NATIONAL QUALLITY FORUM

Moderator: Sheila Crawford
March 21, 2014
10:00 a.m. ET

Sheila Crawford: The Health and Well-being Q&A session for our measure evaluation process. Also on the line are my colleagues Adeela Kahn, Ashley Morsell, and also Kaitlynn Ector-Robinson and also Robyn Nishimi. And before we get started, we wanted to do a roll call of the committee. So I guess we can just go down in line. I'm not really sure who is registered. We knew a lot of people were available for today's call, but – please hold. I'm sorry?

OK. Is (John Abbott) on the call?

(Tom): Hello?

Sheila Crawford: Hello. Can you hear us?

(Tom): Yes.

Sheila Crawford: Hi, (John).

(Tom): No. This is (Tom).

Sheila Crawford: (Tom). (Tom), how are you?

(Tom): Good.

Sheila Crawford: Sorry about that. And I know, (Sarah), you're on the call as well.

(Sarah): I am.

Sheila Crawford: Hi (Sarah).

(Sarah): Good morning.

Sheila Crawford: Is (Michael Bayer) on the line? Is (Ron) on the line? OK. (Emilio)? (Jane Chang)? Catherine Hill? (David Quall).

(David Quall): Yes. I I'm here.

Sheila Crawford: Hi, (David).

(David Quall): Hi.

Sheila Crawford: (Margaret Lock)?

(Margaret Lock): I'm here.

Sheila Crawford: Hi (Margaret).

(Margaret Lock): Hi.

Sheila Crawford: (Patricia McCain)? (Amy Minute)? (Jacqueline Malin)? OK. (Amir)? (Marcel Solis)? OK. Jason Spangler?

Jason Spangler: I'm here.

Sheila Crawford: Hi, Jason.

Jason Spangler: Hello.

Sheila Crawford: (Mike Sotto)? OK. (Robert Valdez)? And (Arjun)? I'm not sure if you're on the line. Renee Frazier, are you on the line by any chance? (Ronald Ing)? OK. And (Chisa)? OK. I'm not sure if I did not mention somebody's name who's on the line and my apologies, but if you can just let us know if your presence?

(Katie Sellers): This is (Katie Sellers). I'm here.

Sheila Crawford: Hi (Katie).

(Katie Sellers): Hi.

(Caroline Gillman): And (Caroline Gillman) is here.

Sheila Crawford: Hello.

(Caroline Gillman): Hi.

Sheila Crawford: OK. And I'll turn it over to Adeela.

Adeela Kahn: Hi everyone. Thank you again for all of those who called on Monday. Again, that was very short notice for us to kind of try and get everything together. But from what we understand, it went well and I just wanted to thank you all again for those of you who are calling back.

So, today, what we wanted to do was just provide a brief SharePoint overview just to get you familiar with where the documents are located on our site. And then with that, we'll open up the floor to any questions. If there aren't any questions, then we thought as a project team we'd like to take some time to look at one of our example documents which is the "What Good Looks Like" and really just kind of use that document to map back to the NQF as your criteria. And again, if there are any questions then we'll go ahead and answer of those.

So, I'll go ahead and get started. We're going to be using the Webinar today. So if you look at your Webinar screens, there is a link that says Committee Homepage. That's the link to our committee SharePoint site. So, if you all could just open up the SharePoint site, you can see here on the screen share that I have up that we have our main page up.

So at the top, you'll see we have some general documents, the committee guidebook. We have our measure evaluation criteria guidance which was updated just this past year. We have the measure information which is "What Good Looks Like" which is the document we'll be going over today. We also have the entry of criteria algorithms which we've actually pulled out from the steering committee guidebook to just make it a little bit more acceptable. We

have a phrasebook here which has a lot of the common vernacular that we use at NQF. It can be kind of TBS to learn all of it. So we've compiled it altogether into a really handy phrasebook for anyone who is interested. And then we also have our workgroup assignments and the document that's on the web is the most up-to-date. I know we've had several people move around in groups to their availability. So if you do notice that there is a mistake, please just e-mail me and we'll get that fixed right away.

Underneath the general documents, you can see the measure documents. These are the measures that we've gotten in the project and they're separated by a subtopic. So we have our community level indicators of health and (inaudible) and basically if you just click on the title of the measure, it'll (inaudible) and that folder contains the measure information form which is this document here that got the (MIF) written at the end of it.

