NATIONAL QUALITY FORUM

Moderator: Medicare Grouper
June 3, 2014
3:00 p.m. ET

Operator: Welcome to the conference. Please note today's call is being recorded. Please stand by.

Ashlie Wilbon: Good afternoon, everyone. Thank you for joining us and for being patient. We wanted to make sure that everyone was able to get on the phone line and in the webinar OK.

Male: Yes. And looks like I'm on my computer, so I'm good to go.

Ashlie Wilbon: OK. For those of you that are dialed in to the speaker's line, if you could mute your phone if you're not speaking, that prevents a lot of the feedback and the background noise from interfering with the call. Thanks.

So, again, thanks everyone for joining us. This is the call to discuss the comments that came in for the draft report for the Episode Grouper Evaluation Criteria. So, we're going to start out with a roll call to see who from the panel is on the call. This is Ashlie Wilbon, I'm one of the NQF team members. I'm joined by Beth Carey, who is the Project Manager now for this project. I'll let her introduce herself, and I believe Ann Phillips is on the line, as well as Taroon Amin.

Elizabeth Carey: Great. Thank you, Ashlie. So, as Ashlie said, my name is Beth Carey and I'm the Project Manager. So if you need anything, just send me an e-mail or give me a call and I'll try to help you out. I'm going to do a roll call and I apologize in advance if I mispronounce your name, just let me know.

So, we have on the line our co-chairs, Kristine Martin Anderson and Joe Cacchione. And then I'll read off the list of our committee members.

Stephen Bandeian, Stephen are you on the line? OK. David Bodycombe?

David Badycombe: Present.

Elizabeth Carey: Thank you. Francois de Brantes?

Francois de Brantes: Here.

Elizabeth Carey: Thank you. Dan Dunn? OK. Nancy Garrett?

Female: (Present).

Elizabeth Carey: Present? Was that a present?

Ashlie Wilbon: I think that was an echo.

Elizabeth Carey: Jennifer Hobart?

Jennifer Hobart: Here.

Elizabeth Carey: Thank you. David Hopkins?

David Hopkins: I'm here.

Elizabeth Carey: Thank you. Jim Jones?

Jim Jones: Yes I'm here.

Elizabeth Carey: OK, great. Thank you. Marjorie King?

Marjorie King: I'm here.

Elizabeth Carey: Thank you. Mark Levine?

Mark Levine: Present.

Elizabeth Carey: Thank you. Jim Loiselle?

Jim Loiselle: Here, present

Elizabeth Carey: Thank you. Thomas MaCurdy?

Thomas MaCurdy: Yes I'm here.

Elizabeth Carey: Thank you. Jelani McLean?

Jelani McLean: Here.

Elizabeth Carey: Thank you. David Mirkin?

David Mirkin: Present

Elizabeth Carey: Thank you. James Naessens?

James Naessens: I'm here.

Elizabeth Carey: Thank you. David Redfearn?

David Redfearn: I'm here.

Elizabeth Carey: Thank you. Andrew Ryan? OK.

Tamara Simon? OK.

Christopher Tompkins?

Christopher Tompkins: Yes. I'm here.

Elizabeth Carey: OK. Great. Thank you.

Ashlie Wilbon: So, thanks, Beth. So, now that we have an idea of who's on the call, I just wanted to get things started. And the purpose of today's call is to give you guys a sense of what the theme for the comment that came in. We also send along with the memo, with the information for this call, we provided you an Excel spreadsheet of all the comments and some themes that we've pulled out

from those various comments provided the commenter name, organization and also some proposed or draft responses that NQF staff pulled together on behalf of the panel.

So, we hope that you had an opportunity to review those. And if you have any specific input on draft responses that we've provided to the individual comments, we'd encourage you guys to send those via e-mail to the team. We'd really like to use today's call to focus on some of the more broad themes that tend to weave themselves through many of the comments and use that – and focus the committee's discussion or the panel's discussion today on those themes.

So, in the memo, you'll note that we have pulled out some of those comments for you or some of those themes for you and post a few questions in response for those themes to get further clarity from the panel on what additional items or topics and things we might want to address in the report to further clarify for the audience on where the panel landed in terms of their recommendation and guidance on those topics.

In terms of the summary of the comments that we received, we didn't receive very many but still a good number, about 40 comments from about 12 different member organizations. And those again varied on various topics and – which you can see reflected in the comment table.

So, on that note, I'm going to hand it over to Kristine and Joe and let them lead us through on the discussion on the themes that are outlined in the memo. And again, if there's any additional items that the panel members identify either within these topic areas or additional theme, we encourage you to suggest those as well and we can figure out how to squeeze those in if there's time.

Kristine Martin Anderson: Great, thanks. And this is Kristine, and Joe and I have an opportunity to just talk for just a couple of minutes in a prep call. And one thing I want to call your attention to as we go through these major themes and issues for discussion is that our purpose is really not to figure out what a good

grouper looks like, so much as it is to figure out what's the right way to evaluate groupers.

And so, we might modify some of the questions that you see on the major themes and issues for discussion to make sure that they're centered on our role as coming up with evaluation criteria. I think it's very pertinent for our first one – Joe and I are going to kind of take turns on here. For the patient-centered episodes of care approach, as you know there were some a lot of dialogues that we've had as a group about that. And then in the draft report, the acknowledgment of that dialogue was there and commenters did come online and comment about either supporting the patient-centered episode of care approach and also acknowledging challenges for using that approach or provider profiling. And then the draft report describes provider-centric episodes as well.

So I want to modify question one because we really don't want to get in to how do you use a grouper, right? At one approach or another and say, does this group agree that whatever the evaluation criteria we would have, that it would have to allow for groupers that may come in for endorsement that may taker either a patient-centered or provider-centered approach or some other kind of hybrid in between. So, from a starting point of view that our criteria should be applicable to both of those types of approaches.

Thomas MaCurdy: That sounds very reasonable.

Kristine Martin Anderson: OK. And then ...

Thomas MaCurdy: Kristine …

Kristine Martin Anderson: Yes, go ahead.

Thomas MaCurdy: Kristine, this is Tom, I mean, obviously I agree with that. I mean I do have some concerns about the way provider-centric and patient-centric are kind of described in the report. But I can send an exclusive comment if you wanted about that.

Kristine Martin Anderson: That would be great and I know that sometimes, you know, I'm sure we don't have the wordings exactly right. So, that would be great. But there's another question which is, is there any kind of – is another type of approach that we want to make sure that we were able be open to beyond the two described in the report?

Is there a third way that this could be thought off?

Male: You mean procedure-based?

Kristine Martin Anderson: Right. Could there be procedure based? Could there be – you know, whatever the group thinks to make sure that we're making sure that when we test out our criteria, we're thinking of the kinds of groupers that we would want submitted.

Male: And maybe Kristine we could just leave and so that the document should reflect a flexibility that if something – a new method of – a new different type of centered approach that we don't quite understand yet, becomes available that this grouper tools need to be flexible enough that they may need to account for that.

Kristine Martin Anderson: OK.

Male: Some integrated delivery network, something that may look and feel different, you now, is it patient-centered, is it provider-centered but is there something new and I think just having making sure that we live language and it create some flexibility and it doesn't pin us into one or two or three columns.

Kristine Martin Anderson: OK. And ...

David Redfearn: This is David Redfearn, I think it would be worthwhile to mention specifically procedure-focused episodes. Those are being used more and more now, often times in term of specific specialties or super specialties what they talk about highly-specialized physicians that focus on just a few particular procedures. That model works pretty well for those kind of providers if you want to do provider profiling. So, I think its worthwhile mentioning that at least.

Kristine Martin Anderson: OK. At least acknowledging that there can be disease or procedure-based. So in addition to talking about what you might be thinking about in terms of whether you're looking at care from the perspective of a patient or from the perspective of a provider, also allowing sort of the fact that there could just, you know, very unique and capsulated situations where you just want – you want to know an episode about something specific like a procedure. Is that what you mean?

Mark Levine: This is Mark Levine. Would the fact that the – should an endorsement be considered for a grouper that does not specify what its intents and purposes?

Kristine Martin Anderson: That's a good question, Mark. How would we evaluate something for whether or not it meets our criteria if we don't know its intent or purpose?

Male: Yes. So, one would think then that one of the criteria for even consideration of endorsement would be the need to specify what the intent is.

