NATIONAL QUALITY FORUM

Moderator: Sheila Crawford
April 15, 2014
2:00 p.m. ET

Reva Winkler: Good afternoon everybody. This is Reva Winkler from National Quality Forum. Thanks all for joining us. Today, we're doing a Q&A session for our Surgery Steering Committee members because it's time to begin working on measure evaluation. And we want to be sure that you have all the tools and know where to find all the information.

So this should be a very informal session. So for those of you who are already on the line and hopefully a few more of your colleagues will join us. Feel free just to jump in if you've got any particular question.

The operator just said there were five of you signed in. I'm just curious, who's on the line, just shout out a name.

Frederick Grover: It's Fred Grover from (Denver).

Reva Winkler: Hey, Fred.

Lynn Reede: Lynn Reede.

Reva Winkler: Great, Lynn, thanks.

Larissa Temple: Larissa Temple.

Reva Winkler: Hi, Larissa, how are you?

Larissa Temple: Great.

Reva Winkler: Who else is on?

Collette Pitzen: Collette Pitzen.

Reva Winkler: Hi, Collette, how are you?

Collette Pitzen: Good, great. Thanks.

Reva Winkler: Thanks for joining us. Anybody else?

AJ Yates: This is Yates from Pittsburgh.

Reva Winkler: Hey, how are you? Great guys and hopefully, more of you colleagues will join us but again, like I say, this is an informal session as we start working. So what we wanted to be sure is that you all were familiar with the tools that you'll be using to evaluate measures. So the first thing we've got and if you're viewing the webinar, you'll see that we are looking at the SharePoint page for this project.

If you're not viewing the webinar and have access to your computer, you may want to go to the SharePoint page. If anybody is having trouble accessing the SharePoint page, please let us know. This is where all the information's going to be and it's going to be real important so if you can't get a hold of it you'll be in real trouble.

I want to point out something that we've just posted and that's the very last documents highlighted. I don't have a clicker. And that is the workgroup measure assignment. For this project, we've got approximately 30 measures to review and that would be a real incredible workload for each and every one of you to do 30 measures. So what we've done is broken the measures and the committee into four groups which we call workgroup.

And so, that document weighs out the measures in each group and the folks assigned to each group. Do you have any questions or issues about the assignment? Feel free to get back to us. We just posted that today.

So, our first workgroup call is on May 1st. So we've only got two weeks and the purpose of the workgroup call is to give the committee members a chance to have an initial discussion about the measures and the criteria to prepare for the in-person meeting.

The measure developers will be on these calls so you can ask them questions. They can begin to hear some of the issues you might be raising and prepare themselves to respond at the meeting. But really it's an opportunity for you to ask questions if you're not sure about what the information in the materials are or to ask questions about the criteria and how it might apply. Or if there's any, you know, anything that you have any questions or concerns about.

Doing this preliminary review really does help prepare everybody for the in-person meeting and keeps that going smooth. The in-person meeting has a heavy workload. Our agenda of 30 measures will be quite intent so we really want to do as much preparatory work ahead of time as possible to make that happen in a timely fashion. So I point everybody to the workgroup assignment document and again, don't hesitate to get in touch with me if necessary if you have any questions about it.

So, essentially what we're going to be asking you to do with these seven or eight measures in your workgroup is to review the document and begin to think about how you would respond according to the criteria. And we're going to talk – review the criteria a bit today. And please if you have any question, this is really what this hour is for.

So, I just want to scroll down a little bit and show you that we are posting the measure documents below that sort of initial box. And you can see the number of the measure 113, 114, 115 and 119 – scroll down.

What we're going to be – we're going to look at one of this in detail in a little bit but if you open these up, you'll find that there are sets of documents that you will need to review. And so, why don't you open the one we talked about, (543). No, it's …

(Off-mike)

Reva Winkler: That one. We're just opening one of them, (453), just to show you that this is one of the measures and there are a couple of documents. We'll open this up a little bit more later but you'll see that there is a main document, it's a Word document that has all the measures submission information. And sometimes the developers have provided expert document appendices or extra tables, often spreadsheets of codes and they will be in there too. So, everything you need to know about the measure should be in this document set.

As I say we're going to come back and revisit this one in a minute but first thing I'd like to do though is to review just a little bit with you the importance of the Committee Guidebook. The Committee Guidebook is really your main reference and resource for doing the measure evaluation. And I want to take you to the section on measure evaluation – yes. OK, which is section 6, page 31.

And this begins a few pages of fairly detailed information about NQF evaluation criteria. Scroll down a little bit. And we go through the measure evaluation criteria in the order that we will be evaluating the measure. And what this section of the guidebook does is it gives you the detail about NQF measures evaluation criteria. It gives you a lot of the reasons and rationale for the criteria.