We also have the evidence form and the measure testing form and these three forms are really the main forms that you're going to be using for your evaluation. Anything else is supplemental that the developer would like you to look at, but what we really, really want you to focus on is what's written in those three documents.

We also have – some of the developers have also provided code tables and other like risk variables and their – in an Excel sheet for you to take a look at also.

So, going back to the main homepage, underneath the measure – all the measures are listed here underneath the measure documents. Underneath that, you can see that we have our meeting and call documents. And here, this is a great resource, what we're going to be doing is we have all the agendas loaded onto the SharePoint page that are ready for your viewing. So if you ever can't find the dial in information or anything like that, you can actually just click on the call and it should be loaded in there.

Once we have – we have recordings and transcripts of every call. And so we have that ready, we'll post those and post the transcript as well.

Does anyone have any question?

Male: No.

Adeela Kahn: OK. So …

Male: Yes.

Adeela Kahn: … some of you have your preliminary surveys too soon. So I just want to show you where you can do – where you can turn those in.

So, on the left hand side, you can see there's a link that says "Committee Preliminary Evaluation Survey". So you're just going to click that and you'll click the Respond to Survey button up here. Once you do that, the form will come up and what you'll do is you'll just pick the measure that you've been assigned to. So, each of you either have one or two measures. You'll pick the measure that you've been assigned to and then you'll just answer and there are three text boxes for you to provide any kind of feedback about the measure and we want to try and relate any feedback back to the criteria.

So, to say the, you know, if you feel that the evidence is weak to support the measure focus then you would note that in the one – the box that says 1a, evidence that support the measure focus.

So that's all I have for the SharePoint overview. If there are no questions then we can go ahead and go through the What Good Looks Like document.

Male: OK. I'm good.

Adeela Kahn: All right. So let me just pull those up for everyone.

So, this is just a guidance document that we developed here at NQF to try and really help our committee members understand how we want them to evaluate the measures using our criterion and our algorithms.

So, we came up this document to really – it's a good developer resource as well. We give it to the developers to try and just to make sure that the bar is kind of equal across the board. So, this is just an example of a process

measure here. These are illustrative examples and hopefully that any of these are actual real submissions. Some of them have been modified just an FYI.

So, the measure title for this measure is women with urinary incontinence who will receive pelvic floor muscle training. So the way to use this document is there are some instructions here but there's also some really important key points about how you're supposed to be evaluating this.

So, for example, we have here, so this is a measure of a process and the process is pelvic floor muscle training urinary incontinence in women. So the evidence should go back to this and the measure should be tested with those kind of measure focus in mind.

If you scroll down, there are some key points listed here. So, you want to see the NQF guidance for rating quality, quantity and consistency of the body of evidence and report from the evidence taskforce which is available on our website and it's linked here.

A systematic review of the evidence is the scientific investigation that focuses on specific question and use of explicit pre-specified scientific methods to identify, to select, asses and summarize the findings similar but separate studies and this can include a quantitative synthesis depending on the data available.

And the body of evidence includes all the evidence for a topic which is systematically identified and based on pre-established criteria for relevance and quality of evidence.

An expert opinion is not considered empirical evidence but evidence is not limited to randomized control trial.

And there's variability and evidence reviews, grading systems and presentations of the findings, however, the information should be reported as requested in this form so that the steering committee can evaluate it according to the NQF criteria.

So when you're looking at – when it says to briefly state or diagram the linkages between structure process, intermediate outcome, and health outcome. And finally, include all the steps between the measure focus and the health outcomes.

So one key point here that we have is just to indicate a causal pathway and to not discuss the evidence and the item but just to present the linkage.

So we have here the developer of this measure provided a nice diagram of pelvic floor muscle training release to increase strength, endurance, coordination and muscle activity and urine leakage.

Pelvic floor muscle trainings maybe prescribed to increase strength, endurance and coordination on muscle activity or timing to suppress the urge. So they've indicated what the causal pathway is.

So when you're looking at the evidence, this is where the algorithm comes in. So the developer of this measure says that the source of this system review for the body of evidence that supports this measure is the clinical practice guideline which they've checked off, which means that they have to complete sections 1a.4 and 1a.7 and they've also checked off another systematic review in grading of the body of evidence which in that case they'd have to complete 1a.6 and 1a.7.

