteristics that occur during the episode that are part of the underlying pathophysiology as opposed to a patient safety issue that arises related to, or an occurrence that arises due to an actual complication of what happened. And we had some discussion around that.

And decided that rather than being proscriptive about what they want them to do, we would want the submitters to provide a narrative so that the reviewers could understand their logic.

We talked about basic principles. And one of the principles that we -- and it really ended up with at the end was that whatever the clinical logic is that's used the system should have been tested for reliability, perhaps with a set of claims

data related to various diseases that then the episode grouper can be run using that set to look at the results, similar to what Medicare did when they did their bid for an episode grouper.

We also talked about other -- that we felt that the clinical logic should undergo some sort of face validity testing, perhaps using panels of expert clinicians.

I'm trying to remember what else. Did I miss anything else within the testing? The reliability and validity testing.

We basically felt that rather than say here's what you should do for your clinical logic, show us, help us understand the logic behind your clinical logic and show us that it's valid and reliable. That's sort of where we ended up.

We also felt that the clinical logic should be patient-centric clinical logic, but that it should be understandable and usable by providers and by patients.

So, it's always patient-first logic about how you think about these clinical things.

And we listed the things that we had talked about before. We re-discussed attribution but we just -- they reminded me that we'd already put that to rest.

And from the rest of the group -- I was the only one that had three cups of coffee so that's why I got elected to speak. Did I miss anything?

MS. HOBART: The third category we talked about was handling of the data that fed into the model.

So, currently that's probably largely claims-based but that might evolve over time into more clinically rich information.

But whatever the source was that there should be criteria around having clear documentation about the specifications, about how the model is expecting to see the data.

There should be some sort of data profiling capability that would identify when the data that the model is receiving is not in the format or content of what's expected. So people could make sure there wasn't a mismatch of the actual model and the data that was going into it.

And in general that there would be those types of feedback processes to make sure that you were using the model correctly in terms of the data that you were submitting.

DR. KING: And we also, we forgot, we also talked about the fact that the data should be able to be used for performance improvement by providers.

MS. HOBART: Yes, so there was a usability component too I think with the clinical logic. Not only -- there needed to be a high-level story that people could understand the derivation of the clinical logic and what the framework was.

And then there needed to be a drill-down capability that you could see the services that were bundled into the episodes that would give you a path for action.

So, was there anything else from the rest of the group?

MS. WILBON: Other group members, feel free to chime in if there's something that you're -- you're doing a great job. But if there's something.

So, I'll just kind of summarize some things that I think were really good that the group came out with in terms of principles around some of these topics.

Marjorie mentioned there should be the ability to drill down for clinical improvement.

We talked a lot about kind of the use of clinical severity levels and how the specificity of the episode in terms of the type of I guess the broadness or the specificity of the episode may be indicative

of how severity levels are used.

So, I think Chris' example if I can use that about heart failure patients, if there may be different levels of severity of heart failure patients, that if your episode is just heart failure that you may want to use some type of severity levels to kind of further differentiate and further specify the type of heart failure patient that is being measured.

Also, some discussion about the sensitivity of triggers and that there are different types of trigger codes and flags, including -- some of them may be clinical. They could be site-specific in terms of whether it's outpatient or inpatient.

They could be potentially clinical data, or claims data-based, or procedure-based, or what have you.

And, let's see. That there needs to be some kind of recognition that you don't want your episodes to be so sensitive,

opening to everything so that you're creating kind of these false episodes that don't really have any meaning, but that there's -- the episodes that are being open are still clinically relevant to the population of the data that is being put into the grouper.

DR. CACCHIONE: We also thought there would be some consistency on the triggers.

But on the end mechanism, or the end events, that that might be variable depending on the end user who might choose or it may be the contractor, somehow that those would have some variability.

Whereas the triggers would be very consistent. There would be a transparency and a consistency to the triggers in terms of both how they're used and then the risk profile that might be -- the prospective risk profile for these triggers.

DR. BANDEIAN: Can I ask a couple of questions? Now that I've stopped looking at the EKG.

I think I just heard you say something to the effect of defining things to try to prevent false positive episodes from opening falsely when the condition was not really there.

I understand that that's an issue. There are multiple ways of addressing that question. And so I'm not -- I mean, I think that the issue is not so much how the episode is defined, but rather what protection is being taken against the possibility of false episodes being opened.

And I don't think that that concern is necessarily one that should govern the identification of the condition episode. Because again, as I say, there are countermeasures so to speak that -- so that's sort of point number one.

MS. WILBON: I think the group

generally agree with that. It was just a recognition that that's an important consideration. Yes, we're on the same page.

DR. BANDEIAN: Okay. Point number two, and maybe I -- because I've been multitasking and being bad looking at my phone, did you talk about sort of how one would think about how one would define a condition? What is a condition? And is that up there?

DR. KING: We didn't talk about it because -- we talked about it but we weren't explicit about it. I mean, basically -- again, I'm on the cardiac work group for the CMS one so I just figured that they all worked that way.

DR. BANDEIAN: Okay. Because --

DR. KING: We did talk a little bit about surgical ones versus medical ones versus do you get it out of claims data. We talked about registries, would that be an appropriate way to get --

DR. BANDEIAN: No, I mean like what are the basic units of analysis.

