NATIONAL QUALITY FORUM

Moderator: EENT Group
May 1, 2015
1:00 p.m. ET

OPERATOR: This is Conference #: 83496056.

Welcome everyone. The webcast is about to begin. Please note, today's call is being recorded. Please standby.

(Shaconna Gorham): Hi. This is (Shaconna Gorham), and Vy Luong and Reva Winkler and we like to welcome you all to the EENT Standing Committee Call for the Measure Evaluation Tutorial number one.

(Judith), I know that you're on the line. Do we have other committee members, I'm sorry, on the line also?

(Jackie): Yes. This is (Jackie).

Female: Great, thank you.

(Shaconna Gorham): Hi, (Jackie).

(Rich Madonna): Yes, (Rich Madonna).

(Shaconna Gorham): Hi, (Rich).

(Crosstalk)

(Shaconna Gorham): I'm sorry. Repeat your name.

Seth Goldberg: Seth Goldberg.

(Shaconna Gorham): Hi, Seth. Great. Anyone else?

(Dan Herrnstein): Yes, (Dan Herrnstein) is here. Hi.

Female: Hi, (Dan).

(Shaconna Gorham): Hi, (Dan).

Female: Hi. I'm sorry. Is it (Tammy)? Great, thank you.

(Tammy Bradam): Yes. (Tammy Bradam).

(Shaconna Gorham): Hi, (Tammy).

(Tammy Bradam): Hi.

(Shaconna Gorham): So I just wanted to remind you all that since you are attending this call that you do not need to attend workgroup call number two, Q&A call number two that will be the same information covered in this Q&A call as we cover next Thursday.

With that introduction, Reva, would you like to start?

Reva Winkler: Sure. Thanks everybody for joining us. This is a very – meant to be a very informal session and our first priority is to deal with any questions that you may have about the evaluation process and NQF's measure evaluation criteria.

Certainly, if you don't have any questions I'm prepared to kind of go through the criteria again and highlight some of the things that I think bear repeating. But again, I really encourage you if you do have any questions just jump in and don't hesitate to interrupt me.

So we're moving into the process to prepare the committee for meeting in person in June. And during the month of May, we will be holding four workgroup calls where it gives you a chance to talk about the measures in sort of a preliminary fashion the steering committee member – steering committees in the past have told us that this is a particularly valuable

experience when it's the first time for a new committee doing this for the first time that they sort of get a chance to get their feet wet before the in-person meeting.

And so, in order to – for the workgroups, we need to be sure everybody's comfortable with the process as well as the evaluation criteria. Earlier this week, Vy sent out the actual measure information or told you where to get it on SharePoint as well we've broken the group into four workgroups and we've broken the measures up into four buckets and so each group handling four or five of the measures. So those assignments, you should have that information.

Today, we're not so much going to talk about those specific measures because we really haven't had time to look at them. But we do want to talk about the evaluation process and NQF's criteria. We went through the criteria briefly on the orientation call and we hope that you've had a chance to perhaps look at the committee guidebook which does talk in a little bit more detail about the criteria. But today, we really want to be sure that any questions you have about those criteria and what we're looking for in your measure evaluation is addressed.

So just to start off with, I'd ask anybody are there any specific questions you hope or addressed in our conversation today?

(Tammy Bradam): This is (Tammy Bradam). I have a couple of questions.

Reva Winkler: Sure.

(Tammy Bradam): Some of them aren't even related to the measures, it's about our trip in June.

Reva Winkler: Yes.

(Tammy Bradam): Can we ask that question now …

Reva Winkler: Sure. Absolutely.

(Tammy Bradam): … or later or e-mail?

Reva Winkler: No, go ahead and ask. Go ahead and ask.

(Tammy Bradam): So we need to book our airline using the Concur online, right?

Reva Winkler: You've received those …

(Tammy Bradam): The hotels …

Reva Winkler: Yes. You've received that information from our meetings department?

Female: Yes.

(Tammy Bradam): Yes.

Reva Winkler: OK.

(Tammy Bradam): And then hotels already been taken care of, so just basically at this point we need to be booking our hotel.

Reva Winkler: The airline, your flight.

(Tammy Bradam): I mean the airline, yes. Sorry.

Reva Winkler: Yes.

(Tammy Bradam): And then the meeting is at 3:00 on Thursday.

Reva Winkler: That's correct.

Vy Luong: So (Tammy), this is Vy from NQF. Yes. That's all correct. And I think that – I just got your e-mail with all of the questions. I can definitely reply to this …

(Tammy Bradam): OK.

Vy Luong: … offline? How does that work?

(Tammy Bradam): That's fabulous. Thank you so much.

Vy Luong: OK, great. Thanks.

Female: All right.

Reva Winkler: But I think the question about arranging travel is probably something others are interested in. So I'm glad we talked about it.

Vy Luong: Yes. And if anyone ever – have any question related to NQF travels, please feel free to shoot me an e-mail and I will work with you to get all of your answers answered – all of your questions answered. Thanks.