So, right below here, they've indicated the clinical practice guideline and I won't read that off for you but we really don't want anyone to summarize or paraphrase or shorten the guideline, we'd like to have it verbatim just to make sure that the guideline is again going back to the measure focus and not something more distal to the measure focus.

Specifically going down to the systematic review, this is kind of where the algorithm really go – comes into place. So, if you'd look at the algorithm which I'll pull up right now, the first question for looking at the evidence. So I'm trying to make this a little bit bigger so it's a little bit easier to read. So does the measure assess a performance on health outcomes or patient reported outcomes?

So the answer to that question would be a no because we – as we noted before, this is a process measure that we're looking at. So we would move down to the algorithm into – from box one to box three.

So for measures that assess performance on an intermediate clinical outcomes process or structure is that based on systematic review and grading of the body of empirical evidence where the specific focus of the evidence matched, what is being measured.

So, in this case, the developer did provide a systematic review. So we would again go to box four because the answer to that question is yes.

So, is the summary of the quantity, quality and consistency of the body of evidence from a systematic review provided in the submission form?

A systematic review is a scientific investigation that focuses on a specific question and uses explicit pre-specified scientific methods to identify, select, assess and summarize the findings of similar but separate studies and it can include a quantitative synthesis depending on the availability of data.

So let's go back to the look of What Good Looks Like and see if that's what they did.

So you can see here, they provided the citation and so they did a Cochrane Systematic Review. This systematic review identify the quality of evidence based on the risk of bias. The system for determining risk of bias was explained in the Cochrane handbook.

Let's go down.

So, the findings from this systematic review is we're looking at 1a.7. Again, here are some key points listed for you to keep in mind. But again, 1a.7.1, the information in the following questions in this section is based on the Cochrane Systematic Review. This systematic review addressed pelvic floor muscle training for women with urinary incontinence in comparison to no treatment, placebo, or sham treatment, or other inactive control treatment.

So the grade that was assigned for the quality of the quoted evidence with the definition of the grade is what we're asked for next.

So again, it should include both the grade and the definition of the grade. Again, not all grades are on a number or letter scale. And the grades for quality of the evidence are different from grades for the recommendation although the two are related. So we want to make sure that they have the appropriate grades.

So for this measure, they said that the overall grade of the methodological quality was not assigned in the systematic review, individual study quality was graded on the scale of for risk of bias. And then provide some explanation below. And below is the definitions and the other grades.

And the time period was 1999 to 2008, the quantity and quality of the body of evidence was 14 randomized control trials and the overall quality of the evidence across all the studies and the body of evidence and we're looking at 1a.7.6.

So, I won't read that off but overall this developer did a pretty good job filling out the – filling out the measure evidence form. So, if we go back to the algorithm, the answer to this question would be a yes. They did provide the quality, quantity and consistency.

So, this is where kind of your judgment where we're looking for you to kind of give us some more insights. So, does the systematic review include, when we're looking at quality of moderate or high, the quantity is that – sorry, is the quantity moderate or high, is the quality high, is the consistency high?

A high certainty that the net benefits are substantial, does it say that the high quality of evidence that benefits clearly outweigh undesirable effects and if measuring inappropriate care, we would like to see moderate or high certainty of no net benefit or harm outweighing the benefit.

If that happens and you would raise the measure as high, if the systematic review has quantity of low between anywhere of low and high, the quality is moderate, the consistency is moderate and high then – and – or there's

moderate certainty that the net benefit is substantial or moderate to high certainty that the net benefit is moderate then we would rate that as moderate.

And again, similarly the information for how you would rate it low would be if the consistency is low or controversial moderate or high certainty that the net benefit is small or that the benefit or harm outweighs and or that there's no net benefit, or if there's low quality of evidence. In that case you will rate it as low.

So, that's what you would be doing for all of the measures. We would be looking kind of at those boxes we've outlined for this algorithm.

So if we were to do an outcome measure. For an outcome measure, you again would start at box one and asking the question doesn't measure asses of performance on a health outcome, you would say yes.

We actually don't require an evidence-based for outcome measures because it's difficult to gather evidence on outcome measures. So, you'd basically be asking yourself this question on – in box two, does the steering committee agree that the relationship between the measured health outcome and at least one health care action whether that's the structure process intervention or service is identified, stated or diagramed and supported by the rationale, in which case it would just be a pass or no pass. So, the outcome measures are a little bit easier to evaluate in terms of evidence.