DR. KING: No, I don't think so.

DR. BANDEIAN: So, because this then actually relates to, I don't know, not necessarily drill-down but sort of actionability, usability, as well as perhaps to validity as well.

Let me give you an example. This is probably the one. I'm trying to remember.

So, just as an analogy if you look at some of the DRG labels. I shouldn't be throwing rocks at them, but if you look at some of the DRG labels it's a little unclear exactly what is clinically in that DRG category.

And actually, in terms of episode groupers of when I was in the commercial health insurance world and we were using groupers I remember one occasion where we showed some grouper results to one of our

specialty committees and the doctors on the committee said well, that's interesting. Now tell me exactly what that condition category is.

And so for things to be clinically meaningful, actionable, et cetera, it would seem that it would be at least helpful, not necessarily something that you would require for the Good Housekeeping stamp of approval, but it might be at least helpful that it would be reasonably clear from reading the condition category label to clinicians exactly what it is that is within that unit of analysis.

DR. KING: We didn't talk about it at that level, but we did talk about it in the context of face validity, perhaps that you would show, and Chris may have to help me out, but that you show clinicians the result of your episode grouper.

You show them the clinical characteristics of that person who had that

claims data -- not the clinical. You show them the claims data from that one individual person with that episode grouper to a clinician and they say yes, that makes sense, or no, that doesn't make sense as sort of a face validity level.

But Chris, do you understand his question and can you help me out here? You were in our group.

MR. TOMPKINS: Well, he probably understands his question too and can restate it if we need to.

I think what he's saying -- let me make a stylized comment here which is that being a non-clinician but observing a lot of clinicians over the last few years there is an art to determining when do you say that there are distinguishing characteristics of a condition that make it different from some other condition, and furthermore, there are distinguishing characteristics within a condition that need

to be identified and kept track of whenever you're trying to purportedly make constructs that are useful for comparison.

And part of the clinical logic of an episode grouper is to have an inventory of definitions as to what constitutes those conditions. And presumably somebody can articulate how that came about. I mean, that's a nice aspiration.

But even more concretely and more immediately the episode grouper should have in fact an objective definition of what that condition is which consists of what we call trigger codes which are individual diagnosis codes, et cetera.

So that anybody who is viewing supposedly the outcomes related to heart failure can actually trace it all the way back to say what set of diagnostic codes would have qualified to call this person a heart failure case in the first place.

DR. KING: Which is why we left

the first part very open-ended, so that people could explain that. As opposed to being proscriptive and say you've got to have specific definitions here. But you have to explain sort of how you got there.

As a clinician I understand that there's an infinity of heart failure patients. It's a continuum within many, many, many little subsets within there.

DR. BANDEIAN: I don't know if this is helpful or not, but let me -- I could give an example or two.

My sort of favorite example would be talking about an ankle fracture. It's fairly simple.

So, should that be a condition episode? Should it be fracture of the lower extremity as a condition episode?

And then even further let's talk about ankle fractures. And there's a difference between a simple lateral malleolus fracture versus a trimalleolar

fracture.

And there would be very, very different resource use implications as well as very different implications in terms of complications.

So it's not just -- it's partially a clinical meaningfulness/understandability/usability concept, but it also goes to the underlying issue of whether the entity that is being looked at actually is homogenous in terms of the expected resource use and the expected rate of complications.

DR. CACCHIONE: We had a lot of discussion around this about when do you -- because there is, you know, your example is a good one.

We talked a lot about a pneumonia, an outpatient pneumonia versus an inpatient pneumonia. And then a pneumonia with a parapneumonic effusion.

And so we talked about should

risk-adjusting or severity judgments, or do we specify populations based on severity prospectively, or does everybody go in as a uniform population with a fractured ankle? And is that taken care of on the back end through risk-adjusting and through whatever data you're collecting through the bundle?

We deferred to Chris on this a little bit because he has more experience, but we really did arrive at the fact that we thought that there is some sort of prospective risk profiling, risk stratification, something.

And that using that risk stratification as best we can up front to identify what bundle, is it 1, 1A, 1B, or 1C around this ankle fracture is something that we believe -- we thought was one of the clinical constructs that needed to be in place in a bundle.

And that needed to be sort of transparent to who is -- by the bundle-maker

or whatever the author to say this is how we use this construct.

MS. MARTIN ANDERSON: One thing that I'm getting a little bit mixed up in this conversation on is it seems to be a mixture of some software features which I think, you know, we can have opinions on, but I'm not really sure whether or not it's evaluable for NQF.

Some principles around some things that you need to know, have submitted about an episode and how the clinical logic is done so that you can make it -- evaluate.

And then also some preferences maybe on how the actual episode grouping happens. And I think it's this third category that makes me the most nervous. Because I don't know that we can superimpose how someone in the -- should do an episode.

But I do think this issue of this is an area where you want some transparency in the application process. And if we could

translate that into this should be explicitly noted in terms of how are you handling --

DR. CACCHIONE: -- done prospectively, retrospectively, but it needs to be transparent.