Reva Winkler: OK. All right. So in terms of our measure evaluation criteria, these have evolved over time and they do seem to be at times fairly detailed. I'd like to direct you to two documents that we want to focus on. Vy has one up on the Webinar and I think she sent you the link. This is our sort of basic foundational guidance document with the criteria. It's actually was updated. This was newly posted as of yesterday.

The updates in it from prior version of 2013 really don't affect the measures that you are evaluating. So for all – if you've read some of the earlier guidance and criteria, it still applies. These really won't have any – the changes really don't apply to the measures we're looking at today.

So I really would encourage you to consider this document which is posted on SharePoint, in the committee general document section as a really – as a resource. If you're unclear about the criteria, please refuse this as a reference. It's not readable like a novel but it's certainly an excellent reference document when you're trying to understand the measure evaluation criteria.

And the other document that I think is very useful as you're beginning to do measure evaluation is the document we call What Good Looks Like. And what this is the staff put together examples of what we believe are well – the types of responses and types of information we expect to make it straightforward for the committee to do your evaluation. And these are meant as guidance for the organization submitting the measures, but also for you to see, you know, what's expected. So you have something as a basis of comparison.

So these two documents, I think, are particularly helpful in getting you ready for looking at the measures that you are specifically assigned.

So are there any questions about finding those two documents? Hopefully you've got them. We're going to be showing parts of them on the Webinar.

OK. So like I say, in the absence of questions I'm quite capable of filling the air time.

So, Vy, on the guidance criteria document, would you scroll down to where we start talking about evidence on, I believe it's page five.

Vy Luong: I would do that now.

Reva Winkler: OK. So when it – evidence is really one of the most important discussions that the committees have about measures. One of sort of the NQF values is that the measures we endorse are evidence-based. And the criteria to determine evidence has evolved over time and in response to the needs of various stakeholders and particularly end users.

And so when we look at measures to be endorsed by NQF we need to keep in mind the context. Measures are used in a wide variety of reasons for a wide variety of organizations. NQF's specific purpose is to endorse measures to be used for accountability purposes, as well as quality improvement. But it's the accountability purpose that we need to remember because that does put the need for measures to be strongly evidence-based, solidly reliable and valid, and just really very, very strong in order to be – to make comparisons among providers whether it's hospitals, health plans, clinicians. The goal is to be able to make them valid in comparisons. So keep in mind that that's the purpose of the measures NQF is looking to endorse.

So evidence portion of it is indeed one of the most important of the criteria to evaluate the measure. And the criteria that NQF uses depends on the type of measure that it is. We see mostly outcome measures and process measures within our portfolio. There are other types of measures such as structural measures, efficiency measures that combine both a quality and a cost component. We don't see many of those. We see measures that imply efficiency but not true efficiency measures. There are patient reported

outcome measures that are based on patient reported outcome assessments. We also see measures of appropriate use.

But in our – this particular project, we really have basic process measures and outcome measures. And so it's important that everybody's clear on the difference because the evidence requirement for the outcome measures is different than it is for the process measures.

So does anybody have any questions or do you feel very comfortable, you understand the difference between the outcome and – outcome measure and a process measure? Are there any questions on that?

Seth Goldberg: Yes. This is Dr. Goldberg. It appears that the (inaudible) and were actually appropriate used criteria. They were – Yes, they were just (inaudible) for two different diagnoses.

Reva Winkler: Yes. You know, they are in – there is an appropriateness element to them but I don't believe the measure specified that they use the formal appropriate use criteria. So, you know, I agree with you that there is an appropriateness element to them. Those are process measures. The question of did you do the right thing or not. So we'll be treating them as process measures for the evidence evaluation.

And the reason it makes it different is that a health outcome is considered fairly self-evident. I mean from a stakeholder perspective what everyone cares about is what happens, in other words the outcome. Typical outcomes include mortality, complications, functional status outcomes, things like readmissions or a proxy for bad things happening probably. And so there are variety of outcomes that are basically fairly self-evident because we're looking for the relationship to patient outcomes.

So from an evidence perspective, very little evidence is required except to explain a rationale between some structure process, intermediate outcome or something that's actionable that can affect that outcome. In other words, is it actionable on the part of the health care system to affect that outcome?

And so it isn't really a full evidence review as you would expect and what we expect with process measures. So the real focus on the evidence review tends to be around process measures. And so process measures, we are really looking for a systematic assessment of the body of published studies that relates that process of care with the outcome to the patient. And so we will be spending most of our time around evaluating evidence for the process measures and much less when it comes to the outcomes.

Are there any questions on that …

(Judith Lynch): Yes, this is …

Reva Winkler: … distinction?

(Judith Lynch): This is (Judith). I just wanted to clarify something like audiological evaluation no later than three months that would be process?

Reva Winkler: Yes. That's a process measure.

(Judith Lynch): OK.

Reva Winkler: Thanks. Any questions from anybody else on the differences?

OK. So, Vy, let's scroll down a little bit more and look at table one here. And it talks about the type of evidence required for each of the difference measures. And we mentioned the health outcome and the patient reported outcome. This is simply just a rationale. But if we scroll down to process, what we're talking about is looking at all of the evidence, not just selected-picky-choosy-the-one-I-like evidence but really a systematic review of the body of evidence.