Are there any questions?

OK.

Male: On that example that you showed, the one I'm long enough to figure out how we might – how I might grade it anyway when you use it. I sense that which what you showed and what I was able to read I might – for example when you said moderate or the middle at such …

Adeela Kahn: Yes. OK. So that's good. It looks like – I mean, I think we at the table are wrong and kind of nodding our heads with agreement …

Female: Yes.

Adeela Kahn: … that we would probably do the same thing.

Female: Just a question …

Male: But then I guess I'm understand – I guess I'm understanding what you want to do then.

Adeela Kahn: Yes.

Female: No. I'm really good actually. Yes.

(Margaret Lock): Is our screen supposed to be changing or is it – my screen has frozen on what the cover page of What Good Looks Like.

Adeela Kahn: OK. Why don't you try refreshing your browser?

(Robert Valdez): I'm sorry. I should have introduced myself. This is (Robert Valdez) speaking.

Adeela Kahn: Great. Thank you.

(Margaret Lock): Refreshing did work. This is (Margaret Lock).

Adeela Kahn: OK. Great. So we can go on to looking at how we would evaluate both the testing portion of the measure submission form now.

So again, if we – we have an example of What Good Looks Like in terms of testing.

So we have some examples here. When you're looking as the testing form, again, this is basically an – and what we're showing is an annotated version of the testing form. It's just got extra key points for the committee to kind of understand.

So for – the developer has to check off what type of measure it is and we just want to make sure that they only check off one type. Sometimes we do get some testing forms that check off both outcome in process and really you can't

have both, you can only pick one. So, just to make sure that that's correct on the measure that you're going to be looking at.

So, scrolling down, so when you look at question one, the data sample that's going to be used for testing this measure. Some of the key points we'd like you to keep in mind is that the data types and level of analysis checked below should be consistent with what's located in the measures spec form? And that's the measure information form.

The sample used for the testing should be representative of the entities whose performance will be measured and patient serves. So, ideally, we'd like, you know, something broad, a cross cutting. Sometimes we get measures where, you know, there's only four test sites and sometimes the committee members are kind of, you know, weary of that like how is that representative if there's only four test sites or if it's only located in one geographic region that somehow comes up especially when you have something like, you know, infectious disease or something that we know is dependent based on geography.

And then the sample size again should be positioned for the statistical test that we're asking for. So for this example, the developer checked off that they're going to be using a clinical database and registry data and that's how the measure is specified and that's the measured – that's the data that is used to the measure as well.

So, what they provided to us is – what data set that you used was the Medicare home health outcome and assessment information set. So the OASIS data set specifically from March to June of 2009, the measure is specified to measure the performance of a hospital facility or agency and it's also tested at the hospital facility or agency level.

So, that's also an important point to know that they have to consistent. The measure is expected for our facility levels and it can only be tested at the facility level. If it's, you know, sometimes we do get examples where they've checked off that it's, you know, a health plan and they only provide testing at the – if it's – you can have a measure that is specified for a health plan or a

facility but that means that the measure has to be tested at both of those levels also. Sometimes what will happen is that you've checked all hospital and health plan but you only provided testing at the hospital level. So you just have to make sure that those two are consistent.

So, here we ask the developers to provide some description about the data set that they used. So – or the measure entities that were measured for their testing. So, in this example, they had 20 home health agencies representing various types of locations and sizes, the agencies are recruited to a national state associations and they were selected based on the capacity to do the reliability study with at least 20 to 40 patients and representing various ownership location and size.

So, they had four private, four profits, six private not for profit teams. Six private non profit, one health department, five hospital based, and two visiting nurses association. They are located in two states Arizona, Montana, New York and Texas and then they have the size of those the abilities as well. That's it.

So, how many and which patients were included in the testing and analysis? So, they had 20 to 40 patients for agency for a total of 500 and then they have below some characteristics listed out. And this is – again, this is what we're looking for. A lot of times, this is not what we get. So you just want to make sure that what you have is the right information and you have enough information to determine what was – what they use for their testing.

So going on to reliability testing, this is kind of where the algorithm comes in. We have an algorithm for reliability. We just switch back to the algorithm.