Kristine Martin Anderson: OK.

Male: Actually, Mark, I think that's a really good sort of overarching theme that should sort of help govern these different center approaches, the intended use.

Kristine Martin Anderson: OK.

Male: There's some language in there – to the staff, some language in there that would reflect or have that as an overarching theme and that all of these different centered approaches then could fall under that depending on intended use.

Kristine Martin Anderson: OK. And just to the staff. Let's take that under consideration because I do remember we had some dialogue in our in-person meeting about stating intent. So, let's just check on that current criteria and make sure we're covering that.

Ashlie Wilbon: Yes. There's – hi, this is Ashlie. There is already some language in the report that specifies that that would be information that the developer would need to

submit. So I think we're covered but we can definitely provide maybe some more discussion about that.

Male: So, I think this highlights something that, you know, these things are sort of arbitrarily bucketed but there is a lot of crossover when we're going to involve these different topics as we start to think about questions one through five or whatever we have, one through five. In a lot of these, there's a lot of overlap.

Ashlie Wilbon: Right.

Kristine Martin Anderson: Yes. And I think for question three, if you all would agree with this, I think we don't need to have that conversation, right? I mean, if there is a discussion, the kind of conversation we maybe need to have is would the criteria – and we might not to do this offline, right? If you look at the criteria as it's currently drafted, does it cover all of the scenarios that we think that episodes might be built around?

So, if whether you had in your mind, it was a patient-centered approach to the criteria, look liked they're correct. Take that hat off, put on a hat on for provider-centeredness, do they still look like it asked the right questions? Put on the, you know, a very specific procedure hat, does is it still look like it has the right questions? As you do your evaluation and you're able to send some comments in, it would be very helpful if you could try that exercise and make sure that we are being inclusive enough in our criteria that it would work. And that we don't – we are not expressing a bias through the criteria.

If anyone has anything specific to share about, you know, what their views on that right now, that'll be great. If not, I think that's something you guys could – we could all take as a homework assignment and make sure that it's working.

Joseph Cacchione: I think the intended question, an intent question raised earlier as you start to think about all of these overarching questions, questions one through five, I think probably permeates all of those. So maybe there is – I think it is addressed in the paper. So, that intent really is starting to look a question all five different topics we're going to cover today. That intent is sort of – leads through all of those. So, think about it. We should make sure in terms of

staff, I think it is in there, we just – it does relate to all five of these when you start to look in …

Ashlie Wilbon: OK. Joe, do you want to take the next one, starting on the topic two?

Joseph Cacchione: Sure. Yes. So, demonstrating validity of the grouper, and this is broken down into really three areas. It's sort of chronic episodes and about – and with the co-occurring conditions or co-morbidities or other episodes that might be crushing over as limitations of the current data and how that data is being used. There are some address in letter B about the ICD-10, and then risk adjustment which there was – from the provider side, there was lots of discussion around how risk adjustment will be done whether the data is limited in what it can do and whether there should be – and whether the validity of the risk adjustment using the current methodology in this first in this is going to be adequate.

So, open up for comment there. If we start with the letter A, are there thoughts around defining better definition of the chronic episodes and how things are going to be handled with the co-existing conditions or co-occurring bundles that might occur.

Does everybody agree with the – just sort of the – what the staff have outlined here in the response? I mean do they have …

Kristine Martin Anderson: I can offer you my thoughts which is that in general – this is sort of a general thought, this is Kristine, I think it's always relatively risky in advance of knowing intent to try to tell somebody how they should or should not or even if they should or should not risk the severity adjust episodes.

So, I think we should address, one, is there risk adjustment or not and justification for that risk adjustment and why it's appropriate given the intent sounds right to me. But restricting how – whether you should or how you should risk adjust to me seems a little early for this science.

Joseph Cacchione: Yes. Me, too although, that, you know, there's lots of ample – there's lots of data. I mean I don't think that we should specify. I think that when you get

back to the overarching theme of validity, that we need to be sure that whatever methods are being used, you know, that there is some – that the grouper that is in question has demonstrated the ability whether dependent – despite the – or whatever the methodology is, that they have established some validity of their risk adjusting methodology. Because that's where you're going to get the most pushback from providers. It's on the risk adjusting methodology.

Mark Levine: Yes. But the flipside to that is that if there's enough clinical specificity to the grouper how it differentiates more complex care, then it's really independent of the risk associated with the patient. Meaning (a simple fractured finger) …

David Redfearn: Yes, this is David Redfearn. Some of the grouper and some tack on patient risk course at the end and some do both, I don't think we're ready at this point to pick and choose between those methodologies.

Joseph Cacchione: Yes, I don't think so either. I think that the question really has – whatever grouper is available or whatever grouper we're discussing, has there been any assessment of the validity of the risk adjusting methodology that's being used. So, the first speaker asked, can you give me an example of a condition you started down the road of a broken finger?

Mark Levine: Yes. I mean just to differentiate it. If there's such a thing as a simple broken finger, how would that differentiate for a 20-year-old healthy male as opposed to a 68-year-old female with diabetes and hypertension. How would the risk of that patient impact treating a simple fractured finger? So, the risk might not always be associated with the care being provided to treat.

Joseph Cacchione: OK. Again, it's going to get back to intent, right? So, that if somebody – depending on what the intent of what this is being used for because that – if for a simple thing like a broken finger, a hip fracture in an 18-year-old, clearly it's going to be different with no co-morbidities versus a hip fracture in a 92-year-old.

Mark Levine: Right. Then we're both arguing our points, I think.

Joseph Cacchione: Yes. So, I would agree with you there. Risk adjust …

Mark Levine: All right.

Joseph Cacchione: … so – well, let me push back and say – so what you're saying is that risk adjusting does not always need be to be done depending on the condition being treated?

Mark Levine: I think that would be a fair statement, yes.

Joseph Cacchione: Yes. I guess in certain circumstances that maybe true, I would say that in the vast majority of cases, I don't think that that's true, but I don't have any data to support that.

Francois de Brantes: So this is Francois it's – I think we're arguing something that's irrelevant anyway. The point made in the comment is that the field, right, so, this is the feedback from stakeholders and NQF members is strongly suggesting that they would either individually, collectively or as groups have a hard time voting for the endorsement of a grouper that doesn't contain some level of severity or risk adjustment and that whatever methodology is used has to be transparent and understandable and so on and so forth.

Typically, severity models if you have – severity models impact the expected resource used depending on what the data suggests. So, if the zero variability in the management of patients with broken fingers, then your severity model is going to have no effect. And so, it's not question if do you need to use it or not use it. That's the purpose of a severity adjustment model is that it looks at the variability of the expected costs given the makeup of patients as evidenced by diagnosis codes and other indicators. So, there's no variability and there's no adjustment.

Joseph Cacchione: If there – that would be true, but trying to get a clinician to buy in this, you know, what you're talking about is a group of statisticians in the backroom figuring out, you know, factors for assessing risk of a population when that's not necessarily clear. My point still partially is correct in stating that there needs to be clinical differentiation is more important in analyzing and comparing care performance than just simply referring your risk factor in there.

Francois de Brantes: Yes, yes. I would agree with that. I agree with that.

Kristine Martin Anderson: OK. So, really, it looks like the comments looking for guidance and clarification on the approaches and implications, et cetera. So, it may be that in the – the question's really about guidance, so it sounds like we're recommending some guidance that says that there be justification provided for the – and a full description of the risk adjustment methodology?

Male: Yes. I think that's a good summary.

Female: Yes.

Kristine Martin Anderson: And we're back to you, Joe.

Joseph Cacchione: Yes. So, yes. So then the question five is – so I think we address that question five, are there threats to the validity or episode groupers that should be specifically addressed? I think we addressed that and then questions, fixes how should threats to validity be considered during evaluation episode grouper. So, I guess what stuff – the question here, how should threats to the validity – can you explain what you – I'm a little fuzzy on this question.

Ashlie Wilbon: Sure. This is Ashlie. So just looking for qualification on whether or not there should be specific assessment on how adequately this threats to validity have been addressed by the developer when they submit their grouper or whether or not there should be additional criteria to assess the, you know, the rationale or their description of how the threats to validity have been assessed for the grouper. Is that help or?