NQF has been evaluating measures for more than a decade and over that timeframe, a great deal of quality measurement has evolved and changed. And in response to all of that evolution, NQF evaluation criteria also evolved. So if you've worked with us in the past perhaps and most likely, our current evaluation criteria is likely to be different than what you can recall. So it's important that you know what the current criteria are.

The main criteria haven't changed much, importance to measure and report encompasses three sub-criteria around the evidence for the process of care and the outcome whether there's a performance gap or opportunity to improve and how high a priority. We also have some composite measures in this project, so it will – the number one criteria will also include assessment of the quality construct for why that composite was put together the way it was. So we'll go into some of this in a little more detail.

The second criteria is scientific acceptability but essentially, what it is is testing the measure for reliability and validity. And so, we'll talk a little bit more about what's expected there.

Feasibility is another criteria and that really has a lot to do with the data source, the burden involved in collecting data, in reporting data. And then usability and use is another criteria. How are these measures being used or if they're new measures, how are they envisioned of being used? What's in the impact on quality improvement? And whether there are any potential unintended consequences.

So, particularly for our measures that are undergoing their maintenance review, those that have been endorsed previously and sometimes it was maybe only three years ago, but some of these measures have been endorsed for 6 years, 8, years, 12 years and so the question is, do they still have value in NQF portfolio of measures?

Another thing that we will be looking at is related to competing measures and in fact we've got a couple of those too. And, you know, it adds a lot of confusion and chaos when there are multiple measures measuring the same thing. So we really want to understand the value of each measure and whether you can determine the best measure among competing measures. And certainly, if measures are related we want to see those measures be harmonized or aligned so the specifications can work together and the folks out in the field who are measuring and collecting data are not having to look at distinct and unique sets of specifications for each measure and retool it each time. That is really something that's problematic for implementation.

So these criteria really haven't changed a lot but the underlying sub criteria, the underlying things that we're looking for to evaluate against the criteria really have evolved and become a lot more precise. So I'm not going to read this section of the guidebook to you but I do – it's important that you do read it because if you don't you probably will not evaluate the measures very well.

So we're going to scroll down. And you'll see that a lot of these information – we spent a lot of time left on writing this out trying to really provide the rationale for all of the criteria so that you'll have a clear understanding of why we're using the criteria we use to do.

We also have sections that talk about evaluating new measures that may have not been put in play. And therefore, we don't have a great deal of experience with compared to previously endorsed measures which should have a track record that we want to see how well they've done, what's in their impact, have there been any unintended consequences.

So as we go down and look at the details of this, we also put references and you'll see the links of what good looks like. And this is a document that's in your document set that's really provide some examples to most – for the measure developers to explain what it is we're looking for – the kind of information that really speaks to the criteria that makes the committee's job straight forward in being able to look at this information and the criteria and say yes or no, it needs it or not.

So steering committee members have told us that they've find this document valuable as well. So we make reference to it and link to it. And I would encourage you to take a look at that as well.

So as we're coming down the first criterion is an important to measure in report. We talked a little bit again a lot of the background. And we talked about NQF preferences for outcome measures that has been a real evolutionary aspect of measurement where we are seeing a lot of outcome measures and in this particular topic here in surgery, we have a large number of outcome measures.

And outcome measures have their own strength and weaknesses, some of their own methodologic challenges. And so, we will be looking at those. But as you can see, NQF preference for measure does put outcomes above process measures and above structural measures. And we talked about why outcome measures are particularly important to us.

So when we look at the initial criteria, we provide some key points as to really for you to pay attention to that you really understand what it is we are searching for in the evaluation of these measures. So the first one is evidence. This is the biggie.

And evidence is something that is – it probably varies on different people's ideas in what evidence is. But NQF is looking for empirical evidence. We are actually looking at the body of evidence which means all appropriate and pertinent published studies around a topic area or process of care. And so, we are – we look at the evidence around the quantity of studies, the quality of those studies and the consistency of those results. Now, for – that's very important for process measures or intermediate outcome measures and for structural measures because we ask the developers to diagram the relationship between a process and a structure in your intermediate outcomes to patient outcomes.

And that relationship is what we're looking for in terms of the evidence. What's the evidence? How good is the evidence? How strong is the evidence that that process of care can get you improve patient outcome?

So evidence is important it is not limited to randomize controlled trials. We're looking at all of the studies. We know that some areas of health care are studies much more in-depth than others. And so, you will find there will be some measures that have just pages and pages of documentation of the evidence and somewhere it's relatively limited.