MS. MARTIN ANDERSON: But it needs to be transparent. And we can articulate you need to be -- this is a question you need to be able to answer about your grouper and justify and test.

DR. CACCHIONE: David, do you have a comment?

MR. REDFEARN: I'm interested in the idea of prospective risk. Because most of the risk adjustment with groupers that I'm familiar with you calculate the risk on the same time period that you do your grouping. So they're happening at the same time.

Now, I do know the PFE model will allow you to go back a year before you

actually start your grouping process to calculate the CRG risk score that you want to apply. Is that what you're talking about?

MR. DE BRANTES: This is Francois. I can interject here. We create prospective budgets in our grouping system, so that's a pure prospective adjustment of the patient's historical cost. But you're actually doing a prospective budget, so you're severity-adjusting and estimating what the future cost is going to be. Which by the way is essential for payment purposes.

MR. TOMPKINS: If I may, I may get this wrong but I think I understood your question.

I think Joe used the word "prospectively" but in a different sense. We did talk about the fact that we want to make sure that there's clinical homogeneity. And that can be done he said front end, back

end. And I think he was using front end and he said prospectively with regard to the taxonomy of the episode definitions themselves.

In other words, you can -- to use Steve's example you could have four different episodes that define four different types of ankle fractures, and if you did it that way then you need less risk adjustment at the back end because prospectively in the taxonomy you have already absorbed, or you've already accounted for by stratification a lot of the severity.

Now, Francois' comment also ties back to maybe something you said too. But it is -- and was part of our conversation too, namely when you are defining the risks that you want to control for when you're making useful comparisons across providers, to what extent do you want to limit yourself to information that is, quote,

"prospectively known" that is before the episode begun or before the period of accountability begin.

And Francois was saying that in many instances people prefer to define the risk, quote, "prospectively" so that the consequences of clinicians actions and inactions and so forth are part of the end result and are not adjusted for in mid --

MR. DE BRANTES: Adjusted for, right.

And again, I mean I think to the comments earlier I'm not sure we should be deciding which approach is best or worst more than encouraging the developers to be clear about what approach they've taken and then delineate its shortcomings if there are any.

DR. CACCHIONE: I'm not sure where you left off.

DR. KING: As far as I'm concerned I'm done.

(Laughter)

DR. KING: He summarized it. Basically it's explain yourself and convince us that this works and that it's valid and reliable and all that.

Did we miss anything? I sense everybody running out of steam. Did we miss anything from your perspective?

MS. SIMON: We didn't elaborate a lot about this in the group, but there was a little bit of a distinction made between diagnosis and procedures. Procedure is a very clear trigger event for lack of a better term.

But I think part of our discussion around this example of pneumonia and really stratifying out the outpatient pneumonia from the inpatient pneumonia from the pneumonia with parapneumotic effusion was trying to come up with a comparable trigger for a diagnosis code for lack of a better construct. Does that make sense?

DR. CACCHIONE: The only thing I would add would be the use case and I think you hit it a little bit. We did try to break it down versus -- the provider-centric versus the patient-centric.

So we talked about the uses of the clinical logic using a more provider-centric use for this with regards to public reporting, payment, comparisons, whether health plans or networks would be using that and performance improvement. We did brush over the attribution.

And then on the patient-centric side it's really about the patients having the ability to compare cost and out-of-pocket cost as well as provider quality.

And so we do think that the uses of these tools will have a little different implications as we start to look at if they're being designed more in a provider-centric or a patient-centric way.

But there is some convergence of

all these things, especially as you start to look, the cost item, the quality item. But they do -- and they do converge, but there are some subtle differences.

MR. HOPKINS: That's sort of interesting, that last point. I'm trying to think of the role of NQF in monitoring or arbitrating how insurers present cost information to their members. Is that something that NQF would be involved in?

MS. WILBON: Can you repeat that question? I missed the last few words of your sentence, sorry.

MR. HOPKINS: I think Joe was suggesting that one obvious use of episode groupers is, and we talked about this earlier, enabling patients to see or health plan members, let's say, to understand what costs they're facing when somebody has recommended a procedure or treatment regimen to them.

And maybe they have some choices

of providers, and maybe those choices in part are related to cost.

So my question was would NQF ever be wanting or needing to interpose itself in that process of determining whether the health plan is providing that information in an appropriate way to its members. I can't see that. But that's really what that use case was if I understood it.

DR. KING: I don't think so. I think that was just a codicil.

Unless the insurer was going to say the average Doctor X cost $10,000 to replace your knee or whatever, and they have a deductible. That would be the only way -- they're not going to. I couldn't see --

I mean, it's really about making sure that the measure is valid, reliable, that the grouper is a good grouper that can really give a good estimate of cost that's not going to encourage -- that's going to discourage fly-by-night companies from

developing them and selling them to insurance companies. That's kind of my thought.

DR. CACCHIONE: Maybe I was misunderstood or I probably misspoke. I don't see it being that granular, at that level.

But I do think that the use of a grouper tool if it is patient-centric should have the ability for -- if it's done in a way that is transparent and is useful can aggregate costs around an episode so a patient can understand what a total episode of cost is.