And this are evidence criteria, it was developed along in the same timeframe that the Institute of Medicine issued their report on evidence reviews and the types of evidence reviews necessary for good clinical practice guidelines. Those two companion pieces really were instrumental in guiding NQF's development of our criteria. We wanted to be consistent with what was happening out in the rest of the world.

And so we really want to look at those systematic reviews that discuss the quality or the quantity consistency of the body of evidence that whatever process is being measured leads to desired health outcomes in the population of being measured. And so we really want to – look, as an example for instance, evidence at the use of a certain drug results in whatever, reduced hospitalization, lower mortality, lower morbid events, whatever it may be. We're looking for the evidence that says that that is the right thing to do.

Now, when it comes to process measures, money – many of them are based on clinical practice guidelines from societies or others. And those guidelines are typically based on evidence review. As the Institute of Medicine's report indicated, however, it's a wide variety of approaches, methodologies and the quality of those evidence reviews that underlie many clinical practice guidelines.

However, as a result of that report we are seeing an evolution in the development of clinical practice guidelines towards more methodologically solid evidence reviews to support those guidelines. And currently, the National Guideline Clearinghouse requires that that – that a systematic review methodology as indicated by the Institute of Medicine is required to put a measure in the quality measures or the guideline clearinghouse. And also, they indicate which measures previously in the clearinghouse have been updated to conform with the IOM guidelines for evidence review.

So we're seeing this transition to solid methodology for supporting guidelines with good evidence review. And so you will see that the criteria will accommodate guidelines that are assumed to have had a reasonable evidence review, however, if we can have the specifics of the evidence review that's even better.

So are there any questions about the – what we're talking about on systematic reviews of evidence for the process measures?

No? Everybody's clear on what we mean by systematic review? Because it – these are the kind of things you're going to be considering when you do review your measures.

(Todd): This is (Todd). I do have one question. Can you …

Reva Winkler: Sure, go ahead. Yes.

(Todd): … hear me? So a clinical – a well-published clinical practice guideline is adequate or you expect us to go to the primary literature and reevaluate …

Reva Winkler: I do not expect you to do any extra work but read the paper that's put in front of you. We have asked the developers to provide that information in a series of questions that they respond to. So they should be providing that information to you, OK. We do not expect the committee members to do the primary research at all.

(Dan): So. This is (Dan). So, a follow-up – because that's what confuse me. So if I'm doing – whatever example, pharyngitis, whatever measures are under that on – in – on the website, I just figure out, that's the ones I should be looking at. I don't need to be looking for my own on primary literature.

Reva Winkler: No. Absolutely not. That would be far more than we could ever expect you all to do. So the developers have provided information. And as you look at each of the measures in the measure worksheet, you will see information that's provided by the developers and you will also see the preliminary analysis that was done by staff, by us, me primarily. And in it, we will help guide you through, you know, how you should look at the information provided by the developers with the criteria in mind.

So you really should not, unless you feel compelled of your own, need to look beyond that document, OK? So we really want you to evaluate the information that's been provided to you.

(Dan): Perfect. Thank you.

Reva Winkler: OK? All right. So, one of the things that committees have found to kind of help them out is if we can provide them sort of a decision pathway. And if you go down to algorithm one on page eight, for those of you following the Webinar, if you got the document that's great. This is really meant to help the committee with the information in a decision logic fashion.

So if you look, the first question asks, "Is it an outcome versus a, you know," or asks, "Is it an outcome?" If the answer is yes, there's a simply one-step question of do you agree that there's a relationship between the outcome and some process or a structure that can affect that outcome? If it's a yes, you pass. If it's a no, you don't.

If it's not an outcome measure, and by the way we do have two or three outcome measures in the group so that that will apply for a couple of the measures. However, the vast majority of them are processed measures, so it's not an outcome measures so I'll put you down in to box three. And so box three asks about performance of the either a clinical intermediate outcome process and is it based on a systematic review of the body of literature? And this could be a guideline.

So if indeed the evidence is based on a systematic review and one of the questions that the developers respond to is, is the evidence based on a systematic review? If so, what's the source of it? And there's a checkbox around things like clinical practice guideline, U.S. preventive services taskforce, something other like a Cochrane review or an evidence-based practice center review. And so they're going to tell you what it – where their systematic review came from.

There are occasional measures that weren't based on a formally performed systematic review of the literature. And they would – so you would follow the pathway down to the nose and the developers should provide some empirical evidence that you would just follow that algorithm decision. But the vast majority of them have an evidence review typically based as part of a guideline. And so that will take you over to box four, and this is where you'll start answering the questions. Is the summary of the quality, quantity and consistency of the body of evidence provided to you? Do you have, you know, it was however many studies and the studies were considered good or not so good or whatever.

And if you do have that level of detail then you would move to box five and look at the conclusions of the systematic review of how, you know, how many

studies, how consistent? Is there a high certainty that the benefit is substantial? You'll see the criteria listed in the 5A, 5B, and 5C boxes and that would drive your rating.