Now, as I said, we're looking at the evidence between the relationship of what's being measured and the patient outcome. We'll, if you're dealing with an outcome measure, the relationship is pretty self-evident. And so, actually, four outcome measures requirements for evidence is relatively limited. It really is an assessment of whether there are processes of care or actions that can be taken by the health care providers to influence those outcomes.

And that's really is it. We're not looking for the whole body of evidence, quality, quantity and consistency of evidence. We're more talking about outcome measures. So it really is important to be sure, you know, we're

working with the process measure and outcome measure as you start evaluating information provided for evidence.

Also the information provided is – that provided by the developer in their submission. They have an electronic web-based portal to submit the information and we just put it into a standardized format for the committee to review.

Sometimes developers put in a lot of extra information that isn't necessary. So just keep your eye on the criteria in terms of what is necessary even if they tell you a lot of other stuff. OK?

So that's the – so evidence – we're looking at studies. So recently again with input from previous committee members and other stakeholders we've tried to find some tools to help the committee evaluate evidence. And so, this year is the first time we've actually have developed these algorithms to help committee members think through, to have a look at the evidence.

And so, we hope that these will be useful for you. And then, we ask you to please try and use this but give us any feedback on how useful you do find them. If you noticed, box number one asked whether it's an outcome measure or a process structure other type of measure. If it's an outcome measure, you go to the right box number two and then you just, you know, the question we talked about is there a relationship between the measure, between the outcome and at least one health care action. If yes, then it has the criteria. If not, then perhaps not.

So outcome measures are very straight forward. And we've got a large number of outcome measures. So evidence may not be that significant in criteria when we're looking at outcome measures. When we look at scientific acceptability, that's where we're going to spend a lot of time evaluating outcome measures. OK.

So box number three is where you end up if indeed it's not an outcome measure then it's something else. And so, the question there is, for those other type of measures, is the evidence provided by the developer. We are not expecting you to do any independent evaluation of evidence. We want you to

just look and see what's been provided – just read what's been provided by the developer. Is it based on a systematic review and grading of the body of evidence other than just selective study?

So that's a question for you to look at information provided. If it is a systematic review, then you move on to box four and if not, you would go down to the lower box below – I think seven. OK. But we'll go back up to box four.

And so, when it comes to systematic reviews, what are the most common sources of evidence or measures is clinical practice guidelines. We're in a bit of a transition as we're moving for clinical practice guidelines seeing more based on systematic reviews. Some organizations that do guidelines do very in-depth systematic reviews according to the guidelines published by the institute of medicine, but others, not so much.

And so, there is a great deal of variation in the guidelines that are out there. And we acknowledge that we're in this kind of transition zone so we have to be cognizant of that. So if the information behind the guideline is complete and gives us the full description of the body of evidence that describes the quality, quantity, and consistency – super – or if the developer has presented another systematic review that's published such Cochrane review or sometimes you'll see them as an independent publication.

Then, the question is do you have that information on quality, quantity, consistency. If so, you would move into box number five and you look at the conclusions of the review to determine whether you would rate the measure on evidence of high, moderate, or low as described. Slide down a little bit.

If not, if you don't have those details, then we understand that often guidelines won't have them. Then if no, it's great for the evidence or recommendation indicates that it's high quality. If that's the case, then you can still rate it as high as the moderate.

And so, this is how the close sheet is meant to help you kind of find your way through the evidence criteria. You can also see that if indeed there wasn't a

systematic review done either as part of a guideline or some other type of review, it walks you to the other step about other types of evidence.

If indeed and it happens that's extraordinarily rarely, there really – the measure does not meet the evidence criteria as laid out here but the committee feels that it is a very, very important process of care. None the less, even though it has no evidence base, there is an option for an option that's laid out here on the orange boxes. So you do have that option as well.

So this is how the – this algorithm is meant to help you and we really are looking for any feedback from you to determine how helpful it is. So this is simply just say a tool to help you apply the criteria that otherwise described above the need to guidebook. So that's evident. And I'm going to stop here and see if anybody has questions at evidence.

OK. Then, what I want to do is go two one of our measure examples. So let's go to measure – yes – 453. So if you're watching the webinar, we've got – we've pulled up the major information form for measure 453. I just picked the process measure because – as an example to talk about today.

Now this form is simply a way of structuring the data that as measure developers have put into our system. And so, it's in a standard format. Do you know where to look for things on all the different measures?

So measure information form starts out with the brief measure information. And as the title the number, who's the steward, equips the brief description. It also includes the developer's rationale. Why did you make this measure? What were you hoping to accomplish?

