NATIONAL QUALITY FORUM

+ + + + +

EPISODE GROUPER EVALUATION CRITERIA
EXPERT PANEL

+ + + + +

THURSDAY
FEBRUARY 6, 2014

+ + + + +

The Expert Panel met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 9:00 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*
DAN DUNN, PhD, Optum
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
KAREN PACE, PhD, MSN, Senior Director, Performance Measurement
ANN PHILLIPS, Project Analyst
ASHLIE WILBON, RN, MPH, Managing Director, Performance Measurement
EVAN WILLIAMSON, MPH, MS, Project Manager, Performance Measurement

* present by teleconference

AGENDA

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

9:06 a.m.

THE OPERATOR: Welcome to the Medicare Episode Grouper Expert Panel meeting.

Please note today's call is being recorded.

Please stand by.

MR. WILLIAMSON: Good morning and welcome to day two of the Episode Grouper Evaluation Criteria meeting.

We want to welcome you all to the second day. We had a very productive first day and we got a lot of issues out on the table that we're going to continue to discuss today.

We have a new member in the room joining us, Dan Dunn. He just stepped out. Okay.

We went through a disclosure process yesterday that we want Dan to go through as soon as he is finished chewing. But we will have him introduce himself, and

then, we have four bullet points up here on the slide that I will remind you of.

We need to disclose any involvement in the development of an episode grouper system, personal financial arrangements or affiliations with a specific product or service based on a product, investment in specific products by organization, and employment by or other affiliations with organizations, companies, or other entities that own, develop, or use episode groupers.

So, Dan, we'll have you take it away. Just press the Speak button here.

MEMBER DUNN: Okay. Good morning.

Dan Dunn, Optum, Senior Vice President, Business Solutions. And I have been involved and continue to be involved in the development of Optum's episode treatment groups and procedure episodes grouper. I actually have equity in United Health Group, which owns Optum.

MR. WILLIAMSON: Great. Thank you very much.

At this time, I will turn it over to our Co-Chairs, who will help us kick off the meeting.

CO-CHAIR MARTIN ANDERSON: Great.

Well, everyone looks a little refreshed now, right? That was a really great day yesterday. I think it was really tiring for all of us by the end, but we got a lot out on the table and I think a lot of things to work with for the next iteration for staff.

Just a couple of reminders for today, partly for Dan, who wasn't here yesterday. So, it is just to give you kind of our guidance for how we're interacting.

And then, also, just a reminder for everyone else. We are all here as individuals. So, it is important that, even if we have strong, passionate views about a particular solution or approach we're taking on a product, we are really here to get

diversity of opinion and have people share their expertise, and not really to win an argument over what is the best way to build an episode. We're not really working together to define the best episode grouper. We're working together to try to figure out how would you evaluate an episode grouper and how would you know a good one when you saw one. So, thanks for that.

The other observation is that we have diversity of expertise. And as often happens in committees, just keep in mind that some people will naturally speak more than others, but we will be looking, as Co-Chairs, for people when they put their tabletop up. When they want to make a comment and they haven't really said as much, we'll make sure to call on them first, so we hear from everybody. So, I just want those of you who have more to say to know that we are not passing over you. We're just trying to make sure that we hear from everyone.

So, with that, did you have any other comments?

CO-CHAIR CACCHIONE: No. I think we had a nice dinner last night and there was some conversation about the output of the Committee and were we successful.

I think that we are charged with something that is in a very short period of time what I think is a very difficult task. That being said, I don't think that when we walk out the door today, that the task is nearly done, nor did I think the staff or Christine and I thought we would be done at all.

They see this as more of a brainstorming and really directing and starting to hone-down thoughts on where we are going to go with this. I think there is still going to be a fair amount of work to do, and staff is going to have to winnow this down, take our thoughts, come up with a product. And then, we will have to react to that

product. I think that is where a lot of sort of some of the heady work.

But this has been very enlightening. I consider it a more of a brainstorming and really trying to get a lot of ideas out there. And then, we will continue to refine this process over the next six months.

I think Steve asked me yesterday how did we think we did. I don't think we were surprised by the product that we ended up with yesterday. And I don't want to set people's expectations that we're going to finish this today, just because this is our face-to-face.

MR. AMIN: So, if I can offer a few reflections from yesterday as well, and then, I'll turn it back over to Evan to walk through a little bit of how we are going to structure the agenda for the rest of the day.

So, ultimately, our objective and the charge of the group is to identify what

information developers of episode groupers would submit to NQF, how NQF would evaluate said grouper, and then, potentially, other considerations in terms of use cases and things of that nature that we might want to consider.

I think what was clearly challenging in the conversation that we had yesterday was that we spend across the continuum in terms of different options for different components. We had a lot of debate around different options for different components. But, again, I think the most important thing that we want to identify is what are the elements that would need to be submitted to us.

I think once we go through where this Adjustments for Comparability Group went, we are going to ask that we really follow that structure again and come to some sense of consensus around what are the most clear components that we would need to have

submitted. And then, what are some basic criteria that we could use for evaluation?

And I know that framework has been challenging to some in the room. You know, moving directly to criteria at times, and then, really spending a lot of time around potentially user options or things that are not necessarily even within the grouper's control, for lack of a better term.

So, hopefully, it sounds like many of those people who had some concerns, we have had some conversations at the end of the meeting and over dinner last night that I think we are in a better place today. But, really, that is where we need to get to by the end of the day today.

So, if there are any questions about that, we can talk about that, but I will turn it over to Evan to talk about what the agenda for this morning will be and, then, a very important session this afternoon.

MR. WILLIAMSON: Thanks a lot,

Taroon.

So, as I'm showing you here on this screen -- let me screen-share it for everybody who is not in the room.

We are going to be removing one of the agenda items we had on the original agenda for today, which is the finalized principles and proposed criteria. Again, as we had mentioned earlier, we are still in the brainstorming process and we don't want to really get into the finalization of anything at this point.

So, rather than have that, we are going to continue our review of the principles and proposed criteria, starting with the adjustments for comparability group that did not present yesterday. So, we will start with that.

And then, we are going to revisit the construction of clinical logic and try to find, first of all, some of the similarities. I think some of the groups came up with a lot

of the same principles regarding reliability, validity, some of the things that each group discussed in relation to their module, and then, try to find some areas where we have some disagreement or differences of opinion in those sections.

We will have lunch around noon. And then, we have two hours devoted for the afternoon for the implications of episode grouper endorsement. So, we really want to dive into that section, get a lot of ideas on the table for what it will actually mean when NQF or if NQF evaluates groupers, and what that really means in the whole landscape.

We hope to wrap up by 2:30 and get everybody on their way today.

So, that is kind of how the agenda is going to lay out today.

So, with that, we will go ahead and start right away into the review of principles and proposed criteria. We will go to the risk adjustment or the adjustments for

comparability section and pull up those slides.

We designated Jelani as our spokesperson. He happened to leave the room for a time. And so, with nobody else volunteering, he was volunteered.

(Laughter.)

So, go ahead and take it away.

MEMBER McLEAN: That's funny.

So, I am definitely going to need some help. I was just talking to David this morning, and we were talking about how some of the things slipped our mind from overnight about what we talked about.

(Laughter.)

So, feel free to jump in.

One of the things that we definitely -- well, our topic at a very high level, as you know, is about risk adjustment, case-mix adjustment, what data is input, and what do we get from that.

One of the top things that we

discussed was inclusions and exclusions and transparency specifically with that. You know, what type of data are you feeding into or required to define your cohort, so to speak. And then, once we do that, you know, transparency in what the results are from your test case, a lot of things that you guys talked about in other sessions, about having a test trial and test reporting, we honed-in on as well.

And then, we moved on to, I believe we moved on to that section of a topic of reliability and face validity. And some of the things you talked about, about is it accurate from an external standpoint, does it represent what the original intent for the grouper was, what the original intent was.

And then, man, I've got a couple of things.

MR. WILLIAMSON: You can go back to the risk adjustment. So, we did the inclusion/exclusion.

MEMBER McLEAN: Inclusion --

MR. WILLIAMSON: Yes.

MEMBER McLEAN: And then, we went through the risk adjustment portion.

MR. WILLIAMSON: Yes.

MEMBER McLEAN: Do we want to talk about the risk adjustment portion?

MR. WILLIAMSON: Yes, sort of that. Yes. This was all kind of stream of consciousness yesterday. We were just getting all of our ideas. So, you can see we bruised a lot of paper here on the wall, and we tried to capture as much of that as we can on these slides. So, we will try to talk through it.

MEMBER REDFEARN: One addition about the inclusion and exclusion that we talked a little bit about is the sensitivity of the model and about what happens to the results when things are included or excluded. So, that model sensitivity to this kind of variability we thought was important.

MR. WILLIAMSON: I think the data

fallout was something that we talked a lot about, about range of potential fallout.

MEMBER McLEAN: Yes, the data fallout was definitely where I was going with the transparency. You know, we talked about maybe even having some test funnel report that shows from the beginning at each point the breakdown of data that potentially they lose in the test file.

Because one of the big concerns with groupers is, from the end-user, you always say, "Well, that's not the data I put in. Where is my data now?" So, therefore, we talked about the transparency and that potential loss of data because of the various reasons with the groupers.

Then, we did, you're right, we did move to risk adjustment. We had a debate about should it even be included into a grouper or should it be excluded. Should it be their risk adjustment after the grouper is complete?

And there are various reasons for both sides, from both perspectives. I think we included -- and I can't remember specifically the wording around it, but there were two different types of risk adjustment we talked about. It was actually more of Mark's wording around the population type of risk adjustment, and then -- what did it say? -- the patient risk adjustment, the patient-type risk adjustment.

The population risk adjustment could take place, we felt, within the grouper, and that was fine. But when we get more specific around the patient, there are certain attributes that a grouper in the claims can't account for that you may have to do outside of the grouper, looking at the result.

And then, from there, we went through --

MR. WILLIAMSON: Stratification.

MEMBER McLEAN: We went to where?

MR. WILLIAMSON: Stratification.

MEMBER McLEAN: Stratification, yes. I can't see it.

MR. WILLIAMSON: It's around the corner, but it should be up on the slide here.

MEMBER McLEAN: I don't want to move away from the microphone because no one can hear me then.

Stratification. So, we talked about stratification of the population. And one of the things we talked about specifically with that was, ideally, a grouper would stratify itself by just your inclusion criteria.

For example, if you look at certain groupers, they say you have to have certain diagnoses and procedure code combinations within that. But there is concern, does that stratification suffice for the output or the end-users' desire? And I think the conclusion that we came to is that is more of an end-user responsibility than it is the grouper's responsibility. The

grouper's responsibility is to specify, and going, again, back to transparency, specify what those requirements are. So that the user in the beginning knows what's going on and what they actually have to feed the grouper. So, I think that is where we stood with stratification.

We were actually all over the place. We finished kind of early with our original topics, and we moved to some of you all's topics.

So, then, we moved to -- what is this, Evan?

MR. WILLIAMSON: We talked about scoring, just that the developer would have to describe that they provide an output that enables scoring to occur. They provide the file for the end-user to create a score.

But we are discussing, is this part of a grouper system as opposed to the grouper itself? Yes, it was really drawing a line about where that handoff is between the

grouper and, then, the use of the grouper by the end-user.

MEMBER REDFEARN: Can you go back one slide?

MR. WILLIAMSON: Sure.

MEMBER REDFEARN: I would like to comment about the conditions and sub-conditions one, too, which I think is interesting. I mean, this is an example of why you might want to stratify your output.

So, the requirement is that the grouper would support the end-user's ability to do that, not necessarily that it is built into it. Because we talked about a few examples of which there is sort of a built-in heterogeneity of the condition, but the groupers typically put them into one category. And if you are looking at that particular condition specifically, you might want to drill down and split that episode apart into different types of members, so it can make more sense and it is most logical.

MEMBER McLEAN: So, I think I remember what we were talking about here, Evan. We were talking about to what you said, the grouper just allowing the ability to do different types of use cases with the output. So, not necessarily that the grouper would provide a score. The grouper will provide some output, but it is not that we are measuring it on its scoring ability. We are measuring it on does it provide the capability for the end-user to do certain things they are expecting the typical user would want to use and desire from a grouper standpoint.

And I think those base criteria have to be set. We don't, obviously, know the answer to that now, but we said the grouper would need to be able to report and provide information that it can and it is capable of doing those things. So, I think that is where we were going here.

MR. DE BRANTES: Hi. This is Francois.

Can you guys clarify on that? I mean, why would that be a requirement?

MEMBER McLEAN: Well, Francois, why would it be a requirement that a grouper be able to put the user in a position of flexibility?

MR. DE BRANTES: Well, why would it be a requirement for a grouper to provide certain outputs?

MEMBER McLEAN: Go ahead.

MEMBER LEVINE: I believe our discussion led to the fact that the grouper needs to express what it is capable of accomplishing. So that the end-user knows upfront what the capabilities are in using the output of the grouper.

MR. DE BRANTES: Yes, that makes perfect sense, Mark. But I thought you guys were going more in a direction of saying the grouper has to have certain types of outputs.

MEMBER McLEAN: Oh, no, no. I apologize for not being clear on that. It is

more about, once the grouper expresses what its capabilities are, then it is very clear to the end-user what the output would be to allow those capabilities.

MR. DE BRANTES: Good. Thank you.

CO-CHAIR MARTIN ANDERSON: This is Kristine. I just wanted to jump in, too, because we started to talk about this a little bit yesterday. I think we will have to tease it apart a little bit more today in our bigger session. We started it in our bigger session yesterday, which is trying to get more clarity of when we're talking about preferences for functionality, right, we're trying really not to be prescriptive about what a product should do.

But there are people in the room who have preferences for functionality, and they state them, right? So, what some of the sorting will need to be, was that really just a statement of preference for functionality? Is it something that we need to know in order

to be able to evaluate the grouper, right? And in some cases those overlap. Or is it something that is really just an option that is not related at all to this evaluation process.

So, we will have to sort through that, and that is part of what we are going to be doing today. We are saying, does it affect criteria? Is it information that needs to be submitted for endorsement? Or is this really just a statement of preference? I think we are going to get them all out on the table and, then, we will have to sort through all that.

MEMBER McLEAN: This goes back to the reliability and validity at every discussion, I think, at every topic that everyone had. You know, I think everyone is on the exact same page, that it has to be able to express its reliability and face validity for external validity to the end-user and the actual attempt and goal of the episode

grouper. So, I don't know if I need to elaborate much more on this. We had a pretty extensive discussion about this yesterday.

MR. WILLIAMSON: Well, I think there was one thing that we just wanted to highlight. As far as disclosure of limitations, that is when we talked about there was the example of cancer, cancer episodes, where if it doesn't have staging information, it is not going to produce -- the result won't be as reliable or be as valid there. And so, something where you are able to disclose the known limitations of the grouper was something we talked a lot about.

MEMBER McLEAN: Oh, yes.

MEMBER BANDEIAN: Are you expanding further on the validity or is this the time to talk about validity?

MEMBER McLEAN: Go ahead, yes.

MEMBER BANDEIAN: Okay. so, one way to think about this is sort of what I -- and I may have the words wrong. So, please

excuse me. Sort of construct validity, when an episode is constructed and we say that it is a case of pneumonia, is it really a case of pneumonia. And, No. 2, that we include in that episode all of the care that is related in some manner to the pneumonia.

But another type of validity might be -- and again, I'm not necessarily going to have the right language -- but might be when comparisons are made across providers and the score of Dr. Smith is 1.3 times expected and Dr. Jones is .9 times expected, and therefore, Dr. Smith seems to be more costly than Dr. Jones, is that a valid conclusion?

And so, that I would sort of say is kind of an analytical validity point. So, I am assuming that one would really need to have tests and further definitions for both types of validity.

MEMBER McLEAN: I will go back to your point about the construct. I will start with that, the construct validity.

I would presume in the example you gave around pneumonia, if pneumonia was for some reason the creation of some sort of episode, usually, with groupers, what triggers the event, the trigger event, will define what procedure and diagnosis combination.

Now I think that is where you go with the transparency around the inclusions and exclusions. And to some degree, you're right, you have to assess it. There is some validity around that. Is that a truly a case of what we're saying? But, for a test report, I don't know if that is something that you would put in a test report, so to speak, just from my perspective in looking at it.

But I do see where you're going as far as trying to ensure that the diagnosis that we are seeing and the type of condition we are seeing is true, but can you put some sort of statistical value around that? I'm not sure.

MEMBER JONES: Yes, we talked a

good deal about that. To your second point, that is a large concern, in that when you are facing out and you are saying Dr. Smith is better than Dr. Jones based on this calculation, we were talking about things that could have an impact on the power of that comparison. In other words, what is the likelihood that those differences are due to chance or, more importantly, what is the likelihood and can it be disclosed and quantified of the impact of not having a drug file included in the grouper, for example, or any other data that may have fallen out during the load process?

You know, we didn't solve this problem, but we talked a lot about how it would be very valuable to have some sort of guidance around that. In other words, how do we assess the impact of fallout beyond a certain acceptable level?

MEMBER REDFEARN: First, I would say that you're talking about something that

is sort of post-grouper. You're using grouper information to make an inference about the efficiency of a physician or physicians.

When you are looking at that ratio and you want to get a sense about whether that ratio is accurate or not, there are strategies that you can use. And it would be nice if the vendor talked about that. I don't think it is absolutely necessary, but you can use statistical techniques to determine confidence intervals or something like that.

You could also repeat the measurement across time and say, well, if you add three months of data to your analysis and you look at the same two physicians, do the relative positions stay the same across time? So, there are techniques you can do that, but I think that is post-grouper. That is what I would call using groupers as a foundation for cost efficiency.

MEMBER BANDEIAN: Well, yes and no. I mean, I understand what you're saying

about post-grouper, except that I would, then, say, well, why are we even talking about doing risk adjustment in this? Because risk adjustment is really the purpose. Risk adjustment, I think, is to try to help to ensure the comparability of episodes. And, yes, there are other techniques like stop-loss and exclusions that can be used to help to make one more confident about Dr. Smith versus Dr. Jones.

But, for example, I will just use my sort of simpleminded example. Suppose that we were talking about hip fracture, and the patient happens to also be morbidly obese. It is almost certainly the case that more rehabilitation would be required because of that coexisting condition.

And so, to my mind, I think that that is one of the sorts of things that risk adjustment might potentially address. And so, while I understand that there are a lot of other post-grouper calculations that kind of

go into this, it does seem as though -- I mean, let's just suppose that there were no adjustments for coexisting conditions.

MEMBER KING: Steve, I just want to cut in here for a second because, from the perspective of the group, we're not arguing about a grouper should and shouldn't do around risk adjustment at all. We didn't really have that conversation.

The conversation was, what kind of information should need to be submitted in a form? And as it related to any type of adjustment, the conversation was, if the grouper itself has a way of stratifying the risk, patient risk, through the grouping mechanism itself, that that needed to be disclosed, right?

And we recognize that the measure level, you know, when you are looking at output from groupers, there are often more risk adjustment as you are doing these post-processing services. And we said that's not

really what -- what we really wanted to make sure that the grouper disclosed was how it was handling the topic of risk, whether it was doing it by stratifying episodes and making some episodes marked as more higher-risk groups of patients than other episode or whether it produced some type of a case-mix number that was associated with the episode that would allow someone to do something else.

But what we really were trying to get at was we thought this was an important area of disclosure of what the episode grouper does for the purposes of reviewing for endorsement.

MEMBER McLEAN: Yes, I would add to that, one of the things that we tried to stay focused on is we are not telling the groupers what it can and can't do, like you said. But, in bringing it all together, if you're going to say that you can do this -- this goes back to the transparency in the very beginning -- what does your grouper do? And

then, saying, well, if this is what your grouper does, then give us a report of the reliability of that, the validity of that.

So, if you are going to do a risk adjustment, you don't have to explain your methodology. Give us a test report that you have tested the reliability and the validity of your risk adjustment methodology with some sort of data, so that we can present it to us and we can say, "Yes, this is good" or "No, this does not meet the criteria."

The same thing for your inclusions and exclusions and the funneling that we talked about. All those things, if you are going to say these things that you're going to do within your grouper, then provide a report, that is really what we are saying. We are not assessing what you can and cannot do within the grouper.