Now, whether it translates into the NQF needs to be the entity that sort of discloses that. But at least around the idea around an episode -- or a payment model like the episode payment model that it at least allows the consumer for comparison purposes to understand that there is an episode cost that far exceeds the event that

occurred, but there is some episode cost.

So I guess I don't know that -- I guess I don't know -- I would defer to the NQF folks to say. I mean, I don't think NQF wants to insert itself there, but I do think that there is a role for NQF to establish the principles of how an episode tool would work to -- and how it should be constructed to allow an apples to apples comparison. Mark?

DR. LEVINE: Yes, I think it would be pretty difficult and dangerous for NQF to get into endorsing the uses of groupers. That's up to whoever is using the grouper to be able to do that.

I think our role is to look at are the groupers being constructed in a way that enables them to be used in a reliable manner.

MR. AMIN: Just quickly on the NQF note. So, the challenge that we have here is that in typical performance measures

we endorse measures for quality improvement and performance accountability applications broadly. We don't distinguish between accountability applications.

I think one of the issues that we need to explore further, and I know that our subgroup had a lot of discussion around this, is that there still is this overarching fact of the use of the episode grouper may in very clear ways influence the design of the grouper itself.

And so bringing in the use case obviously has to be front and center, but I will say that it's not NQF's goal of this effort to impose itself in particular applications, particularly in commercial applications which I'm not even sure that we quite have the leverage to do anyway.

But it's more to say let's understand the use case in a more detailed way, especially if it has implications for the actual design of the grouper itself.

MR. TOMPKINS: I'm not sure if we want to belabor this or play it out. Because if I understood the question it could be framed something like this. Is NQF indifferent to the use of a grouper -- one grouper versus another.

Let's say one grouper studiously collects likely complication costs associated with an elective procedure and the other grouper doesn't.

The health plan might prefer to use the grouper that does not if it turns out that it's the cheaper providers who have the higher complication rates.

So that if they're trying to give information to the members about which ones to choose that they're very selective about the costs that they're revealing as part of that bundle then it might lead more people to use the cheaper one, whereas there's been a research base showing that consumers are very sensitive to avoiding providers who

have a high share of costs that are related to complications.

DR. MIRKIN: I can't speak for NQF, but I think as far as -- and certainly there's no -- there's nothing that requires insurers or anybody else who aggregates data to use NQF-endorsed measures.

But in fact I think NQF-endorsed measures have become sort of the standard for health insurance or anybody who's going to report on those kinds of areas.

So I see no reason why NQF can't more or less endorse principles for a construction of an episode grouper which I think is what we discussed in our clinical group. Not saying this is the only way to do it, but if your grouping methodology follows these principles and this is something that we think is appropriate for -- it can be endorsed. So I guess that would be up to NQF. You can't force anybody to use it.

The other thing, as far as use case I don't think we were discussing, and the rest of the group correct me. I don't think we were discussing use cases as something that would be required for NQF endorsement, but it was a way of packaging the various criteria.

If you have -- this is your intended use case, and I guess if a developer said we want to use it for all these things then they would be -- the whole set of criteria would be applied against that. I think that's how we were looking at use cases, correct?

MS. MARTIN ANDERSON: Just to make sure I'm getting this right, you're articulating a desire to make sure that the use case disclosed. And then I think you all took it a step further and said and maybe NQF could have a set of principles that would be important to support that kind of a use case. Right? But not a required.

So it's how to walk that line of not telling them exactly how they have to group it.

DR. MIRKIN: To Mark's point, I don't think the group ever was suggesting that NQF would say here are appropriate use cases. But more or less you disclose what use case if you are going to limit yourself to a particular use case and then here are the criteria that NQF would say are appropriate to use to evaluate your grouper for that particular use, be it reimbursement or different kinds of things.

DR. BANDEIAN: This is a different topic from what just has been talked about so I can either stop or continue on. I mean, it's within the clinical domain.

MS. MARTIN ANDERSON: Mark, is yours related to this topic or a new topic? And Jennifer?

MS. HOBART: I'll just quickly close on this. I mean, I think it would

just -- the health plans would benefit from having what expertise there is to help guide. Because there is a lot of both market and political pressure for these transparency things. And it's complicated to figure out so we would like, at least some health plans would like to leverage the expertise without saying it's a mandate on either side if one needs to be done.

And also, to have some consistency of engagement with the provider community so on both sides we aren't going crazy with 50 different approaches.

DR. BANDEIAN: I see that there is some mention of complications there. And again, maybe I missed the further discussion. And we talked a little bit about that in our group.

But complications can add a lot to the cost of a condition episode. And so I would think that one would want to know what the mechanism is for determining

whether a complication should be assigned to a condition.

And in cases where -- I mean, and what's the logic and what is the mechanism, if any, of resolving some ambiguities as to what actually caused the complication and what do you do if it seemed like two conditions might have caused the complication.

And I'm not sure if you'd necessarily want to quite go down this road, but it does seem almost in my mind that addressing complication costs is a requirement in the sense that the comparisons of cost can be extremely misleading if complications are not considered.