Slide down, you see just basic of the numerator and denominator and exclusions just to get a flavor of who's being measured and what's being measured. And then, very important information about what type of the measure it is. This is a process measure. But data source that this measure is specified for and has been tested for, and what is the level of analysis, who's being measured.

In this case, we're talking about a process measure based on looks like a combination of claims and medical records at the facility or hospital level. And so, this is a – I think it's a well-known (steep) measure. So that's how it's described.

So this gives you the basic information. So to – you have a context of the measure. We talked about it in previously endorsed. When was it last endorsed? So you can see it's been awhile for this one and so it's definitely due for its maintenance review. Then the information provided by the developers plays into this form in order that we – that helps you go through the criteria.

So we start out with the first criteria, importance to measure and report. And we start out with evidence. Now with – from a lot of feedback from the developers they've found that the submission system doesn't allow enough formatting flexibility to answer the questions the way they want it.

So what we're trying out is a Word document attachment that gives them that flexibility to put in things anywhere they want. So for evidence you'll see that there should be a link attachment that's really just at the end of this form so it's not in a separate place, it's just at the end of the form.

We have it blank. It went out there. There we go.

And so it takes you to this Word document that's attached. So this is all one document, this is an internal link. But it's just a way of helping navigate and so the evident questions are in this attachment. What we've done is provided in these boxes some instructions to the developer on how to fill out the form and then the second box is just a quick review for you all, two of the criteria. So these are just kind of reminders, go down.

So here we asked them to reiterate what type of measure it is. And this one is a process measure and as we slide down some more, if it were an outcome we ask them to answer question 182.1 and that would be the end of it. That would be all they would need. But instead this is a process measure so we go down to 183 and we ask them, what is the source of the systematic review?

And see this one match your algorithm so you're in box 3. So they tell us that it's from a clinical practice guideline. So they're expected to complete section 184.7. So we slide down to what we've talked about, the guideline. And the guideline reference citation is there. And then the guideline recommendation for this particular process of care, so they tell us the page number it's on and they quote it directly and give the rating that was assigned to this measure.

OK, so this tells us really what the guideline recommendation is and the evidence grade. So as we slide down we still don't necessarily know what their grading scale is so we ask them to provide the grading scale. So in this case it's a 1D process of care evaluation or evidence and that it means a strong recommendation supported by low quality evidence suggesting net clinical benefits or harm.

Now I will say that the criteria is focused more on the evidence and less on the recommendation. But – so you will – this is where your judgment is going to come in to play on whether that's going to be good enough. So they provide more information on this grading system. Again, just to help you interpret the result. And so we then ask if the guideline is evidence based rather than expert opinion, because that's what we're looking for, an expert opinion would not qualify, then they should complete more information in 187 to tell us what it said.

So we go down to 187 and this is again the description of the evidence itself. And in 187.2 they've given us a link which I was looking at this morning. It's a massive document with evidence tables that address large different types of questions around infection control. But one of them is the use of urinary catheters and so the section is there. These are the evidence tables. Then down further as we scroll down they do tell us that the timeframe, 1995 to 2007.

They tell us what the studies are as a way of the description of how many, you know, to minimize control trials. Several systematic reviews, I mean so we've got multiple studies, it's not just one. Again, the overall quality of the evidence, we ask them to describe, you know, and they talk about moderate quality evidence suggest the benefit of shorter versus longer (inaudible)

duration based on these (inaudible) risks, you can read it. And so this is the detail of the evidence that really addresses the quality, quantity in terms of the consistency of the evidence.

And so you will take a look at those – the details of those results and the criteria and determine how you would want to rate it by moderate or low. And so this where you find the information in this particular measure is in the evidence attachment.

OK, so that's evidence, so does anybody have any question what we've gone over so far?

Collette Pitzen: Reva, this is Collette.

Reva Winkler: Hi Collette.

Collette Pitzen: Hi, this is a question I guess I have my developer had on too. But I know it's going to crop up as we're going.

Reva Winkler: OK.

Collette Pitzen: Just some clarification, as they're going through, you guys use language about the entire body of evidence and having a rating associated with that when more frequently within the guideline there's ratings that are associated with particular actions.

Reva Winkler: Yes.

Collette Pitzen: Go ahead.

Reva Winkler: So I think it depends on how it's laid out. Those actions are the processes of care typically and it's that process of care and its relationship to outcomes which is the evidence we're looking for. So, you know, we do you – Ideally, you want it broken down so that you know the studies that specifically address that process of care.

So, you know, given the sort of changing environment we have around using evidence to support guidelines, I mean we have to do the best we can. But we really are looking for the studies that say that doing this gives you better outcomes for patients. And some measure that's going to be lot easier than others and we know that. So we'll just do the best we can.