Because I agree with you, there are groupers that do risk adjustment, and that's perfectly fine. That meets some end-

users, maybe as yourself, it meets their business need. But there are also users that would not choose that grouper. They would use one that doesn't do a risk adjustment, prefer that, and do like what David is saying he would do, do an additional risk adjustment post-grouper, because they feel like that is more applicable for their business need. Does that make sense?

MEMBER BANDEIAN: This is a difficult topic, No. 1. And, No. 2, I'm still kind of reorienting my thinking to thinking about how NQF thinks about things, which I have never done before. So, I understand that NQF has a special mission and a special approach to things.

Maybe let me just try this on. Maybe this is what you are saying. So, let me try some words that may be helpful, at least for me.

Perhaps you might ask the vendor to explain their thinking about how one makes

sure that episodes are comparable, so that valid conclusions could be drawn from them, and just have them describe their methodologies and approaches to address that.

MEMBER HOPKINS: It is really a wide-open question.

MEMBER BANDEIAN: Well, I am trying to figure out what the difference is between -- I mean, in other words, to me, I would actually ideally like to see some fairly-rigorous statistical testing of a large number of condition episodes. So that in a standard prototype implementation, recognizing that, yes, the end-user can do all different sorts of things, but in a standard use case implementation I would like to see statistical results and have a real solid understanding of whether -- let me put it to you this way:

At the end of the day, the doctors of America will be a little distressed if they are being scored and, then, the methodologists of the world come around and say, "Well,

actually, there are all these problems in the report that is telling you that you're high cost."

So, again, I'm backing up and saying, I don't know where the NQF responsibility begins and ends. Actually, I guess, well, I'm here, so I'm trying to help.

But it does seem to me that at the end of the day the doctors of America are going to be really upset if they feel that the system is not actually giving valid information.

CO-CHAIR MARTIN ANDERSON: I think this is a matter of semantics. I think the one thing we all agree on is that there ought to be disclosure around how this topic. Some people phrase it as that the episodes are comparable. Others will say, how are you handling risk and, if you are, disclose it.

The one thing I will say about NQF's processes, what I have learned over the years doing this is that you can't possibly

anticipate really well what is going to come in the future in terms of what someone might submit as an episode grouper in this case for endorsement, right?

So, it is possible that someone could define, for a purpose of some new payment methodology, that the episodes that they are looking at are heterogeneous in some level, right, and there are some analytics that happen after the fact, right?

We just want to keep the language so that we're not boxing-in the methodology. So, however that is done, let the staff work on that. But I think the point is the same, which is, whatever it is you're doing, either to make the episodes homogeneous or whether you're to account for risk in an episode, we think that is an important element of an episode design and we should find a way to have that in the criteria as it relates to validity and, also, to testing. Is that fair?

MR. DE BRANTES: Yes, and if I

could make just -- this is Francois -- two comments?

So, the first one is, if you look at quality measures that have been endorsed by the NQF and other measures that have been endorsed, they are measures, period. So, you've got a numerator and you've got a denominator, and groups have agreed on the composition of the numerator and denominator. And so, each one of these measures outputs a number, most of the time a percentage or a value.

Groups on the outside, whether they are health plans, provider organizations, medical specialty societies, others in the industry, have used those endorsed measures and brought them together in baskets for various scoring purposes. Those scoring purposes have never been reviewed by the NQF, nor is it the job of the NQF to review those scoring mechanisms.

So, that is very analogous to

you've got a grouper that is designed to create a number because you've got claims that are assigned to a specific unit of accounting, which is an episode. And therefore, it creates a number.

What happens after that with that number, how it is used, how it is assembled with other numbers to create a scoring mechanism is not within the purview of the NQF.

And I would submit that the second point I want to make is that, if the NQF becomes highly prescriptive in what it asks developers to submit for these grouper endorsements, no one will submit anything.

MEMBER LOISELLE: Okay. This is Jim Loiselle.

I want to add to that, if we were to go that route, then, again, as a vendor that embeds ETG and others in our solutions, would that require a vendor, then, to also reach out to NQF? It is just a much broader

potential process than just looking at the grouper itself. What happens to the downstream analytics, results, measures, calculations, that is for the individual user to decide what you do with the group detail, once it has been completed.

MR. AMIN: So, I just want to clarify. This is Taroon from NQF.

The goal of this exercise is not to get into the reporting components or how scoring would potentially be used in various different health-plan-type or for CMS scoring purposes.

The one question I do have for this group, in particular, is it sounded like there was a combination -- you described scoring in some ways as how the observed-to-expected value is calculated. To me, that sounds like the result of the grouper number, as was described by this, which would potentially have some ramifications that would include risk adjustment or some things like

that.

Is that what you are referring to or not? No? No. Okay.

MEMBER McLEAN: Yes, I would just take note, I guess. I am really interested in -- I just want to make sure you understand what we are trying to say, as far as our goal is not to tell a grouper what they can do. I think that would be very complicated to do. I don't think you want to even get into that business.

And with providers, working with the designation program that measures providers, I definitely understand where you're coming from, Steve. However, the groupers we use, it is not their responsibility to make that explanation and that evaluation. It is our methodology. It is our responsibility to make that explanation and communicate to them how they were equitably measured and evaluated against each other in their case mix.

But we don't come back to the grouper and say, "Hey, the grouper did it," because it's not the grouper's responsibility. It is our responsibility. It was the grouper's responsibility to disclose to us what it could go and what that number actually meant that they gave us.

From that standpoint, it was our responsibility to take that number and our interpretation of that meaning behind that number and do a risk adjustment and a case-mix adjustment for our evaluation of a provider.

A lot of groupers, they are not doing it at a provider level. They are giving you a population. They take the dataset for the entire population, and here is your result. So, they are not trying to compare providers in that sense.

MEMBER HOPKINS: While we are in the reliability and validity topic, the concept of testing has come up a number of times, and I'm troubled by this, at least the

way we have been thinking about it. This sounds like we think that it is fairly straightforward to, quote/unquote, "test" an episode grouper.

And the other thought on my mind is, you know, these groupers have been tested widely in the market. That is different from most of the measures that come to NQF, some of which have never really been used before they are endorsed.

So, I'm just sort of wondering, you know, what weight can we attach, if any, to market testing? To me, it is very significant. It has met the test of countless health plans, providers, purchasers, you name it, the customers of the folks who use these tools. And the tools have been adjusted and tweaked to sort of meet the objections that have been raised over time. I just don't understand how NQF is going to test these episodes.

CO-CHAIR MARTIN ANDERSON: Yes, so

our group had this discussion about the market and acknowledged widely that for existing groupers, just like any existing product that is out there actually used in the marketplace, or at least purchased and, then, variable use by the purchasers, there is a feedback loop and there is a refinement that occurs that meets the market demand.

And we also acknowledged in our group that oftentimes it is the market who slows down the advancement, right, that tells the developer, "Don't fix it because you're going to mess me up for all the things I've implemented with it. So, I don't want you upgrading every time."

Now it is different from code sets, which they often want the groupers to be accurate. But, for functionality, risk adjustment, ICD-10, fill in the blank, they don't always want it updated. But that is a market dynamic that takes care of itself.

At some level, you would ask the

question, if that's the case, they're deeply embedded, what is the value of a commercial grouper even seeking endorsement, right, because what does it get? That is something that we started to talk about and stopped.

But, then, we also said, analogous to the measure situation, there will be new groupers that have never been used in the market, right, just like we are having with measures. And what kind of testing would be adequate testing for endorsement in that circumstance? Exactly analogous to measures, where we struggle with this topic. I know NQF struggles with that topic, too, and the Steering Committees struggle with it.

So, I do think there is value to market testing, but I don't think NQF has ever wrestled to the ground how do you intersect usability and use with the testing requirement.

MEMBER HOPKINS: But, if I hear you right, there would be weight attached to

market tests, right?

CO-CHAIR MARTIN ANDERSON: I think we have acknowledged that. I don't know whether we are to the point of saying "Wait" to anything, but we acknowledged this is an issue.

DR. BURSTIN: Let's build on that. Kristine is absolutely right. This is an issue we have talked about for a long time, David. And I am glad Karen Pace is here, our methodologist, if she wants to add anything.

But, you know, there is a lot of testing in a perfect laboratory and, then, there is testing in the real world. And measures, particularly when they come up for maintenance, we do request that we have information about how this is playing in the real world.

So, I think, increasingly, we would love to have measures come to us that have been out in the field and have some real experience behind, rather than ones just

tested sort of in a more perfect environment.

MS. PACE: Yes, I would just add that, even when measures have had widespread use, the expectation is, then, demonstrating reliability and validity should be relatively simple because you have lots of data to run some reliability and validity testing.

So, I think, you know, it goes both ways in terms of the performance measurement, that at least to date we haven't viewed, just because it is in use, that you don't have to provide any demonstration of reliability and validity.

MEMBER HOPKINS: I just think you have got some very big challenges in designing the test for an episode grouper.

MEMBER LEVINE: Stepping back a little bit, this discussion makes me think about the goals for this whole activity. I wonder how this sheds light on, if you are evaluating a performance measure, and the performance measure is framed in terms of an

episode. If NQF has certified whatever sanction, the particular episode grouper system, does that mean that the evaluators of that measure, then, can completely ignore the particular framing of that episode for that particular performance measure?

And I would argue no. Because, as I think I mentioned yesterday, I mean, when you look at an episode grouper, it is not monolithic. You have to look at each definition of each episode, and is that appropriate to this particular instance and particular use, particular measure, or not?

So, I am not sure what insights you are going to get from that or how much help it is going to give. I think one of the discussions we kind of had yesterday was, in fact, perhaps what we are really trying to do here is help the NQF folks, committees and others, who are trying to evaluate the appropriateness of episodes in a particular context, to give them some guidelines, so that

they can make sense, does the episode work here or not?

MR. WILLIAMSON: And I will add from our group here, we did discuss that in our group. I think some of the options we brought up, I guess we will discuss later during our implications for endorsement.

You know, do you create additional criteria for episode-based measures? Are you evaluating the episode grouper? What kind of guidance do you provide to the Steering Committee when they are evaluating episode-based measures? So, I think those are all things that are on the table for this.

MEMBER DUNN: And I apologize, I wasn't here yesterday. But have you spent time more sort of clarifying or defining what you mean by reliability and validity?

And just as an example, and you can think of statistical reliability to the ability to discern differences versus noise. You can think of the fact you are running the

grouper for five different sets of data, and you get the same result.

The validity, to Steve's point, construct validity, clinical face validity, actually, the fact that it has been vetted in the market, which is kind of a squishy validity. Anyway, I'm sorry if I missed it, but have you -- because I think that is going to be important if you include those as criteria.

CO-CHAIR MARTIN ANDERSON: That was raised in our subgroup, and, Karen, I think you provided an answer, described how NQF thinks about validity and the ways that NQF thinks about it.

MS. PACE: Right. So, we have defined reliability and validity in terms of performance measurement. We were talking yesterday specifically about validity. The same thing applies to reliability.

NQF's criteria allows looking at that for the data that go into a performance

measure or in the grouper and at the performance score level. So, for example, with a performance measure, you may be looking at the reliability of the individual data element, or the validity. So, say you're using claims data. A validity test might be looking at claims data in relationship to medical record abstraction, if you consider the medical record abstraction kind of the gold standard or the authoritative source.

And you could do sensitivity and specificity. You know, is the data being pulled, pulling the right cases? Is it missing any of the right cases, as a validity test at the data?

When you're talking about the computed performance score, the score that is given to the provider on a particular topic, the validity tends to get more at a conceptual level. Does that performance score perform in a way you expect? So, if someone is doing well on low readmission rates, how does that

relate to maybe a process measure about discharge planning? What is your expected correlation or relationship, and is it performing in that way?

So, validity of that performance score is really about how confident you are in the conclusions you make about quality based on the score that you have given that provider.

CO-CHAIR MARTIN ANDERSON: And there was a discussion, also, yesterday again about how would you go about testing validity, right, and across groupers? One idea was put on the table that perhaps there should be a standard dataset that all the groupers ran, and you were able to look to see whether or not the results were as expected.

And we talked about the fact that the problem is there isn't a right answer. The question is, you know, we talked about, is there a logical answer, right? So, at least can they explain why it is that their grouper

performed the way it did on a dataset?

And again, this was just ideas thrown out there. Or whether, then, the groupers just do that on their own, as the measure developers do, and then, they submit some set of results that says this is how our grouper performs.

It is highly difficult here because we are talking about so many different episodes and different types of episodes that are all included in one grouper. So, how would you do it or what line would you draw for the developer to be submitting that a committee could reasonably say, "Yes, they've done the testing and we agree that it appears valid."?

MEMBER DUNN: I think you are right, there is some question of the threshold, as to what is enough or whatever. But there is also the question of what the criteria is. Because I can think of clinical validity. I can think of putting a valid and

homogeneous unit of analysis, which is more of a statistical concept rather than clinically it makes sense.

So, I just think when you get to the point of defining these things, maybe this is a followup, but I think you need to be more structured on these concepts about what you mean.

CO-CHAIR MARTIN ANDERSON: If there are specific ideas people have for criteria, I think that is part of what we are trying to get at, right?

MEMBER DUNN: If I could just say one more thing? I think we kept talking about risk adjustment. Again, I have only been here for less than hour, but I think maybe the term is better to think of it as risk assessment, because risk adjustment is taking sort of a measure or risk assessment and applying it in a measure. So, case mix or risk-adjusting a provider's results using results that that grouper can tell you in terms of the level of

risk or the category of risk that something falls into. Just sometimes I think we mix those two birds, and it can be troublesome.

MEMBER McLEAN: So, we also talked about feasibility. And this was an interesting topic. Like I said before, everything ties back to the very beginning when we talked about the bigger concern with groupers is, you know, the volume of data that you start with and end with. And in the feasibility, do I have enough sample size for what I need to do?

And this goes back to even groupers being able to accommodate many conditions and many groups all at once. And it performs at different levels within each one.

For example, a grouper may look at hypertension and it may look at diabetes. It make look at COPD. And it may perform as far as different levels, based on its criteria, and it may give a different output. It may

give 30 percent for one, for hypertension, 75 percent from the episode for diabetes, and so forth and so on.

So, we need to be able to figure out, is the grouper and the algorithm, what is going on within it, feasible for the analysis in the end, and the number of episodes that I start with? And that goes back to that final report that we talked about before, being able to provide in their testing -- this is the one thing I do think is tangible in their testing -- to show at each step that you have, each high-level step that you can explain and put a definition around, can you show, when you are testing, the proportion of data that is being lost in your test runs? And obviously, you will run more than one test run just to make sure that it is repeatable and it consistently happens.

But I think that is very important. I think the group agreed that that was very important for the end-user to be able

to understand, and it is very important to say that this grouper actually is working from a value standpoint.

Because Kristine brought up very well that you could communicate that to the plans, I mean not to the plans. Excuse me. You can communicate that to the end-user, but at the same time my response was that you can tell a user that they will lose data due to membership eligibility, continuous enrollment. But the number that they have in their head would never be as large as what it really is. They will think they will lose 10 percent of their episodes, when, in reality, they may lose 40 percent of their episodes or 50 percent of their episodes due to various reasons.

So, a test file or a test report demonstrating that range, so to speak, would be very helpful from an end-user standpoint. And I think the NQF can evaluate that and say, "Well, this is good. This is a good range.

This is an acceptable range for an episode grouper."

MR. DE BRANTES: Hi. This is Francois.

Can someone expand a little bit on that because I'm not getting this particular point?

CO-CHAIR MARTIN ANDERSON: So, Francois, I think the conversation was about there are certain, for each grouper, there are certain events or circumstances that would cause data loss, right, whether it is whatever requirement you have for continuous enrollment, whatever. Another example is maybe clean periods. Another example was, what if you are missing pharmacy data? So, there's multiple different ways that the developers know when they are working on their grouper that there would be an exclusion of a set of claims.

And so, I think the comment within our Working Group was, part of knowing whether

or not it is feasible to use a grouper is at least taking each of those risks and saying, "What is the range of claims that one might lose in this circumstance typically?" And I don't think there was an answer on how someone would get to that because the datasets all differ. But it was just this sense of this was an "aha" for folks after they had started to use this grouper, that they found that it wasn't feasible or usable in the end because it turns out that they didn't have a dataset that would support that particular grouper.

So, it was really about transparency in terms of what are those impacts on feasibility or usability of the grouper system. And that came from MD users really more than developers in the room.

MR. DE BRANTES: Okay. So, just playing it back, having clarity on the impact of potential missing fields or other data elements that can impact the total number of episodes that get retained in the outputs?

CO-CHAIR MARTIN ANDERSON: That's right.

MR. DE BRANTES: Okay. Conversely, I think the group, though, the developers need to be also potentially clear about the risks involved in taking in episodes that might end up by being fragments.

So, for example, eligibility is a good point. If you relax enrollment eligibility rules, then you can end up with folks who have episodes that cover four months as opposed to nine months or as opposed to twelve months.

CO-CHAIR MARTIN ANDERSON: Right. The same idea.

MEMBER McLEAN: Yes, that's a good point, Francois. That is the same idea. I agree with you. We agree.

MEMBER MIRKIN: So, this is Dave, Dave Mirkin. There are so many Daves here.

So, I have a question for the group. Are we saying that -- I think this is

what Kristine said -- is that what is important, and Francois said, is that there be transparency to the potential users on these issues, not that they are going to be criteria for NQF endorsement. Am I correct in saying that that was what the group was thinking?

CO-CHAIR MARTIN ANDERSON: I think the group did not differentiate, sort of starting out early, between what are the things that they felt like we needed to have versus what should actually be in criteria. So, it started with here are some things that ought to be disclosed, right?

So, there were no thresholds discussed that say, "Hey, if you had a circumstance like this," for instance, a grouper that really doesn't work without pharmacy data, well, we all know they all work less well without pharmacy data.

So, it was really a matter of disclosure. Really, we didn't get into whether or not there was an appropriate way

you could use that in a criteria of saying this grouper is not feasible. I mean, again, they are all going to have loss of data, and some of it is to protect the user.

MEMBER MIRKIN: Right. Because I just think it is okay, as a developer, I think it is okay to have -- I think most of us who are developers, this is what we provide to somebody, or at least they demand from us before they purchase our tools or products. But I guess new ones, maybe it is good to specify that.

But I think it would be almost impossible -- I mean, I look at cost/license fees; it depends on what you buy, right? I mean, I am just saying, if you buy a billion products, if you buy it with consulting, if you buy it with -- I mean, I just think it is one thing for us to be able to say, you know, you need to be able to disclose everything upfront, so there aren't hidden fees down the road, which I think most of us who are

developers already know that that is the kiss of death, right?

But, anyway, that was it.

CO-CHAIR MARTIN ANDERSON: I think we are going to get at those issues in the future, right? Don't we have another -- we have a whole market implications conversation.

But, normally, within feasibility for measures at least, because, again, measures are not software systems. These are different. I mean, they can be embedded in software systems, but they are not, in and of themselves, a software system.

There is a requirement that there be at least disclosure if there is a cost to somebody who adopts an NQF-endorsed measure. I think we are going to wrestle with that a little bit later today, of how parallel is that when you are talking about software systems where, in general, there will be fees. I mean, even open-source systems are not free.

Nancy?

MEMBER GARRETT: I just wanted to add a little bit to that cost/license fee and the feasibility discussion. You know, one of the reasons we're here is my understanding is that eventually NQF is going to be asked to endorse the CMS publicly-available grouper.

And from a user perspective, cost is certainly one aspect of feasibility, but, also, how accessible is it really to the public to be able to use? So, for example, I work at a small safety-net provider. We downloaded the code for the CMS HCC Risk Model, and we implemented it with our Electronic Health Record data. And it was free to us, other than our internal resources. We really wouldn't have been able to invest in a more expensive market system. Those are really aimed a lot at health plans.

But I encourage us to think more broadly, both about the data sources -- you know, we have been talking a lot about claims data, but providers need to do this work, too.

And so, can the groupers be built to handle EHR data? But, also, feasibility from a perspective of, is the information there in a form that is easy to use and implement?

CO-CHAIR CACCHIONE: David Hopkins.

MEMBER HOPKINS: I am a little troubled by the discussion around data and missing data because it doesn't strike me as a property of a grouper. It is the property of the database to which it is applied.