Joseph Cacchione: Yes, I don't know, what do people think of this? This is sort of like, I mean do you see this – I mean validity by – I mean it's almost – a validity in and of itself calls into question these things that are threats to it, right? So, there is something that is – there's a degree of validity depending on what the variables are that go into that validity measure.

So, I don't know it's almost – I don't know am I missing it here? It's almost repetitious when we talk about what are the threats to validity – validity accounts from the threats. Right?

(Crosstalk)

Male: (Inaudible) approach to this, this kind of reflect what the stakeholders have said. They've given some specific concern one being outliers, one being multi-morbidities, one being different the ophthalmologist talking about these specific things. We could put in the criteria of some keywords that how hard is your method addressed concerns about multi-morbidity. How does your method concern or address concerns about the effect of outliers. And that would be one way to do it and just get more clarification from the developers.

Joseph Cacchione: Yes, I think that, you know, putting in the recognition of outliers at the minimum because clearly these people sort of fall outside of – when you start to look at risk adjusting measures, there are certain things that there's just not enough data on but we all know that it is a variable that would add to or would confound the episode.

But, you know, validity is really based – should be based on the data that has been collected. But there are certain patients, for example, in bypass surgery, we know that cirrhosis is a very prominent risk adjustor, but we can't account for it because there's just not enough data out there. And so, so these outliers I think need to be addressed here. But I think the other areas are really accounted for in any validity model anything that can be counted. It should be accounted for in the validity model.

Are there further thoughts on that?

Kristine Martin Anderson: So if there are a list of such things, which it sounds like there are, it might be a good idea to forward them onto NQF so that they, so that in the criteria we can consider listing the types of concerns that you want addressed in the validity testing.

Joseph Cacchione: Yes. So, in an example and some of the things the work we've done, you know, excluding people with chronic act or on active cancer could, excluding

people with HIV, I mean excluding people with the, in procedural variable or procedural things, excluding people with dialysis, they're not easily accounted for because there's not enough large data sets to account for these in a validity model, but they are – they clearly are somebody that may need to be excluded or and maybe there's a list of those things that you could account for.

Kristine Martin Anderson: OK.

Joseph Cacchione: OK? All right. So, if there are – are there other comments on the second topic and questions four, five and six? Again, I think I just – in summary I would say that the intent again here – intent what these are being used for and the validity really of that intent has to be sort of overarching. So, I think it is an overarching theme for all of five of these different issues that we're going to take up here.

Kristine Martin Anderson: OK, great. OK. Moving on to number three. This one's I think we need to sort of tease out a little bit on this call if we can to help the team. Many of the commenters as you might have noted would make a comment that might have been really more geared toward overall group or logic, the resulting episodes or the results on measures that are applied to those episodes. And I think we need to be clear about what we're asking for in the evaluation criteria and then how far we're going in that scope.

So, something they sort of taken together, the first question is how it – what are the differencing characteristics, what's really just a grouper, you know, than versus what is an episode and then what is the measure so we can be clear about the definition of those three things. And then we should talk about what this process, this episode grouper evaluation process should – how far should it go and how this evaluate – at what levels or levels should we be evaluating.

I think that in our last set of meetings, we said, OK, all three of these are important, but what about the evaluation? So, let's start with that first question on what is a grouper, what is an episode and what's a measure?

Next we talk about a grouper definition and our in-person meeting.

Ashlie Wilbon: Hi, Kristine. This is Ashlie. I'm going to – we're going to screen share the image from one of the diagrams from the report that we tried to illustrate this in case that helps or if people have suggestions for how to better clarify that maybe through this picture that might be helpful.

Kristine Martin Anderson: OK.

James Naessens: This is James Naessens, I'll go ahead and make an attempt. So a grouper is a methodology in associated software to assign episode. And episode is that combination or a group of services that a patient receives over a specific period of time and a measure is a cost or utilization or quality assessment for a group of episode.

Male: Hi, I would just want to modify that the first one the grouper a good definition but instead of saying software, a thing to define the process that would fit either a software or a set of instructions.

Kristine Martin Anderson: And I think we have – I think the diagram does it for me in terms of separating what they are. I think it doesn't the answer the question how far we're evaluating in our criteria. It doesn't call out what a measure is. And it shows what a grouper might be if you actually went into all the line in a different place. It shows what an episode is, so an episode is a specific output of a grouper and then measures are layered – or ways so that you can look at various outcomes within an episode.

I guess my assumption had been we are only evaluating through the creation of episodes and that measures themselves would have to go through the measure process but I don't know what other people are thinking about that.

Male: Yes. That's was I thought and I think for us to start beginning down into that measure level I would be a little bit worried about that we go beyond scope of what I understood we are to do here.

Male: We might want to send this report that the measures are specifically addressed through other NQF channels?

Male: Yes. I think that's a good idea.

Male: I guess you could say that measures are judgments that are made from the output of a grouper. The grouper can – is a technology to define bundles of services that can be looked at together and their use potentially could be a measure.

Kristine Martin Anderson: Anyone else's ideas, I think we don't want to – we don't want to fall into words nothing in a group but if we have ideas to give guidance to staff around, the separation of these three ideas, I mean do we agree that you should really separate the concept of a grouper from, you know, evaluate – can you evaluate a grouper without evaluating its results and episodes or vice-versa? I'm struggling with how you could separate those two for endorsement.

What do you say, I endorse you but only for the following five episodes or do you endorse the grouper at large and all of its resulting episodes or do you have to look through them episode by episode to endorse them?

Male: Let's get back to validity though, I mean really I'm starting to look and I almost think that you have to – when you come back to validity, you know, if the methodology is sound, they have to know to demonstrate the sound of the or the validity of the grouper and that is deriving a what is essentially a valid episode, you know. And I think that, you know, I don't know, I don't know that we know enough about it to stay that in all cases, that if it's sound in one area it may not be sound in another. I don't know that we know enough about it, yes.

Jelani McLean: Yes. This is Jelani. I would actually agree with that. I don't know if we know enough to separate a group from these measures and the output. And we don't know enough tools to make an assumption that an episode created for hypertension. The validity of that is the same as the validity of one episode created for diabetes just it goes the same grouper. And so forth and so on.

Christopher Tompkins: This is Chris. You know, I think it's pretty basic question we don't have to tell NQF at this point what a measure, but a grouper is a means to an end. It's not an end in itself. It's a – the purpose of the grouper is to automate a process that – and make explicit a process that is either implicit or hand-built

measure by measure. So, the idea of moving into the space of creating their grouper and therefore evaluating grouper is a little bit analogous to moving from the retail level to the wholesale level.

A criteria for success may not change, but the process by which the clinical logic and episode construction logic to use some jargon are made explicit and automated that formulates what the grouper is. May I – in the in-person meeting, I asked – I did flip this question around and I said how could NQF or anybody else evaluate a resource use measure unless they knew all the criteria that were used to select records and select services and select cost by time and by their nature and so forth into the so-called groups.

So, the grouper is just an automated – it automates the process for doing that. I would think that when NQF considers evaluating groupers again, it's a means to an end. They say the grouper developer says, "We want to produce the following resource use measures." And it might be tens, it might be dozens it might hundreds. And we think that this resource use measures are useful in the same way that they would be useful if we had hand-built and hand submitted each one of them one at a time.

So, as useful measures, they need to be evaluated on all the criteria that NQF or the uses for evaluating and determining whether a measure is useful. But by moving enough to what I'm calling now just for the second a wholesale method, it expedites the process because it allows NQF – well first of all it allows that developer to not hand-build hundreds of measures. But by automating the process, a measure developer makes it possible also and (expedite) for the evaluators to simultaneously evaluate a consistently applied set of logic.

Now, one of the questions the evaluators might want to do is convince themselves first that the logic is consistently applied. But that's – so, that's a step. But if the logic is appropriately and consistently applied, then it allows the evaluators a very expedited way to judge that across the whole diverse set of output or a resource use measures that the grouper produces and simultaneously looking at it from the other point of view. If you have the task of evaluating hundreds or dozens of resource use measures, you can take

advantage of the fact that the logic it's used to produce those is now explicit and available in the grouper itself. And you can backtrack and examine the validity of that logic and inconsistency of the logic.

So, I think that the space we're moving into is a way from the retail level where developer say here, here's a resource use measure, what do you think of it to saying here, we have found a way to produce hundreds of resource use measures, what do you think of them?