But this is also a complicated a topic and it's hard to get it right. So I don't know whether you address that, or any folks --

DR. CACCHIONE: We did talk a lot

about complications versus comorbidities versus expected complications. And we thought there was a real continuum there. And we understood that there was a -- yes, there was quite a bit of variability in terms of the cost as it relates to those things.

And we had a fair amount of discussion. I don't know that we concluded anything, we just really spent a fair amount of time talking about that there were comorbidities, there were expected complications and then complications that were a defect in care.

And that there was definitely a gradation in terms of the expense of care or the cost of that care. And there are implications. I mean, I think that everybody understands that there are implications that you would be paid for comorbidities but you might not be paid for unexpected complications. But we never got

that far down the road on that.

DR. BANDEIAN: If I might just elaborate just for a second if that would be helpful.

So, comorbidity I think has an impact on the expected cost of a condition. So, if a person has a hip fracture but also happens to be morbidly obese they're probably going to have a little difficulty with rehabilitation and the rehabilitation would take a little bit longer. So that to me is kind of a risk adjustment issue.

Now, complications also have some risk adjustment implications, although I think as Chris Tompkins just indicated you might not necessarily want to risk-adjust for the complication because that would be kind of giving the folks a little bit of a pass on the fact that the complication occurred.

So, actually really what I'm focusing in on is a concern that if we do

not have a mechanism for taking into account complications, you know, a post-operative wound infection, et cetera, it might appear as though episode A is less expensive for the same condition than episode B. But actually if you factored in the complications it was twice as expensive.

And so in terms of again the validity concept of when we compare two episodes, putting aside the issue of comorbidity for the moment, and one is more costly than the other is that actually a true statement that makes sense.

And I would submit that if we do not have at least some mechanism of taking into account the complications that are directly related to that condition episode we might be making a very misleading judgment.

So, if one --

DR. CACCHIONE: So that's part of the clinical construct. The complications

need to be somewhere, and that -- we needed to have those complications that are failures and those complications that I guess are less about a failure, more about something that is a known complication.

But yes, we did account for that in the discussion and we think that needs to be part of it.

I think we ought to stop with this. We only have 45 minutes left and we have two other topics to go and we're sort of beating this. And so let's go onto the -- did I miss that?

MS. WILBON: If the group feels like we're done with this and there's not anything else to add then we can move onto the next topic.

DR. CACCHIONE: I wasn't aware that --

MS. WILBON: No, that's fine. I mean, I think we were thinking there might be more discussion, but I know people are

kind of petering out.

DR. CACCHIONE: Well, if there are more discussion points or there are other questions. I think, David.

MR. HOPKINS: I don't understand what episode grouper would not incorporate complications? I can't imagine that. So, it sounds more like an academic issue than a real one.

MS. MARTIN ANDERSON: I think one thing I'm trying to catch up on is this concept of what really is a principle. You know, what is a principle statement.

And if a principle statement for an episode grouper was something like the episode grouper must be able to account for complications. You know, I don't really know what --

MR. DE BRANTES: You can't say that. I'm sorry. This is Francois. You can't say that without then having to define very, very, very clearly your concept, i.e.,

NQF's concept of a complication.

MS. MARTIN ANDERSON: Yes, I wasn't actually trying --

MR. DE BRANTES: And so is the complication a natural progression of the disease which is how some people define a complication? Is it an error? If it is, what type and how do you bind it.

MS. MARTIN ANDERSON: Yes, I think that what I just heard them say is that they're saying the complication can be all of those. And I was just trying to paraphrase and give an example of is that what we mean when we say we want to have principles. And if it's just so obvious --

MR. DE BRANTES: The principle should be you explain what you're doing.

MS. MARTIN ANDERSON: But you have to tell them under what dimensions do you still need explanation.

So there's a couple of different things. One is some principles. The other

is what information do you have to submit. If this just falls under the category of the kinds of things you should have to submit and explain in your application that's what I'm trying to get at.

When we said there are some principles that came out of this clinical logic workgroup what are those principles as compared to -- and do we have clarity on that? If we do then we can follow up later. Versus what are those things that just need to be submitted.

DR. KING: Yes, the principle is you need to explain. Not that you need to define your condition, your complications, but you need to explain it, you need to address it.

And another principle is that you need to pass testing criteria or validity criteria for your measure. And we kept going back to the test cases, or what did we call them, sample cases or something.

So, if really what an episode grouper is is about defining cost of care then you should be able to have the person who had the hip fracture, got the anemia, got the myocardial infarction. And you should still come up with the same cost of care regardless of whether you call that myocardial infarction a complication or part of the disease of that poor old person who happened to have silent coronary disease, got anemia and had the heart attack.

But, so if you go to the testing part of it, the validity, the reliability testing part of it with the test cases then who cares how you explain your clinical logic and what you call a complication versus what you call just a natural progression of disease.

Did I explain that sort of? That's basically what we cardiologists kind of came up with.

MS. MARTIN ANDERSON: Were you

saying that the test -- are you articulating what needs to be tested, or a particular methodology for how to do the testing? Or were you saying both?

DR. KING: We landed on test cases, but there may be a better methodology for doing it.

DR. CACCHIONE: I think we said it was iterative, you know, that it was going to be continually refined. But I don't think that we got -- answered your question specifically.