So, here we have algorithm number two, guidance on evaluating scientific accessibility of measure property, guidance for evaluating reliability and this includes eMeasures. We actually don't have any eMeasures in this project so we don't need to worry about that.

But here, the first question is, are the submitted specifications precise unambiguous and complete so that they can be consistently implemented?

And we're looking for definitions, value and codes with descriptions, logic, and again for eMeasure if it's in an HQMF or QDM.

So, going back to the What Good Looks Like. Actually, that question would be answered using the measure information form. You're going to make sure that all of the nothings unambiguous there that all of those data are readily available and there's nothing ambiguous there. So the answer to that question would usually be yes.

And so, the second question at box two was what the empirical reliability testing conducted using statistical test with the measure as specified? So going back to the What Good Looks Like document.

Again, we have those key points listed here for you to take a look at, just kind of as a reminder of what to keep in mind when you're evaluating the measures.

So, we have here at 2a2.1, the level of reliability testing was conducted using critical data elements used in the measure and you can do that using inter-abstractor reliability, data of element reliability. And then we have the performance measure score. So, we're looking for like a signal-to-noise analysis.

So, example one, listed below is an example of data element reliability and I'll leave that up there actually for you all to read.

So, for example one, they used inter-rater reliability was assessed for the critical data elements used in this measure to determine the amount of agreement between two different nurses of the same patient.

So, patients were randomly selected from the (pace) plan visits for start or resumption of care and discharge assessments for each day of the study.

The first nurse assessed the patient on usual visit. The second nurse visited the patient and conducted the same assessment within 24 hours of the first assessment so as not to confound differences in assessment with real changes in the patient.

So, the data analysis includes a percent agreement and a Kappa statistic to adjust for a chance agreement for categorical data.

So, this is exactly what we're looking for when we say describe the method of reliability testing and what tests were described. You should be able to understand in plain language what the developer did and how they did it.

Similarly, we have – they've described them performance measure score, testing that they've done, they used a data binomial model, and they used the methods outlined by the technical report of reliability of provider profiling, the RAND tutorial which we often sight here at NQF as one of our favorite documents to give to people.

So, if we go to the next section, so we're asking here for each level checked above what was the statistical results from the reliability testing?

So here, you can see that there was – that they provided the Kappa statistics here for the percent agreement for – between the two nurses.

And for example two, the – all this little – or the kind of goal here, for example two, they provided the signal-to-noise results here. So, you got the reliability statistic as well as the competence interval.

So, going back to the algorithm. So was empirical reliability testing conducted using statistical test with the measure specified? So, we would answer yes.

Was the reliability testing conducted with a computer performance score for each measure entities? The answer to that question would also be yes.

Was the method described inappropriate for assessing the proportion of variability due to real differences among the measured entities? Again, here is where we're asking like do they do a signal-to-noise, did they do a random with half correlation or any other accepted method? So, the answer to this question would also be yes.

So, based on the reliability statistic, again, this is where kind of your judgment comes in. Was there high certainty in that case where you would rate the measure as high or is there moderate certainty? In that case, you would rate this as moderate. If there's low certainty, then you would go to the box seven right here and where it's asking you, was there other reliability testing reported, if not, then you would rate it low. If yes, then you would go here to box number eight. So, in this measure, they actually had – they had a good Kappa – they had good Kappa scores.

So, wait, I'm sorry. So, we're looking at – for the algorithm, we're looking at this performance measure scores. So, I would say the reliability is pretty good. We have a range from 0.98 to 0.99.

So, in this example actually, they provided – have they not had such good reliability, we would have gone back to the algorithm and they did provide – in box eight, it's asking, was reliability testing conducted the patient level data elements that are used to construct the measure performance? In which case, we're asking here if the – I know, Robyn, if you want to kind of explain what we mean when we say data element reliability? I'm a little bit fuzzy on that.

Male: Before you ask that question, I guess I'm struggling with the other ones that I have that checked off that they did reliability or their reliability testing was conducted, and yet, in the next section 2a2.2, they refer this to the next section on validity testing to answer their questions about reliability testing which makes me think that they didn't do any reliability testing here.

The section is really short. They just go right to validity testing. So, all of this, you know, from one through eight, I mean I – I guess I'm just saying no on all of these. There's – or do I have to try to decipher from their validity testing if they even touch on – I don't know if they totally misunderstood the difference between reliability testing, validity testing or …

Female: Could you …

Male: … what, I still have to read …

Female: I'm sorry.