And on NQF side is the other side of the coin, instead of having convening 100 committees to evaluate 100 measures, they say why don't we have a committee that looks at all these measures simultaneously by taking advance of the fact that the logic is now automated and explicit.

And as somebody mentioned it could be suffered probably is with that scale. But at that point emphasize the fact that it's really just a way of automating and making explicit of process. If it doesn't change the nature of the work of developers have in the first place and it doesn't change the nature of the work that the evaluators have in the first place either.

David Redfearn: This is David Redfearn and I mean when I was on the measure or the cost measure thing a couple of years ago. We didn't have one episode measure from (OPTIM) on ECGs that I – we rejected – the committee I think rejected the measure, didn't approve it was because there were some questions about the underlying logic. So, in fact, NQF is evaluating episodes on episode by episode basis.

Kristine Martin Anderson: Yes. I mean I think part of this is a workload decision on NQF point of view that, you know, trying to endorse episode by episode starts to create a process that looks like what it's like to endorse measure by measure. Does this group feel like it's possible to endorse at the grouper level or does it need to come down to the episode level?

(Crosstalk)

Male: Since logic is based …

Male: Come back to what I said earlier.

Joseph Cacchione: This is Joe Cacchione, I'm just not sure we know at this point. I think that if we can – at the beginning we can endorse a grouper. But, I'm not sure that we can endorse every – it doesn't mean that if we endorse a grouper that every episode that comes out that grouper is endorsed.

Christopher Tompkins: Taking the question from the opposite direction again this is Chris. I think it's possible for NQF or some other group of scientists to invalidate a grouper. That is to say a grouper submitted and it – so it misses the mark. It's making decisions that are really not, that are without merit or it's making bad decisions or whatever. I mean it's possible to say this that NQF has found the grouper to be invalid for the following reasons and that in itself (inaudible).

I think it remains to be seen, because it's never been done. Whether NQF can in fact go to the process and flip it around and say we endorse this as being valid, I mean that's sort of late in the game to be making that comment. But I suppose you can set the standards for invalidity and I suppose if the grouper passes all those tests meaning that is not found to be invalid then by default you could say that's under all the available scrutiny this grouper has past the validity checks and reliability checks and is suitable for the purposes as stated by the measure developer.

And it's – but it's not different I suppose my prior comment is. It's not as if it's a logically different scientifically unrelated undertaking as NQF implicitly and measure developers implicitly already have to do. It's moving enough to the wholesale level and saying, OK you can – we come up with consistent rules. And see what are those are with themselves passing the validity checks and the reliability checks.

Male: In other words the minimum criteria rather than national criteria.

Kristine Martin Anderson: Yes, I think a lot of the groupers allow some kind of toggles or flexibility that if you had to get into approving all the outputs, every time that grouper was trying to be used in any way the user or the developer would be stuck in another cycle of approval. That may not be advantageous to the, you know, overall usefulness. And I think if NQF maintains the approval of the

measures through a measure cycle, you know, if you want to actually get a measure endorsed lots of measures are used but are never endorsed. Then there's at least another test to say are the judgments being made based on this Grouper seemingly valid. That's kind of where I fall out.

Male: Agreed.

Ashlie Wilbon: Hi, Kristine this is Ashlie. I just have a clarifying question that maybe – that might help a little bit for me here to get things straight. In the last – in the report and then at the meeting, the committee talked about the evaluation kind of looking at episodes in terms of determining the degree of homogeneity and whether or not, you know, the claims and the – or patients that landed within a specific episode where homogenous and really represented. What was intended to be measured? Would that be something that would – if the evaluation says the grouper level that the panel would recommend is not done or does that still remain something that would be important to evaluate as part of the grouper?

Kristine Martin Anderson: Anyone have any view point?

Christopher Tompkins: I'm not sure how that question is different from what we were talking about before …

Kristine Martin Anderson: Yes.

Christopher Tompkins: So maybe, maybe my – this is Chris my response is not quite on target so I admit that in advance. The new word there was homogenous and it's – that relates directly back to what we we're talking about in a prior part of this conversation namely validity. In social science, validity ask the question or makes the assertion that you're measuring what you purport and say that you're measuring.

And so, for example, if, if you say that you are measuring the cause related to a set of patients who have the following condition with this – with the, with the following severity something like that. Then, that's valid if in fact you have used to create that cohort or if you have adjusted for all the potential other factors that would make that assertion invalid. But if you have adjusted

for all the other confounding assertions that would or factors that would make that valid then in fact, by definition, you've created a homogenous cohort. That's why I didn't see how it was necessarily separate.

So, if you believe and I think we deal as a group in generally that it's not a very good practice to measure something invalidly. That is that you say you're measuring this, but you're really measuring that because there are things that are driving the results that are unobserved or not adjusted for. I think we accept that and having accepted that then we have in the principle accepted a criterion of homogeneity with respect to the cohorts or the patients or the circumstances that are being judged.

Ashlie Wilbon: OK. I think I got that. Thank you.

Kristine Martin Anderson: OK. We move on to topic four.

Joseph Cacchione: Topic four, this is Joe again. So linking quality to the grouper – linking quality to grouper episodes or measures. And again, some of this is sort of crossing over into the measures – measure stuff. But again, if you read the section there, the paragraph above question nine and 10, it really – it's how do you link quality to these groupers although they're not designed effect of the grouper has not been for quality. It's going to be important. They are important signals for the success of this again coming back to intent. What are these being used for.

And so I guess question nine is, you know, is there a criteria or a method that we should be judging these groupers on that or is it just enough to say the quality measurement needs to be integrated into episode groupers and they need to account for some measure at the end of these groupers of quality that – so that we avoid some of the unintended consequences that some of the commenters have identified. So I'll stop look for input.

Jelani McLean: Hi, this is Jelani. So this goes back to the question of intent in the beginning. What I'm certain with is it's hard to have an intent without any consequence. So – And the consequence is in my mind the measurement. And that's why I'm having a hard time separating the episode grouper from the measurement.

Quality measurement, how is measurement, whatever it may be that is a consequence of the beginning intent and that's what we're going to evaluate or instead the criteria to evaluate these groupers from. So I don't know I'm directly answering your question well that's why I'm struggling with is how can we separate in the end goal or the end measure that was specified by the developer of the grouper for the intent? How can we separate those? That's I'm sure with.

Kristine Martin Anderson: I think it's a completely reasonable that someone could design a grouper with their intent just to get a handle on costs, right? And just say somebody might want to create that.

Jelani McLean: Right.

Kristine Martin Anderson: I don't know that – my view is we should not create a requirement that that organization that develops a grouper for that purpose needs to insert anything else in particular in it around quality or whatever. If it's intended for that use and it's evaluated for that use, you know, if it's endorsed, it's endorsed, you know, as a grouper that does that. Now the problem is that right now NQF endorsement list on the measure side is not specific to intent. So once something is endorsed it can be then used for other purposes and other high stakes purposes whatever.

And they're really – they can still say this is an NQF endorsed methodology and it might have been submitted as its intent with all cost. It gets endorsed and then they start reporting out, you know, quality measures on it. It was never looked at from that perspective. But …

Male: Or worse if they don't report quality measures on it and it has consequences from a quality standpoint.

Kristine Martin Anderson: But NQF has traditionally stayed away from, you know, making endorsement tied to intent which is sort of why were in a weird place but you can actually evaluate the methodology without understanding its intent. But you're not endorsing it just for that use. You're actually endorsing it that it meets the criteria based on the intent they submitted.

Male: Yes.

Jelani McLean: All right. So, how can you evaluate that it actually met that intent without looking at the end of the result of it is what I'm trying to …

Kristine Martin Anderson: Yes.

Jelani McLean: … I guess came with the clarification. I agree with you that it has, you know, typically been that way. And that's why we're in a bit of a challenge here. But I guess I'm struggling with what is the end result that we actually evaluating them on because an episode is just a group of a compilation of activities within a given timeframe or a given – I guess let's just stick to timeframe. But there is some end view of that and that is ideally in its most simplistic form of measure. And whether this cost quality or whatever happened within that timeframe.

So therefore what do evaluating to say that the developer's intent was actually met. And that goes back to the validity of it under reliability and all those other questions we've been talking about.

Kristine Martin Anderson: Yes.

Male: Is there a way to in this section, I mean I understand what you're saying Kristine because I understand it. I'm just sort of the – it's the unintended consequences of the unintended uses of these things if they get a blanket endorsement that can be used in a way that doesn't account for this. And I'm trying to figure out how do we put a safeguard against that?