MS. MARTIN ANDERSON: Okay. So we still maybe have some work to do to say what information would have to be collected as part of an application for endorsement, what are anything that you would tease out as principles that are just information you're going to share, or you know, require to be shared. And then if you have any requirements.

It would be relatively unusual to

specify exactly how something needed to be tested, but I do think we have to wrestle to the ground how do you determine validity if -- on the clinical logic.

MR. AMIN: Our construction logic group touched upon some of these very same topics. So maybe what would be helpful is to walk through a little bit of where we were. Because this is not by any means complete but I think it raised some of the same concerns that you're raising right now.

And I think some of us walked away wondering whether we really have -- whether we're suggesting testing approaches, or we're actually looking at criteria.

So, this may be helpful or it just may add some more complexity to what we needed to do tomorrow. But I'd rather at least put it on the table and then maybe we can address it together tomorrow.

So, I got nominated as the spokesperson.

(Laughter)

MR. AMIN: Although Tom was a quick second so he's going to help out.

So, what we talked about were -- I mean, I have some overarching statements that the group had. And I think our group was probably the most -- well maybe, I don't know, I don't want to speak for everybody but had some of the more innovative approaches here that really suggest that NQF rethink its typical endorsement process and how it relates to episode groupers.

And suggesting that potentially our consensus development process of convening panels may potentially not be the best approach for what we're trying to achieve here.

So again, I want to just tag this because I know we can't have a full discussion of this today but I think we need to give this some thoughtful consideration tomorrow.

But overarching, you know, what I really tried to push the group to work with in the framework and ultimately we push back was that we wanted to first discuss what were the components that we would want submitted, how would one evaluate those components and how would those components potentially vary depending on the use.

And we talked about -- our charge was to look at issues of concurrence of clinical events, complementary services, hierarchies, missing data, things of that nature.

And ultimately what the group came back with was the two really important things that we would want submitted is an understanding of how conditions -- identifying conditions of an episode, essentially the trigger and end mechanisms of the episodes within the grouper, and then the methods by which the claims are assigned to an episode, and all of the steps that

would be required to have services assigned to an episode.

That would include how these tiebreakers are managed, statistical inferences, and all of that should be transparent.

A concern that David raised which was a valid one was around whether even this would be feasible in an environment where we would be evaluating hundreds potentially, depending on the grouper, of individual episodes.

And so that was -- those were the components that we could at least agree on. We discussed a number of others but those were the only ones that we could agree were components that should be submitted for evaluation. Those are quite different than what our charge was so that was by design in some ways.

The second component that we were looking at was how one would -- what we

would use to evaluate this. And essentially we came back to essentially I think what your clinical logic group came up with which is essentially validity testing was the dominant testing approach that we would want to look at and essentially was the dominant criteria.

And one potential approach, again, this is not the criteria, but one potential approach that one could use for validity testing would be to develop -- using a validated data set to use a set of scenarios and follow where the claims were assigned, understand the service assignment episodes and then put the episode grouper in a potentially more complex environment to understand how the groupings changed.

So again, we talked about the two elements that we would want submitted. And then the dominant approach to actually doing the testing was validity testing.

And then on the next slide I

think we discussed at a high level -- Evan, if you can move me to the next slide -- maybe Tom, you can give us a better example of what a complex environment is for moving the episode grouper to a complex environment.

MS. MARTIN ANDERSON: So put the claims in a complex environment. What's that mean?

MR. MACURDY: The sort of thing I had in mind was to first start out with, just to be very concrete, suppose you start with 25 claims so that the group would be able to get a handhold on how various claims got assigned in that sort of world.

And then take that and maybe put it in an environment where there's 250 claims and see how things get reassigned. Because you find very often that they do get reassigned and the question is why.

They can be assigned to a different episode type because they're now

classified as a complication, or that sort of thing.

But the biggest challenge I've always had with these sorts of things is to be able to -- I think it's better to start somewhat simple so you get a grasp of what's going on and then build the more complicated.

So once again, it would be a case of something that's manageable and then putting it in a more complex environment and then see what happens. And then be able to address why it did what it did.

MR. AMIN: So, a few other components that we discussed. Let me just lay this out for the group.

Steve really recommended that we lay out essentially what we mean by validity in this environment. And there were at least two that the group agreed upon which is that the person actually had the condition, and then the services assigned to

the condition are correctly assigned.

And then there are other examples in terms of identifying or determining high- or low-risk conditions as a potential, another area for defining validity.

One key takeaway for me as we were discussing this, and I think this has a lot of implications to how we think about the criteria validity is that there may be no -- the group felt very strongly that there may be no right or wrong output. But the intent of the grouper needs to be clear.

Which in itself is a little bit challenging to think about in terms of validity, at least the way I conceptualize validity which is that there is some truth, or some right that you're trying to move toward. So, that was an interesting characterization of where the group landed and I think has some very clear implications for how we think about criteria.

And oh, so we explored the

question about use. Again, I think adding to the complexity here was this overarching issue around use, and that the groupers may vary based on use and may require flexibility in defining parameters.