If indeed you don't have the details of the evidence review and some guidelines do not provide them, they just give you a recommendation and they may say it's based on A level evidence but there's nothing to tell you what that evidence really is, then you would follow the path down to no. And then it asks, "Does the grade for the evidence of recommendation indicate, you know, high quality evidence, high certainty, strong recommendation?" In other words, is it good? Because presumably the organizations creating the guideline is making the recommendation based on strong evidence. And so you would rate it and the highest rating for that when we don't have the details of the evidence review would be moderate but moderate is still a passing grade.

So you can see how this algorithm can be helpful in sorting through the information. Also, in the preliminary analysis that staff has done is then to guide you through this too and so I help you by saying, you know, "This is a process measure go to box three. This does have a summary of the (QCs) so you can go to box five to figure out your rating." or, "You need to go to box six to figure out your rating."

So I've tried to give you a little bit of guidance so that you're not having to overly spend too much time on the process and can get to the meat of the – of what we're really looking from – is what your rating is.

So are there any questions around evidence and the use of the algorithms? We've gotten good feedback on algorithms from other committees who've used them. They felt it was helpful in sorting through the information.

(Judith Lynch): Reva, this is (Judith) again. The preliminary analysis, is that found under each measure?

Reva Winkler: Yes. It's in each measure worksheet.

(Judith Winkler): OK, thank you.

Reva Winkler: Thoughts or questions from anybody else? Really? I can't believe I'm being that clear.

All right. I'll trust the silence that you're all with me. So, that's essentially the – what it is about evidence. So if not, evidence will move on to other things. I will mention that we are looking for things to be evidence-based, very rarely or unusually a committee may feel so strongly about a measure that it does not have strong evidence that they wish to make an exception. But again, I think this requires a great of discussion by the committee in consideration for whether you truly want to recommend a measure that is not evidence-based as something that shifts that providers should be held accountable to. So – but that is an option, though I just want you to be aware. But again, it's something we don't expect to be used very often.

OK. There's additional information in this guidance about evaluating the quality, quality, and consistency of body of evidence which I'll let you peruse at your own. But we'll go down to page 11 of the guidance and talk just briefly about performance gap. And this is another important subcriteria because the opportunity to – for improvement is important to – for measures to function as drivers of quality improvement. If everybody is already performing at 100 percent it is not worth the resources to collect the data, do the analysis and report the results just to say everybody's doing grand. That is not the purpose of the measures that NQF is looking to endorse.

We are looking to find the tools that will drive additional improvements in quality. And so the opportunity for improvement is an important must-pass criteria. And we are looking for specific data from the measure if it's a measure that's been endorsed and been used than there should be some data on what's happening and particularly even data overtime for trends to see if it's effective.

Sometimes we see measures that are – seemed to have very high performance. And before we assume that there is absolutely none, no opportunity for further improvement, we do want to consider things like the distribution of scores, what's the range? We also want to look at that number and representativeness

of that data. I mean if you're only measuring three or four entities it may not be representative of providers in general.

We may want to look if it's available of data on disparities because perhaps the opportunity for improvement is not necessarily overall but perhaps for certain subpopulation.

And so also, the sizes of the population at risk, so how big of a denominator are we talking about? Could an improvement from 95 percent to 96 percent really be – affect a large number of folks? So those are the kind of things you need to think about in terms of opportunity for improvement when you're looking at the data that's provided.

So any questions about opportunity for improvement or the performance gap criteria?

OK. I will tell you that often this one tends to be consternating for committees because the question is, you know, how high should we really expect performance measures to go? Is 95 percent really a pretty good level and there's not much room to run? Or maybe 100 percent where we need to keep driving to? And that's often a conversation that committees find themselves in.

OK. So those are the two criteria under importance to measuring report. And passing both of those criteria is required for endorsement.

So any questions before we move on to reliability and validity? Is this – Is what I'm saying making sense or is everybody totally befuddled?

Female: I guess just processing, how it will work with the …

Male: Yes.

Female: … one that I've been assigned to.

Reva Winkler: OK.

Male: Yes, I agree. I think it makes sense on a large scale. I get to look at my …

Reva Winkler: OK. Well, the other thing that's possible for you guys is now that, you know, we finish today, take a look at your measures. If indeed you've got some questions, shoot them to us or feel free to join us on the next Q&A call and we can talk them through also. But we do want to answer your questions, so don't hesitate to ask them.

Male: Right.

Reva Winkler: OK.

Female: Let me ask – Let me go ahead and ask this question, though, because some of the measures are continuum. So you need to have AHAP and then B is a result of A and C is a result of B. Because it's sort of a …

Reva Winkler: Right.

Female: … continuum of care. But if we've done – we're doing great with A, we're not doing so great with B and C still has a long way to go. Do we need to endorse – I know I'm talking in circles, do we need to endorse A?

Reva Winkler: Well, that's the question …

Female: I mean, or do we have the …

Reva Winkler: … research.

Female: Or do we need to find the disparities even though it may be small but it actually could impact the latter?

Reva Winkler: Yes. Actually, you're raising a question that's absolutely a terrific question for the committee to consider when you're doing your …

Female: OK.