David Hopkins: Yes, Kristine. This is David Hopkins. This issue keeps coming up from various panels that I am involved in. And it is a big issue but the fact is that NQF leverage is through endorsement. And all it can do is endorse measures or endorse concepts or episode groupers. It has no control over use.

Kristine Martin Anderson: Yes.

David Hopkins: And I just think, you know, we can make statements about what we've considered to be responsible use, but we can't control it. And let me just say

one other thing in terms of this specific recommendation. It seems to me that the time has come for us to restate this concept of linking quality to cost and make it two-sided. The statement is absolutely right. The focus these days is on value. And why are we reporting quality measures without cost?

Kristine Martin Anderson: David, are you proposing that every grouper must be able to handle quality and cost reporting?

David Hopkins: No, I am agreeing with you.

Kristine Martin Anderson: OK.

David Hopkins: That's not – There can't be a requirement of episode grouper. So it's just a requirement on the user to do …

Kristine Martin Anderson: Right.

David Hopkins: … appropriate linking of quality measures.

Kristine Martin Anderson: Yes. Which NQF does not have purview over, right?

David Hopkins: Correct. But I'd also like to see their concept become two-sided. And that's sort of a big – bigger issue with NQF.

Kristine Martin Anderson: Right.

Male: I guess I'm stuck on – I understand the sort of task in hand. I understand that we probably need to focus narrowly. Maybe I'm being a little too stretching the bounds of what were – but I still have great reservations to separate the two or separate this. And I would say the converse is we shouldn't be reporting cost measures without quality measures either, you know, because I think that that occurs as well.

And so I'm trying to figure out what we can do. So question 10 is what guidance should developers have when seeking to combine episode-based utilization measures with quality. So is there guidance that we can do that could help to – help guard against this or are we over stepping the bounds of what we're trying to achieve here in terms of creating endorsement or some

guide to endorsement toward a grouper. Because I think that's the minimum we should do is we should at least give guidance on how quality should be woven into these episode groupers.

Christopher Tompkins: Well NQF is already separately addressing this question. I'm not saying it comprehensively addressing it, but it's the last few comments – this is Chris again. These last few comments have pointed out these measures don't stand alone as (the vacuum). And Kristine your original statement is correct. I mean it's a grouper can be designed and used simply as a resource use measure development tool. And it doesn't necessarily – it's the fact of – it doesn't have to itself integrate quality measures.

The measure developer for a grouper doesn't have to take on the added dimension of quality measures but as the more recent comments now in this call have indicated what the field really needs is a coming together and I mean literally coming together of the quality and the resource use measures. And NQF is partially any way tackling that in other ways and in other committees and forums.

I think the – therefore the incremental question I supposed for this group is, is there anything special or different when it comes to applying that concept of linking quality and resource use measures in the grouper space as might be different from, you know, what I called earlier the retail space where you're taking an individual quality measure and an individual resource use measure and you're linking them for a purpose of measuring efficiency or value is all the guidance about that already in place if you handle it at the individual measure or is there something new, special, different, incremental when it comes to doing into the grouper space. Definitely the answer is no.

Male: I agree with an extent but I would tell you that it assumes an even uptake of what NQF is putting out there as output. And I am concerned that somebody may choose to use NQF episode endorsement for their grouper methodology that have little to do with the measure applications from what – the map or the output of the map or some other part of NQF. And so that – you can't just assume that everybody is going to sort of fall in line with whatever NQF is putting out there. So I think we need to draw some attention to quality and

cost measure within the episode. Unless we can somehow guarantee that there's going to be an even uptake of all the things that NQF has outputted.

Kristine Martin Anderson: You know I hear what every one is saying. And I too have heard this argument quite often. My sense though is as you're coming up with evaluation criteria for episodes we could only go one of two ways. We either say the episodes, you know, just state your intent and we can evaluate for the criteria against the intent or we can try to tell them that some kind of intent needs to exist if they're going to do cost, they have to do quality or et cetera. And I'm not hearing a lot of support for the latter even though people are frustrated with the current environment. So I don't know that it would be reasonable to say every episode grouper must be created to be able to reflect the quality and everyone must be created to reflect.

Male: And I'm not opposing, Kristine. What I'm – I think question 10 taken out of and that was a question is there needs to be some guidance on quality.

Kristine Martin Anderson: OK.

Male: The word quality needs to somehow be woven into this. Now, whether we say they need to account for in every episode grouper, I understand that episode to be just a cost thing.

Kristine Martin Anderson: Right.

Male: But a cost tool.

Kristine Martin Anderson: Yes.

Male: Quality needs – the word quality needs to show up in here somewhere.

Marjorie King: Yes, this is Marjorie King. I've been listening, you know, with sort of my provider hat on. This is a really good conversation. I think providers would buy into a debt or if it included the quality. And I wonder if it couldn't be just be some sort of, you know, modular or section in the application it says how would you propose to link your episode measure or whatever your cost measure is outputted by this episode grouper with an existing NQF quality

measure to demonstrate or to measure value for a particular disease. I was thinking about heart failure. You could do a cost measure about episodes of heart failure. And you could link those to quality measures, re-admissions for people with heart failure mortality that sort of thing. Maybe you could have almost like an essay question to look at intent. And gain that maybe a really stupid idea. But providers aren't going to buy into this unless they – it is linked in some way to concepts of quality.

Mark Levine: This is Mark Levine. I think it goes back to our previous discussion of what are we evaluating. Are we evaluating episode groupers, episodes or measures? If we're evaluating episodes and measures then I think this discussion is very pertinent but if we're evaluating a grouper. The grouper I think needs to pass standards for its validity, its reliability and what – and to define clearly what it is that it expresses as its product. And then people have the option of using it in one way or another the danger that we get into I think and that we have lived through in the last generation are perhaps good groupers that produce episodes that are then used for purposes for which it was not intended and then wind up sullying the perspective of the grouper itself which was perhaps never intended to accomplish what the use flows.

But if our job is to look at criteria for groupers that's different than the criteria for an episode, an episode might lead to include both cost and quality. And measures I think they are pretty good criteria that has evolved through them, but I think that's different then and we need to discriminate the criteria for groupers.

Kristine Martin Anderson: OK, so maybe a way to come out on this one would be that we would ask in the criteria as we're evaluating it explicitly as we're talking about intent to ask about intent for quality and intent for cost measures. And then as maybe we can have a part of our evaluation around and intended uses to have them describe how they are going to consider side by side use of cost and quality data. Because I mean they could still say not applicable we don't intend to do that. But at least then it's explicit that the question is being asked.

Male: I think that's a reasonable compromise

Male: So Kristine how is that going to work in practice. I mean if you've got NQF review panel looking at a cost measure.

Kristine Martin Anderson: Yes.

Male: And, you know, the developer either has an intent or doesn't have an intent to include quality measures or instructing somebody else to who could do it. The panel is going to kick it out.

Kristine Martin Anderson: Well let me give you an example, right? So let's just say you've created a grouper. The grouper was intended for measuring cost of care, right? It says that in the evaluation that grouper has been endorsed. Then in a measure committee a measure comes in that's going to measure, you know, cardiac quality or something based on the grouper. And a grouper is a grouper that was designed for just cost, right? And the developer has said look we designed this for cost. We did not anticipate nor test this, you know, for use in quality measures. Then the panel I guess has a decision to make, right? Around is this sort of off label use of this grouper acceptable based on whatever they're reviewing on scientific susceptibility and validity and everything else?

It doesn't tell them they can't approve it. But it at least flags the issue and they have to evaluate the scientific acceptability of that particular measure given that it's built off of a grouper that wasn't designed for that. I mean that happens today, right? But we don't really have as much transparency or – and the grouper themselves were never endorsed.

Male: Right.

Kristine Martin Anderson: So you're right in the sense, David, we never can seem get away from the fact that the Steering Committee still with guidance have to make decisions. And there's no absolutes on you can or can't make this decision. But it would make it a little more explicit.

Male: Yes, and I'm also thinking about the, you know, the current active project that at NQF which is examining cost and research use measures. Are we now

saying to that whole committee don't look at those measures in isolation from quality.

Kristine Martin Anderson: I don't know. This – That certainly not in our purview of what we're looking at here, we're looking just that episode, right? So I don't really know – I don't even know if they're based on an episode grouper. Probably not all of them, right?