And in terms of what could be expected of the vendor or the grouper, I think, you know, all of the strictures around make sure that your data are consistent and complete, and all of that, should be provided as part of the user interface with the system. But I don't see how we could consider, you know, what data might be fed into it. You can feed any data into these things, and they will produce an output. That is a user use issue.

CO-CHAIR MARTIN ANDERSON: The

point that the group was making is, is the grouper articulating what elements of its design will throw data out, right? So, for instance, now some of them have user-controlled toggles, right? So, it's all over the place.

But where are the places in the design where data, those episodes will be kicked out or those data elements will be kicked out, if it doesn't meet a certain criteria?

MEMBER HOPKINS: So, for example, is it the case that, generally, where you feed the database where there is some missing pharmacy data, is the grouper supposed to know that?

MEMBER McLEAN: No, no, the grouper is not.

Can you hear me?

They are not. The grouper will -- for example, we go back to the pharmacy data; we can go to the member eligibility issue.

There are very valid reasons why things could happen, but there are also reasons why groupers throw out episodes.

For example, if they are looking for a condition and a procedure code combination, but they may drop certain episodes due to certain outlier-type episodes that are higher-risk because they want to normalize the populations. That is something the grouper has embedded in its code. It is the data you're feeding it. It is not an issue with your data. It is more about that combination.

And so, that needs to be disclosed, and then, maybe a potential range in their testing. Because a lot of times the summary page that most of us read about groupers, the one-pager, it doesn't specify that. And so, it needs to be disclosed, so that you can say, "I understand that I do have this range of population." For a physician or someone, they can see, well, that range seems

reasonable about this type of patient or this type of condition. And then, they may understand the output a little bit better. And I think that can be disclosed.

We talked about data-sourcing. I think where we are going here is what type of data that is needed. Am I correct, on the right path here? I think this might have been when I stepped out.

MR. WILLIAMSON: We listed that on our feasibility. Again, it goes back to what Nancy had mentioned earlier, just about are we going to implement EHR.

MEMBER McLEAN: Different types of data sources that may be available --

MR. WILLIAMSON: Yes.

MEMBER McLEAN: -- to use groupers? I think this might have been when I stepped out. But I definitely can relate to her and this topic about you don't want to build groupers, I mean, you don't want to build criteria that is focused solely on

claims because things may evolve. We expect groupers to evolve. We expect them to get better. It is also a competitive market. So, they are going to build on that.

And what Nancy alluded to is that providers are using this as well, and they want to use other data sources. And I am sure payers would like to as well, and the other stakeholders would like to.

And so, there are data sources. They are not saying the grouper is just by data availability at this current moment or just haven't been built that way so far, but they may be. And you don't want to build criteria that limit what a grouper can bring and what a grouper can do.

And I think this was the key point here, and I think it is very important. I don't know the answer and the solution to that. You know, how do you look at something that you have never seen before? But I do think it is very important and you have to

take serious consideration into that.

CO-CHAIR MARTIN ANDERSON: Yes, not all these sentences make sense, but we did totally --

(Laughter.)

This is what happens at the end of the day. I think that literally got lifted off of a slide, I mean, a thing on the wall.

I know the one thing we really talked about here was this current or planned use, right? Who is currently using it? What does your installed base look like? To your point, David, about where is this actually actively in use or where is it planned to be used.

MR. WILLIAMSON: And one thing I will add to that, and it is on our white sheet over there that we kind of list on here, the one think we talked about was the quality signals. You know, should an episode have occurred? And it is something that might be outside the scope of the usability and use,

but it is something we wanted to at least cover as far as an episode of prevention, an episode of care management. You know, how are those considered in this?

CO-CHAIR MARTIN ANDERSON: There was also a conversation we had about gaming.

MR. WILLIAMSON: Yes.

CO-CHAIR MARTIN ANDERSON: Again, we didn't know what to do about it, but that there is a susceptibility to gaming, if you know it is going to be used particularly in a payment scenario, where you can start to see severity float up, et cetera. They were adding codes. People talked about what happens when you add additional diagnosis codes and how it changes the severity. All these are covered in the literature. There was no answer to that. It was just a comment that was made around that this is a real issue.

MEMBER HOPKINS: Could you say more about the second bullet?

CO-CHAIR MARTIN ANDERSON: I think that was Jelani's point about planning beyond claims data.

MEMBER HOPKINS: I mean, is that something that a vendor of a current system built on claims data is supposed to provide some statement about how you do chart data or something?

CO-CHAIR MARTIN ANDERSON: No, again, these were just comments people had about usability and use into the future. These were not, you know, shaped into criteria. It was just various conversations.

MEMBER McLEAN: No, I don't think it was directed at the developer and what they need to submit, but more about NQF and how they need to design the criteria to be able to account for different types of data sources, different types of groupers that may account for different types of data.

Because not only are we looking at how we are going to evaluate the development

of the grouper, but, also, how is NQF going to set the framework to evaluate these groupers. And various groupers, various types of groupers may come in.

And what we were thinking about was the future, you know, and are we going to be prepared for the future. It may happen next year. It may happen in five years or ten years, but we don't know. But the question is, are we prepared for that?

MR. WILLIAMSON: It was also mentioned about the episodes of prevention or care management. It goes back to the payment scenario where, if they are being paid on episodes, how does prevention and care management fit into that framework? Again, not in a criteria sense. This was toward the end of our session where we were just trying to get everything, get all of our ideas out.

MEMBER BODYCOMBE: Can I add to that? You know, I would argue that we already have a healthcare system that emphasizes with

activity. So, we pay for people to do things, which is a very non-preventative thing. It favors specialists. It favors extensive care.

Once we implement a payment system based on episodes, let's grind up those episodes and have as many as possible. Right now, attribution systems favor specialists. The specialists are going to get lots of money out of this. They are going to be very happy, and it is going to be counterintuitive to what we are trying, what might really help drive costs down, which is prevention and avoiding episodes entirely. So, that is part of that point.

MR. WILLIAMSON: And again, yes, this was framed, this discussion was kind of framed around our current usability and use criteria. And one of those is progress towards achieving the goal of a high-quality, efficient healthcare system. So, I think, again, the link isn't quite there on these slides or even, I guess, in our discussion,

but we just, again, wanted to get these discussions out on the table, you know, these things out on the table.

So, move on to the next one, which we were kind of wrapping up with where we had a discussion about acute versus chronic episodes. And then, Mark brought up the idea of a claims analysis and how that differs from actually an episode, where if you do a 30-day lookback on an operative episode, is that actually part of the episode or are you doing claims analysis?

So, I think this is the type of stuff where we started really getting high level about defining an episode. I will let somebody elaborate on that, but I think that this was more, again, of just trying to get some of these topics out on the table.

MEMBER DUNN: So, is that criteria or that just the discussion topic?

MR. WILLIAMSON: Just a discussion topic.

CO-CHAIR MARTIN ANDERSON: These were just discussion topics, things that came up.

MEMBER DUNN: Okay. Thank you.

MR. WILLIAMSON: I guess we will kind of try to wrap this up here.

One of the last things that we wanted to talk about was the maintenance, and something that we really hadn't talked about before was -- and again, David Redfearn brought this up -- is the cost of implementing a new version of a grouper. I know we kind of touched on this earlier.

But the process for keeping, you know, if we do endorse groupers or if we do collect this information, what is the process for keeping that current? Does it go through an annual update process? Do we need to set minimums for the length of time? What if there are new guidelines? How do we handle that process is something that we will need to discuss. I guess that kind of falls under our

implications for endorsement section. But is there a difference between logic and features of a grouper that cost a lot of money to implement versus a mapping table or something that can be implemented just as switching out a data file or something? So, again, just more topics that we want to discuss beyond the submission elements and criteria, but things about how you keep this current.

CO-CHAIR MARTIN ANDERSON: I think this is a big deal because maintenance of endorsement today for a measure, because it is narrow, right, either someone can raise an issue and say this measure no longer reflects current practice and it needs to go through an immediate maintenance cycle or not be endorsed while it is being fixed, et cetera.

But, when you have got something like an episode grouper that is trying to use the totality of all of the data, well, there would be an emergency maintenance all the time, if it was required to stay up-to-date

constantly.

And so, how to think about a framework for maintenance of endorsement for episodes needs to look very different than a framework for measures. And that is really something that we didn't really delve into in detail to say, hey, this is an issue.

MEMBER REDFEARN: The specific example is NQF might have some motivation to encourage constant improvement for these models that are risk adjustment, that are clinical logic. And there is a value to that.

But, then, if the vendors do that, and the vendors roll the products out to their customers, they are going to get pushback from the customers. And I can tell you, I was on the product enhancement evaluation team for Optum and Symmetry product. When they would talk about a major new release coming out, you would get WellPoint, United, Aetna, Cigna, Humana, all standing up in the room and saying, "No, please don't because we can only

implement one of these changes like every two to three years. Literally, there is so much complicated IT process, there are so many downstream processes that depend on this source, when you change it, we have to change everything."

So, the practical thing is the vendor comes out with a new version, and nobody adopts it for two years. So, there is this dynamic in which you want to improve the product and make it better, but it is not going to get adopted.

MR. WILLIAMSON: Okay. Do you want to break? All right. I taketh and I giveth back; there we go. Come back at 10:30, 10:40. All right, 10:40.

(Whereupon, the foregoing matter went off the record at 10:18 a.m. and went back on the record at 10:42 a.m.)

MR. AMIN: All right, we're going to get started.

So, in terms of a time check, what

we are going to try to achieve in this next session before lunch is to try to understand where there are areas of consensus in terms of the charge of the group, which is around the submission elements that NQF would want to see, if it was to evaluate an episode grouper and potential criteria that would be related. Obviously, there's a lot of differences of opinion here, but we wanted to at least capture some of the areas where it appeared to be that there was some consensus.

So, we are not going to be updating this list as we go. We just wanted to walk through it at a high level. Maybe, Kristine, you can help me with this. I will sort of leave it with Kristine to kind of walk through some of these elements. And then, we can talk through the actual criteria.

Again, the goal of these submission elements is not to predefine what should be in these elements or preferences, but what type of information would we want

developers to submit.

So, I will turn it over to Kristine, unless, Evan, you have anything else. Oh, he's not in the room.

Kristine?

CO-CHAIR MARTIN ANDERSON: Okay. So, this is a summary that Taroon and the rest of the staff put together while we were talking. So, there they are showing their talent again.

They just pulled together what are the submission elements. So, this is meant to answer the question, what would you ask a developer to submit as part of the form at a high level, though the details can be worked behind the scene?

So, just to go through it quickly once all the way through, and then, we will come back and talk about each major bullet.

So, inclusion and exclusion criteria was really around information, how the data is used as it feeds into the system.

We can talk about where it really needs to fit in this.

Exactly how services are assigned to an episode. So, a full description of the steps and the tiebreaker logic and any statistical inferences.

How the grouper is handling the issue of risk, the assessment issue that was discussed.

This transparency element on fallout of data. How is data lost? How can data be lost? What are the elements of the design that would push data out or episodes out?

How the product is maintained. So, it is not really keeping endorsement current, but the process of keeping the episodes current.

What the current and planned used. Who are your users and what is the planned use?

And then, information on testing

broadly. And we will get into in a few minutes what we might mean by the criteria for adequate testing.

So, first, are there major categories that are missing? Mark?

MEMBER LEVINE: I wonder if a section on input requirements, that if you want the grouper to work, here's what you have got to feed it with. Sort of like the Little Shop of Horrors, it constantly needs --

CO-CHAIR MARTIN ANDERSON: Yes, so what's the data that is required to run the grouper?

MEMBER LEVINE: It constantly needs to be fed.

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER LEVINE: Right.

CO-CHAIR CACCHIONE: I'm sorry. Is it enough to say, what is the input required? Or do you have to disclose what the inputs are into that?

MEMBER LEVINE: If you want an

output of "X", you've got to supply it with "Y". And so, the grouper will take it from "X" to "Y" or from "Y" to "X", but you've got to be explicit in terms of what the grouper needs in order to give you a given product.

MEMBER DUNN: Actually, I agree with that. I would maybe add to it. It gets a little bit into your fallout of data. But there is a sensitivity of all these methodologies to how complete your input data are and such. So, I think there should be some sense of sort of required elements, maybe optional, and then, for any given element, some of the key things you've got to make. You know, obviously, complete diagnostic coding.

CO-CHAIR MARTIN ANDERSON: Maybe the data validation rules or the --

MEMBER DUNN: But it is more this sensitivity of the results to the inputs.

CO-CHAIR MARTIN ANDERSON: To the input? Okay, I think that does fit under the

transparency on the followup, yes.

MEMBER DUNN: Okay.

CO-CHAIR MARTIN ANDERSON: We could expand that to its overall, the sensitivities to the data.

MEMBER HOBART: So, I think it is basically the same thing. But under transparency, I think if you could add the end-to-end sort of data flow. So, what does a user need to put into it? What data profiling does the tool, then, provide? Then, what logic is applied to inclusion/exclusions to the data, and how is that documented? So, I think just putting that all into the start-to-finish data flow would be a way to handle it. Incorporate all that.

CO-CHAIR MARTIN ANDERSON: Okay. Other additions?

MEMBER DUNN: Dan Dunn again.

So, where is the validity and reliability? Is that on the next slide?

CO-CHAIR MARTIN ANDERSON: We will

talking about it on another slide. Right now, it is only a little smidgeon of it is into that, and how is the grouper tested and test results. So, we will be talking about validity and reliability even more on the next page.

MEMBER BANDEIAN: So, by testing, you mean what validation has been performed, what reliability testing?

CO-CHAIR MARTIN ANDERSON: Right. Reliability and validity, and both clinical validity and, also, construct validity, base validity, right.

MEMBER DUNN: Okay. So, that is under your tested, the last bullet?

CO-CHAIR MARTIN ANDERSON: Right. Testing and test results.

MEMBER MACURDY: So, I guess there are two things I wanted to mention. One is -- and this was discussed yesterday as well -- is there is often kind of a presorting or pre-organization of the data that is done. It is

kind of the same as inputs, but it is often done kind of without the group. You know, it is like it is separate. It is not just plain claims; it is grouping of claims. And there is a whole variety of ways to do that.

And that is something that has got to be kind of done as a separate step. When you say "all these steps," it is in there somewhere, but I think it is worth fleshing that part out.

And then, the other part, I mean, if we are trying to get something that -- as I expressed yesterday, one of my biggest concerns is I really know how difficult it is for a group who is not very familiar with a lot of the -- even if you are familiar with the intricacies of how to get a handhold on what the grouper is doing, and does it do something I would expect it to, I think you should have something there that has scenarios that maybe it is not provided by whoever is submitting it, but is provided by the group.

So that there is some intuition built up as to, does it work the way I expect it to work in a simple case to start with? And then, put it in a more complex case. Something that gives whoever is doing the evaluation assessment some handhold as to what it does and why it does what it does.

And I can't overemphasize that. If you don't, I mean, those of us that are even familiar with this, it takes forever to really get into these and get a sense of what is going on. So, you've got to give people as much of a chance as possible, so that there is a simple case they can think of first, and then, try to make it more complex.

So, anything that could be done there, and while you kind of put it under here, this is going to be a case where it kind of has to be done in a couple of stages, where whoever is doing the assessment has to first kind of think about it for a while and, then, figure out this is the sort of case I would

like to look at, or something of that nature.

But it is something that is motivated by the Committee, not so much motivated by --

CO-CHAIR MARTIN ANDERSON: Right. So, what you're describing is something quite different from what NQF does, but you have been making a good case for it, which is that, instead of the normal process, the current process, which is that the submitter would send in their results of testing where they are assuring the Committee and proving to the Committee they have tested and gotten adequate results themselves, you're also saying that the Committee will have difficulty with that, and the Committee would be better served if they were to ask all of the groupers that are under consideration for endorsement to do something common --

MEMBER MACURDY: Yes.

CO-CHAIR MARTIN ANDERSON: -- both something simple and something more complex,

that the Committee could, then, evaluate the results of?

MEMBER MACURDY: I mean, an example would be where there are a couple of very simple scenarios that members of the Committee are comfortable with. And that may vary depending on the composition of the Committee. And then, what is handed to the group that is doing the submission is that scenario in a more complicated environment. And if it doesn't pop out the way you expect it to, then ask the question, "What happened?"

Something that gives a very concrete anchor, so you can say, well, if this scenario didn't pop out the way I expected with 25 claims, it is different with 100 claims or 500 claims. Why is it different with 500 claims?

You might even, then, have, yes, whoever is doing the submission then run it with just the 25 claims to see if it works then, something of that nature. But something

where whoever is doing the submission doesn't exactly know what everybody is looking for, so that the Committee does and they don't. And that's going to help because it anchors; it gives you an anchor. That is what you are really looking for, some kind of anchor.

CO-CHAIR MARTIN ANDERSON: I think the NQF staff has heard this. So, I think that they will be weighing between something like that or just saying, "Look, the developer would have had to do it." So, in this case, you know, for the role that you're playing at CMS, CMS would just submit what you did, right?

So, the question is, what do you really want to require of the developer? I think you guys have captured these two different options, right? Okay.

Okay. Others?

MEMBER DUNN: Dan Dunn.

Are we looking at missing elements? Do you want to talk about --

CO-CHAIR MARTIN ANDERSON: We are looking for comments, anything you want.

MEMBER DUNN: Comments? I think the fallout data one could be a lot more specific. I guess, what is even fallout? So, one could be that services didn't group and they weren't able to find an episode. But that gets tricky. Because if you have a comprehensive clinical breadth grouper like there are on the market, what drops out are things that just you couldn't find a clinical spot for them anywhere, which is interesting and important.

But more important is, did things not group that actually should have grouped, and did they group to the right place? If you had sort of a diabetes, heart failure, joint degeneration of the back alone grouper, what are you going to do with what falls out? Because there's going to be 90 percent of the records aren't going to find a place to go.

That would just be one comment.

What that means depends on the context you are putting it in.

Second was I just would distinguish that between sort of fallout of episodes. So, episodes that were validly created and, for whatever reason, were either not included in output -- but just as a note, a lot of these systems we are going ours, including ours, will not throw anything out, but they will actually mark it as incomplete or a financial outlier. And it is up to the user to decide what to do with it in measurement.

So, I just have a little trouble with that bullet in general.

MEMBER MACURDY: I mean, Dan, yesterday we talked about those kinds of complications a fair amount. And even just the exposition you went through, they are even more complicated than that, as you well know.

To try to figure out how to handle that, do you have a suggestion as to a

systematic way to kind of organize how you do that? Because I'm not sure I do. Because there are so many different combinations and so many things that can happen.

MEMBER DUNN: Yes. Make a distinction first between records that didn't group. And again, you need to understand the context of why they didn't group. It could be because there was no place for them to go.

MEMBER MACURDY: Yes, I think that is less of a challenge --

MEMBER DUNN: Okay.

MEMBER MACURDY: -- than the ones where you will change the environment a little and grouping gets changed.

MEMBER DUNN: Actually, I would have that one -- I think reliability came up. I put that one more in the kind of predictability of the outputs, given the methodology rather than fallout.

MEMBER MACURDY: Well, I don't know if I would call that reliability. I

mean, your grouper will do that. If I have 25 claims and I have 200 claims, the claims will get regrouped, often for very logical reasons.

MEMBER DUNN: But I don't know why you would drop the claims and --

MEMBER MACURDY: No, they are not dropped. They are grouped to a different spot. They are grouped in a different way.

MEMBER DUNN: Right. So, maybe that is less, again, I think that is less fallout data than the --

MEMBER MACURDY: Well, I think that is what they mean by "dropped" here. I don't think they mean that it went -- I mean, if you are looking at a particular episode, not the whole grouper but a particular part of it, it could be dropped in the sense it went somewhere else. It doesn't mean it is dropped because it never got grouped.

CO-CHAIR MARTIN ANDERSON: I think we get the issue, right? And so, as they try to write it up, we will all get a chance to

edit and see if it meets our needs.

MEMBER DUNN: I agree transparency is important and letting people know why something did. But, if you have a criteria that someone was trying to make a judgment based on what fell out, either episodes or claims --

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER DUNN: You just need to understand.

CO-CHAIR MARTIN ANDERSON: Exactly, and I think right. This is more, I think, about understanding how the grouper works than about setting a criteria that says it can't drop data like that. Okay.

MEMBER DUNN: Right.