And so one of the challenges here is that if you're actually looking at validity and the grouper in a lot of ways has flexibility or the parameters are changing what exactly is it that you're looking at.

And the group felt pretty strongly that the developer should specify the use of the grouper in the evaluation and the parameters and at least the range of those parameters so that we can get a sense of what exactly it is that we're evaluating and that's being tested.

So, maybe I'll turn that back to you in terms of kind of how that resonates with where the clinical logic group was. But I think it still raises some more macro

questions around what exactly is it that we're expecting to be submitted. What is the criteria and what are rising to the level of principles. And I think those are sort of consistent.

MS. MARTIN ANDERSON: Just a couple of quick questions, clarifications. What do you mean by the word "parameters" in this context?

MR. MACURDY: Pretty well every grouper -- a concrete example would be you can often vary the criteria for what starts, you know, what's an open period that starts or ends an episode. But pretty well the groupers had various kind of parameters you can set.

MS. MARTIN ANDERSON: Okay.

MR. MACURDY: And for the most part the way that's been generally handled is to have the -- whoever's submitting the grouper to give their recommended parameters to begin with and then you can see how they

vary.

MS. MARTIN ANDERSON: The user-controlled options that are in the beginning.

MR. MACURDY: There's usually quite a few.

MS. MARTIN ANDERSON: Yes, okay. We talked about that too.

So, if you go back to the previous page I think the one thing that it sounds like you're highlighting but strikes me as the second bullet said validity is services assigned to the condition are correctly assigned but there is no correct.

So, what do we do about that? I mean, you know, so in a sense the bar is can you explain it credibly but not necessarily does it work correctly because we can't define correct, right? That's what I'm hearing. I'm just repeating. I'm not --

MR. DE BRANTES: Well, this is Francois. So, I'm not sure you can't define

some level of correctness.

So, to Tom's point, if you have a preset claims database that has very specific profiles of patients.

You know, for example, that they have certain conditions. You know that they've had certain procedures done. You might know that some of those procedures ended up by having complications such as infections, et cetera. So, you know all of that ahead of time. That's your base claims data set that you transmit to the developer for testing through their grouper.

If what comes out is markedly different from the picture painted of these patients a prior then something is wrong. There's an inconsistency that they should at least be able to explain.

So there is something right about does the patient have -- is there sufficient evidence in the data set that a patient has pneumonia, or is there sufficient evidence

that the patient has diabetes. And if the answer in that claims data set would lead anyone to say yes, but the grouper comes out with no, then there's a discrepancy that needs to be explained.

Similarly, if there are certain services that are very clearly misassigned because -- and Steve's example was an X-ray for an ankle fracture that ends up by being dumped into an episode for pneumonia. Obviously that's wrong, that's just basically wrong.

So, there is some -- there are right answers in some circumstances, but then there's a fair amount of gray area.

An example of a gray area is if a lab test was done for a patient who has two conditions should the lab test be assigned to both conditions or just one. That's a subjective design that at some point someone needs to justify the reason for that single assignment, or for a double assignment.

So, the areas we think that are fairly clear-cut, something is just clearly badly assigned, an episode that should have been triggered is not triggered. And then there's the rest of it which is you're not going to be able to stand up to any kind of gold standard because it's a function of the subjective decisions made by the developers as they designed their grouper.

MS. MARTIN ANDERSON: Okay, thank you. So, I just suggested to Taroon I think maybe we should avoid the word "correct" right? Something services assigned to the condition -- or logically assigned. Just acknowledging that there can be more than one correct that's a design -- that you could at least logically explain. Steve.

DR. BANDEIAN: The exact word, adjective, et cetera, should be.

I think, first of all, I think Francois summarized the discussion extremely well.

Just as another example, just to kind of make the point. Suppose that we have a glycated hemoglobin has a lab test and suppose that it is assigned in the grouper to the patient's hypertension episode but the patient actually also has a diabetes episode. So, it would be really a little hard to imagine why a glycated hemoglobin would be used for hypertension but it would be pretty easy to understand why it would be used for diabetes.

So those are the sorts of things which looking at output one could have clinicians look at and make a judgment of whether the error rate there, whether it would cause an error or not. Whether things are sort of looking basically okay.

And it's actually pretty easy. You just would take all of your hypertension episodes and see exactly what services are being associated with it.

If there's a lot of stuff that

doesn't really make sense related to hypertension then you might at least question it.

MR. MACURDY: Just to emphasize the point further. Francois' point is there's a real challenge comparing across groupers because they do different classifications in what you call an episode. And one's not right and the other one's not wrong, they just have a different organization scheme.

And even within the same grouper you can -- by making the case kind of more complex with more claims something can get reassigned and you look at it and it's reasonable in terms of the way it were assigned. Something that was an episode by itself becomes a complication to another episode.

So, even within the same grouper it can be quite plausible in terms of what the assignment is. And that's what's meant,

that there's not really a right or wrong answer.

MS. GARRETT: I just wanted to elaborate a little bit on what Taroon was saying about that we feel that the typical process might not really work for this.

So we talked about the fact that really episode groupers are software, they're software products. And so there's not a numerator and denominator that's set in time that you can evaluate and then that stays the same for 3 years until the next endorsement process.