Reva Winkler: … final recommendations. That's an – it's an absolutely perfect discussion point for the committee.

Female: OK.

Reva Winkler: There's no automatic answer. That's something for you all as a group to discuss and decide.

Female: OK, thank you.

Reva Winkler: Sure. Pardon me. All right. Then let's move down and talk a little bit about reliability and validity. These are the scientific acceptability of the measure properties and we're talking about specific – specifications of this measure with the codes, with the definitions, with the – all of the information it takes to calculate this particular measure. We want to know will the result of this measure, as specified, give us reliable and valid results? And so we are looking for some demonstration through testing that that's the case. And so that's what we're looking for.

So under reliability, there are really two types of things we're looking for to support reliability. The first one is that it's well-defined and precisely specified. These days, we usually are seeing very well-defined and precisely specified measures that are using codes that – standardized coding that's existing. For eMeasures, we're seeing them in a standard HQMF format that's actually required.

And for the most part, I think developers are very good at specifying and defining their measures. But this is an area that we really count on your expertise to look at those, to be sure that those definitions are accurate. And if there are the needed definitions so that everybody interprets the measure specifications in the same way so that it'll be implemented in a standardized fashion and the results will then be standardized and comparable. So that's the purpose of it.

We are in the process of transitioning from ICD-9 to ICD-10 some time in the near future. And so we have been requesting both ICD-9 and ICD-10 codes from the developers and they are – they give them to us.

eMeasures, again, are specified in HQMF which is the standardized format. We do have six eMeasures in the project. We have an internal technical review of eMeasures for the HQMF for the value sets and the quality data

model, so that we will be providing you a report of the finding from that technical review. We don't expect you to do that. So if you feel like some of those things rather foreign and your knowledge of eMeasures may not be that great, don't worry about it, we will be providing a great deal of the information to you.

We don't have any patient reported outcome performance measures nor any composite. So there are some additional guidance for that.

So any questions about specifications? There is a section in the measure submission that really goes into detail, the questions that are asked of the developers to provide us a full specification. We expect to have complete details so that someone could take that information and go implement the measure and needing nothing else. So we do expect to see all of the details.

So reliability testing is an important part of the scientific acceptability evaluation. And reliability is really looking at the precision of the measure. And so we're looking for demonstration though some type of testing that the data elements are repeatable or producing the same results with the same population, same period of time.

We allow testing at either the data element level or at the level of the measure score. Typically, the type of reliability testing is different depending on which of those two they choose. There is a preference for testing of the measure score because actually that's what we care about. We want to know whether the actual measure result is reliable, though not absolutely required.

So is everybody clear on what we mean by the data element versus the measure score? Is there any questions about the – what we mean by those two different things?

Male: Yes, can you clarify what you mean by that?

Reva Winkler: Sure. The data elements are really the building blocks for calculating the measure. So if the measure is about, you know, a group of folks with a certain diagnosis, diagnosis is one of the data elements, there's probably an age

requirement. So an age is a data element. And say – may be a population that's taking a certain drug, so the drug name would be a data element.

If you're measuring whether they have, you know, a certain laboratory test and a lab test or the lab test result will be a data element. So those are all the building block pieces that are pulled together with the calculation algorithm to create the measure result which is the answer to the, you know, to the measure, whatever the measure is about. If it's the percent of patients taking beta-block, you know, patients with coronary disease taking beta-blockers the measure result will be 89 percent, or whatever. So those are the differences.

And it's important that the data elements are reliable as well as the measure score is reliable in terms of the results. Did that explain it?

Male: Yes.

Reva Winkler: Everybody clear there? OK. As part of the preliminary review, I also sort of tell you what the testing was done. It was done on the data elements during this kind of test, just to help make it clear so you're not struggling to figure that part out.

Reliability testing of the data elements are usually a comparison of several abstractors, because usually the data element testing is done with measures that are abstracted for medical records whether they are electronic or paper, doesn't really matter. But the question is, will the same, you know, how – what's the concordance between multiple abstractors in pulling the data out? And there are some data elements that are notoriously difficult to find. And so sometimes you'll see low scores on the inter-rater reliability on certain data elements and that will be important in understanding how reliable the entire measure could be.

The measure score is usually a value – reliability is usually evaluated by a signal to noise analysis, looking at the data from multiple entities and determining how much of the value of the result is really a good strong signal versus noise and this often is related to the number of measured entity. So if you have a low number of small – very few measured entities to scatter maybe so much that the noise is very great. It may be that the reliability is not very

good below a certain number of observations. Let's say that you need a minimum sample size of 10, 25, 50 whatever. And the data may show that you – the reliability gets – becomes lower as you have smaller and smaller sample sizes.

So those are the kinds of test results we'll be looking at. And the developers have done a pretty good job of providing you with some – with easy to understand information. And the preliminary analysis we did is meant to guide you through it and point out the things that are important to focus on.

And so, again, as an aid, there is an algorithm for reliability testing. And, Vy, if you can go down to that. We'll just show you then kind of walk you through that one briefly. And you'll see that the first question is about the specifications. Are they precise and ambiguous, complete so if they can be consistently implemented? I mean think about it in your own personal situation? Could you take that information and implement it in your – in where you are? If not, that's going to be a real problem. There's no way that measure could be reliable.