CO-CHAIR MARTIN ANDERSON: So, why don't we go to the next slide? Do you have control? Oh, Evan? Okay.

So, then, we tried to narrow down the dialog that we have had around each of the categories that are in the criteria. So, for

example, there is importance, then scientific acceptability, which includes validity and reliability, feasibility, and then, usability, right? So, we took that construct and took the conversation from yesterday and today. We will do each one one-by-one, okay?

So, under importance, the only area that importance was really raised was in saying, if you're trying -- this is the scoping-in, right? You're trying to say, should NQF evaluate this submission from a developer? And on the measure side, it is defined as an important-to-do.

I think it was Francois who threw on the table, well, maybe it could be on the number of lives covered or the number of dollars spent that is covered by this grouper.

We wanted to raise the question, should we even have an importance criteria? Because if you do something like that, what you would exclude would be narrow groupers. Let's say somebody developed a pediatric-only

grouper or a grouper that only worked for cancer care, or fill in the blank. Would we want to filter those out and not have them evaluated by NQF because it doesn't cover a broad enough population? Or would we want to say we actually don't want this kind of criteria and any kind of grouper can come in?

Mark?

MEMBER LEVINE: Would it be appropriate to require the developer to identify the use cases for which is intended?

CO-CHAIR MARTIN ANDERSON: As a way of determining importance? And are there use cases that would not be important?

MEMBER LEVINE: That would depend upon the user. But at least the user would know what they are getting.

CO-CHAIR MARTIN ANDERSON: This is now how NQF evaluates a grouper, right?

MEMBER LEVINE: Yes, but what I am suggesting is --

CO-CHAIR MARTIN ANDERSON: Okay.

Oh, so you're going --

MEMBER LEVINE: -- that one of the valuation criteria would be an expression of what this grouper can be used for. It is not appropriate for pediatric cancer or it is appropriate for -- you know, we also talked about the two basic approaches to grouping, which is sort of population-based and, then, provider-based, and an expression of which approach is being taken, or if it is flexible enough to be able to do both, depending upon the use case, et cetera.

But I think the developer needs to express as part of the evaluation criteria what this grouper is intended to be able to accomplish and what it can't.

CO-CHAIR MARTIN ANDERSON: Okay. I think that would actually be on the first page, too, right? So, we didn't have anything on the first page around things they would submit that would tell us anything about the purpose of the grouper. So, I think we missed

that there.

And then, what is sort of being raised here is, could NQF evaluate that and say this grouper is appropriate for that purpose? And I think that would still probably come out across all of the other criteria.

We can take that and figure out what to do with that?

CO-CHAIR CACCHIONE: Tamara?

MS. SIMON: I just wanted to say I think it is important to include children. I am stating I think what is the obvious. But I would hope that NQF would be wanting to endorse pediatric groupers, for instance, if they're really interested in promoting the health of large populations of people, which I believe you are.

MEMBER MACURDY: So, I guess I have concerns about both of the criteria you have up there. And it really relates to the point that Mark brought up.

I think a lot of the sort of groupers that NQF is going to be asked to be involved in are going to be provider-centric groupers. And you can imagine a provider having not very many lives and not very much money. It is a lot to their practice, but it is not a lot to any other broader scope.

So, I mean, almost any criteria you use there, I think it is going to be somewhat restricted. I think there is going to be tons of those sorts of things. So, it is not just going to be pediatric groupers. It is going to be really very specific.

MS. WILBON: So, I just have a clarifying question. This is something that came up in our group, the Clinical Logic Group, yesterday. I thought I had it straight, but as you guys continue to talk, I am getting confused again.

When you guys say "provider-centric" versus "patient-centric," are you saying that the logic within the grouper is

based -- the claims are grouped based on the provider or that it is based on the patient episode? And then, at the end the results come out and you use it to determine the performance of the provider? I am just trying to understand, is the provider or patient focus in the logic of the grouper or is it how the grouper is being used?

MEMBER MACURDY: No, it is in the logic of the grouper. So, the notion would be you could have two different providers deal with the same sort of illness for a patient. And what one would deem as services that a provider might be held accountable for would be different in those two circumstances.

I mean, these are going to be used in Value-Based Purchasing. That is really coming at CMS. It really is coming. That is why NQF, I'm sure it is one of the reasons why CMS is kind of anxious for NQF to set up this process, because these are coming like in spades. I mean, it has been mandated by

Congress for a while, and there is not that much time before it happens. And it is going to be basically true in every payment system.

We are developing those sorts of Value-Based Modifiers, and they have already been implemented for hospitals. But it is going to go across the board. It is physicians or --

CO-CHAIR MARTIN ANDERSON: Yes.

MEMBER MACURDY: So, it is going to be pretty involved. And that is going to be a main kind of --

CO-CHAIR MARTIN ANDERSON: So, I think this scope question here really gets to, you know, when you look at the way we have even started to define episode and episode grouper, which are also being modified, there really isn't a scope difference that would let you say, you know, if someone were to put together something that might be more like a bundle, right -- so, I used the examples of the Geisinger Total Care Program, right? So,

they are going to take a period of time post-cardiac surgery.

If someone wanted to call that a grouper and submit it, you know, should those types of submissions go through this process, right? So, if we don't bound it at all, anything can come in. And how does NQF decide is that a grouper or not? Because our definitions won't tell you.

So, we want to really think about this issue. I don't think we can solve it today, but do we want to bound this?

MEMBER MACURDY: Well, that is one reason I mentioned before that I see continuing between a bundler and a grouper. It is true there are bundles that go across different kinds of providers, but the kind of logic you use for those is really not very different.

So, there really is a complete continuum, and there is going to be a continuum. That is kind of where a lot of

activity and action is.

And so, I could see a case to be made where on an ACO it is called a bundler because all the services are supposed to be done in that ACO.

The only place you really have the coincidence of the two is when you do managed care because, then, the patient-centric and the provider-centric are the same.

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER MACURDY: But if you look at ACOs, ACOs come in kinds of varieties --

CO-CHAIR MARTIN ANDERSON: Yes.

MEMBER MACURDY: -- in terms of services they cover. Or physician groups, I mean, they come in all kinds of varieties in terms of the particular kinds of categories and services they cover.

CO-CHAIR MARTIN ANDERSON: Yes.

MEMBER MACURDY: And kind of a way a lot of the payments are going is to kind of pay them kind of based on a performance of

whatever those services are.

CO-CHAIR MARTIN ANDERSON: Okay.

MEMBER MACURDY: And those vary by a lot.

CO-CHAIR MARTIN ANDERSON: Right. Right, and I think that's --

MR. DE BRANTES: Yes, if I can just for a second make a -- hi. This is Francois. Just a comment.

I don't think these things are mutually-exclusive.

MEMBER MACURDY: No, I didn't mean to suggest they were.

MR. DE BRANTES: Okay.

MEMBER MACURDY: No, I was even saying the opposite, that there is such a continuum, that they are not for sure.

MR. DE BRANTES: Okay.

MEMBER MACURDY: Yes. No, I didn't mean to suggest that even a little.

MR. DE BRANTES: Because any grouper you can use for -- and again, this

gets us to the use more than the grouper itself -- but you could use it for multiple purposes, as long as the outputs have enough detail that you can resort the data.

Yes, if it fancies you to evaluate anesthesiologists just based on what they do in the hospital, then, fine, you should be able to do that with any output from a grouper.

CO-CHAIR MARTIN ANDERSON: Thank you, Francois.

Mark?

MEMBER LEVINE: At the same time, if the grouper is being used to evaluate a group of providers, it has got to be an output that is understandable and usable for the purposes of quality improvement, which is a very important bottom line.

And I would think it is one of the most important criteria that we can come up with for the eventual acceptance of a grouper by the provider community. It has got to be

evolved in partnership with them, so that they feel a sense of ownership of the criteria that are evolving. And it must support an appropriate level of clinical drilldown to enable clinicians to understand what they can do better in order to improve. And those are enormously-important bottom lines.

MEMBER MACURDY: The term that is often used is it has got to be "actionable". That is kind of the word I would use, is that providers --

MR. DE BRANTES: Yes, but who is going to decide? I think, again, you guys are getting into details of uses of the output that are completely subjective. And so, who is going to evaluate whether something is actionable?

CO-CHAIR MARTIN ANDERSON: Yes, go ahead, Mark.

MEMBER LEVINE: Yes. I think that is one of the purposes that we need to get the developer to express what their intent is in

the use of a grouper. And if the intent is to use it for provider performance improvement through Value-Based Purchasing or some other mechanism, that it must support an appropriate level of clinical understanding, clinical buy-in, and opportunity for usability and clinical utility.

CO-CHAIR MARTIN ANDERSON: Okay. So, we'll take that under consideration for the usability area and think about how to walk that line, right, where we're not -- NQF doesn't play user, right? So, at some level, it has got to be, is it acceptable? Has it met some criteria? And we have to figure out how to deal with that.

So, I get the issue. So, we'll capture that, right? Okay.

Let's move on from importance. But I didn't hearing anyone saying for sure that things should be cut out, right?

David, do you have an idea there?

MEMBER HOPKINS: I was actually

going to be radical --

CO-CHAIR MARTIN ANDERSON: Well, go ahead.

MEMBER HOPKINS: -- and say that the groupers, at least the ones I'm familiar with, should get a pass on importance.

CO-CHAIR MARTIN ANDERSON: Well, don't you think they would already?

MEMBER HOPKINS: Huh?

CO-CHAIR MARTIN ANDERSON: We're actually proposing there not be any importance criteria.

MEMBER HOPKINS: So, you are saying the same thing?

CO-CHAIR MARTIN ANDERSON: Take out importance.

MEMBER HOPKINS: Okay.

CO-CHAIR MARTIN ANDERSON: Because we can't think of a way to bound it right now, right?

MEMBER HOPKINS: I would leave it that way.

CO-CHAIR MARTIN ANDERSON: We might come back to it. For the next six months, we can change our mind.

MEMBER HOPKINS: Because that didn't make sense to me.

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER HOPKINS: Okay.

CO-CHAIR MARTIN ANDERSON: But this was something was raised before that said it had to have a certain scope in order to be considered important.

MEMBER HOPKINS: Yes.

CO-CHAIR MARTIN ANDERSON: And we're saying, well, maybe not so much. We can't anticipate what is going to come in the future for something that might be defined as a grouper for managed care, carveout for cancer, as an example.

MEMBER HOPKINS: Okay.

CO-CHAIR MARTIN ANDERSON: So, we are not going to a priori eliminate it by saying there is some criteria that has to

cover the full population.

MEMBER HOPKINS: Good.

CO-CHAIR MARTIN ANDERSON: So, if over the next six months you all come up with a good definition for importance, then we can add it. But, for now, we don't have one. Is that fair? Okay. One knocked down. Okay.

Now we are into the hard one, the scientific acceptability. The team took a stab at describing how validity might be assessed. And so, here are some things up there. Why don't you all look at that? And then, we will take comments on that.

(Pause.)

This is, again, what they have to demonstrate that they have tested and give results to the Committee, so the Committee could determine whether or not they thought they did it adequately.

MEMBER HOPKINS: So, this is back to raising the issue about, you know, for how many different conditions do they have to do

this? Because each condition will have those questions associated with it.

CO-CHAIR MARTIN ANDERSON: You say for every condition?

MEMBER HOPKINS: It is not feasible.

CO-CHAIR MARTIN ANDERSON: What's feasible?

MEMBER MACURDY: Well, I mean, the problem you run into is the safest way to do this is to make each experience for each patient to be their own episode and it is homogeneous.

So, the real problem in these, and if you look at all the groupers, I mean, they will have a way of expanding the number of episode types even, but that is a challenge all by itself, is to have to put all kinds of experiences into, say, 500 buckets is pretty hard, and the groupers do it in a different way. And how homogeneous they are is a real challenge.

So, it is not that I don't think you should have it there. I just think it is almost like one of the core problems in developing groupers. You know, you could have two different groupers who have different buckets. In fact, they tend to. That is why it makes it hard to compare them.

And then, if you really press them to say, "Well, can you make this more homogeneous," well, they can break up the buckets more and more. And then, pretty soon you have 3,000 buckets and you've got two people per bucket or two people per bucket per provider, something of that nature. And that is really where it is like, okay, well, great, you've got them homogeneous, but you can't use them.

And so, that is the balancing problem here. I think everybody recognizes that that would be nice to do, but that is really the tradeoffs that the people who develop groupers face.

CO-CHAIR MARTIN ANDERSON: But you are agreeing that they at least have to discuss how they actually tested homogeneity and what they did about it?

MEMBER MACURDY: Yes. I mean, it basically is saying, how do you define the episode? I mean, what are you going to call a bucket?

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER MACURDY: So, they are kind of the same thing. So, once again, it is not that I object to what is there. I mean, that's fine. It is just operationalizing I think is where the real -- I mean, any grouper is going to do -- anybody who is going to give a description of the grouper is going to have this. It is just --

CO-CHAIR MARTIN ANDERSON: Have the right three buckets. Now we just have to deal with how do you operationalize it. But this is the right three buckets. Someone said give me something broad that just says what's

the real threats; what do they perceive the threats to validity to be, and how would you address them. And there was a very specific, you've got to at least deal with the clinical validity, right? And then, also, the construct validity -- are the three we heard.

Are there more? Dan, I think you had something to add?

MEMBER DUNN: Maybe first a clarification. What do you mean by threats to validity, meaning what we feel are the holes in the -- well, I'm sorry -- what the developers feel are the holes in the methodology? So, you are expecting them to say, "We think this works well except for" X, Y, and Z?

CO-CHAIR MARTIN ANDERSON: In yesterday's discussion, somebody on this side of the room proposed that there be a systematic way of having the developers talk about where there were risks in validity and how they address them.

And so, if we can't get more specific than that, then we shouldn't have it on there.

MEMBER DUNN: Again, so validity and the application or validity and the --

CO-CHAIR MARTIN ANDERSON: Results. The validity of the results.

MEMBER DUNN: You mean where the developers don't -- for example, with a diabetes episode, it is built the following way, but the developer thinks this part of it isn't going to work well? I just don't know how you're going to get folks to comment --

CO-CHAIR MARTIN ANDERSON: I don't think we meant that specifically. It was more it is almost like their philosophy and how they even approach the groupers, right? So, how they may have handled the differences in the risk profile of people that have diseases where there is a great variability in the overall risk within an episode, right?

So, they handled it by breaking

them into separate episodes. They handled it by assigning -- I don't know. These are examples people used yesterday.

So, I don't know, Steve. We can strike it. If no one understands it, it should not be on here.

MEMBER MACURDY: Actually, I would recommend you strike it. Because, I mean, you can easily put that under the other two in some way.

CO-CHAIR MARTIN ANDERSON: Okay. We've got some --

MEMBER DUNN: Maybe the recommendation would be think of -- Tom was touching on this a little bit -- there are certain challenges inherent in episode grouping. You are often left with making a choice on which way to go or the other.

If you could identify those and ask folks to comment on how they approached it and the pros and cons of what they did, if you make it more tangible, I think it could --

CO-CHAIR MARTIN ANDERSON: I think we could follow up on that and put that in what we would ask them to submit. I think that might be a good idea for how we get more clear on what we want, other than all the steps in the process.

MEMBER DUNN: Because if you open wide up -- I would somewhat limit the list of real challenges here and tradeoffs and I think get to the real -- the decisions people make and how their grouper works along those can really tell you a lot about how well things are going to go and how they can be used.

CO-CHAIR MARTIN ANDERSON: Okay. Jelani?

MEMBER McLEAN: Yes, I wouldn't say strike it. I would say rewording it. Because if you do any research, you have limitations to your research. And I think that is what Dan is getting at, is having them express their known limitations and, then, how they addressed them. And just wording it that

way.

I think what is striking, what is alarming is the threats related to the comment, I think. But, really, what you are trying to say is everything is going to have some sort of limitation. And then, how do you address those known limitations when you build your grouper?

CO-CHAIR MARTIN ANDERSON: I think we can put that in the previous section, though, what they should disclose. It kind of reminds me of our public reporting process, why we have to tell people why our revenue might not be achieved.

David?

MEMBER REDFEARN: Well, the concrete example, Dan, is that Optum has expressed the opinion that, because we typically don't have staging, cancer staging, information in administrative claims data, you have to be cautious about interpreting episodes for cancer, particularly if you're

trying to build case rates and stuff like that.

So, that is the developer saying, "Well, here is a limitation of the underlying data that limits the validity or usefulness or utility of these types of episodes." That is the one specific example I know of in this area.

CO-CHAIR MARTIN ANDERSON: Okay. Mark?

MEMBER LEVINE: Which is another of saying, should we be looking at validity and reliability in terms of the entire functioning of the grouper as a system or are we looking at validity and reliability for each of the episodes that are produced by the grouper?

And I think that is a very important distinction because, you know, you could have an overall picture that looks pretty okay, but it doesn't function in the areas that you need, that you think are

important.

CO-CHAIR MARTIN ANDERSON: And this is criteria for validity. So, are we saying that the Committee should evaluate the performance of each grouper for every type of episode?

Now you know there will be another standing committee that works continuously.

(Laughter.)

Steve?

MEMBER LEVINE: That is really a workflow issue.

MEMBER BANDEIAN: At least for a sample of conditions, if not all of them.

Actually, I have a question and maybe several comments, not surprisingly.

What is exactly is the difference between clinical face validity and construct validity? So, I am a little confused about the language. I mean, and I can, instead, propose an alternative, but I am just a little confused exactly what the words mean.

CO-CHAIR MARTIN ANDERSON: There was a case made amongst this group that you have to think about both clinical validity, right -- in other words, some evidence that the reasonableness of the episodes in terms of acceptance from the clinical community is there, right?

And then, the question is, can you also test, statistically test, your episodes to say whether or not they actually show properties that would say that they are, in fact, homogeneous as intended?

MEMBER BANDEIAN: Okay. Let me try this and see if this is useful either as an amendment, a supplement, or a replacement.

One test of validity or one concept of validity would be -- probably have a few tests, subparts to it -- but the basic question is, have we captured the cost of a condition accurately, to the extent that that can be done?

And so, that actually has a first

preliminary question, which is, was there actually even really an illness here? Because you can often have claims that will say pneumonia or ankle fracture, or whatever, and actually it was just a ruleout or just some sort of fluke information.

There also are issues about whether we are accurately distinguishing between one case of pneumonia versus two cases of pneumonia. And then, finally, whether all of the costs and services associated with the care of that pneumonia are captured in some way, shape, or form.

And I would say not just the cost of the pneumonia, but perhaps to some extent some discussion sequelae of the pneumonia. So, if you did not include the fact that the person had sepsis as a sequelae of the pneumonia, it would be a little bit of an incomplete representation of the total cost of the pneumonia.

So, I'm not quite sure what the

right word for that is. I guess I thought that was actually a construct, validity concept, meaning, is this thing representing a real condition episode and is it containing everything that relates to that one patient's condition?

Now the second question sort of merges into risk adjustment a little bit. But let's assume for the moment that all of the episodes that have been constructed represent genuine illnesses that we have accurately distinguished between one case and two cases of pneumonia, and that we have included all of the costs that really are reasonably associated with pneumonia.

So, then, the next question is -- and this kind of mirrors some of the earlier discussion, but I am trying to stay away from the use case a little bit -- are there ingredients in the system so that one can say this case of pneumonia is comparable to this case of pneumonia? And so, therefore, it

would be reasonable to compare the costs of the two.

So, that, for example, might entail a couple of different things. It would be maybe we need to distinguish between a community-acquired pneumonia versus one that is related to a person being on chemotherapy or a fungal pneumonia or some such. So, these are different types of pneumonia that have different costs and risks associated with them.

But the other part would be, you know, does the patient have comorbidities or sociodemographic factors that would affect the cost of the pneumonia. So, recognizing that there is sort of a use case issue, I would at least wonder -- I think there is a question of, is there enough in the system so that one can actually make a valid comparison potentially, theoretically, a valid comparison between case one of pneumonia and case two of pneumonia?

So, those would be the two types of things that one might want to look for.

CO-CHAIR MARTIN ANDERSON: Let me try this a different way. You're a committee. Now we're all a steering committee. And the developers have submitted their test results. Are you looking for two things or one thing?

This is just to everybody. When you're saying it is valid, is there a distinction between you say, hey, it's clinically valid. They have proven that they have done all this testing with their clinical panels, and their panels have accepted this particular grouper as valid. Our Cardiac Committee signed off on it. Go all the way down the list, right?