And it's complex software. There's lots and lots of elements.

So, one of the criteria we suggested is that there be an iterative process for improvement of that tool. And for the next version to be released and that there's clinical input into that process. And that the developer can demonstrate that that's part of the tool that they're

bringing forth to be endorsed. So we're not freezing it in time. So, that's something that I think is a little bit different.

And we also talked about the complexity of this. And having an expert panel spend a day and a half on this we might not get what we really need in terms of understanding how that grouper works and if it's going to work well enough.

So, we sort of threw out NCQA has a process for certifying vendors that do HEDIS rates. So certifying that they actually know how to take in claims data, apply the right algorithms and produce HEDIS rates. Maybe it's something more analogous to that than a typical endorsement process. So that's another thing we talked about.

And then one other unrelated point is we also talked a bit about how we've been really assuming all day that we're talking about administrative claims data.

But I hope that we can be a little more generic in the language that as EHR data becomes more readily available that that can potentially be a source for episode grouping. Because there's lots of potential use for that within providers and a lot of clinical richness that isn't found in claims data there. So, another point that we talked about.

MS. MARTIN ANDERSON: Just a quick question on this testing methodology. I want to go back to that for a second.

So, is it a reasonable assumption that if you came up with a set of data and you had them run through it that if had you chosen different data you would have made the same endorsement decision? What's the risk of teaching to the test and/or what's the extensibility of that kind of testing? You've probably done it so tell us.

(Laughter)

MR. MACURDY: Well, I wasn't

suggesting that you have the people who developed the grouper to do the scenarios.

And one of the challenges you'll find in the grouping is that you can change -- you can actually change the order of claims and it changes grouping. Now, it's fairly slight, it happens, but it affects the hierarchies.

So there is some arbitrariness that occurs there that you just kind of can't handle. And all the groupers have that sort of problem. So I mean, I meant you literally can just resort the claims and you get a slightly different kind of reorganization of the grouping.

Certainly if you add more claims, even in episodes that are not related to the one you're after you can get a new grouping. So there's a lot of sensitivities. And to try to go through every one of those scenarios is a real challenge.

What we were trying to do is

figure out -- was to give whoever's doing the evaluation a chance to be able to get a grasp of what's going on and say does this look like first round reasonable. And I think that's about all they're going to be able to do at that point. Because it is a real challenge.

MS. MARTIN ANDERSON: Okay.

DR. BANDEIAN: I'm not sure that there was consensus on this although there may have been, I'm not sure. And Tom or others can say whether there was or not.

But in addition to the sort of vignette or scenario scenario where there are sort of by definition a relatively limited set of scenarios one could also run through the system a few million people and look at aggregated results.

So one could look at aggregated results, for example, for diabetes, or for pneumonia, or what have you. And look at results that seem reasonable, things that

look reasonable, things that don't look reasonable.

Now, again I'm not entirely sure that there was or wasn't consensus about this. You know, you could look at how long the pneumonia episode was lasting. If you saw a lot of pneumonia episodes that were lasting 180 days it might raise an eyebrow.

And you could also -- picking up on what we were just talking about a few minutes ago you could look at all of the types of specific services that were being assigned to the episode. So, if one saw a lot of cardiac stress test being assigned to diabetes it might again raise a little concern, especially if the person already had a coronary episode.

MS. MARTIN ANDERSON: The model I think that you all are proposing for consideration is a model that looks a lot like software certification programs. So, whether it's how Meaningful Use is done,

it's how all these various software certification type programs where it's independent testing that is -- usually induces cost on the developer. The NQF wouldn't necessarily be doing all that.

But I get the model now so thank you for explaining it.

MR. HOPKINS: If you did that then you obviate all the discussion repeating itself over and over and over again in steering committees when a measure comes that's based on an episode grouper.

MS. MARTIN ANDERSON: Maybe.

(Laughter)

MS. MARTIN ANDERSON: I know we have to get to public comment. So we have like a minute before public comment. So Tom, you want a last word?

MR. MACURDY: Yes. I just want to note that it's not as straightforward as just doing software certification because there's a lot of judgment calls that are

involved.

But I think the main thing is, and it gets back to Taroon's earlier point, that the usual kind of process is going to be somewhat of a challenge. And you are going to want some analysis done here. Exactly, a process. Exactly.

MS. MARTIN ANDERSON: Okay. So now Evan, are you doing public comment?

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

OPERATOR: Okay. To ask a question please press * and then the number 1. At this time there are no questions or comments.

MR. WILLIAMSON: All right, well I think that wraps up our agenda for today. We want to thank you guys for your attention. We know that was a long day and

I think we got a lot of work done that will help tomorrow as we start to really move towards some recommendations and what we're going to put in the report and move towards. Kind of herding all these cats that we kind of have running around right now.

But we'll be convening again at P.J. Clark's. Our reservation is at 6 but feel free to head over at any point. And thanks again and we'll see you tomorrow morning bright and early again.

(Whereupon, the foregoing matter went off the record at 4:43 p.m.)

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Episode Grouper Evaluation Criteria Expert Panel

Before: NQF

Date: 02-05-2014

Place: Washington, D.C.

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