Collette Pitzen: Yes, great thanks. So just to clarify, we're looking at drilling down specifically perhaps within a big guideline and finding that evidence for the specific process if we're …

Reva Winkler: Absolutely, absolutely because if you go back, scroll back up to what is it, one point – the very beginning where we talked about the relationship to patient outcome. Essentially, what we're looking for in evidence is this process of care, you know, has this impact on patient.

And so yes, here if you look at 183, urinary catheter insertion for surgery. Timely removal leads to deceased risk of infection. So what we're looking for is the relationship between catheter use, timely removal and decreased risk of infection.

Collette Pitzen: Great.

Reva Winkler: So that's the evidence we're looking for not everything else that could possibly apply to urinary tract infection prevention.

Collette Pitzen: Perfect, thank you.

Reva Winkler: Sure, any other questions from anybody?

AJ Yates: Yes, this is Yates. The form that you just showed for the urinary tract or the (Foley) removal (skip) measure, is that form similar to the tool that will be used for collecting our thoughts for being sent in before each of the workgroups?

Reva Winkler: Similar, I mean that's more of a questionnaire to give you a place to jot your thoughts down. The questions will be really around, you know, evidence. What are your initial thoughts, gap, or opportunities for improvement, that

kind of thing. So it will go through the criteria. Yes, that's the way we're going to do the evaluation.

(Wunmi Isijola): And AJ, this is (Wunmi). On the SharePoint page on the left hand side, there is a section that says surveys and at that point in time you'll be able to provide your thoughts on each measure priorities for the workgroup.

AJ Yates: OK, and you're showing me that now?

Reva Winkler: No, not right now.

AJ Yates: OK.

(Wunmi Isijola): We can pull it up really quick.

Reva Winkler: When we get there. OK.

(Wunmi Isijola): On the left hand side, that's where if you see straight from the left – Yes, that's where you'll be able for each individual measure.

AJ Yates: OK, they'll be laid out there for each of us – for each of the groups?

(Wunmi Isijola): Exactly. Yes, you'll just pick the measure. Exactly.

AJ Yates: OK, thank you.

(Wunmi Isijola): No problem.

Reva Winkler: Sure enough, so any other questions about evidence for this particular example? OK, then what we're going to do is we're going to go back to the form. OK, and we talked about evidence and now we're going to go down to one 1B which is the balance gap and on these two we asked them to provide scores for you to just measure.

Now if it were a new measure, quite possibly they may not have any really significant amount of data because it hasn't been used and that's OK. We would then ask for some indication from the literature or some data source

that hey, there is a quality problem here. So we really wouldn't expect to see a lot of details data on use to the measure because it's new.

But a measure like this has been around for a while, yes. Some significant made data has been presented if you can see the sort of relatively recent results at the national level. They do provide in an attachment greater detail but you can see the results over the quarterly results from 2012 and 2013 and the overall national level. You can also see progression and little bit of an improvement for each quarter.

They do tell you how many hospitals are captured and how many patients are captured. So this gives you a sense of what current performance is in use of this measure. So the question to you all is, is there opportunity for improvement in this measure? We certainly have seen over time that measures can be victims of their own success where at the beginning, there is a wide variation in performance but with the use of the measures and appropriate incentives, we see that gap essentially disappear as measures become topped out.

This is not unusual for successful performance measures that have been endorsed for a while. So, you know, we want to know from your perspective, your thoughts on whether indeed there is further opportunity for improvement that you balance the positive measurement against the information provided and that's why we turn to you all for your assessment of that criteria.

Frederick Grover: What's been your experience when you do discontinue a measure for that reason of regression and that of a decrease?

Reva Winkler: I don't know, Fred, that we've had a lot of experience. One of the things about this we've – most of measures that's been retired have been for CMS's IQR program and when CMS retires a measure they stop data collection. Now, they've only done that in the last couple of years and I haven't heard about them, you know, going back and checking again but perhaps we'll get a better sense of it maybe from some of the more registry-based measures that continue to be collected and can be, you know, calculated even if the focus is not on that specific measure.

Well, we don't have lot of data on that at this point.

Frederick Grover: I'll be interested in looking at it so the (old) Hawthorne effect.

Reva Winkler: Yes, well one of the problems is there are so many new measures but there's only a certain capacity for measurement out there.

Frederick Grover: Right.

Reva Winkler: And so the desire to replace measures that really just don't have a lot more room to run with newer measures that address newer areas and are more robust is high and so we just make room for them but it's very fair question and, you know, we probably have to think a little bit more about the best way to approach that.