Is there that plus something else that says, in addition to that, we have done the following types of statistical tests, you know, either comparing the results we received to what is known in literature around what the total variability is in cost of care for this

type of condition? Are there two types of things you are assessing or is there only one thing you are assessing, which is that, hey, if the clinical community says this is valid, and these guys can prove that to me, I'm good? So, where are they? Where is everybody on that?

Chris?

MEMBER TOMPKINS: Well, unfortunately, it is all the above. But, for example, there are what we call inputs to the episode grouper, which should say this is the list of codes that we say, when looked at together, constitute a condition with a certain label and a certain definition.

And then, when we say that there are certain services that are, quote, "relevant" to that condition, then the clinicians review these. Is this a fair representation of that condition? Is it leaving something out? Is it including too much, et cetera? Are these services really

plausible, and so forth, like that? So, the validity of the inputs.

And then, as you ascribed, there is another what you call criterion validity, which is to say you have external criteria, such as, if you are going to trigger heart failure conditions, then is there a prevalence rate that the episode grouper expresses in the population that rings true or corresponds with expectations or what is observed in using other means or other ways of calculating things like prevalence rates?

And then, similarly, if you are constructing an episode that says this purports to be heart surgery, then you expect there to be knowing what the DRGs is, knowing what the procedural or the professional bills are. So, you have a sense of how expensive that is, and that has been calculated in various ways, and you expect the grouper to approximate that, not to be wildly off, right? So, that is the second type.

And then, in the middle, which I think is more of this face validity part of it, although the inputs is part of that, is a mechanism in which you actually mechanically walk through what the computer is trying to emulate. And what we are trying to emulate is to say that, if you see a medical history, relevant information about a patient, a physician can look at that and start to say, "I see what the conditions are. I see when they started. I see what the treatment patterns were. I see when the condition resolved. And I see when there is another condition that probably exacerbated the first. I see another condition that probably resulted by way of complication from the surgery or complication from the treatment or non-treatment of an illness."

And you can tell the story, the clinical story, of what happens with a patient. So, you show them the chronology of the services. The physician can recreate what

the medical history or what the status was at any point in time. And that is what the computer is trying to emulate. Of course, it is trying to emulate it over hundreds of conditions and millions of patients, and the complexity of the comorbidities and the overlapping and the rest.

And so, in the middle there, in addition to purely the inputs, here are the codes, or purely the outputs, here is the prevalence and the average cost, you have this corroboration that the computer is, in fact, emulating the clinical logic that physicians would apply, understanding each patient's medical history.

CO-CHAIR MARTIN ANDERSON: Thank you. That is very helpful.

Jelani, did you want to add more?

MEMBER McLEAN: No, actually.

CO-CHAIR MARTIN ANDERSON: Marjorie, do you want to add more?

MEMBER KING: Just as a clinician,

I would want more than just attestation that you've got a board of clinicians looking at it. You also need testing.

CO-CHAIR MARTIN ANDERSON: Right, but there is a test process. Right.

MEMBER KING: You know, we need evidence. Yes.

CO-CHAIR MARTIN ANDERSON: And Mark?

MEMBER LEVINE: Yes, I was going to pick up, I think, similar to what Marjorie was saying. The quality of the peer review I think is something that needs to be defined. Is there adequate clinical review of the grouping methodology, the clinical logic? Does it make clinical sense? Have the right people looked at it in order to come up with the groupings that are being used?

CO-CHAIR MARTIN ANDERSON: I would argue that NQF might figure out a scoring, a high, medium, low of how much. I think, right now, if you are comparing to literature, you

know, or whether or not it is based on solid evidence, et cetera, it gets sort of rated. But I think in this case the evidence ratings might be something like, is it just a clinical panel that went and reviewed it? Was there, in addition to that, corroboration with external data sources that say that the outputs look right? And in addition to that, what is the level of validity testing that was done with clinicians on the output level, not just the input level?

Okay. I could imagine something that might shape itself on that.

David?

MEMBER HOPKINS: Just one more piece on that because that sounded right to me, but is it condition-specific, what you just said? I'm sorry, but we have got --

CO-CHAIR MARTIN ANDERSON: I think the question is, for all the episodes that you create, did you do this for all of the clinical conditions?

MEMBER HOPKINS: Okay. You can describe the process --

CO-CHAIR MARTIN ANDERSON: Right, right.

MEMBER HOPKINS: -- and the peer-review mechanism.

CO-CHAIR MARTIN ANDERSON: And what was the scope of what you did in terms of the --

MEMBER HOPKINS: And you did review the literature for 500 different conditions?

CO-CHAIR MARTIN ANDERSON: Well, you know --

MEMBER HOPKINS: Uh-hum.

CO-CHAIR CACCHIONE: I don't think there is a way around it not being condition-specific. I mean, frankly, we can prescribe how it should be done, but, frankly, I mean, it is going to be condition-specific because, otherwise, there is no way around it. You can't compare apples and oranges and be

effective.

MEMBER HOPKINS: No, but, I mean, I can see how it would be logical to have your clinical panel somewhat specific to an area. Cardiovascular, so we had a cardiovascular panel. They reviewed all of those episodes, and something about that.

CO-CHAIR CACCHIONE: I think that that is even difficult today because there are electrophysiologists. There are heart failure specialists. And unfortunately, the literature, keeping up in your own subspecialty area is often difficult, let alone keeping up on the broad field of cardiovascular disease and saying that this is adequate.

And, listen, it is going to fall apart; these bundles will fall apart if the clinicians don't believe that they are meaningful.

MEMBER HOPKINS: Well, what is your validity test?

MR. DE BRANTES: Right. So, this is Francois.

I mean, I guess, again, we need to be cautious because -- Joe, was that you saying --

CO-CHAIR CACCHIONE: Yes, that was me, Francois.

MR. DE BRANTES: Yes, yes. So, expand on that and think about it. The burden, therefore, for a developer would be not just to have, say, a cardiovascular working group, but to have a working group for congestive heart failure, a separate working group for ischemic heart disease, a third working group for essential hypertension, and so on and so forth.

The cost of doing so, the burden of doing so would basically negate anyone bringing their grouper through for endorsement. I am just being realistic, guys. I mean, you're talking about five years' worth of development time to put all of these

working groups together to review every single code and every single episode definition.

MEMBER LEVINE: I believe that there is actually some legislation before Congress at the moment that is going to call for exactly that kind of review, open and public review of things.

And to make Marjorie's point again, once a peer group reviews a condition or a procedure or an episode of one nature or another, one of the essential feedback loops is back to that group. After you construct the logic, let's run it through the program and see what it is that it actually does accomplish. Does it accomplish what the group wanted? And is it now at a level of peer acceptance? Not only peer discussion and recommendation, but it needs to go through another phase of actual acceptance.

MR. DE BRANTES: Mark, certainly legislation can apply to Medicare in any which way it wants, which that is the prerogative of

Congress. Fortunately, Congress doesn't have that same prerogative on all of the rest of the industry.

And I am just voicing a concern that, if you establish that level of burden for any grouper to go through the process of endorsement and to demonstrate that it has an ongoing process to have that review done continuously -- because, to Joe's point, it is almost difficult for anyone to follow up on the changes in the evidence in their own medical specialty -- you're killing off any potential for an innovator to enter into the market.

CO-CHAIR MARTIN ANDERSON: I think we have --

MR. DE BRANTES: That would be the net result, is that no innovation will occur because you are stacking the deck towards the multibillion dollar companies.

CO-CHAIR MARTIN ANDERSON: So, I just want to make sure, as we add comments on

here. I think we have the range of opinion around how much testing could happen between every clinical condition, multiple points in time in the development, and then, also, maybe that's not so feasible. So, I think we have that range.

Is there anything else that is in addition to that anyone wants to add?

MEMBER LEVINE: I think there is a relationship to use case, and there may be different criteria for different use cases. If this is being used on a national level to modify physician payment, for instance, that is a very important high-level use case that must be specifically addressed in the criteria. Are there criteria that set that bar may be different from criteria that would set other bars?

CO-CHAIR MARTIN ANDERSON: Okay. Thank you.

Can we go on to reliability? It's not over, right?

So, the only example captured so far I reliability was really how does this grouper perform across data sources, different size data sources, multiple different time periods, and that could be length of the data period or over time. These were ideas that were thrown out in terms of thinking about reliability.

Thoughts on reliability?

Okay, Tom?

MEMBER MACURDY: Well, I mean, I definitely don't think an episode should go across, you know, be required to go across multiple data sources. In fact, I could even make a stronger statement that often groupers are dependent upon the particular payment source or payment rules that are in a payment system.

Let me not use Medicare. Let's suppose I use Medicaid.

CO-CHAIR MARTIN ANDERSON: You didn't mean sources? You meant like, if you

have a claims dataset, have you tested it against multiple different datasets, not just one you developed in one and, then, just pushed it? I mean, you did your testing and your development --

MEMBER MACURDY: So, if you started with the universe of Medicare data, you mean subsets of the Medicare data or --

CO-CHAIR MARTIN ANDERSON: Or a different time period of Medicare data.

MEMBER MACURDY: You are saying the same data source?

CO-CHAIR MARTIN ANDERSON: He meant databases. So, in other words, if you developed it on one, have you at least tested it on a different data source or a different data source with a different period of time? So, if you develop it so that you at least can talk about how it performs in different data sources, databases, not data sources. Datasets maybe I should say.

MEMBER REDFEARN: I think testing

these groupers across time makes a lot of sense because they tend to be easy to do. But the complication is, then, you have to keep in mind what version of the grouper are you using. Okay? Because coding can change underneath you as you go across time, is the grouper updated?

So, if you run a grouper that was developed three or four years ago against current data in which there is new coding, new medical practice, and things like that, it might not do very well.

And so, you have got the underlying data is changing, and you sort of want the underlying model to change, too, to keep up with these kinds of things. So, I like that across time. That is really nice, but it is complicated. It is difficult.

CO-CHAIR MARTIN ANDERSON: Steve?

MEMBER BANDEIAN: Yes, if I understand the basic idea of reliability, it is kind of like we know this car is moving.

I'm a policeman, a state policeman, and we know it's going 65 miles an hour. And I've got my radar gun and I do 10 measurements of the care, and the issue is, do I always get 65 miles an hour or what is the spread around that?

Having said that, it is picking up, I think, really on what David said. It is a little hard for me to see how one can have -- but maybe Tom is going to -- anyway, it is a little hard for me to see how one does reliability testing in this context.

Now, ideally, one would be able to say, "Dr. Jones does too much of X in time period one and time period two." And you would see some persistence of the pattern over time.

That strikes me as setting a really high barrier because of all of the factors that go into the calculation of Dr. Smith's score. And so, I could easily imagine that the system is actually functioning really

well, and in time period one Dr. Smith has a good score; in time period two he has a bad score, or vice versa.

So, I, myself, am having a little trouble understanding how to make a meaningful reliability test, but, again, maybe I am missing something.

CO-CHAIR MARTIN ANDERSON: Tom?

MEMBER MACURDY: Yes, I have the same sort of question. I think the best way for me to kind of express my concerns is with a particular example.

I mean, I can have a sequence of claims that are the same for two beneficiaries or two individuals, and you get a different grouping. And often, when you drill down, it is reasonable that you got a different grouping. Now is that reliable?

I mean, you guys just want to add yet a third enrollee and, then, a fourth, et cetera. But I guess I am not sure. You know, reliable, it sounds nice; across multiple

databases sounds fine. But I am just not sure operationally what that means or what kind of criteria you guys would use to make that judgment.

Because, like I said, I can just take a case where I will do one individual, two individuals, three individuals, and you can often get different answers as you go across, though, simply because the circumstances are different and it is reasonable that it happened. And then, yes, you have to understand each one of those scenarios. But I don't know; is that reliable?

CO-CHAIR MARTIN ANDERSON: But I presume it shouldn't be happening that you run the same data through the episode grouper multiple different times and get different results, right?

MEMBER MACURDY: Well, you can re-sort data and get a different answer.

MEMBER LEVINE: Yes, and the

sequence in which you present the data can also influence it.

CO-CHAIR MARTIN ANDERSON: Maybe we can think about a different way. Maybe at least that should be disclosed, right? So, maybe there is no an easy way. So, high reliability of an episode grouper sounds unlikely at the moment, right, from what I'm hearing?

You're saying you can't, for instance, say, "I developed this system, and if WellPoint runs it and, then, United runs it, they can at least reliably assume that they have gotten comparable results."? You can't prove that, right? Is that what you are saying?

MEMBER MACURDY: Yes. I mean, a good example there is WellPoint may have a different kind of payment system or the way it provides services, the way it registers services, compared to another plan.

I mean, it is the same problem you

run into when you try to do drug safety studies and formularies have different step therapies. They are not the same thing, and it is a challenge.

So, what a procedure or a set of services means in WellPoint is different than what it means in another one. That is kind of fine, but there is a context there.

CO-CHAIR MARTIN ANDERSON: I guess I am just having trouble with saying that we can't say that episode groupers could be reliable. We need to figure this out.

(Laughter.)

MEMBER MACURDY: I just did. It is not that I want to say that's why --

CO-CHAIR MARTIN ANDERSON: Right. We have to wrestle this one to the ground.

Jelani?

MEMBER MACURDY: I am just saying the planning it is a challenge.

MEMBER LOISELLE: This is Jim. Sorry to interrupt.

To a certain extent, we are still somewhat arguing over what you do with the results, and that WellPoint and Aetna, or whomever, if they are fee-for-service, one needs to look at how the grouper processes the claims or detail and/or assigns the clinical categories. What you do with it, whether it is a cost or a servicing issue or an analysis issue, that is post-grouper. That is not what a grouper does.

MEMBER MACURDY: I disagree.

CO-CHAIR MARTIN ANDERSON: No, my examples wasn't that. It was just to say that, normally, you would say that this program works if you change datasets. You would want to be able to say that for reliability, that we have confidence in this grouper if you switch datasets.

MEMBER LOISELLE: Correct.

CO-CHAIR MARTIN ANDERSON: That's what we are after here, not that the results that they would get would be somewhere

comparable, but how do you know that it doesn't only work on the dataset it was developed for?

MEMBER LOISELLE: Correct. Correct. My point was you can't use those values, and you mentioned like cost, as the reason to determine that.

MEMBER McLEAN: I think, if I could interject here, yes, I agree with Kristine. You know, a grouper has input requirements. They went back to it and they said, "This is how you have to submit the data." Whether it is one plan's data or another plan's data or another user's data, you have to conform to that grouper's input requirements regardless.

Now I agree, the output is different for your business. That is your business. And so, that is what you have to be able to understand, how to interpret the output.

But what we are doing here is

saying, does the grouper work? And does it work based on your input requirements and does it do what it is supposed to do?

And I think you have to be able to say that. Otherwise, certifying on a national level, groupers would be useless if you can't make that determination.

CO-CHAIR MARTIN ANDERSON: We are saying that the developer must be able to demonstrate reliability. Now what we have to struggle through is what would we accept as such a demonstration.

MEMBER McLEAN: Right.

CO-CHAIR MARTIN ANDERSON: Okay.

MEMBER BANDEIAN: At least in terms of what I'm familiar with -- and, obviously, there are lots of things that I'm not familiar with -- but in terms of what I am familiar with, the logic is actually fundamentally deterministic. And so, if you rerun it 10 million times, you will get exactly the same results 10 million times

because it's a computer; it doesn't make mistakes, and the logic is deterministic.

Now it may be that some of the systems are not deterministic, and that would be interesting to see, to take the same exact -- and to Tom's point about sorting the data, again, at least that would be interesting to test because, obviously, if the sort order changed the result, that would be a reliability issue that would, I think, be something that would be noteworthy.

But now I did have an idea, which is vastly overly-difficult to do. It would be interesting to see, okay, if we go back to what I was saying previously about the pneumonia case -- that is to say, did we properly identify the episode of pneumonia; did we properly identify the costs associated with the pneumonia, et cetera, et cetera, that type of validity concept -- I think it would be interesting to take the same system and run it off data from health plan A, B, C, D, E, F,

and G and see if we are still seeing that the pneumonia episodes that are being constructed are valid, given our validity criteria.

Now that strikes me as a huge job, but it would be important to know that in this test set of data, where the validity looks good, it may not be really representative of all possible uses.

CO-CHAIR MARTIN ANDERSON: Yes, I hear you. And I think there is a parallel here to what NQF is doing in looking at e-measures, right? And the way that sort of came down for the short-term was you had to have tested your e-measure in at least three different EHRs, as an example, right? It assumes the software itself would be reliable, but that you had at least tested it in different places.

But I don't think we are going to resolve this. I will take any more comments on this topic. We may have to come back to it.

David, I know you had -- Jelani feels good. David and Dan will get the last two words on reliability for today, not forever.

(Laughter.)

MEMBER HOPKINS: Never forever. It goes on.

I really like the way you framed it. And again, in a practical sense, I am a physician and I contract with, let's say, two health plans for a lot of my patients. And each of them is doing this kind of analysis and coming up with either similar or dissimilar results.

And what can NQF do, if anything, to assure that these results are comparable? That is a tough one.

Do you know where it takes me? It takes me to all-payer claims data. This is not an NQF endorsement issue. It is, again, a use issue. I keep coming back to use issues distinct from endorsement of groupers.

MEMBER DUNN: Yes, maybe just a comment. And actually, I think Steve's comment was on the right track and you followed it, Kristine. But it is almost more of, is the grouper robust, meaning it can produce an expected result with validity across different scenarios. Because at this point, the same set of data, it is going to produce the same result again and again and again.

But can it produce a valid result across different nuances, which could be different health plans? It could be different cases, and so on. I think that is probably the more important point. And maybe reliability isn't even the right term.

CO-CHAIR MARTIN ANDERSON: That was a good discussion.

Anyone --

MS. WILBON: I just have a question. Is that feasible? I'm just saying, if we were to kind of have a criteria and say

like what our high bar, not saying that that would have to be the mechanism to test it, but if we were to consider that a high bar, is that feasible for a developer to do?

MEMBER DUNN: A good question. Well, obviously feasible. But the bottom line is it is not -- see, three different health plans, you run three different pneumonia patients through. To really tell whether it worked well across all, you would have to actually go through and clinically validate, do a methodologic. Did it do what it was supposed to do, given the methodology and, two, does it make sense? I think you have to get to that extent to be able to assess that.

MEMBER MACURDY: I think I can answer it best by saying, suppose you have an episode grouper that works well for Medicaid in South Carolina and that's it, only in South Carolina. And I would use California as an example, but there are 58 Medicaid programs. So, there isn't a California program.

Okay, if it just worked well in South Carolina, it doesn't work in any other state, would you say it's not valid or it is not reliable? I mean, Arkansas is developing its own right now for its particular program. And if it works well for that program, that seems fine.

CO-CHAIR MARTIN ANDERSON: Okay. And then, there's always this other question, which we don't have to get into in detail. But there will always be organizations that choose not to be endorsed, right, that choose not to get whatever they are doing endorsed because they are doing it internally, or whatever. So, the question is, if you want national endorsement, what should the bar be, right?

MEMBER DUNN: Two seconds.

CO-CHAIR MARTIN ANDERSON: Yes, yes.

MEMBER DUNN: Okay. Tom's point is a good one. If something is built for a

specific purpose, then how valid is it when applied outside of that purpose? That is a test of its usability and reliability and robustness.

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER MACURDY: And I can't imagine a state wanting --

CO-CHAIR MARTIN ANDERSON: Right, right. And if they want it, then they may want -- once it is endorsed, it is assumed others will use it, right? So, you don't really necessarily want to endorse something that only that one developer is only going to use, right? And, you know, in theory; whether or not that happens in practice I don't actually know.

So, let's go on to feasibility. So, the only two topics --

MR. DE BRANTES: Yes, I have some --

CO-CHAIR MARTIN ANDERSON: I'm sorry. Is someone on the line?

MR. DE BRANTES: Yes, this is Francois.

I am still going to make the case of you don't want to stifle innovation. Well, I and a lot of other people in the country would not want NQF to be responsible for stifling innovation.

CO-CHAIR MARTIN ANDERSON: Francois, can you tie that back to the conversation? You mean by requiring, by making --

MR. DE BRANTES: By making requirements --

CO-CHAIR MARTIN ANDERSON: For endorsement?