And then the second – but most of the time the answer will be yes and we'll go down with empiric testing, reliability tested conducting using statistical test with the measures that's specified. In general, the answer to that one is going to be yes, though I will say that for the eMeasures there is an allowance for empiric validity testing at the data element level to account for reliability. And that's the arrow going over to box three. And I will point out to you when that applies. So we don't expect you to figure that one out. It took me a while to learn to figure it out, identify it myself.

So as we go down, you know, to box four that, yes, empiric reliability testing was conduct, it asks "Was it conducted with a computed measure score?" And if the score is what was evaluated, you go off to the right, yes. Then they ask, "Was the method appropriate?" And I usually give you an indication of that in the preliminary analysis. Most of them are signal to noise and the LSC is done quite well. And then you will go to box six to determine based on that result how would you rate the measure? And that will be up to you to evaluate the testing result.

If it was not on the measure score, in box four your answer is no, you'd go down to box eight where it's asking about reliability testing conducted at the data element level. And this is where you're looking at probably inter-rater reliability and the questions will be, "Did they test all critical data elements? Did they provide the statistical analysis of the agreement? And then based on those results in box 10, how would you rate the measure?" So as you can see, moderate is as high as you can get if all you've done is data element reliability testing. That reflects the preference, again, for measure score testing.

So, again – but the combination of the algorithm and the guidance we've provided in the preliminary analysis. I think hopefully this will take away some of the scariness of this that I know most committees look at this the first time and kind of go, "Oh my goodness". So we've tried to help you and provide as much guidance and – as possible. But feel free to contact us with any questions. We're happy to help if you're finding that you're just confused and not real sure about what this means.

So we're going to stop right here, because that's were reliability. And any questions at this point from anybody?

Male: Have the developers actually done a work on these measures?

Reva Winkler: What do you mean? They provided the information.

Male: Yes. The information is now available online on the SharePoint site.

Reva Winkler: Yes. That information is in the measures worksheets. You're going to have a main worksheet for every measure. It starts out with an identification of the measure so you know what it is. Title description, who developed it, why did they develop it, what's the basic numerator, denominator is. Something to get you oriented on what we're talking about.

Then you'll see the preliminary analysis that's done by us. And when that's, you know, once you're through with that set of questions will also – part of the preliminary analysis will also include any comments we've received in a pre-evaluation comment period that's currently ongoing. We solicit any

comments from folks to know if you've ha d any experience with the measure, any issues with the measure, if it's work for them, not working for them, you know. Just to see if there's something we can find from the field. Those will be included.

Once that information, the last comment section is there, you'll see just pages of information about starting with evidence and then gap and the specifications, that is all information provided by the developer and you have it all.

Vy Luong: Reva, I'm actually going to just pull up a worksheet just so the committee members can see. So that's being pulled up …

Reva Winkler: Yes, that sounds good. So you see, I think – yes, Vy just pulled one up on the Webinar. And as I said, this is the worksheet. So the first thing, brief measure information is strictly – this is the information entered in to our data system by the developer and it's just sort of a standard set of information to get you oriented as to what's being discussed.

Then if you scroll down, Vy, you'll see the preliminary analysis starts and we go through this criteria in order providing you with just a basic reminder of what the criteria is about. You'll see the highlighted underscored evidence, that's a link to the evidence details in the – further on in the document and this is all the information provided by the developer. To scroll back, as you go down there are a few more links we'll put in and then some questions for you to think about in terms of doing your evaluation.

Then you'll see in the shadowed area the – I'm trying to guide you through the algorithm so that hopefully you'll be able to spend your time and efforts on thinking about the measure information and the – how you evaluate the information rather than the process of getting to the rating. Again, you'll see similarly with the information around opportunity for improvement, we've given you information and questions.

And so if you go ahead and scroll all the way down and you'll see that there's a place for putting comments in. And if we get any comments from the pre-comment period we'll put them in there for you to kind of be aware of what

folks in the outside world may be thinking about the measure. We don't get lots of comments, but occasionally we had some that are very pertinent and very important.

So, again, you'll see the sections on reliability, scroll down some more, validity, the whole thing. So what we don't try and redo the information provided by the developers, we do try and point out the things that are the most important to use for the criteria. There are several questions on threats to validity around exclusions, risk adjustment, comparison and the information comparison of data sources.

OK, scroll on down. So as we go down past feasibility, past usability. OK, all right, go down to the evidence one. Keep going. All right, starting right here, this starts with evidence. You'll find there are numerous pages to follow. All of this subsequent information is provided by the developer. We provided either through our data systems a series of structured questions or we provide them with Word documents that they fill in. And so all the information from here on out is provided by developer.

Scroll down to the testing document, Vy. And this is where they put the information about how they tested the measure and their results. So, all that's in there for you.

Does that answer your question?

Male: Yes.

Reva Winkler: OK. All righty. So that's why we've tried to help you navigate through lots and lots of information. All righty.