Male: Yes.

Ashlie Wilbon: No. The measure is currently under evaluation, this is Ashlie, are not linked to an episode grouper. So they're standalone measures.

Male: But this concept is limited to episode based cost measures, right? As I've heard the other members speak it is much more generally felt that we need to link quality to cost and I'm marketing vice-versa as well.

Kristine Martin Anderson: Yes, David. I don't think we'd be addressing that here because we're just really focused on the episode grouper.

Male: Yes.

Male: Well there clearly is – these things are not, you know, we're trying to pigeon hole something that is not easily, you know, to just say OK, we're just going to – and I think Kristine I think you're doing a good job of trying to keep it focused here. But these things are hard to just sort of treat in isolation because, you know, a lot of these things that are being addressed in other places are going to spill into this and are going to have some impact …

Kristine Martin Anderson: Sure.

Male: … on episodes and we're trying to keep just an episode focus or a grouper focus. And it's had to just treat these things in isolation.

Kristine Martin Anderson: Right, I agree with that. To stay in our lane so we're meeting our scope of work, but other than that I understand that they overlap and we should take what's the best way to consider that. And if people have other

ideas beyond at least asking for intent to be expressed on the groupers and I think we should try to do it.

Male: Yes and I think that's probably where we should land. I mean, I think otherwise we – and listen as much as I am so passionate about the quality side in linking both cost to quality and quality to cost. And I think that to stay in as you described it stay in our lane, I think really asking people to describe as part of what the intended use of this grouper is and put it out there. And that would be probably a good compromise to at least account for so the quality is either left on the sidelines or not – or its at least accounted for in some way.

Kristine Martin Anderson: Yes. OK.

Male: So maybe we move on to number five then.

Kristine Martin Anderson: OK. So I think the question says it all, how should clinical guidelines be used to develop and define episodes of care for episode grouper. There were a number of comments that said, "Hey, one of the things we care about the most is that these episodes are clinically sound and therefore built off of established clinical guidelines." I don't think our current criteria addresses that directly. And you can look at the draft report and some of the comments that are considered there. How should we be addressing that in the criteria?

Male: Isn't that part of the validity discussion?

Kristine Martin Anderson: I think so yes, that's where it was it.

Male: It is.

Male: If there's a conflict between the …

Male: Kind of measurable.

Male: Right, it's not a measurable validity. So I think that's part of the issue here is there's no – I mean, and I think as a core layer to these things – these episode tools need to evidence changes or guidelines change and all of a sudden you need to do something on somebody and these episode tools – episode or the

grouper tool as well as the episodes need to be updated based on the most recent guidelines. And I think that otherwise you're going to get a lot of provider pushback here.

Christopher Tompkins: This is Chris. Actually this seems to strike me as potentially two different questions at once. One is and maybe I'm thrown by the word guideline. And on the one hand it's possible to construct an episode grouper that is intended to define what quote ought to happen in the way that the word guideline is usually used. That is to say that, you know, this service should be provided followed by that service. And that can create in the sense a clinical gold standard against which, you know, various practices and occurrences can be judged. And if that's a contemplation here then that probably should receive some discussion. Its own category.

The second way in which this maybe guideline is referenced just a little bit less intrusive in the sense which is to say that and again I wouldn't use the word guideline for it. So maybe I'm straying some other point of this question is. But in the – to Mark's point in the part of the validity of the grouper because its part of the steps of the grouper are to decide – determine for example which services and cost ought to be assigned to any given episode.

And the logic for doing that can rest on for example what could be called relevant services. That is you got a service that's done. There's a certain maybe nebulous diagnosis code on it and there's a procedure code on it. And the grouper is asking the question should this service be assigned to this episode or to that episode or to both. And there has to or there is or I supposed there needs to be a clinical – a deliberate clinical decision to say that that service relevant to that episode therefore it is appropriate to consider assigning that service to that episode.

So anyway, and just in summary, there were tow things here. One is the latter. The direct assignment logic which I think does go back to the validity question and great measure. And then, there's other one, and I'm not sure if the quarterly is about that, whether or not clinical guidelines per se could be use to develop a gold standard of what ought to happen in terms of utilization

patterns inside an episode. And that would seem like a whole new realm of evaluation. ..

Kristine Martin Anderson: Yes.

Christopher Tompkins: … for the grouper evaluators.

Kristine Martin Anderson: So, I struggle with this one quite a bit when I first read it, because, you know, I see guidelines as what is put out there as what should happen, right? And I see groupers as a representation of what did happen. And so, clearly, there is overlap, right? It would at least want to consider the guidelines to make sure that the related services actually get linked together appropriately because it gives some knowledge to the logic.

On the other hand, you don't use them exclusively because if somebody deviates entirely form clinical guidelines and start treating someone in whatever way they treat them for good or for bad reason. I'm not judging whether it was a good decision. You still want that reflected in an episode and that shows the variability in treatment.

And so, I think the question is how do we ask – how we at least ensure that the developer took guidelines into account when they defined the services that ought to be linked together or considered for linking together. You know, recognizing there is window period of all of those kinds of things that might getting away of actually having them end up in the same episode.

Male: So, Kristine let me just add one thing to your comment because you started down the road and I thought you're going to go there. And the other thing that the guidelines don't do and which I think we need to somehow figure out. If they don't address what shouldn't be being done and because in most cases if you look at a set of guideline, the potential use of this groupers is really – some of it's around cost measures and over utilization.

And so, the issue if somebody doing too much within a grouper show that the accounted for and because the guidelines don't say, don't do this test or don't do that test except within various certain circumstances. So, we got to figure out a way that the guidelines can address the overuse issue as well and can be

appropriate and really come back to this appropriate use of whatever services are being measured.

So, guidelines are helpful into specifying what should be, but there is a whole body of appropriateness and overuse that are not addressed in guidelines that I'm sure are accounted for in this.

And I think the best we can say is sort of A is the clinical guideline should be being used. But I'm not sure how to get it this whole issue around appropriateness and the sort of overuse of things, of resources and an episode that need to somehow be accounted for and so, I guess I'll stop there.

Marjorie King: I think a lot – this is Marj, I think a lot of the comments came from the commenters not understanding groupers.

Kristine Martin Anderson: Yes.

Marjorie King: Yes. And the same issue with quality, but on the other hand it's about educating providers about what a grouper can and can't do. And, you know, there are appropriate use criteria out there, but I think we put that in I don't think that will nullify them. I think it's – so, I come what's the clinical – in malpractice what – I forgot what the phrase is that the clinical – on the clinical practice that you're doing versus the guideline. You're really looking at common clinical practice in a grouper.

Male: I guess what the providers are most worried about in this – some of this – I mean, what they're probably most worried about is the intent of this is to pay providers or and that somehow is not accounted for, but every person that has X-condition should have A, B, and C and it's not accounted of for in the grouper methodology. And somehow the grouper is being used for payment or for judging resource use that there somehow it is going to be an unfair representation.

Marjorie King: Right. They're afraid that it's going to drive down cost and minimize the ability to do the higher cost tests on patients when it is …

Male: No. I think what providers are looking for is optimization of what would be appropriate testing.

Marjorie King: Right.

Male: And this is this whole issue over used needs to be addressed not just what needs to be done, but what shouldn't be being done as well within an episode.

Marjorie King: So, I was glad to be …

Male: I'm not sure it belongs here, but there's got to be some accounting for of guidelines and appropriate use.

Marjorie King: Would you advocate then that the report that's generated should – that we should address the concept of incorporating appropriate use criterion guidelines in some way into development of episode groupers or stay away because again it's not pertinent?

(Crosstalk)

Male: This is complicating the issue of wouldn't this actually be taken care by the linking of quality to cost and, you know, I'm thinking quality in a broadest term to include appropriate?

Male: Yes, I think it does.

Christopher Tompkins: Well, this is Chris again, adding to my first category. The first one being is clinical guideline and gold standards. The second being implicit in the grouper which is that it uses clinical logic to determine a relevance for assignment of a service, I'll throw in the third category here which may also be sort of related to this question namely that the logic for linking a service to an episode really can be go on two separate dimensions. One and the focus here is the clinical dimension, but grouping can also occurs simply by coincidence meaning services that are occurring at or in the same window of time or perhaps in the same setting.