Male: … the details.

Female: Could you please show which measure you're referring to so we can follow them off.

Male: Yes. It's prevention dental ceilings for six to nine-year old children at elevated Caries Risk.

Robyn Nishimi: OK. And I'm sorry. I miss – this is Robyn, I was talking on mute. So, reliability – if they have done validity testing at the data element level, so that they have looked at the data elements against what NQF consider is the goal of standard which would be the actual medical records or the claims form, then they don't need to do the reliability testing that you're referring to and I suspect that maybe what has happened …

Male: OK. OK. Thank you. That …

Robyn Nishimi: … at the measure

Male: That would explain it then.

Robyn Nishimi: Yes. So, we try to …

Male: So, then, I still go through no, no, no, no, no through all of this as I'm, you know, reliability testing was not done but it wasn't done because they've done the validity test as you've described it, is that right? So, it's not necessary and not against the measure saying no to all of these things.

Robyn Nishimi: I don't – I think we consider it to be the reliability testing. But I haven't followed the algorithm. Let us get – let Adeela get back to you.

Male: OK.

Robyn Nishimi: I haven't gone through the form to see where that would take you.

Male: OK.

Robyn Nishimi: In terms of reliability on the data elements, what you can do is have two – as the box describes, look at the different data elements, go in, have two different individuals, abstract the data, see if they get the same results where there's agreement, where there is disagreement that would be in a rate of reliability. That would be a test of reliability at the data elements level.

(Alyssa): Yes. And this is (Alyssa). One of the things that, you know, you bring up a good points about where the algorithm will take you based on what is seen in the submission form. But one of the things I wanted to remind the committee is that the developers will be on hand as well. These algorithms are there to guide your decision and it's a binary decision that you finally make whether or not you feel that these measures are reliable or valid. It will be a yes or no vote and there will be discussion around this.

So, while the algorithm is a new guidance document we've put together, it may not bring out all of the new ones and all of the considerations that you may need to, you know, have to make that informed vote. But that's why we have you there to discuss and you're able to ask the developer for any clarification to ultimately inform your final yes or no on those criterions.

Male: OK. Thanks.

Adeela Kahn: Are there any other questions? I'm actually just pulling off our testing taskforce report that we actually didn't post on the SharePoint page, but I think after today's call, we will post it. And it has a little bit more guidance about how we should be looking at the testing that I just want to kind of pull up the section that does talk about the data element of validity is equal to reliability portions. So, just bear with me for a second.

So, I'll actually – I don't want to waste our time here. So let's move on. But then I will send out an e-mail and just update everyone on some guidance where they can find more guidance.

Male: Thank you.

Adeela Kahn: Yes. So again, just going back to What Good Looks Like, they provided an interpretation of what the results mean which is something really important

and we really want the developers to provide an interpretation of what the results mean sometimes. They provide all these results, but we have – because we're a multi-stakeholder organization we have kind of all levels of expertise at the table. And so, we just want to kind of, again, level the bar across the, you know, level the playing field. So, this is why we've asked them to provide an interpretation. And this one, they provided an interpretation of the Kappa score. So, and they provided an interpretation of their reliability testing here. That makes sense?

Moving on to validity. Again, we've got some key points here for the committees to take a look at, which I won't read them off but they're very helpful. Just important thing to note is that we oftentimes have a lot of developers that give us face validity. And we do accept face validity, but face validity is the – with the demonstration of validity and a subject to challenge by group of experts. Face validity is only an option. It's systematically addressed, assessed, and it's assessed for the performance score resulting from the measure as specified. So performance measure distinguish quality, not just the measure concept is a good idea or the data elements are appropriate or feasible to collect. So again, we're really looking at testing of the measure focus.

So, for this example, the developer provided performance measure score validity by doing empirical validity testing which you'll see in example one, and a systematic assessment of face validity which you'll see in example two.

So for example one, the empirical validity testing of the measure score, validity testing of the hospital score on the process measure of timely reperfusion in AMI patients was conducted by correlation analysis to determine the association between the process and the outcome of 30-day mortality. The hypothesized relationship is that better scores on timely reperfusion should be associated with lower scores on risk-adjustment mortality.