MR. DE BRANTES: -- of various aspects for endorsement so unachievable by anyone but huge, established companies, you will immediately stifle all innovation.

MEMBER BODYCOMBE: Now, as a follow-on to Francois, what I have been hearing for the last two days is a lot of

instances of "depends" -- depends, depends, depends. And you know, we all at the start of this session had to talk about our conflicts of interest. And I have a question for NQF. Is "We don't think you should do this" a viable option? Is that acceptable? Or does your contract with CMS say, "Thou shalt produce this and, otherwise, you don't get paid."?

(Laughter.)

MR. AMIN: So, we will have discussion. Yes, we'll have that discussion after lunch.

MEMBER BANDEIAN: And actually, we also don't get lunch.

(Laughter.)

CO-CHAIR MARTIN ANDERSON: Okay, we'll leave that there.

Okay, let's go on to feasibility. So, the two observations made about feasibility, potential criteria, one was around, is the data that is required to run

this grouper generally available, right? So, that's one of feasibility, and this is feasibility that this could be adopted by, an NQF standard could be adopted by others.

And the other one that has come up is cost, which is, are there barriers to access to this grouper that are related to the cost to run it, which is one of our panelists mentioned earlier?

So, comments on those two?

MEMBER HOBART: I just had a question. Does NQF usually have a gait about cost of applications? I mean, I just didn't see it in the criteria you showed the other day.

CO-CHAIR MARTIN ANDERSON: It has to be disclosed, and sometimes committees do take that into account if they have an alternative that does not have a cost.

DR. BURSTIN: The Committees consider the cost under feasibility now.

CO-CHAIR MARTIN ANDERSON: Right,

but it is not a pass/fail, right?

MEMBER HOBART: It is a transparency of cost structure.

CO-CHAIR MARTIN ANDERSON: Yes.

MEMBER HOBART: Is that what we're saying?

CO-CHAIR MARTIN ANDERSON: Yes.

MEMBER HOBART: Okay. That is not to me feasibility. I mean, that is providing the information. I mean, it is a judgment whether it is a cost that is acceptable or not. That is what I am trying to say: do you judge this as too high a cost or is it just you need transparency of what the licensing cost structure is?

CO-CHAIR MARTIN ANDERSON: Every one of these criteria within have judgments applied by the Steering Committee for whether they were high, medium, or low in feasibility or reliability. So, it is possible that a very expensive grouper might be rated lower in feasibility than one that did not have a cost.

MR. DE BRANTES: So, this is Francois.

Just as a point of clarification, you could have and there are some of these applications are so that software is a service. With the software in the Cloud, you load up your data. It does all this wonderful stuff and returns results, in which case your cost would likely be a licensing fee and there might the dance points of additional services that the organization might sell around that.

On the flip side, there might be another software application whose license fee is relatively low or lower than the software as a service option, but would require the organization installing a whole slew of hardware and operating system software and all kinds of other stuff before they can even run the software. So, it obviously is a cost.

And so, how do you distinguish these things?

MR. AMIN: Francois, this is

Taroon.

The way that we have handled that in the past is that we ask the question of, if any user who had access to the data wanted to get the measure results, regardless of which method a developer uses, the software as a service or he actually is buying all of this and putting it in-house, how much would that cost for you to be able to do that?

And so, we would assess both of those types of costing models. And then, the Committee would have to evaluate whether that was undue burden, to be able to actually run the data themselves.

MR. DE BRANTES: I think you would have to be a little bit more precise here, simply because part of this would likely be asking the developer to provide a list of all the technical hardware and software requirements that would be needed by someone who would host the software.

Because you can't make an

assumption. I mean, it is incredibly difficult to price these things out. It depends on multiple factors. I mean, if you are going to rent Amazon space, it is very different than if you are going to buy the hardware and host it internally.

CO-CHAIR MARTIN ANDERSON: Thank you.

MEMBER JONES: What we talked about at our subgroup was perhaps considering the requirement of a very standard way of submitting their total cost structure, as well as I would suggest that they would disclose when other billable hours kick in.

So, for example, if you get output that you did not expect, what is their planned response to that? What is the availability of support there? Because what I have found is oftentimes you will get very unexpected large bills.

CO-CHAIR MARTIN ANDERSON: I think that we are going to hit this again and market

impacts, too.

But, Dan?

MEMBER DUNN: I probably should recuse myself from commenting, but just a few, given we obviously are involved in this piece.

But, one, there is the methodology part of a grouper. And then, there is the software that implements it. Two different things. And then, to Francois' point, it is really hard to disentangle that these systems are usually embedded in other systems that do all the post-grouper parts, as well as all the services around supporting them. So, again, you need to be really precise and people need to interpret the value of everything you get for what you pay for.

The other point is, you know, this tension between wanting it to be better, current, maintained, all the work to do the clinical validation, and so on, that tension between having to spend more money to keep it valid and, also, to maybe potentially have it

endorsed goes right against this idea of looking at the cost of people.

So, you know, I see the budget that it takes to stand up our groupers, and it is not small. Unless you have a public entity financing all this work, I have a hard time seeing how this one -- if people perceive there is value in what they get and how it is delivered, that, to me, is more important than the price, the specific price, that is attached to it.

CO-CHAIR MARTIN ANDERSON: Thank you.

CO-CHAIR CACCHIONE: I would come back to the intended use. And so, I guess I made the case for this idea of this refreshing and this has to be updated. It is really about the intended use. And depending on what the intended use of these things is -- I mean, the DRG system has been around forever. It got updated in increments of DRG, then MS-DRG, and now AP-DRGs. And that worked, and people

sort of learned to adapt to it and there were minor modifications.

So, I guess what I am saying is, depending on the intended use, if you are going to get the provider buy-in, depending on what they intend to use this for, it may require that, if we plan to use it for public reporting and people are going to be judged unfairly based on clinical information that is dated, boy, you're not going to get provider buy-in.

But if you are going to use it for payment and it is not going to meaningful change the payments to providers, then they buy in. They bought into the DRG system that way.

But if you are going to hold people accountable and tell them that they somehow look bad or they are not good providers, you are going to have a lot of pushback by the provider community.

MR. DE BRANTES: Yes, but, again,

I think we need to be careful about, are we establishing criteria here for the CMS grouper? Or are we establishing criteria here for all potential applicants? Many of these groupers have nothing to do with the Value Modifier that CMS is working on.

CO-CHAIR CACCHIONE: Well, but they could be used by private payers. They could be used by private payers to create public reports around providers. In cases now, they are deselecting people based on economic profiles, and they are getting pushed out of the network. So, these things are being used in ways that are disadvantaging providers.

MR. DE BRANTES: So, are you suggesting that that ought to stop unilaterally?

CO-CHAIR CACCHIONE: No, I'm --

MR. DE BRANTES: I'm having a tough time here because NQF so far has never gone down the pathway of saying, "Here is a

measure, hemoglobin A1c, and you measure it this way." And there are lots of people out there who are using those measures, as you know, to deselect or rate physicians based on whether or not they meet a certain threshold of their patients with hemoglobin A1c's less than something or above something else. And NQF has never gotten involved in any of that stuff.

CO-CHAIR CACCHIONE: Yes. I would like to say we can stop here because we have a public comment period.

I would just say that it needs to be fair, both for the vendors as well as the providers.

CO-CHAIR MARTIN ANDERSON: Yes, I think, again, this is an area where we have got the range of opinion. NQF has been down this road before that says, do you really get into different bars for different uses? And so, you've done your job by expressing the range of opinion, and that will get carried

back.

So, do we want to go to public comment and pick up on usability after lunch?

MR. WILLIAMSON: Yes, let's do that.

Do we have any public comments in the room?

MS. McLLRATH: I'm Sharon McIlrath with the AMA.

I just would like to say I am not exactly sure where you should put it, but sort of picking up on the question about different specialties and subspecialties, that when you are looking at whether the data is reliable across different data sources and different time periods, that it would be important to look and see, is it sort of consistent across different subspecialties? Because we have heard that is one of the problems with a lot of what is out there.

MR. WILLIAMSON: Do we have any other public comments in the room?

(No response.)

Operator, can you please open the lines for public comment?

THE OPERATOR: Yes, sir.

If you have comments or you would like to ask a question, please press *, then the number 1 on your telephone keypad.

(Pause.)

At this time, there are no questions or comments.

MR. WILLIAMSON: Thank you very much.

We will now break for lunch and reconvene at 12:30.

(Whereupon, the foregoing matter went off the record for lunch at 12:03 p.m. and went back on the record at 12:38 p.m.)

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

1:08 p.m.

MR. WILLIAMSON: Okay. We will be moving on now. We are going to talk about implications for NQF and market implications for endorsing episode groupers, so really challenges and the path forward.

And so, at this time I will turn it over to our Co-Chairs -- or turn it over to Taroon, who will prime us for this discussion.

MR. AMIN: Okay. So, I know this is the discussion everybody has been looking forward to.

(Laughter.)

So, you know, I will preface the discussion by saying that, as we started this work -- well, the first, the time, the way we are sort of going to structure this conversation.

So, we have until about 2:30 to walk through this. We originally talked about structuring this discussion in terms of

implications for NQF endorsement and, then, having some discussions around the implications for application. So, I don't want to take for granted that -- this may be a new effort for some folks and folks may not really know the full scope of NQF activities to have additional conversation around implications.

So, firstly, I think most are familiar with the fact that NQF endorses measures as national consensus standards for use for broad reporting and accountability applications. However, NQF does not endorse particular measures for particular programs. So, I mean, we ask the question about how a measure will be used, and it could be used in a particular program, but we are not necessarily looking at the question -- I mean, criteria doesn't change depending on whether it is being used for a public reporting program versus a Value-Based Purchasing Program.

So, NQF also convenes the Measures Applications Partnership, which is tasked with making recommendations to HHS in terms of rulemaking, in terms of recommendations of measures that should be used for programs.

In this last report that was submitted to HHS on February 1st of this year, actually five days ago, the MAP reviewed 46 condition-specific episode grouper measure concepts. So, they were submitted to the MAP to provide preliminary guidance on, and they were submitted in the form of measurement concepts, so measure concepts, very much in the typical construct of NQF-endorsed measures.

So, this relates to the conversation that we have been having in a number of ways. The first is that the question of use case was a dominant part of our conversation over the last two days. And typically, NQF does not -- you know, this question of how a measure is intended to be

used is not something that is considered an endorsement, but, more or less, something that is considered under the Measures Application Partnership.

So, in a lot of ways, what we are suggesting here is that these two are blended. And so, the original way that this conversation was structured was to have discussion on implication for endorsement and, then, some conversations around applications. And this conversation will obviously be blended for the next session. So, I just wanted to say that upfront.

So, I want to just walk through -- Evan, actually, going back to the slide -- in terms of implications for endorsement, to walk through some of the questions that we wanted some feedback on and guidance from the group on. And I will walk through the questions and I'll turn it over to Kristine to lead the discussion on.

So, what are some of the benefits

and unintended consequences that may come from endorsement of episode grouper systems, episode groupers broadly, is one of the questions that is outstanding. I think this would be the place to have a conversation of recommendations of whether, to David Bodycombe's point earlier today, should NQF be doing this at all?

Given that many of the episode groupers that could be considered for endorsement are commercially-owned and contain proprietary components, how might the evaluation, the endorsement of these groupers impact the market?

Further, should future efforts seek to align the public and private sector in terms of the use of a single endorsed episode grouper system? And again, that fits under the domain of the Measures Application Partnership which is tasked with the goal of trying to align the public and private sector. Obviously, there is a lot of methodological

limitations to being able to do, from the fact that the data structures/datasets are different. But should this implicitly be a goal of our efforts with episode groupers?

Moving on to the next slide, further, I think we have had some discussions about this to a certain end, but I will just raise it again as another consideration, that we, as staff, continue to consider and be concerned about. It is that, given that many of the episode groupers contain user options that might impact the measure score, is it actually feasible to endorse a national standard that would be consistently, that could be consistently applied across users, given that is the function currently of endorsement?

Further, given the complexity of episode groupers -- and this goes back to at least my breakout group where Tom MaCurdy and others noted -- given the complexity of episode groupers, is it actually feasible for

a multistakeholder group to evaluate them and make endorsement recommendations? The consensus development process by its structure requires the multistakeholder group to be evaluating these products. And is that truly feasible, given the complexity of what is before us?

Further, is it an expectation that groupers -- let's just not use the term "grouper measure" -- let's just say, how do we think the grouper should be used in combination of quality measures and, broadly, should we be expecting a quality signal within episode groupers? Or is that not a function of episode groupers that we would expect? Again, with the assumption that, if you just compare with NQF's position on this, just comparing resource use assumes a stagnant level of quality, which we know is not true.

And finally, what are some of the considerations for the path forward for evaluating and endorsing episode groupers?

So, again, I think it would probably be helpful if we maybe take a step back now to the beginning -- Evan, if we can go back a slide? -- and maybe take a few of the sort of questions and concepts at a time. I know this ranges a number of different topics for people, but, as we move forward with this work, these are obviously things that we want to keep in the forefront to ensure that, if and when we start to look at the endorsement of episode groupers, that we are approaching it in an appropriate fashion.

So, I will turn it over to Kristine.

CO-CHAIR MARTIN ANDERSON: So, we saved the really easy stuff for last.

(Laughter.)

Let's go to the first page. Oh, this is the first page. Okay.

So, why don't we just go question-by-question because there's a lot and they are very different topics. And let's start with

unintended consequences. What might happen if/when NQF endorsed an episode grouper, one or more?

Mark?

MEMBER LEVINE: I think the key is what you were saying about use cases, that not every use case is important for endorsement. But I think there is one overarching use case that is important for endorsement, and that is the public acceptability of a grouper that is going to change, potentially impact the physicians of America and how we take care of people, how we treat our clinicians with fairness and openness. And that is a special use case that is actually required by law to be presented to you for consideration. I am not sure that every other use case is as important.

CO-CHAIR MARTIN ANDERSON: You are noting both the use and the scale of the use, correct? Okay.

Dave Mirkin?

MEMBER MIRKIN: I just want to go back to something Francois said, which is, depending upon -- I mean, I am assuming that there will be a bunch of hoops to jump through to get endorsement. And I think one concern would be to really have an effective, an appropriate endorsement process. I have just been struggling with how and why, other than the public source grouper, why anybody else would go through that. And then, it becomes less meaningful in a broad way.

So, I think that would be, in other words, all this work might just be for the CMS grouper, which is okay, which, then, I think would help focus us. And you don't have to worry about all the applications that might be outside of that.

CO-CHAIR MARTIN ANDERSON: Nancy?

MEMBER GARRETT: And sort of adding to that, I think one possible unintended consequence is that, if the endorsement process is really time-consuming

and onerous, then what is in it for the commercial companies to actually go through the process? And if they abstain, then what are the downstream effects? Okay, then, does that mean that a whole set of cost and resource use measures are not going to have NQF endorsement, and we are going to have kind of two levels of measures in this country? We are going to have endorsed measures and not-endorsed measures? And does this start to water-down the effect of NQF endorsement? I think that is one scenario you could see playing out.

CO-CHAIR MARTIN ANDERSON: So, you are raising the specter that, well, then, what if there is non-endorsed episode grouper and a measure that comes in that relies on a commercially-available grouper that might not be endorsed or in your case is not endorsed then? Does that have implications for whether or not that measure could get endorsed? Okay.

MEMBER GARRETT: But even a little

bit taking it further, will the measure developers even bring those measures forward to be endorsed? Or do they just throw up their hands and say NQF endorsement isn't that important anymore in this space because it doesn't mean anything?

CO-CHAIR MARTIN ANDERSON: In this particular space. Okay.

Others? Francois, I know you have an opinion out there.

(Laughter.)

Okay. No one else in the room?

MEMBER BODYCOMBE: You know, I really appreciate what Mark says. I think it is absolutely right on, and we have got to applaud CMS for taking this whole thing on. I wonder if NQF endorsement is the best modality or strategy for CMS to promulgate their episode grouper, though, just to throw that out as a thought.

CO-CHAIR MARTIN ANDERSON: Okay. Marjorie?

MEMBER KING: As a provider, I would ask, what do the provider organizations feel about NQF endorsement? Because if the provider organizations feel strongly that NQF endorsement is very important for what their members are going to be paid against or judged against, then that is very important. In other words, how would NQF endorsement impact acceptability by providers?

CO-CHAIR MARTIN ANDERSON: Okay. David?

MEMBER HOPKINS: I hadn't thought about this distinction, but I can see how CMS would want to have outside peer review of the system that is being built specifically for it. And I can see why logically it might look to NQF to perform that function. I wouldn't want to not see that happen, but I can't see how, getting back to the comments that were made by others, how we are going to pull all these commercial vendors in and, then, the potential startups for the future into this

whole arena. So, maybe there is an answer, which is let NQF serve the needs of CMS directly in some way, and then, let's just struggle with the commercial vendors somehow.

But, again, I really think we have to answer that question about what happens when you have specific Steering Committees at NQF looking at measures of cost and resource use that are generated by these episode groupers. And can we somehow help them get past some of the basic questions that always come up around the grouping?

CO-CHAIR MARTIN ANDERSON: Can you imagine a benefit, any of you, imagine a benefit to an entity with a commercial grouper being endorsed? What is the benefit to them?

MEMBER BODYCOMBE: If I could, just from a purely economic perspective, it blocks entry in a sense for our competitors. So, being endorsed, theoretically, means your product is -- you know, you have a distinct group of products that basically have the

monopoly, and the new entrants are kind of blocked.

And as Francois' point, the innovators who are really out of the box and coming up with whole new conceptualizations of this are completely blocked out because they're not playing by the rules.

CO-CHAIR MARTIN ANDERSON: Other thoughts?

Nancy? And then, Tamara.

MEMBER GARRETT: One other kind of different flavor of unintended consequence, depending on exactly what endorsement meant, I would worry that it could mean that the episode grouper logic is somehow frozen. And we have talked about that a little bit. This is software that is constantly being improved, and there is an iterative feedback loop. Is that going to be included in the "what does it mean to be endorsed?" Because it was endorsed on a given day in its current state, but we don't want to stop it from changing and

improving.

CO-CHAIR MARTIN ANDERSON: Change your mind? Okay.

Dave Mirkin?

MEMBER MIRKIN: One potential advantage would be, I guess, or maybe it is a prevention of a disadvantage, I think NCQA accreditation would help if plans all went through that, of course, that's quite expensive. And then after a while they decided it really wasn't important in order for them to market, and then all of a sudden it became important again because the employer community said we expect spent at least parts of your business to have that certification.

So, I wonder if that is something that might happen in the future to make NQF endorsement quite relevant for the commercial profits; i.e., if the big employers said, you know, "We only are going to use -- we want you as our carrier to use NQF-endorsed metrics of all kinds." And then, if you didn't have

that, obviously -- I can't imagine any of the major players in that market not, then, going for NQF endorsement if the big employers said, "That's what we want." I would ask the big employers.

CO-CHAIR MARTIN ANDERSON: Tom?

MEMBER MACURDY: Well, I think that has already been said, but I just wanted to emphasize. I mean, I really strongly don't think that NQF should try to align public and private and get a single endorsed grouper. I mean, first of all, I think it is a fool's errand. I mean, it gives part to Francois' point, but a lot of points. But you want a lot of flowers blooming here because this is technology we want to develop over time, and we really do want to allow for innovation.

I don't think it can be done, but I don't think you should even put it down there as even trying. I mean, you are not going to have a single grouper within Medicare. It is going to depend upon the

context. And Medicare, it is a particular kind of payment system. So, it is not going to work in commercial. So, one, I just think it makes no sense.

CO-CHAIR MARTIN ANDERSON: Thank you. So, your comment is on the third bullet?

MEMBER MACURDY: Yes.

CO-CHAIR MARTIN ANDERSON: Chris?

MEMBER TOMPKINS: This might be a little bit off-point. It is just sort of flipping the question backwards.

NQF is or wants to be or should be in the business of evaluating resource use measures, right? And under the rubric of the resource use measures, it came out, as far as I know, simultaneously with the consensus view that the resource measures are looked at from the point of view of patient-focused episodes, right? All right. So, NQF already has some official -- I don't know if that is what you call them -- official public published stands on what your business is and how things ought

to be done right by consensus opinion, correct?