Frederick Grover: Yes. Thanks.

AJ Yates: I have a question.

Reva Winkler: Sure.

AJ Yates: This is basically, you know, sort of a low ceiling question of what you're asking when everybody bumps up at the top.

Reva Winkler: Right.

AJ Yates: And the – is there a place for the process suggesting an outcomes measure that would be a simple way of testing the effectiveness of the process measure, in other words …

Reva Winkler: Yes, exactly.

AJ Yates: … because in this particular case, it's great that everyone reports that they're having near 100 percent compliance with taking the Foley out in the first day but did in fact the number of UTIs for those surgical patients go down for that hospital.

Reva Winkler: And that is exactly the kind of thing we would want to see in evolution of measures. What we are going to do is provide you with a list of all the measures that NQF has that it would be related. Now we do have the coding measure which wouldn't be exactly that question, it would be a subset obviously but – and so I think the recommendation from you all that hey, this needs to be, you know, we need the outcome measure at this point. You know, the process measure may have run as far as it can go and we – but we still need to know if it's having the impact we wanted to have.

So those are the kinds of feedback and suggestions from you all that would be very valuable to say, you know, we're not going to get – we may not get much more out of this one but the best –what we should replace it with is, you know, the true outcome measure.

AJ Yates: Right. And I have one other question that's very pertinent to this particular question. One of the things about registry data, for instance, postoperative UTIs in patients that have had their Foley in for longer than a day is that when services see that – surgical services see that they are outliers, they quickly account for that and they quickly fix it and they will routinely show that their UTIs have dropped and they're now in compliance with getting the Foley out in the first day but the question I have is do you ever consider the diagnostic intensity involved, in other words, do they just stop getting UAs?

Reva Winkler: Right.

AJ Yates: Do they just stop urine cultures and that's not so much an outcomes measure as a diagnostic intensity process measure. I mean if people stopped testing for the problem that they might be measured on.

Reva Winkler: Yes. No, I understand completely and those I think fall into the category of the unintended consequences. That is certainly one of the discussion points the committee has raised about many measures about the behavior, maybe undesirable behavior changes that are prompted by some measures and it's a very fair discussion point and, you know, I supposed that if it was felt to be a really significant problem that may be an appropriate measure to be developed as a counterbalance.

But these are all, you know, wonderful suggestions on how do we go from where we are today with these kind of measures and results to, you know, what's next, where do we really need to go with measurement and you may be right that there needs to be more of that kind of measure but certainly, we want you to contemplate those thoughts and discuss them as a group when you talk about these measures.

OK, so any other questions at this point? OK then – so 1D is around the performance scores and they give you a fairly nice distribution, they give you decile result. So it gives you a pretty good sense of what the national picture is.

The other thing that's important under this section is whether we have any information around disparity and it's not unusual that developers are able to provide a stratified data to determine whether there is a disparity and you can see that they've broken out the measure result by race and ethnicity and give you the result. And that's, you know, very useful data to know whether this measure really highlight disparities of care and I'll leave it to you to decide whether the results they provide demonstrate that or not but that's another important aspect.

It's possible that a measure that seems popped out overall may still demonstrate some significant disparities and differences among population which would be the justification for maintaining the measure as still having opportunity for improvement. So that's why disparity data is highly desirable at this point. So that's opportunity for improvement.

Opportunity for improvement is a very important criteria because measurement is costly and so the value of the information we obtain has to be weighed against those cost and if everybody is performing the same and there's not a lot of information about, you know, who needs to improve and variation among providers then perhaps that balance has shifted and it may not be worth the information gained for the cost of measurement and that's why opportunities for improvement is an important measure or is an important criteria to evaluate.

Similarly, with evidence, both of them are most passed criteria so if you feel that the measure is what we say cupped out with very little room for improvement but yet it's a good measure. It might be a victim of its own success. People have been – steering committees have been reluctant to remove endorsement from those measures because they are good measures and they want that goodness to still be available and so a type of designation for measures that meet all of the other criteria. They test out very high in reliability validity, usability, they don't have any underpinning consequences, they're solidly evidenced based. Their only issue would be opportunity for improvement.

We do have a category called reserved status which essentially is an endorsed measure. Unreserved status means, hey, it's a good measure but, you know, it really doesn't have much opportunity for improvement at this point and used with caution. So that is an option open to the committee to place that designation if they feel it's warranted. You know, it's possible that you have a measure that has very high performance but, you know, isn't – it could be replaced by better measures to make room for better measures or something like that and you may not want to use the reserved status option and that's fine too.