Because different numbers of patients were eligible for the process measures at different hospitals, all analyses in which the hospital was the unit of analysis were weighted by the total number of patients from that hospital who

were included in the calculation of a process measures. Report the correlation coefficient in the percentage of hospital-specific variation in the risk standardized mortality rates explained as indicators of strength and its location. So again, they've described the method of validity testing and what it test.

And for the example two, the systematic assessment of face validity of the performance score. Again, just a simple description, face validity at the measure score as an indicator quality would systematically assess as follows. After the measure was fully specify the group of experts other than those who advice on the measure development were assembled to rate face validity.

There are 20 experts including 15 clinicians, two patients, and three purchasers. We provide – they provided a detailed measure to the experts and asked them rate their agreement with the following statement. The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality. So, then they provided their rating scale below.

Now, this is really important. We really want to make sure that when we do have face validity that all of those elements are there for you to evaluate. Again, they provided examples of the statistical validity. The descriptive statistics, they had an N of 709 hospitals and mean with 54.5 and then they provided some percentile breakdown. The correlation coefficient was negative 0.01. And the percentage of variation in mortality explained by the timely reperfusion was 3.2 percent.

Again, for the example two, the face validity, they provided the number of people who rated the measure. The mean rating was 4.75 out of 5. There's some agreement essentially that the measure will provide a good signal of quality.

Again, they – the developers should be providing some interpretation of the results that they have gotten some the validity testing.

So, for example one, the empirical testing of the measure score, the correlation with 30-day mortality is negative and significant, and it is

hypothesized – it's in the hypothesized direction as hospital performance, on the timely reperfusion goes up, mortality goes down. So again, they provided a good rational or interpretation of what their measures validity testing were.

Again, for 2b3, we asked them to do an exclusion analysis and then we've also asked them to provide some statistical results from the exclusions. So, I won't read over these, but, these are all available on the homepage for you. So, you can take a look at it.

And then we asked for their interpretation for the exclusions and then if there are – if these measures are risk adjusted, then we've asked them to provide some information on their risk adjustment methodology, you know, how many risk factors are in their models and then again, we've – we have some key points here for you to really think about as you're looking at the measures.

So, going back to the validity algorithm which is algorithm number three – let me just pull that up. Again, we're looking at box one or the measure specifications consistent with the evidence provided in the support of the measure?

So we would answer yes to that. We're all potential threats to validity that are relevant to the measure empirically addressed. We didn't go over in detail, but I guess that's the example that we did go over. We did do an exclusion analysis. Once the empirical validity testing conducted using the measure as specified, the answer to that was also yes.

Again, we would only answer no if they only did face – if they did face validity and they didn't focus on the measure focus. They only refer to clinical evidence. They only had descriptive statistics. They only describe the process for data management and cleaning and computer programming or the testing doesn't match the measure specification.

So, in this case, they did do everything we asked them to do and the answer is yes.

What validity test is conduct – what validity testing conducted with the computed performance measures score? The answer is yes.

Again, once the method described and appropriately, was it appropriate for assessing conceptual and theoretically found hypothesized relationship? Yes And again, this is kind of where your judgment comes in, where you would rate it high, moderate, or low.

So again, high certainly your confidence that the performance measure scores are valid indicators of quality. Would you rate it high, moderate or low?

Are there any questions?

OK. That's all I wanted to really go over today actually. We do have our groups beginning next week. And so, we would really appreciate it if you could get us your feedback.

Before the work group call, it would be better so that everyone kind of has it. But we – if you can't get it to us that's not a problem, we just would like it before the in-person meeting and I'll follow up with everyone to put more details about and to kind of turn that in. But if there are no questions, then in terms of next step that we have – we have our work group calls next week.

We'll be sending out travel information soon. We've been working with our travel department and so we're going to get that information out to you hopefully before the end of the month. And we're looking forward to meeting all of you in person at our in-person meeting on the next month actually, April 29th to the 30th.

(Katie Sellers): This is (Katie Sellers) and I do have a question if I could.

Adeela Kahn: Yes.

(Katie Sellers): When you say we should get you our feedback, you mean, in terms of filling out that survey, correct?

Adeela Kahn: Yes. Yes.

(Katie Sellers): And do we do that just for the measure we're assigned or do we do that for all the measures in our work group?