So, we've seen this come up in the real world, for example CMS is bundled care, you know, bundle payment for care improvement demonstration which is, you know, has its – a life of its own. It's an ongoing demonstration. When that was being set up or designed in a sense that's a way of expressing episodes right, you got this hospital episode followed by post acute services, you put the whole thing together that's an episode.

A lot of the debate and lot of the designed parameters discussion had to do with, if you have a time window that for example starts around the time of the admission and then it goes for some period of time you know 30, 60, 90 days post discharge for example. The question then is, are all services thrown into that bundle, i.e. thrown in to that episode because of their temporal relationship to the hospital admission and there was off – the so called all cause readmission is an example of that.

Whereas – And if the answer is yes, every covered service that occurs during this time window is thrown into the "episode" then that's putting maximum emphasis on the temporal relationship between the services. But that can come of the expense of another filter which is if there are a clinical relationship between the services that occur in that time window and extensively what the episode is about.

So, if you're admitted for a hip fracture "episode" and the definition of that episode is all covered services that the beneficiary receives, you know, up to 90 days after discharge, then we know that there are going to be many services that can occur in that time window, for example car accidents or other kinds of things or cancer incidents that has nothing to do with the hip fracture episode.

And I think there's a lot of sensitivity around one perspective that says that the purpose of the episode in that in one way is to select, manage the whole patient not matter what happens versus the more nuanced approach with is usually attributed to episodes which is that we are trying to piece apart the various conditions that a patient – the various needs that a patient might have and exclusively assign services to the episode on a clinical basis informed by the temporal relationship among services.

So, I mean again in your consideration might not commence to the NQF staff although I'm not omitting my comment now to the whole committee here is that we might want to show that in the logic of assigning services under the grouper of an episode grouper, there is this notion of more than one dimension, the clinical dimension and the temporal dimension and the measure developer should say something about the extent to which they are predominantly relying on the clinical relationship between a service center and an episode or predominantly on the temporal relationship or to the extent that which they're trying to combine both of those criteria in the logic that designs services to any given episode.

Kristine Martin Anderson: Well, my summary comment is we need to think about this one more because I think we could tease this out to a point where we could make up a question that would make sense to us given how, you know, much – each of us knows about episodes. But, I think our committee doing an evaluation is going to have a hard time knowing whether or not the answer is a good answer unless we can get very, you know, sort of take it down to another level of detail that will help them understand, you know. Because were saying we'll take into account guidelines, but we don't expect you to derive your whole episode definition out of guidelines and it may not be possible to do that in a question that then a committee member could evaluate.

Marjorie King: So I like the world clinical logic rather than guideline. Maybe we need to sort of to discuss that in the report, the concept of clinical logic as opposed to guidelines or appropriateness or whatever.

Kristine Martin Anderson: Right. Effectively, they're going to have to describe exactly what the lot, you know, how they derived their logic and what the major elements of it are, but …

Female: Right.

Marjorie King: And how it relates to the clinical logic, the clinical guidelines, the appropriateness or – it is hard.

Mark Levine: This is Mark. Again I think all of this is within the grouper of validity, how do you test for validity? One of them as you compare to the established guidelines on that clinical topic.

Kristine Martin Anderson: Yes. But I think the hard part market is how do you compare it, right? Because the output of the episode wouldn't necessarily reflect a clinical guideline if that's not in fact what people are doing in practice.

David Mirkin: This is Dave Mirkin, isn't this discussion less about the grouper and the output of the grouper is then about the benchmarks that would be used to compare the output for a particular …

Kristine Martin Anderson: Measure? Yes.

David Mirkin: … episode. So, I think maybe the grouper could be fine. It really then goes back to the application and how do you develop those benchmarks for assess – adequacy. I guess of performance.

Kristine Martin Anderson: OK. I think that's a fair point too.

David Redfearn: This is David Redfearn, I mean when you think about using these models typically to do a provider profiling, what you do is you compare episode cost to essentially average episode cost. So, there isn't really any presumption in that methodology that you're – the criteria you're comparing against is the best or the clinic – most clinically appropriate is what people do on average. So that just sort of bypasses the issue of quality completely. You're just not claiming that you're making a comparison against the best possible episode cost you're just saying this is what is typically observed.

Kristine Martin Anderson: OK. I think we are – we have a half hour left, are there – staff, are there additional things you want us to discuss?

David Hopkins: Hey, Kristine.

Kristine Martin Anderson: Yes.

David Hopkins: David H. here. I've got a couple of big picture question so. If we are trying can I put this out?

Kristine Martin Anderson: Sure.

David Hopkins: So, actually first I have a process question. If I look back at the record of this group, what I felt was, you know, we all got a draft report back in March I think it was and we submit comments of our own. And then there was some revision but it didn't look to me like much and I don't recall everything a red line that went from the March 18 version to the April whatever version that we have now.

So, I sort of wonder about that. That's a process question. How were our competency evaluated by whom and with what we (saw). We start with that one.

Ashlie Wilbon: Hi, this is Ashlie.

David Hopkins: But this is before the public comment process.

Ashlie Wilbon: OK. Hi, this is Ashlie. So, generally our process is we do accept comments and then staff kind of reviewed those in the context of timeline and all that stuff and makes changes to the report is somewhat basically a staff executive division. We generally in the timeline did not have time to share back a ed line version and give you guys' time to review and then have an opportunity to take that red line version or just any additional changes and feed those changes back in. So, we only took the one round of suggestions and put that out for comment.

We really view this as an iterative process. So, again, we will be offering an opportunity to the committee again. This time which we're anticipating probably significantly more changes to the next version of the draft that will go to feedback in July. So, again, we'll be soliciting for input and we can certainly see whether or not there will be adequate opportunity for a red line version to be sent out and solicit any additional input but a lot of time it's actually due to time constraint.

David Hopkins: OK. Thanks for clarifying, Ashlie. Let me just pose the big picture questions I had there. And number one is I'm still not clear and I'm taking that from

future audiences may not be clear about what is the purpose of this whole effort. Is it to establish criteria for NQF endorsement of episode groupers per se? Or is it to establish criteria for evaluation of episode groupers that anyone should use? You know, not necessarily associated with the NQF and process. I think, it's important to clarify this. Anybody have an answer?

Ashlie Wilbon: This is Ashlie, I'm not sure – so you're differentiating NQF endorsement for …

David Hopkins: Of a grouper …

Ashlie Wilbon: Of a grouper for.

David Hopkins: Yes. Endorsement of a grouper.

Ashlie Wilbon: I guess I'm not understanding your question. Endorsement of a grouper, so let me just give an answer and then you can tell me whether or not that answers your question.

So, the purpose of this effort was to establish to one because we've never evaluated groupers, we wanted to spend some time with the group giving, providing some guidance and a frame work to help others understand exactly what we mean when we say grouper.

So, a lot of the questions that we're asking for in terms around guidance how should this be approached by developers, it's really to support that goal of just providing kind of a baseline understanding and what we mean when we say grouper. What consideration should we – should developers be taking in to consideration when they're developing a grouper.

The next goal was really to think ahead and say if NQF was to endorse a grouper and developers were to submit a grouper to NQF for endorsement. What should they be submitting to us, how do we know that we have the information we need to evaluate the grouper and when we have that information, how do we know whether or not it's reliable, valid, usable et cetera. So, that is really the purpose of the work.

So, in terms of the use of the grouper where we really wanted to keep this effort broad, so we weren't necessarily focusing on developing criteria for a specific grouper or specific type of grouper. But, really keeping that open to consider any type of grouper that might be submitted. Is that help?

David Hopkins: That helps, Ashlie, and then I think there's a question about potential third use which is imagine that an NQF resource use committee was presented with a measure that depends on an efforts of the grouper but specific to, you know, particular condition or outcome or whatever, would they be expected and to reveal all this gory details about the grouper construction.

Ashlie Wilbon: So, that's a great question and I will say that to David Redfearn and some of the other people that are on this panel, the Steering Committee for research use and our very first effort, our very first project in evaluating resource use measures. We did accept some measures from (OPTIM) that were a result of the grouper. And quite frankly that was our first time evaluating resource use measures and we didn't really have very many parameters around the types of measure that we were accepting.