And then, in the process of call for measures for resource use measures, you got some measures which maybe not only did not emanate from a grouper and maybe some of it did, right?

The question that I would ask is, how do you know that the ones that did not emanate from a grouper, how do you know that they are valid? And the reason I put it that way is because the purpose of the grouper is to make logical judgments about how the resources were used, often in situations where there is both a competition for determining how those resources were used and in some cases there is actually a joint production that results from the same dollar.

So, when you say that the resource use for this condition or this procedure was "X" dollars, you have already, at least implicitly, done something that the grouper is

trying to do explicitly, which is to look at all the possible ways that dollar might have actually been used. Sometimes it is for more than one thing. You go to the doctor, and the doctor you for several conditions at once. Sometimes you're in the operating room and you get two different operations which may or may not be clinically-related.

And when you say that the resource use is "X", how do you get there? In other words, how can NQF do the job of evaluating resource use measures unless it is implicitly saying something about the logic that was used to parse out how the dollars ought to be allocated to this episode versus the other?

So, I will just end it on sort of a half-joking, dramatic way. How can you go forward in the business of resource use measure endorsement unless they emanate from a grouper?

CO-CHAIR MARTIN ANDERSON: Thank you.

Jennifer?

MEMBER HOBART: I think even from a health plan perspective, there is probably a continuum about how much alignment is optimal the episode groupers and how much variation you want. So, if you are working on trying to change the whole community behavior, you might to go to a collaborative where you all want to be using the same grouper because you are trying to look at the whole claims for the state. If you are dealing with your providers, you might want something that is most familiar to them to facilitate the interaction? And if you are just doing analysis sort of behind the scenes, you could have a lot more variety and niche things that is helping you figure out places you may want to go.

So, I don't think there is a single answer because, even within a health plan, depending on exactly what you're doing, you are going to kind of need possibly

portfolio tools.

CO-CHAIR MARTIN ANDERSON: Thank you.

Mark?

MEMBER LEVINE: I think, again, we are back to use case and, also, the gray area between what is the difference between claims analysis and an episode grouper. There are many measures that come about as a result of claims analysis that are not dependent upon grouping technology.

The use case that I am most concerned with, obviously, is a publicly-acceptable one that does require regular feedback and updating in order to be acceptable. If Medicare were to come up with a process of using a grouper, and the clinical criteria for grouping claims and judging performance, how often should that be looked at again in terms of changing codes, changing care patterns, and things like that?

I think the physicians of America

need to know and to have standards for how reliable, how valid is the technology that is being used in this particular use case. So, anyhow, enough said.

CO-CHAIR MARTIN ANDERSON: Okay. Anyone on the telephone want to chime-in?

(No response.)

Can they hear us yet? Okay.

MEMBER LOISELLE: We can now, but we missed --

MR. DE BRANTES: Yes, we missed the first half-hour or so of this conversation.

CO-CHAIR MARTIN ANDERSON: Oh, okay. So, I guess what I'm asking is, do you want to say anything about the benefits or unintended consequences that might result from the endorsement of an episode grouper system?

(No response.)

Okay. The second question was around impact on the market. So, how might evaluation and endorsement of groupers impact

the market itself for groupers?

MR. DE BRANTES: Well, it is likely -- Francois here -- it is likely they would, and that kind of gets us back to the unintended consequence. Because, unfortunately -- or maybe fortunately, I don't know; some days I honestly wonder -- there is a tendency of the herds to kind of move in the same direction and to say, because this is NQF-endorsed, that is the way it has got to be, which probably is good for NQF, but not necessarily good for the rest of the country.

And so, I think there is an unintended consequence. Because, depending on how these criteria are a set -- and we were able to hear Tom's comments. So, to the point he made, depending on how the criteria are a set, how these standards are defined, you might prevent anyone who is innovating in this field for being able to qualify for NQF endorsement of their grouper, which would, then, kill off their ability to innovate in

this field. So, I think, to me, is one of the biggest unintended consequences of this effort.

CO-CHAIR MARTIN ANDERSON: Thank you.

Dave?

MEMBER BODYCOMBE: I am really interested to hear more, of course, from the folks who actually make episode groupers. For instance, when CMS came out HCCs that are heavily dependent of what is now the Verisk DxCGs, I guess you might argue, oh, well, now there's this free system out there; everybody will stop using Verisk HCCs and use the CMS HCCs, and that never happened. So, Verisk is still, as far as I know, a very healthy company.

CO-CHAIR MARTIN ANDERSON: Okay. Any more on this?

(No response.)

I know we had one comment already, a "no" on No. 3: "Should future efforts seek

to align the public and private sector in the use of a single endorsed episode grouper?" And Taroon gave a little bit of background around the MAP there, too.

Let's hear additional comments, and Tom will get us started.

MEMBER MACURDY: Yes, I just wanted, on the commercially-owned, I mean, I would have thought that would be something that NQF would want to stay far from. I mean, really, you guys are really going to get in the decision process of saying, if we endorse a grouper, it may become the more prominent and other competitors -- I mean, suppose it is a really good grouper. I just think that those kind of criteria, the conflict of interest is going to be huge. It is just something I would completely steer away from.

I mean, if you happened to think that a grouper that is very commercially-successful is a good grouper, it's a good grouper and move on. The fact that people are

going to see that and say, "Oh, it's a good grouper," and they're going to get a stronger market position, it is like, okay, but I can imagine a set of circumstances where you say, "We're not going to do that because we're going to get a stronger market position." I mean, I just don't know how that conversation would go. Conflict of interest, it would be pretty involved.

CO-CHAIR MARTIN ANDERSON: I'll take any more comments on this particular page before we go on.

(No response.)

Okay. Evan, let's flip.

Okay. Now this is, the first bullet is really about, is this even feasible, right? So, we have heard a lot over the last day and a half about how groupers have --

MR. DE BRANTES: The slides haven't shifted in the webinar.

CO-CHAIR MARTIN ANDERSON: Okay. He will take care of it.

So, the question, "Given that many groupers allow for user options that may impact the measure score, is it feasible to endorse a national standard that could be consistently applied across users?"

So, in other words, there is already a lot of flexibility built into these systems. And when you endorse something, is that feasible? Is it feasible to say -- Taroon, is this question to ask, is it feasible to say that certain options are the right options or the national standard options?

MR. AMIN: One of the goals of NQF up to this point, when we are looking at performance measures, is to actually seek to have a national standard. So, the question here is, given that there are many different user options and, as we discussed, depending on the use case, there might be different variations, is it really even feasible to have a national standard in this area? And

obviously, that would have some implications for what endorsement means going forward. So, it is actually more around that.

MEMBER DUNN: Maybe a clarification. What do you mean by "options"? Can you be more clear? I am trying to think of options that affect the grouping rather than post-grouping.

CO-CHAIR MARTIN ANDERSON: The examples that have come up over the last day and a half are clean periods, you know, time continuously enrolled, a bunch of things that the users can set.

And I think the question from NQF is, could there be a national standard on those types of options?

MR. AMIN: Well, actually, it is, can there be a national standard, given those options?

CO-CHAIR MARTIN ANDERSON: Given those options?

MR. AMIN: Yes.

MEMBER DUNN: I think it depends on the options. Some of those, like clean periods, you probably could. I think some people have a preference one way or the other, but it probably doesn't affect -- maybe we can talk to David here -- but it probably doesn't affect greatly the use. You know, it doesn't vary by use cases. It is almost more of a preference. So, it would depend on the --

MR. DE BRANTES: That is actually incorrect. So, I want to make sure that is clear. The use of a clear period is almost impossible if you are going to use a grouper for bundled payments because you need to have a defined time period for payment.

MEMBER LOISELLE: There are other ways of doing this besides clean periods.

CO-CHAIR MARTIN ANDERSON: So, that was both of you on the phone.

And again, diversity of opinion again, that's okay.

(Laughter.)

MEMBER BODYCOMBE: Just to pose another comment, I wonder how many options the CMS episode grouper has. To make it work, it probably doesn't have any options.

MR. DE BRANTES: It actually has lots of options.

CO-CHAIR MARTIN ANDERSON: Okay. I don't think we need to get into the details of the individual groupers. But I think the question really is, do you even have a standard if there is just inherent flexibility, and when users change their options, it can produce different results? So, it is even a standard even when you endorse a grouper if it can be run many different ways? Is that a standard?

MEMBER MACURDY: So, I would answer the question as definitely not. You can't do this.

I mean, let me go back to the example I was using for Medicaid. So, you said that, if it is one state, it is not

national. Suppose it's two states, three states, four states, states in different regions? So, pretty soon, it is like, okay, is 10 state national?

I mean, there is not going to be a national standard because what you do for one set of states isn't going to be applicable for other sets of states. And if you say, well, it's not applicable for one state, is it applicable for two states? Do they have to be in different regions?

And then, I could get into healthcare plans and the same thing, but I will just stay within government plans. I mean, you can't even do this.

So, I think the answer is there is absolutely no way. I mean, I was going to ask what a national standard was, but if you say if it is applicable, it is like, okay, is it applicable to the Medicaid program? That is a reasonable thing to ask in various kinds of ways.

Like I said, in California there isn't even one; there's 58. So, there isn't even a State standard.

CO-CHAIR MARTIN ANDERSON: Okay. So, let's go to the next one. It has come up a lot.

Is it feasible that you could get a multistakeholder group? And NQF defines their multistakeholder groups as having some purchaser, consumer, provider, research organization, et cetera, those entities all present in the group. Could a group like that evaluate episode groupers and make endorsement recommendations?

MS. WILBON: I would say not necessarily. I think we were trying to pose this question in the framework of our current process. Even for measure evaluation, not everyone is a technical expert on the panel. There may be some methodologists. We have consumers who may be interested in the topic, but have maybe some interest in understanding

how costs are going to be applied in different ways or measured in different ways.

So, I guess it is a broad question on whether or not a multistakeholder group could take on this task, if we were to ask them to do it. It is more complex than most of the work that we have taken on before. So, that is really kind of the added nuance.

MEMBER HOPKINS: If you think about it, you are going to get constant resource use measures that have this buried within them. And those groups are going to be evaluating it. So, how can you obviate that?

CO-CHAIR MARTIN ANDERSON: Nancy?

MEMBER GARRETT: So, I think the answer is no. I mean, I just don't think the current process lends itself to this type of work.

Again, as Dan said, there is kind of two elements here. There is the algorithm and, then, there is the software that applies the algorithm. And for groups to come

together in a day and a half, and even with some advanced technical review, I just don't see that it would work.

And you were describing at the break to me what happened with a couple of the measures last year that I thought was instructive. I don't know if we want to talk about that.

MEMBER LEVINE: Actually, it is going to constructive. I think because we went through this process before with NQF, the idea of the TAG, I think you called them, which had a clinical element, there was a methodologist/statistician group, which actually I thought they asked good questions and it was a good process.

I think the challenge was, when all that rolled up to the Steering Committee, did the Steering Committee have enough background to actually make a final assessment? And my opinion is actually some of them, yes. And David was on that group.

I would say you probably need to give greater weight to what the TAG had to say and not let that be overruled by folks who hadn't had the time to immerse in it. You know, nothing against their backgrounds. It was a very esteemed group. It is just they didn't have the amount of time to invest in what was going on that the TAG did.

So, you would probably need in this case to give more weight to the TAG and have them either explain their assessment or reach back to them in a way that the multistakeholder group can be successful. Because they are going to need to subcontract that work out to different folks who have and the abilities to understand it to make a good judgment.

CO-CHAIR MARTIN ANDERSON: Mark?

MEMBER LEVINE: While what you said is I think very appropriate for commercial groupers and others in the space, I think there is a different case to be made

for the particular use case of a Medicare episode grouper, which I think does require multistakeholder input in order to be acceptable in its implementation.

It has got to hear from everybody, methodologists and particularly from those people who are going to be affected by the output of such a grouper. These people are essential to the proper functioning of the grouper and its acceptance and its utility for actually improving care, which is the bottom line that we need to address.

CO-CHAIR MARTIN ANDERSON: So, maybe if you blend those two comments together, the traditional process may not work well, but could NQF design a process that would work well? Maybe.

MEMBER DUNN: A quick clarification. I apologize.

Yes, I wasn't saying that group doesn't have a role. It is just that you may need to rely on --

CO-CHAIR MARTIN ANDERSON: Right.

MEMBER DUNN: -- the experts to inform that group.

CO-CHAIR MARTIN ANDERSON: I think similar things have come up on other technically-complex topics. So, it doesn't sound brand-new.

Marjorie?

MEMBER KING: Yes, and I just want to go with what Mark said. Mark's group has groomed -- how many hundreds of doctors are involved, providers are involved in clinical work groups through the Medicare groupers?

So, we have a little knowledge is dangerous knowledge, but, given a lot of guidance from a Technical Advisory Panel, I think it could give some meaningful input, not just for the Medicare grouper, but also for the commercial groupers as well. So, I think it is doable to have a multistakeholder process.

CO-CHAIR MARTIN ANDERSON: Thank

you.

Any comments on the telephone?

MEMBER LOISELLE: Yes. This is Jim Loiselle.

I agree with that. It seems, if we are focusing on just the needs of CMS and what their full plans are for deploying this, that in a different forum with the right time, that we could come to sort of consensus on supporting CMS and their activities.

I think a broader question around commercial groupers for non-CMS business is probably, as others have stated, probably more problematic.

MR. DE BRANTES: Yes, this is Francois.

So, that, though, doesn't sound like a, quote/unquote, "endorsement," more than setting up a series of expert panels, independent, maybe convened by the NQF, that would have a role in reviewing and providing input to, but falling short of an endorsement.

Because, again, I think as we said, there are implications of that endorsement which create problems and lots of unintended consequences.

One of the roles that I think NQF could have that would be very helpful for everyone in the field, commercial groupers, Medicare, you know, anyone, is to focus on the definitions of the episode. Because I think of those very much like the definitions of a quality measure or any other measure where you have to define what is the population that would be included in this particular measure. So, in other words, what is the denominator?

As Chris Tompkins mentioned earlier, all these groupers have at least one similar feature, which is they have lists of codes, diagnosis codes and procedure codes, that are assigned to any given episode.

And I think that having really significant heterogeneity in different groupers around the country of those specific definitions is problematic. And so, that is

an area where I think NQF could have a very powerful role in creating needed alignment.

After that, the specific decisions, as we have discussed, around do you single assign a service, do you multi-assign, do you do this, do you do that, what are you going to use it for, how are you going to measure, is a different issue. But at least if we have a common set of core definitions of what is the denominator for a diabetes episode, that would be very helpful.

CO-CHAIR MARTIN ANDERSON: Thank you.

Mark, did you still want to say something? Okay.

So, the next question I am going to rephrase because I think it is hard to understand it the way that it is written there. So, Taroon has given me a new way to say this, which is:

"Would you expect to see a quality signal in an episode?" So, the relationship

between an episode and quality, not just cost, looking forward.

So, David?

MEMBER HOPKINS: So, from the nature of the data that drives the episode grouper, I could see some signals about process quality. I can't see serious outcomes being judged. And I hope when we talk about quality, we keep driving towards outcomes, and that is not in these data, for the most part; some are.

CO-CHAIR MARTIN ANDERSON: No, only financial outcomes is the ones that --

MEMBER HOPKINS: Yes, I mean, I get the complications, things like that, you can sometimes pick up.

MEMBER MACURDY: Wait. I mean, if you have a hospitalization, that is high-cost. If you got a lot of that, it is high-cost. Those are outcomes.

MEMBER HOPKINS: Yes, but they're not clinical outcomes.

MEMBER MACURDY: Well, it depends on --

MEMBER HOPKINS: Did the patient get better?

MEMBER MACURDY: -- what gets counted as hospitalization.

MEMBER HOPKINS: Yes.

MEMBER MACURDY: All hospitalizations may not get counted. Suppose it is one related with a particular kind of service --

MEMBER HOPKINS: Yes.

MEMBER MACURDY: -- that was provided.

MEMBER HOPKINS: Well, what I am talking about is the result of healthcare; did the patient get better? That is what I mean by an outcome.

CO-CHAIR MARTIN ANDERSON: Immediate outcomes. He's looking at final outcomes.

Marjorie?

MEMBER KING: It may not be part of the endorsement of this measure, but it needs to linked to in some way eventually. The example of the hip fracture patient, so the hip fracture patient goes home, doesn't get home care services, doesn't get anything, and is stuck at home. It needs to be linked in some way to a functional outcome or a health outcome or something.

Again, I think it will increase acceptability by providers and it will increase acceptability by patients. An unintended consequence could be driving the cheapest cost at the expense of outcomes, functional outcomes and medical outcomes.

CO-CHAIR MARTIN ANDERSON: Okay. Mark?

MEMBER LEVINE: There are some clinical outcomes that you can get from a grouper, occurrence of post-op infections and complications of various natures, readmissions, and things like that, which are

measurable quality measures, not really good ones, the kinds of ones that all of us --

MEMBER HOPKINS: They are measures of bad quality, though.

MEMBER LEVINE: I'm sorry?

MEMBER HOPKINS: They are measures of bad quality.

MEMBER LEVINE: Well, yes, yes. Yes, they are measures of bad quality.

But, also, we need to evolve a structure -- I am going to call it a "skeleton" -- upon which we can eventually graft the muscles and the ligaments and the organs in order to have a fully functional system that will inform simultaneously upon both resource use and quality in terms of positive outcomes as well as negative outcomes. And it has to start someplace.

So, what I am hoping is that we can evolve such a skeletal system coming from the grouper looking at available information, which is basically claims, but in the near

future be able to add to that quality measures that might come from our PQRS system or from electronic data or from the Hospital Quality Reporting System, a variety of other things. Eventually, who knows? Registries and other kinds of activities that might, then, further populate this kind of data, which is starting with claims data, but might eventually get someplace that would be a better place for it to be.

We can't put it where we want it now. We have got to start someplace.

CO-CHAIR MARTIN ANDERSON: Okay. Steve?

MEMBER BANDEIAN: Actually, I think Mark pretty much said 98 percent of what I would have said.

I do think that the grouper technology actually does have the capability of generating a huge number of meaningful outcome measures at this point in terms of complications, readmissions, retreatments,

even a very simple thing, which is how long does the episode last, meaning kind of that is sort of a mark of how long it took to get resolved.

Depending upon the actual grouper system, we could look at issues such as delay in diagnosis, misdiagnosis, acute exacerbations, disease progression in terms of stage.

And you put these all together and, yes, they are not an SF-36 or SF-12, but these are actually pretty important, as you say, negative outcomes, I suppose. But if you don't see those, what that is saying is that the patient sailed through the process, got better, and everything was fine.

CO-CHAIR MARTIN ANDERSON: Thank you.

Dan?

MEMBER DUNN: Yes, I agree with the points so far. Just maybe one comment, though. It is not always going to work well

to try to think of step one; an episode grouper runs. You have all the instance of an episode, all the service that gathered to it, and so on.

You may have a case, I think, Steve, a few good examples, where you may be able to mine the data that went through that episode and create a quality measure. But in many cases you are going to need to have the quality measure and the episode methodology be separate.

An example could be that a service that is really important falls outside the episode or it could be even collecting together episodes that make sense. So, they don't necessarily go together. It is not always going to work, but it is worth thinking about when they can.

CO-CHAIR MARTIN ANDERSON: Thank you.

Tamara?

MS. SIMON: I finally got to the

comment that I was thinking about making earlier. And I apologize because this is a little bit of a secular comment.

But I would certainly hope that an episode grouper would ultimately be tied to a quality measure. And I guess I am a little bit concerned about some of it. My comment I guess goes back to intended use, as well as unintended consequences.

And there has been this push in the adult world to look at readmission rates, and that has trickled down into the pediatric world, where our providers are getting dinged for readmissions. We have really found in children readmissions -- I know there is the debate in the adult world about how much readmissions are preventable, but they are really not preventable in the vast, vast majority of pediatric cases.

So, I guess I just think these specifications need to be made very explicitly, narrowly, or at least with the

knowledge that they may be broadly applied and have huge ramifications.

CO-CHAIR MARTIN ANDERSON: Thank you. That's a good point.

Anyone on the telephone want to comment on this particular question?

(No response.)

Okay. Oh, David, sorry.

MEMBER HOPKINS: Just sort of a summary of what I think I have heard is two different questions. That last one is the bigger one: outside of the episode grouper, are there quality measures? That is a whole issue that NQF I think is looking at.