Female: And also …

Reva Winkler: So …

Female: … Reva, just to know for the standing committee members, once we get your preliminary evaluation surveys that are due – that I believe they are – the due dates are posted on the workgroup assignments, we would just input it here for

the different criteria so that that way it comes in handy when you're at the in-person meeting and you want to just briefly look at what was discusses by you and your fellow colleagues.

Reva Winkler: Right. We try to co-aid all the information that are inputted to this evaluation process in to one place so that you're not trying to shovel through bunches of paper.

OK. Vy, go on back to page 12 where we talk about validity.

Anybody have any questions before I leave reliability? OK.

Validity, again, is related – but it – distinct from reliability and that validity really ask the question, "Does the measure result reflect an accurate assessment of the quality of care?

And so validity is actually quite a significant property to evaluate for a measure. But I will admit, and it is a difficult one, we don't see a lot of empiric testing of validity though it's happening more and more. What is allowed is a systematic assessment of phase validity and it's usually done by the measure developer or their expert panel. Occasionally, they'll get another group of experts to do a systematic assessment of phase validity. But phase validity is loud though the highest rating for that would be moderate. We certainly would love to see some type of an empiric validity testing similar to data – to reliability, validity testing can be done at the level of a data element or at the level of the measure score.

At the level of the data elements, in order for it to be validity testing you would be comparing the information used for the measure against some gold standard source. So there – the difference is not comparing, you know, how the concordance between two abstracted values but comparing say a coded value against the gold standard value in the medical record. And so we are seeing some validity testing at the data element level for measures.

We don't see a lot of empiric validity testing of the measure score, though I think we – that may be coming more in the future. We – most folks rely on the phase validity. And I will say that frankly, what the committee ends up

doing is sort of another phase validity assessment based on the information provided.

And so validity is an important characteristic. We really do want to look at potential threats to validity, things like exclusions, are the right populations included, are there any populations inappropriately excluded, however the exclusions handled the – could impact the validity of the measure result. Also, we look for any risk adjustment or case-mix adjustment that may be needed for a measure. Those are typically associated with outcome measures.

As it turns out, the outcome measures that are in our set actually have no risk adjustment. And so the question to you all is, do you feel that's appropriate? If very well, it might be. But, again, there may be some issues around the impact of case-mix adjustment on the measure result.

We are also looking for things like comparability of multiple data sources if the measure is specified as a chart abstraction and a claims-based measure. I mean, how do you know the two if used in the same program? We'll give you results that could be used to make comparisons. So there are several threats to validity that we want to look at.

But, again, validity has another algorithm. So, Vy, if you want to go down to the algorithm. We'll just kind of quickly use – look at this. The first one is – The first question is, are the specifications consistent with the evidence? If we've said that the evidence was good then the measure better reflect that evidence and not be about something else. So we just want to be sure that's a match between the measure that we're evaluating and the evidence we've already said was good.

Then the second is we're all potential threats to validity relevant to the measure assessed, exclusions, risk adjustments, statistically significant and meaningful results, how missing data is handled. All of these are specific questions the developers were asked to respond to.

And so if you feel there aren't any issues there then you would move to empiric testing of the measure specified, again, we rarely see that. And so the

answer is frequently no and then with phase validity systematically assessed then yes and how would you rate it?

If empiric validity testing of the score was not conducted then it would be – I'm sorry, you go down to six for this is the validity testing in the computed performance scores for each measured entity fell across to box eight for the ratings.

So, again, it's the same idea to help you through this and the preliminary analysis gives you some guidance there. So, is there any questions about what we're – what we mean by validity and what the criteria are looking for validity?

OK. All right. Like I say, if questions do occur to you don't hesitate to contact us.

In the last couple of …

Male: I have a question about validity.

Reva Winkler: Sure.

Male: I think I got what you're saying, but this is just a verification. So, you know, our measure is appropriate use of – a non-use of decongestants for otitis media with effusion. So I guess you'd have two validity elements. One is when someone says patient had otitis media with effusion and we didn't give pseudoephedrine or decongestant, the first validity measure would be, was the diagnosis correct? And the second one would be was the statement in the chart correct about what measure – the patient actually took?

Reva Winkler: And the real question actually for the measure is when you look at the measure result for that patient or that provide, did it reflect what really happened?

I think one of the things that you have to look at how the measure is constructed because if the measure is defined in the denominator by the – say

the diagnosis codes for the visit, no one's going to validate that. That's how they gotten to the measure.

Male: Right. So we don't – do we know how accurate pediatricians make the diagnosis of otitis?

Reva Winkler: Yes, exactly. And I think it's a – something for you all to discuss. But the question is whether the measure really addresses that or not.

Male: OK.

Reva Winkler: But that's a very fair question to ask.

Male: OK.

Reva Winkler: OK? Sure.

Male: OK.

Reva Winkler: All right. So in our last couple of minutes, I just want to touch briefly on feasibility and usability because I think they're fairly straightforward criteria.

Feasibility, we're looking to – for how much burden and this has to do with data source. Clearly, if it's chart (abstracting) measure there's a lot more cost of burden compared to a claims-based measure that's, you know, purely electronic.