And so it was a very difficult task and we learned our lesson from that effort that we needed to kind of do a little bit more thinking about exactly what we were asking for. And since then we have not evaluated any groupers that are, I'm sorry, measures that – or have resulted from an episode grouper. And one of the things that we would like to do actually is follow on work is to potentially explore that issue and Chris brought this up as well and really differentiating what the differences between a standalone resource use measure and a measure that results from a grouper.

And I think we're still trying to figure that out and what implications that might have for other committees that maybe reviewing cost measures. But at this point, the resource used – Steering Committee have been focused on standalone measures. And I think before we would embark on evaluating measure, measures that are part of an episode grouper. We would have to don some thinking about what that really means. So, hopefully that helps a little bit in terms of context.

David Hopkins: You know, I'm still feeling a third degree of ambiguity about what the purpose of this exercise is. So I'll just leave it there for you guys sort that out for more if you feel like it. But I mean, part of where I'm coming from is if anybody at NQF or anywhere else feels like it's important for commercial episode grouper vendors to come forward and have their tools endorsed by NQF. I'm not seeing that, you know, what it is that would lead them to do that. And if they're, you know, they're not going to do it. Then it seems to me we should be focusing ourselves more on not the evaluation of the grouper per se but what are the criteria that anyone who's using it, a grouper ought to be examining.

And, you know, the other thing that I had raised in connection with at least the commercial grouper is, is did we ever really fully discussed a concept of aiming of the expert clinical advisory groups that are used by those vendors. This really speaks to the validity issue and it speaks the issue of if you're really serious to tell determining the validity of a grouper, then you really talking about the 500 or 600 episodes that are within it. And those, you know, those have been fully examined by clinical expert panel for any of the commercial vendors I'm aware of.

So I raised this question about, you know, would we consider some of kind of deeming concept there so that – one is on asking an NQF panel to go through the whole exercise.

Ashlie Wilbon: This is Ashlie. I'm not sure what data were you posing that question to the other panel members or with that specifically for interest stuff.

David Hopkins: I think that's for discussion by the panel members.

Ashlie Wilbon: OK.

David Hopkins: It's a proposal. Kristine, any thoughts.

Kristine Martin Anderson: Honestly, David sort of not yet for me. I'm a – I'll admit a little overwhelmed even by the task at hand.

David Hopkins: Right.

Kristine Martin Anderson: And so …

David Hopkins: You're doing a good job with it. I can smell that.

Kristine Martin Anderson: You know, so I'd – well, I recognized, I mean, one of the hardest part for me every time I interact with NQF is that this is, you know, you're pulling out one piece of a bigger puzzle and taking a look at it and then trying to stick it back into the puzzle and hope that you didn't changed the shape so much that it no longer fits.. And that's something that I struggle with a lot. But I struggle with more than anything is, you know, where would you really have the purview to reach across projects.

And I feel like most of the time in what I'm doing I don't really have the purview to reach across and therefore, you know, it, it does feel a little artificially constrained. And I think a lot of the things you raised are often about the whole puzzle which I appreciate. I'm just not sure how to fix them.

David Hopkins: I think that's a fair way to look at it. And I don't know how to fix them either.

Kristine Martin Anderson: But you get a lot more times to practice.

David Hopkins: Maybe.

Kristine Martin Anderson: Others have of any thoughts for David?

Male: David, can you rephrase your question?

David Hopkins: The one about dealings?

Male: Yes.

David Hopkins: Yes. Well, so go back to the discussion about validity, right? Clearly one of the key criteria that are – that's supplied by NQF in reviewing any measure anyway and then all the suggestion being that we applied to an episode grouper. So if you were seriously applying the concept of validity then that's a grouper and I think some other supported this out and even in this conversation today, you would ultimately be looking at 500 or 600 episode –

distinct episodes whether within the grouper. And we'd have to look at the logic for each one of them.

My point was that that exercise has already occurred at least with respect to the commercial vendors that I'm aware of. And would we or would NQF really feel the need to have the whole exercise done repeated by some other group of clinical experts or are there criteria we could set that said, you know, and sort of this group that reviewed the episode for X, episode grouper developer did the job. That's my concept of DME.

Ashlie Wilbon: There really is no precedence for DME that's where I sort of struggled.

David Hopkins: Yes.

Male: Well, I guess it's right.

David Hopkins: But if you had criteria, you know, typically what you do is you set the criteria for what constitutes an adequate review and then you deem a grouper that met that criteria they have to demonstrate to do that.

Male: Yes, which goes back to do we need criteria for the testing of validity, does a grouper as part of it's grouping algorithm include robust tests of validity. We don't have to then look at each episode. That's a separate issue of that episode is being used as a measure then certainly it's going to be evaluated for that. But does the grouper itself have a robust process for looking at validity.

Ashlie Wilbon: All right. Well if there's no question we're going to have to deal with the line here right? It's not going to be possible that the panels that are reviewing the episode groupers are going to be themselves reviewing all the results of all the testing and determining whether or not the testing passes, right? So it's much too large that we imagine doing that across all these episodes.

So there's going up to be some trust of the process – that the process results in a good outcome, but then there are also has to be some verifying of the process that resulted in a good outcome and how we draw that line, you know, we'll make a big difference as to how much we've worked – what kind of expertise the panels are going to need in order to approve a grouper.

David Hopkins: Yes. I think you state it well, Kristine.

Ashlie Wilbon: OK. And any other topic anyone wants to cover in our last 10 minutes? Staff or did you get everything you needed?

Female: I think so. I think we definitely got enough to go back and start making some edits to report.

Ashlie Wilbon: OK.

Male: From your point of view …

David Hopkins: Will you guys be reviewing our comments in that process at least? You know, the comments we made earlier or?

Ashlie Wilbon: Sure, we can certainly do that or yes. Or if you in the light of the comments that were submitted by the developers if you feel – I'm sorry, by the public and the member that you – if you feel like there's an issue that, you know, certainly rises to the top, you can forward that and we'll certainly consider it.

David Hopkins: OK. Thanks.

Ashlie Wilbon: So, if no one has any other issues to raise we're about 10 minutes of and I'd like to open it up for the public comment. Operator, can you prompt those on the phone how to make a comment?

Operator: At this time, if you would like to indicate public comments, please press star one on your telephone keypad. We'll pause just a moment to compile the roster. Again, that's star one for public comment.

You have a comment from Amita Rastogi.

Amita Rastogi: Hi. This is Dr. Amita Rastogi from HCI3. I've been listening intently to your discussion. One thing I would like to point out is – and to the last point about which David was raising. The grouper is very different than the episodes themselves. The grouper is a set of rules and logic like Chris Tompkins was

describing which put together criteria which bundles services together in an automated fashion.

So for example, the clinical working groups could define an episode very nicely and see what services should be included, what is correct, et cetera. And that part is vetted by the clinician. But then how does the grouper perform? Does it then – given member has different services being performed to them, how do these services get allocated into concurrent episodes that the patients may have. That is the part that has to be evaluated. Then you're evaluating a grouper.

For example, if a patient has a knee replacement episode, is the anesthesia services being pulled into knee or are they going into some other episode that's maybe open concurrently. And office visits for diabetes in a patient who just had knee replacement, where does it go. Those are the things that need to be evaluated.

The clinicians could tell you, yes, an office visit for a diabetes is important and it's part of diabetes episode. But then if a patient has both diabetes and it's having a procedure for knee replacement. How does the grouper handle that? That is the magic. That's what we have to evaluate.

Kristine Martin Anderson: Thank you.

Operator: And there are no further public comments at this time.

Kristine Martin Anderson: Well, thanks everyone for your time. And please do send in your written comments as well. And I'm hearing the staff is going to evaluate them and what we sent in before and we can look for – do we have a date for when we'll see another revision of the report?

Ashlie Wilbon: I've had to look at the calendar. The dates that I have now is we have on CSAC review on July 9th. Staff was planning on working on edits over the next couple of weeks so that we can get the final or the next version of the papers to feedback, you know, in advance at least a week in advance for them to begin to review. So we're looking at probably the second or third week of June, probably the third week of June. We'll send out some more specific

dates on when that is when we can look at the calendar and figure out when that would be.

Kristine Martin Anderson: OK. Thank you. Thanks, everyone.

Male: Great job, Kristine.

Kristine Martin Anderson: Bye-bye.

Female: Bye-bye. Good job.

Ashlie Wilbon: Thank you, Joe and Kristine, for your leadership through the call.

END