Adeela Kahn: We would love it if you could do it for all of them but we really would ask if you could just go for the ones that you're assigned to.

(Katie Sellers): OK. Well, thank you.

Adeela Kahn: Yes.

(Katie Sellers): OK. Thank you.

Adeela Kahn: Yes. Sorry, go ahead please.

(Margaret Lock): This is (Margaret Lock). I also have a question. So, for the work of one in theory, the survey is due today. Is that correct?

Adeela Kahn: Yes. Yes.

(Margaret Lock): And the work group call us next one Tuesday or Wednesday once, you know.

Adeela Kahn: Yes.

(Margaret Lock): OK.

(David Quall): And just to double check, the three documents that you said were most important, because I see there's two that are relatively similar although they have different dates on them in that MIF document. They're seem to be…

Adeela Kahn: For …

(David Quall): … (inaudible) of that.

Adeela Kahn: Which measure are you looking at?

(David Quall): Prevention dental ceilings over six to nine-year old children at elevated risk.

Adeela Kahn: OK. Let me …

(David Quall): There's a – there's one that has a number 2508 before it and then that has a 3/7 date and then one that has a 2/19 date that is exactly the same except for the 2508 in front of it.

Adeela Kahn: I think that …

(David Quall): Which is the one that I should be using?

Adeela Kahn: I would use the one that has the most – if you look at the modified date – I will look at the one with the most recent modified date. I'll actually go ahead and delete the old ones. That's just a duplicate.

(David Quall): OK. You probably want to do the same for the other one I have because I think that also has the same because I think unfortunately, I printed out the old ones. So now, I'm going to print out the more recent one.

Adeela Kahn: OK.

(David Quall): Let's see. So the other one I have, if I could find it …

Jason Spangler: I'm glad you pointed that out because I printed them out earlier and I haven't looked to see if there have been modified documents.

(David Quall): Yes. Double check that because I'm frustrated. I printed out the old one and start marking that one up and I'm afraid that there's a difference between the two. So, I'm going to have to start over.

So, the other one I have is care continuity dental carries and – I'm sorry, dental services. And again, it looks like there's one that dated 2/19 and one that's dated 3/7 with a 2518 before it. So again, always – so, err on the side of reviewing the one that has the most recent modified date and ignore the other one – I'm sorry, I'm so confused, but if it just seems that …

Adeela Kahn: No, no, that's totally on our end and I apologize for that. Actually, they shouldn't be too much different. If at all different, I think what happened was that we had – they just got added – yes. They just got added twice in different naming convention. So, yes, err on the side of going to the most recent one.

(David Quall): OK. Thank you. And so, the three – so that then ignoring that those – the three – are the three word documents and the appendix is just an FYI.

Adeela Kahn: Yes.

(David Quall): OK. Thank you.

Adeela Kahn: Yes. No problem.

(Margaret Lock): I have a question regarding measures that are – that were previously endorsed. Is there any indication of what has changed in the current submission versus the measure that was originally endorsed?

Adeela Kahn: So, there is a section on the measure information form that does ask, since your last endorsement, what has changed, you know…

Female: I can't think of the specific question number, but does that make any question that (adds) them? We did actually on our reviews of the measures and off the top of my head, I would say about half of them (inaudible) to the questions and I didn't. There were kind of good luck with the draw depending upon what workgroup you're in, you may have the luxury of having those. So I think you (inaudible) but there's definitely a question that you can post to the developers at the in-person meeting is they can just provide at least brief summary of what these changes were for the measures that were previously endorsed.

(Alyssa): And sorry. This is (Alyssa). Just to – this is just a point of kind of process for the work groups and for the in-person meeting. We do have in the steering committee guidebook, you know, just what to expect, what will be expected of developers in the steering committee and I just asked the committee to just look at that so that you can get familiar with that.

We have changed some of our meeting process. It's part of our consensus development process improvements. And so, we just encouraged the committee to look at those.

Adeela Kahn: OK. If there are no other questions, then, well, we can sign off. And if you do have a question or think of anything, just shoot as an e-mail and I'll actually look into that – the data element validity for you as well and I'll send that out to the committee also.

Jason Spangler: Thank you very much. Have a good day.

Female: Thank you.

Adeela Kahn: Thank you, you too. Bye.

Male: Thank you, everyone.

Female: Bye.

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

END