You know, EHRs may come somewhere in the middle but we are looking for an assessment of feasibility, things that have high feasibility will have low burden or readily – data's readily available, typically will be electronic. Something that's already demonstrated, in other words it's in use, the data collection is happening, we have more information about – we'd like to have information about, you know, what cost are of collecting that data, analyzing it if available.

What if it's an eMeasure and this is something that we will check on for you as part of the technical review, we do ask for a very specific feasibility assessment of the data elements. Are these data elements that are typically in

current EHRs for multiple vendors? And so that's and important information for the eMeasures. But again it will be part of the technical review, so you'll get a report out from that.

So that's feasibility. Are there any questions about feasibility? It really is one of the straightforward criteria. Most committees don't struggle with it.

The last one to talk about is use and usability. And this really is something that's more applicable for measures that have – or currently in use, have been endorsed before. For new measures, they may not have much of a track record to really be able to know a lot about them and there's sort of more we hope or we plan these things.

But for those measures that have some track record, we're looking for information about how they're being used. What kind of use are they? They used an accountability application which was the intent of an endorsed measure. We want to know if they're being used, how they're used. And even importantly, if they're not being used, why not? Perhaps there's a lag time for getting it implemented, perhaps there's a plan in place that just hasn't happened yet, whatever. But we're looking to see how well they're being used.

We're also looking to see for measures that have been in use and have some data whether we're making progress with the measure. Is it effective? Is it really driving quality improvement?

And so those are the – And then the last one that is part of use and usability is whether there are any unintended consequences. Occasionally, there will be a measure that becomes notorious for unintended consequences and that usually becomes very well known. But sometimes folks that are closer to the use of the measure, people are being measured, have information that's helpful in understanding whether a measure is having any unintended consequences. So that would be part of use and usability. And so those are the subcriteria under that – in that area.

Any questions about that, because I know we're just about to – our time and I just want to be sure if there are any last minute questions.

Female: I have a question about the online evaluation that (inaudible) for the workshop. I don't understand what that is.

Female: Sure. So I can speak to that. So I am going to screen share right now for SharePoint committee (inaudible), there's documents here for your review. If your scroll down, you'll see the measure documents and you'll also see in the general document your assignments.

If you click on your assignments, you'll see what measures you're assigned to and then you can click on those measures to get to the measure worksheet that Reva is talking about as well as any relevant information that the developers submitted.

So for example, if I click on 0653 measure, there's a measure worksheet, they also provided a sample calculation algorithm. And now if you go back to the committee home, you'll see underneath there's a survey section, it's EENT preliminary measure evaluation. So if you click on that, we would like you to do a survey of the measures you're assigned. And all you'll have to do is click on Respond to the Survey to the left, right here, and once that opens up you can use the dropdown list to see which measures you're assigned.

So if you're assigned measure number two, you can click on that and then you can just fill out your thoughts on the criteria for measure two. And hit Submit at the finish at the very end and we will be able to gather all of that information to put in to the measured doc – the measure worksheets later for you. So, that you'll have that at hand along with your colleagues and their opinions of the different measures.

Does that make sense?

Male: What's the timeframe for that?

Female: So it is on your workgroup assignment. I'm going to pull that up right now real quick for you. So each person has been assigned measures, two to three measures. And for example, I just pulled up the workgroup assignments

(inaudible) for our workgroup – the first workgroup on May 11th from 12 to 2:00 p.m., the online evaluation is due May 7th, it's up top.

As for workgroup number two, your online evaluation is due May 13th and your workgroup call is May 15th. Workgroup number three, your online evaluation is due May 14th and your workgroup call is May 18th. And as for workgroup number four, your online evaluation is due May 19th and your workgroup call is May 21st.

Any questions?

And as always, feel free – I'm sorry.

(Rich Madonna): Yes, I have one quick question. This is (Rich Madonna). So for each of the measures you're having. Essentially, two discussions, how – what is the interplay or interaction between the discussion supposed to be like, or is there any?

Female: Sure. So we …

Reva Winkler: Yes, this is Reva – yes.

Female: OK.

Reva Winkler: Sure. Initially, we're asking into act as independently as individuals of doing the survey. But that's one to benefit for the workgroup is where you're going to come together and discusses them as a group and sort of get a, you know, the issues and questions you may have of each other and discuss may be things you think differently. And so it's sort of a preliminary review that you'll work together as a group.

(Rich Madonna): OK. That's very clear.

Reva Winkler: That makes sense?

(Rich Madonna): Yes, (basically).

Female: Any more questions? OK. Hearing no more questions, we're a little bit over time. We want to thank everyone who participated on the call and we look forward to working with you.

Female: Yes, and feel free to e-mail us with any questions pertaining to travel logistics or the workgroup assignments and the workgroup call dates or any other relevant questions you may have.

Female: Thank you. Have a good day.

Male: Hey.

Female: Thank you.

(Rich Madonna): Bye.

Reva Winkler: Thanks everybody.

Female: Bye-bye.

Female: Thank you.

Operator: This concludes today's conference call. You may now disconnect.

END