OK, the next criteria under number one is high priority and really, this is a certain (assessment) of whether it meets the national quality strategy, whether it really has a large impact in a large number of patients or a high severity or a high cost condition and, you know, for the most part, it's only a few measures that ever really stumble on this criteria but nonetheless, it's the most passed criteria as well. So those are the three sub criteria important to measure and report. So does anybody have any questions about the meaning of those criteria?

OK, then what we'll do is go on to next section which is around scientific acceptability and measure properties. And really what we're taking about here is reliability and the validity of the measure.

Now reliability really has two aspects. One is that the specification so we really ask you to take a careful look at the specification to be sure that they're

unambiguous that any terms that need to be defined are clearly defined. That the appropriate codes are included.

That the appropriate calculations algorithm is included and that all the information is provided so that someone could implement the measure in a standardized fashion. So that we can end up with comparable results. So if you see as we go to the section of the measure information form, we talked about different characteristics. So particularly for the maintenance measures I want to draw your attention to S3 that describe any changes for the specifications since it was last endorsed.

And so there are, you know, a few little things that they have done to this measure. It doesn't look like anything really significantly large. But it's important that you at least take note of that. Also, the rationale for those changes, some folks have changed measures in a significant way that's good. And in response to say feedback or something else. So you really want to why that measure is evolving the way it is. The specifications talk about the numerators statement which you've really seen.

But then also the time period for the data collection and then the numerator detail and this is where we want to see all the definitions, all the codes, all the every nitty-gritty, little detail you need to implement this measure. Now depending on the data source it maybe lots and lots of codes and so the developers will also put that in a Spreadsheet attachment. And that maybe another document in your document set. But if there should be codes and there are not codes, that's something you should be aware of and bring it forth.

So numerator detail, then denominator details very similarly and as you see they have appendix A table five for the ICD-9 CM code. Denominator exclusions. This is another area we really ask the committee to pay attention to, are the exclusions appropriate, are any populations excluded that perhaps shouldn't be. And so we really want your thought and expertise in evaluating the appropriateness of the denominator exclusion. Did somebody have a question?

OK as we can now, we talk about other aspects of the specifications whether there's any stratification done, whether there is any risk adjustment for outcome measures, this was going to be a big section. For process measures not much, typically, not only do they provide information within this form but there's usually an attachment that goes into a lot of the details with the measure testing. As you see we scroll down, there's more information about the type of score it is.

The interpretation of the score, the calculation algorithm how you would systematically go through the calculation. Let's see. Whether sampling is allowed, this is typically if it's already been implemented and how the measure has been implemented, whether it's a sampling strategy or not. And then how missing data is handled because how missing data can really affect the measure result. So their policy and specification for missing data should be included.

Again, for the details on the data source whether there's a specific data collection instrument, like most people are familiar with the type of data collection that the IQR project uses, that CMS joint commission used in the (cart) tools. But registries, data collection tool, a various kind, survey tools, all of it would be detailed here. And then again very important the level of analysis.

Tools is being measured. In this case is the facility or hospital level measure. So we're not measuring individual doctors, we're not measuring health plan. This measure is for hospitals. And then the care setting is indicated as well. We ask for a lot of these details because it helps folks to search on measure characteristics to identify measures that they maybe looking for, for use in their program.

Just as we had with the evidence, measure testing particularly is important to be able to tabulate data and make graphs and all of that. So again, we're using another Word attachment to allow those developer's flexibility in presenting their data.

And so this attachment again like the other that was an internal link at the end of this entire document and similarly in the boxes, we've got the instruction for the developer for completion and then the second box is just as reminders to the committee members. So you've got a ready reference for looking at what that criteria is.

So as we slide down and I know, I see where we might be short on time but I just want to give you a quick overview of this and then on our next Q&A session if you have any questions, we can certainly talk about it in more detail.

So we ask for the data that is used for testing of the measure, what type of data is it, we really expect that the measure is specified for type of data source, that's the data source that's tested. So NQF endorsement only applies to data sources and levels of analysis that has been tested, OK?

So they do – so they talk about, we ask for description of what data did they use so this is all from the QIO clinical data warehouse, lots and lots of cases, they do describe tools in it and they give you the data for calendar years in 2012, the specification for the testing. Then if you're going to have nice description of the measured entities including the facilities' adaptations and they have included an appendix with a table on page 13 and it really is a very nice detail of who all is included in the testing.

And so then we ask them for reliability testing how was this measure tested. Now, they don't haven't responded anything on reliability and this is where – there's always this double asterisk. Measures can be tested for reliability or validity at one or both levels either at the level of the measure score, how reliable is that measure results and we see a lot of those, this doesn't happen to be one, but we see a lot of those or it can be evaluated at the level of the data elements, how good is that data element.