Inside the episode grouper, what we have heard is there are, in fact, markers of quality, usually bad quality, that you can use a grouper to identify and diagnose and, hopefully, use for quality improvement. That is all great stuff.

I would just point out that, from a measurement point of view, it is sort of

redundant to look at the resource use and say, gosh, there are a lot more resources going into the care of this patient. Why is that? Because they had complications that shouldn't have happened, or whatever you guys cited. That is an issue in itself.

And then, what led up to it is kind of under the surface and important to understand if you are going to do quality improvement, but I am not sure it is an independent signal.

CO-CHAIR MARTIN ANDERSON: Helen?

DR. BURSTIN: Just a quick build-on to David's comment. I think that is a really important consideration. I think that in this day and age the idea that we would be putting a lot of emphasis on quality measures emerging from claims only without thinking about a more hybrid approach of bringing in increasingly the available good clinical data coming off of electronic systems or patient self-reports, it is just I think we would want

to see this as part of that bigger system, but I don't think we want to go back to the days where quality signals are completely driven off claims alone.

CO-CHAIR MARTIN ANDERSON: Great.

Okay. So, now the very broad question: "Any other considerations, if you think about the path forward, to evaluate and endorse episode groupers?" These are maybe some of our summary comments, anything you feel like you haven't shared that the Working Group, whose full-time job it is to take into account all of these wonderful comments should have on their minds.

I have one I will just offer, which is that we are talking about -- I think the word "path" is right. You know, we are talking about going potentially down a path of endorsing what is effectively software.

And there are some little pieces of that in the past, say around risk adjustment, which is usually represented in

software ultimately, and more and more so as e-measures come onboard we start to look at implementation of measures in software.

I don't think that is something NQF should take lightly. It may have future implications for how you think about endorsement. And the systems, right now, we can separate the measure from the system it is embedded in. And in this case, it is hard to do that, to take this algorithm and divorce it from the system it is embedded in.

So, there may be some other thinking that NQF could do about where are we headed that this might be just a signal for. So, on the one hand, you could say maybe we just won't go down this road right now; we are not ready to deal with that, but it doesn't mean it is going away. There will be another instance of embedded measurement and software that will take a different form. So, maybe thinking about that more broadly might help decide what to do with episodes.

MR. DE BRANTES: Well, Kristine, this is Francois.

Just to emphasize your point, and so that the rest of the Committee members can think about these types of unintended consequences, last year the NQF endorsed the HealthPartners Total Cost of Care measure, which has embedded in it the Johns Hopkins Severity Adjustment Formula. And the measure was adopted with that severity adjustment software specifically.

And today you have got communities around the country who are doing some pilots and testing this measure, and they have all been handcuffed because the endorsement included very specifically the commercial software for which there is a license fee, and as a result of which they feel that they can't move away from that particular software and using the HealthPartners Total of Cost of Care, although there is no evidence whatsoever that that particular risk adjustment software

any better. And with all due respect, because I know someone from Johns Hopkins is here, but there is no real true evidence that a severity adjustment model is significantly better than another one.

So, I just think we need to be very cautious about this. Because when you do get into the business of endorsing more than just a measure, but software, it does have unintended consequences.

CO-CHAIR MARTIN ANDERSON: Dan?

MEMBER DUNN: I am not disagreeing with Francois' point or yours, Kristine, but maybe it is a point of clarification. I think you need to make a distinction between methodology, which ACGs is a methodology. Some of these groupers are a methodology in software.

Because even with HEDIS measures there's probably a hundred different software systems out there that take the same exact methodology from HEDIS. So, just in terms of

terminology, I think you should describe it, if it is a more global methodology or something that you are picking one piece out of, but it is really not -- NQF shouldn't be in the business of endorsing software I guess would be my thought then.

CO-CHAIR MARTIN ANDERSON: Yes. Yes, I get your point. There can be distinctions there. I think just that they are embedded.

Largely, it would be hard, though, to take a measure you can do by hand, like literally you do it; you write it all out; you could do it by hand. It would be very hard to do episodes by hand, right?

MEMBER DUNN: I think it depends on the system.

(Laughter.)

Yes, I'm serious.

CO-CHAIR MARTIN ANDERSON: You're smarter than me, yes. Not by hand, sorry. I hear you. But I hear you. The language is

important is the point you're making.

MEMBER DUNN: Yes.

CO-CHAIR MARTIN ANDERSON: And I think that is a very good point.

MEMBER DUNN: Because you could endorse, say ETGs, for example, as a methodology, and say there was someone else who took it and built it into their own software. That is a whole totally different thing, right?

CO-CHAIR MARTIN ANDERSON: Thank you. Right, right.

Your point is that we're not actually endorsing an application, a software application. It is really the methodology. I get that, yes.

Chris?

MEMBER TOMPKINS: Maybe we are reacting a little bit to the tying together you did sort of emphatically about software. There are some things that sort of we all have in common, like we're human and with similar

anatomy -- (laughter) -- and healthcare conditions, and so forth. I mean, these sort of rise above groupers.

And so, part of what might be -- again, just in the spirit of that last bullet -- you might want to just break down the problem and say we can evaluate some things, but we would prefer not to or we can't get into other things.

So, just to be a little bit sort of parochial, we think of and the people that I was in the Clinical Logic Group with yesterday sort of got this from me, which is to say that we have a software application. The software application embeds in it rules, generic rules, and it offers some options.

But all the clinical intelligence resides elsewhere for the most part, outside the software. In other words, what is a condition? And I have sort of said this before. What are the diagnosis codes that clinicians agree on are a useful, clinically-

meaningful concept of that condition? And then, within that, what are various what we call subconditions that might differ in terms of etiology or manifestation or relevant services or complication risks, and so forth?

And then, on the procedures side, what do people actually do and bill and pay for? And to what extent are they done jointly? Or to what extent is one done subordinate to another, et cetera, like that?

It could be possible that that's what, as humans, we want some consensus around. And I think Francois said it in passing earlier, to say that you probably don't want groupers that are just interpreting human anatomy or healthcare differently. Why start off on something that is so basic for which there actually might be a reasonable consensus? And there might be a reason to have various perspectives meld down into a consensus about what it is.

And then, the software that reads

that can have a lot of different -- it could be written in a different language. It could be written, you know, as a large enterprise, and all these variables we talked on the software side may be left to the industry or other users in terms of how to interpret the information that essentially clinical and essentially has to do with how healthcare is provided and paid for.

CO-CHAIR MARTIN ANDERSON: Thank you.

Steve?

MEMBER BANDEIAN: A whole elaboration on what Chris just said. Some of the clinical logic I kind of personally regard as not matters of opinion, but kind of matters of fact.

And one of the sort of big sort of problems I sort of see in the general world is, if you were to ask somewhere, where is a listing of all the ICD codes that are symptoms that are recognized as being caused by

condition A, B, C, D, or E, to my knowledge, that does not exist anywhere.

And so, part of the logic that is being constructed is really sort of almost a medical encyclopedia that sort of displays more factual information rather than opinion information, that actually is pretty useful for a broad variety of purposes. I mean, that may be irrelevant for your question, but I was just amplifying on what Chris was saying.

CO-CHAIR MARTIN ANDERSON: No, I think that is helpful. So, what you are saying -- and I think it was said in a different way a little bit earlier -- is perhaps there is value in separating out the inputs, the clinical inputs, from the routines that they are used for in software, I mean just as a consideration.

The one thing that comes to my mind, though, is that, then how do you do that without getting all the way to best I class, right? Because would you recognize multiple

different inputs for the same clinical conditions. Just two parts.

And then, secondly, what, then, would be the impact, say, of payment rules? Because data often look a certain way because of the way the payment system is structured which it operates in. Would they be similar? And enough limitation on that variability that you actually could come up with a comprehensive definition across multiple different payment data sources?

MEMBER BANDEIAN: So, actually having worked on a fairly wide range of data sources, and seeing how they work, I actually like the Medicare data. It has got some really nice features.

So, yes, there are some major problems in implementing some logic off of something that is not Medicare data. However, this highly-abstract concept of let's get some medical facts out there is really not so much affected by that.

Again, I am just simply saying, for example, what are the signs, symptoms, complications associated with various conditions? Yes, there are business implications and competitive implications about that, and I have no opinion about any of the above. But I am saying that some of the content of the grouper, I would assume -- I can't, obviously, speak to all groupers in the entire universe -- but some of the content of the grouper really is, hopefully or ideally, things which clinicians with knowledge of that subject area would nod their head and say, "Yes, that's right. That's just what we all know. We just don't have it in machine-readable form."

CO-CHAIR MARTIN ANDERSON: Chris?

MEMBER TOMPKINS: Well, it depends on just how far you want to take it. I mean, I started with the very basics, right, humanity? And then, you say, okay, let's go from here and decide when the necessity or the

compelling reasons for consensus start to drop off.

But just picking up on what you said, this medical encyclopedia, which is the way Steve characterized it, could have either footnotes or separate chapters that say that, when you are articulating the procedure codes through a certain type of billing system, these adaptations or modifications are notable and acceptable, something like that.

I'm just saying that it shouldn't be a barrier to the fact that there could be different administrative systems out there. It shouldn't be a barrier to that more compelling, in my view, benefits of consensus that have to do with areas of extreme commonality that might become a useful encyclopedic reference from which NQF could say, if you're using the encyclopedia as written, then we endorse that aspect of it. And if you deviate from it, so note the deviations.

CO-CHAIR MARTIN ANDERSON: Thank you.

Taroon?

MR. AMIN: I was just going to say, are we --

CO-CHAIR MARTIN ANDERSON: That was our last question.

(Laughter.)

MR. AMIN: Well, first of all, we wanted to add one into this implication.

CO-CHAIR MARTIN ANDERSON: Okay, go ahead.

MR. AMIN: And we also have the Measures Application portion.

CO-CHAIR MARTIN ANDERSON: Oh, no, I meant of this section. So, I turn to you when the section is over.

MR. AMIN: Yes. Oh, okay. Okay. Great.

So, there is one additional question that we wanted to add into this implication-for-endorsement section, and one

of the questions Ashlie brought up, which I think is a really good one. So, we talked a little bit about the implications of endorsing essentially a methodology that would be updated on a fairly regular basis. The current endorsement/maintenance infrastructure that NQF maintains assumes that, as guidelines are updated, measures get re-reviewed to ensure that they are -- or as the evidence gets updated, that measures are re-reviewed and we have an ad-hoc review process. But, typically, measures are reviewed on an every-three-year basis.

What type of maintenance infrastructure would be needed to support a continued or appropriate continued endorsement of an episode grouper that we could describe?

MEMBER LEVINE: I could envision a process where CMS might publish these measures are due for reconsideration this year. Here is a list of them. Here are their specifications. Please comment. And just

make that an open and public input into the definitions that are being used to define the episodes, et cetera, and the clinical logic that is underlying their utility. And that can be done, I think, on a regular and recurring basis in an open and a public manner, similar to rulemaking.

MR. AMIN: Okay. If there are no other comments, I just want to remind the group, separate from endorsement, but related, obviously, in very clear ways, NQF convenes the Measures Applications Partnership who is tasked with providing HHS with guidance in its pre-rulemaking activities. And this February the MAP reviewed 46 condition-specific measures that were the result of the episode grouper, and the MAP specifically conditionally supported these measurement concepts, pending NQF endorsement, recognizing that these measures were critical to the Value-Based Payment Modifier.

So, there are some questions here

that we have that I will just run through, and then, I will turn it back to Kristine again.

The first is, what is the implication -- so, I will say this is we are departing a bit from a general conversation about episode groupers and going to specific conversation about the Medicare grouper and its particular application for the Physician Feedback Reporting Program and the Value-Based Modifier.

So, the questions we have here are: "What are the implications of NQF endorsement on the use of the episode grouper for the specific applications; for instance, the Value-Based Payment Modifier?"

"Is there specific implementation guidance that the Committee would offer future MAP Work Groups in the evaluation and selection of episode grouper measures for the Physician Feedback Program and the Physician Value-Based Modifier?"

Third, "The MAP currently

evaluates episode grouper by selecting individual measure concepts within it; for example, a lung cancer condition episode. Can individual episodes within a grouper be selected for implementation of the Value-Based Modifier without considering the cost assigned to other co-occurring conditions for an individual patient?"

And finally, "What are some of the considerations for the path forward in selecting episode grouper measures for specific applications?"

CO-CHAIR MARTIN ANDERSON: I just want to verify some common knowledge. Does everyone here know what the MAP does? Can you raise your hand if you know what it does? If you do know?

(Show of hands.)

Okay, that is a very small group to have the conversation. So, Taroon, I think you're going to have to give them more context on the scope of the MAP and why they are

looking at this issue.

MR. AMIN: Okay. So, maybe Mark can help me, too, because there is a lot of this pre-rulemaking activities that I am actually not that familiar with, either.

But, from my understanding, the MAP -- actually, Helen, why don't I turn it to you?

(Laughter.)

DR. BURSTIN: I am sure you could do it.

So, the MAP is a group that is a multistakeholder partnership as well that is charged by HHS with providing input to CMS, in particular, on pre-rulemaking. So, when measures are up for particular programs -- should this measure be part of the Value-Based Purchasing Program for hospitals, should this measure be part of PQRS for physicians, for example -- that group is, then, charged with looking specifically at whether that particular measure would be beneficial to that

program, and not necessarily looking at the scientific acceptability of the measures in the way we have been talking for the last two days.

So, more so, does it fit within the context of that program? Does this offer an important piece of measurement that would help improve that program? So, a much more programmatic view, it is a recommendation to CMS about whether that measure should be included in the various programs for rulemaking going forward.

CO-CHAIR MARTIN ANDERSON: Clear? Did that help? Okay. So, now we will go back to the questions.

So, Taroon, I just want you to help me with one thing.

MR. AMIN: Yes.

CO-CHAIR MARTIN ANDERSON: When you say, "What are the implications of NQF endorsement on the use of an episode grouper for specific applications" -- yes?

DR. BURSTIN: Having sat through this discussion at the MAP, for example, part of what -- and Corette is here from CMS as well -- so, part of what came to NQF as part of the review for the MAP this year were several of the episode-based grouper measures. And so, the question is, it was difficult, I think, for the MAP to evaluate those measures without a lot of context, without a lot of information, frankly, about how they would be constructed.

So, I think the question here is, how would the endorsement process for those potentially be helpful to the broader process of thinking about which measures are appropriate for which applications, pay-for - performance, penalties, et cetera?

MR. AMIN: And maybe the goal here is just to really address maybe Question 3, yes, Question 3, which, effectively, the MAP was given the 46 condition-specific episode measures. And so, there is this broader

question I think we have sort of alluded to, had some conversation around. But, essentially, they are selecting individual measure concepts from the episode grouper. So, the main question is, can individual measures within the grouper be selected for implementation without considering how costs are associated with other co-occurring conditions for an individual beneficiary or patient? Well, in this case, the beneficiary.

CO-CHAIR MARTIN ANDERSON: Chris?

MEMBER TOMPKINS: That is a variation of the question I asked before. So, I will just take out the phrase "within a grouper". Can individual measures be selected for implementation without considering how costs were assigned to other co-occurring conditions?

That is what I am saying, is that as soon as you say that here is a resource use measure for a condition, you either have a grouper that has in front of you a hopefully

consistent, logical basis for doing that or you have a standalone measure. And it is left to the imagination of the reviewer as to how the costs are individually assigned to that condition versus other ways in which the same dollar may have been used.

MR. AMIN: So, when you asked the question the first time, I was debating whether we should go down this road. I think it is obviously a very important question.

I think the difference in the way that NQF has looked at that, those two types of resource use measures in the past, is the complication of measures that are outputs, essentially, of an episode grouper is that the decision logic or the tiebreaker logic, as we have discussed here the last two days, is not always clear, transparent, or understandable, meaning that when there are potentially two co-occurring conditions, how the individual dollar is either single-counted, double-counted, or just attributed to another episode

that is present for the patient is not always clear. When you are looking at just an AMI sort of measure that is coming out of a grouper, you might not know that that cost went to, for example, the diabetes measure.

CO-CHAIR MARTIN ANDERSON: And Chris is making the point that that is at least discoverable, if it is an episode grouper, and it is not discoverable if it is just a standalone measure, right?

MR. AMIN: And that's fair.

CO-CHAIR MARTIN ANDERSON: Right. So, the same issue is present is your point, right?

MEMBER TOMPKINS: It seems to me it is worse if you don't have the grouper --

CO-CHAIR MARTIN ANDERSON: Right, because it is not discoverable.

MEMBER TOMPKINS: -- because you don't have that articulation of the rules that were used.

CO-CHAIR MARTIN ANDERSON: Right.

Mark?

MEMBER LEVINE: If there were an NQF-endorsed episode grouper for Medicare resource use, then that would obviate and shortcut your having to review individually in great detail each one of those 46 measures. You could, instead, concentrate on whatever other aspects of the measure that you wish, but you would be confident that the construction process that led to the measure met a certain degree of standard, stakeholder input, peer review, a variety of other kinds of things that would be useful.

So, I think it would make your review of the individual MAP measures to be much more efficient if there were the presence of an endorsed Medicare episode grouper.

CO-CHAIR MARTIN ANDERSON: David?

MEMBER HOPKINS: I am thinking exactly like Mark. So, MAP is in the business of providing advice to CMS specific to their use of measures for the Medicare program.

CMS is also developing an episode grouper. It seems to be very logical that NQF have a mechanism for evaluating the grouper, you know, as a grouper. Take all the great ideas that have come up here and apply those in that process, and set up a process to do that.

Medicare is not using commercial groupers. So, why bog it down with that? Why bog NQF down with that? It is a separate issue.

CO-CHAIR MARTIN ANDERSON: Okay. Yes, I hear you.

And my take on the specific question around without considering -- it seems to me that the MAP would want to understand what the rules were, right? So, they know at least what it is they are accepting as an episode relative to other common conditions that would be co-associated with that episode.

But, to David's point, if they

were comfortable with the grouper, over time that would get easier, right, because they would come to understand how the grouper operates, and it would be transparent, but for this purpose of specifically to advise Medicare.

Something else you want to get out of this part of the conversation?

We have exhausted Taroon.

MR. AMIN: Well, we still have Evan.

CO-CHAIR MARTIN ANDERSON: Yes, I know. Okay. It is time for Evan.

MR. AMIN: I will turn it over to Evan for the next steps.

MR. WILLIAMSON: Okay. We'll see if I can get through this.

So, at this point, we will take all the feedback we received over the last day and a half, almost two days; we'll say two days, the last two days, what feels like four days. And we are going to summarize that into

a report that we will provide back to you for comment.

We are going to have two post-meeting calls. They will be needed. We are anticipating to use both of those. I think there is a lot of reacting to what we put together that we are still going to want input from the group on before this goes out for public and member comment.

So, we are anticipating to be able to put that report out by March 24th. We have a post-comment call on May 14th where we will consider all the comments we receive. If we receive a lot of comments, we may need to schedule additional time, but I am hoping we can just use that two-hour block to get through all the comments.

And then, following that, we will adjust the report as necessary and get our CSAC and Board approval by July.

I guess, at this point, are there any questions or comments, concerns, feedback

about the meeting? Anything you want to share with us?

Again, we will take some comments right now, but if something comes to you after the meeting and you want to send us an email or provide feedback, we are happy. I know last night someone mentioned that they were thinking about something that they wanted to send afterwards. Please feel free to send us any post-meeting feedback, whether it be about the meeting itself or about the content area, anything that we are going to be putting into the report. Please do that.

So, I will open it up. Are there any final comments, questions?

Everybody just wants to get out of here?

(Laughter.)

CO-CHAIR MARTIN ANDERSON: I would just like to thank all of you. Really, really great input, and you made this job a lot easier.

MR. AMIN: And thank you for the Chairs, obviously.

(Applause.)

MR. WILLIAMSON: At this point, I will open it up. Are there any public comments in the room?

(No response.)

Operator, can you please open the lines for public and member comment?

THE OPERATOR: If you would like to ask a question, please press *1.

(Pause.)

There are no questions at the moment.

MR. WILLIAMSON: Thank you very much.

This ends the call. Thank you.

(Whereupon, at 1:55 p.m., the meeting was adjourned.)

C

F

G

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: 02-06-2014

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