And the idea is that if you have reliable data elements, you can put them together and come up with a reliable score. Now, reliability of the measure score is more meaningful and so you can raise a measure that test out well on measure score reliability is high, if they only look at the data element and

reliability at the highest level, rating is a moderate. Because there are still some questions about putting those together.

So when it comes to reliability testing at the data element level and that's what's going to go on here, there is a set of circumstances where if the data that's being used to calculate the measure has been assessed against the gold standard, and in our case, that's going to be the medical record, that data element evaluation will count for both reliability and validity.

And so that's what we've got going on here. So we didn't say anything about reliability because what they do down here under validity testing is tell us about their validity testing of the critical data element, they have compared the data that's abstracted versus their sort of gold standard abstractors if you will.

And so they tell you what the critical data elements that they tested are and then they tell you how it was done and then they give you the results as part of – OK. And here you see the comparison of the two types of abstractors, one is the hospital abstractor, the other is the clinical data warehouse gold standard abstractor and they compare them. And so the percent agreement and the capital, when they can be calculated. And so this is an assessment of the reliability and validity of the data elements.

We still don't have an assessment of the reliability of the measure score or the validity of the measure score. This is sufficient to pass the criteria but only at a moderate level. So I'm just looking to see what else is on, they tell you about the (cap) of statistics. Additionally, things that can influence and be a threat to validity are how the exclusions are specified and so we do ask for an analysis of exclusions and typically what we see in frequency analysis, and so that's what they've given us here, right, OK.

And the percentages and again, we're looking for your thoughts and input in terms of how much of an impact, whether it's appropriate or not that these exclusions will have on measure results and then they talk, they give you more data on the distribution of hospitals and how that – that exclusion pan out.

In this case, there is no risk adjustment but it's indeed as we'll see with many of the outcome measures, we are looking for the details of risk adjustment, it

will be started in this section and this gets to be quite long and detailed when we've got risk model to evaluate and we'll take a look at one of those next time.

The other thing we ask for is identification, specifically significant meaningful differences. I think in this particular case, they're a little bit confused. Validity test is not the same thing as to whether there are statistical differences in performance results.

So the question is, you know, can you differentiate among providers by looking at the results and they really didn't respond to that question very well. And so I think that's really the end – we asked them about details on the missing data and they tell us what their policy of missing data and how they handle it. And some folks – some developers will give us a lot of tabulation of how much missing data is found in their measures.

OK, as you'll see here, OK, keep going. And so these are just extra little details. So that's the rest of the testing attachment.

So like I said, for risk adjusted outcome measures, there's a lot of detail on the testing and development validation of the risk model in that attachment as well. So that's probably going to be much more of a focus for the outcome measures.

So we're almost at the end of our time and I hope that we provided a bit of an overview with an example of a measure, what your tools look like. Are there any questions at this point?

OK.

Frederick Grover: I think you've done a good – it's just the taking in.

Reva Winkler: All right. Well, between now and we have another of this Q&A sessions a week from Thursday and replay it would be great if you could take a look at some of the measures, we've got some of them up now and we're going to put more up over the next day or so. So it should be pretty much all up. And really take a look at the criteria and look at some real examples so that you

can see how you might go about evaluating the measures and then bring your questions next time so we can talk about them. Because if you've got a question, chances are your colleagues have got the same question. So this is an opportunity to share your learning and really be sure that everybody has a common understanding of the criteria and the process for evaluating the measures.

All right, but in the meantime, if anybody has any specific question, don't hesitate to get in touch with any of us here in NQF. We're happy to talk with your one on one as well. If any of the tools aren't working for you or can't get access to SharePoint or any of these things, please let us know and let us help you.

Any last minute questions from anybody before we close of?

Frederick Grover: When do you want us actually to start analyzing or reviewing the things that were assigned after the next call?

Reva Winkler: Is the survey active yet?

(Wunmi Isijola): No.

Reva Winkler: Fred, what we'll do is let you know, we have to get the survey active and let's try and get that done by the end of this week. And so then you'd be able to put them in.

(Wunmi Isijola): And we'll send a follow-up e-mail so that you're aware of the (accessibility) of the survey.

Frederick Grover: OK. Great.

Reva Winkler: All right, guys, any questions from anybody else?

All right, folks, well thank you very, very much for your time. And like I say, if you have any questions at all, please don't hesitate to get in touch with us, we'll be happy to help.

Frederick Grover: Thank you.

Female: Thank you.

Female: Thank you.

Reva Winkler: Take care.

Female: Bye-bye.

Male: Bye.

Female: Bye.

END