NATIONAL QUALITY FORUM

Moderator: Sheila Crawford
May 16, 2013
12:00 p.m. ET

Jessica Webber: So welcome to Workgroup 2 Call for the Behavioral Health Phase 2 project. This is the National Quality Forum. I'm Jessica Webber, I'm the project analyst. And I'm joined here with my colleagues Angela Franklin, Elisa Munthali and Lauralei Dorian.

Just please note that the call is being recorded and transcribed and we'll let you know when these are available.

So, the purpose of today's call is to allow workgroup members to evaluate the submitted measures based on the NQF criteria to determine if they're suitable to recommend for endorsement as voluntary to become substandard. There will not be anybody on this call. This is just a preliminary review in advance of the in-person Steering Committee meeting on July 7th – or sorry, June 5th and 6th.

Also, the measure developers are on the phone to answer any questions or provide any clarifications if you have questions. And we'll have a public comment period at the end of the call.

So with that, I think we're ready to begin our reviews and we're starting with measure 1884 Depression Response at Six Months-Progress Towards Remission. And we have two reviewers marked for this measure, Caroline Carney Doebbeling and Tami Mark.

So, if one of you would like to go ahead and just give us an overview of the measure and your thoughts in the reviews? That would be great.

Caroline Carney Doebbeling: This is Caroline and I'm still in the process of trying to pull up my comments and the measure from the SharePoint site. If Tami wants to start, that's great.

Tami Mark: Sure, I can start Caroline. So, measure 1884 Depression Response at Six Months-Progress Towards Remission is a measure in which adults age 18 or older with major depression or dysthymia and initial PHQ-9 score or greater than nine are examined at six months later to determine what percent had a reduction of 50 percent or greater in that PHQ-9 score. And it applies to both patients who are newly diagnosed and patients with existing depression were identified during the measurement period.

In terms of ratings on importance, three people rated it as important to measure, one rated it not as important and raised the issue of concerns about missing data knowing that almost 80 percent of data was missing at six months and this is a thing that comes up again in the comments.

For those that responded, only 9 percent respond to treatments based PHQ score 9, reduction of 50 percent or greater. I'm also raising issues about using only one screening measure and whether there is regression to the mean. Well, that's potentially an issue. In terms of scientific acceptability of the measure of properties, two voted yes, two voted no.

The comments noted are average score was only 10percent, has the reduction of 50 percent on the PHQ-9 score with a wide variance. Also, it was again noted that there was that high dropout rate, there is questions about the accuracy of the data as reflected in the audit. It was also noted that personality disorder was excluded and it wasn't sure – wasn't clear why access to personality disorders were excluded.

In terms of usability, one measured it high, two measured it moderate and one indeterminant. The note was that it seems to be based on EMR capabilities to submit data, but it's not clear if there – if a low non-response at six months maybe reflecting usability issues.

And in terms of feasibility, one rated it high, two rated it moderate, one rated it low. I think, again, reflecting the fact that it's been – the data, a lot of data seems to be missing particularly at six months.

And so, in terms of the overall assessments, two rated it yes and two rated it no.

So that's my overview. At this point, should we open it up for discussions? Or, Caroline, do you have some comments that you want to raise now?

Caroline Carney Doebbeling: Sure. I don't know if my – I think you did a great overview of the measure, and I have very mixed feelings about this measure as I was reviewing it and I am not certain based on your comments about the other reviewer and what we saw that everyone even interprets this measure the same way.

So that would lead me to have more concerns about the measure going forward. So I think it is fine to open it for comments. I don't have anything else (inaudible) at summary.

Sheila Crawford: So, are there other comments, the rest of the workgroup members? Particularly, we're just starting with importance and discussing with what involves there.

Female: I guess I …

Caroline Carney Doebbeling: This is Caroline. I can time in. I was one of the people who voted that it isn't an important measure. The reason for that is the fact that there are too many people who get lost to follow-up with depression treatment. And so, the fact that this measure is trying to address in a standardized fashion, that issue, I think is a very important issue.

The reason why I have a very mixed response to this is I'm not sure that functionally this is the right way to do that. And the testing of the measure to date and the acceptance of the measure to date, and the fact that it would be very, very difficult to gather the data and less EMRs were in place everywhere

makes me very concerned about the measure itself. What the measure is trying to get to I think is a very important concept.

Dolores Kelleher: No, this is (Jodie Kelleher). I think it's very important to measure and especially at this clinician level and with a follow-up based out, because again there is, I agree – I think it was Caroline. People are not followed in a way that they should be followed and one way to address that is to measure it.

And I'm hoping, I'm maybe jumping ahead here but I'm – in a lot of these measures where there's difficulty in terms of gathering the data, I'm sort of counting on the 2014 increase in measurement tools and documentation of depression screening in the EMR as part of meaningful use, et cetera.

So, just in some ways, it's a near term belief that the measures will be more feasible in terms of gathering data.

Sheila Crawford: OK, any other comments about importance to measure reports?

Tami Mark: I would just as this Tami Mark. I thought it's very important to start to move the field for measuring outcomes and reporting that to consumers and using that as a basis by which consumers can potentially select providers. So I think it's an important effort. My concerns are more with the implementation than with the importance of doing this.

(Anita): Hello, this is (Anita). I would agree with everyone that for me, the importance is definitely there. We have a few trouble in our systems for follow-up but the problem I'm having with this is if they already tried this in Minnesota and had data collected for four years and only nine percent response rate because 50 percent didn't even or whatever, 80 percent I don't think submitted data. I did inquire and they said at this point even though we have (inaudible) and we have an EMR, they would not be able to submit PHQ-9 data.

So, that's the concern of are we putting a measure up that we're going to get really poor results because either we can't have the providers submit the data, and is there something more intermediate steps that can be taken to try to assess for follow-up?

Caroline Carney Doebbeling: I might put that in my comments section I think or at least I meant to if not for this companion measure to this. This is Caroline again. Which was maybe a first step is to look at the uptake of the measure rather than actually picking it all the way to the outcome of a 50 percent reduction in the depression score or at least doing more work on the feasibility side before it would go live for measurement.

I have to address that comment made earlier about being able to compare practitioners. If a measure like this is put into use across all types of providers, there will be very unfair comparisons between providers. The types of providers and the populations, for instance, seen in community mental health centers as opposed to private psychiatry as opposed to primary care, will be very variable because the patient base is extremely variable and so I have significant concern about using a measure like this to rate practitioners against one another.

Tami Mark: This is Tami Mark, just a follow-up on that point which I think is an important point. My understanding is that the measure is being used. I mean I went on the website and they're actually – our ratings for all the clinics in Minnesota based on this measure, they're all very low. And you mentioned that they do risk some kind of stratification to address that issue of difference and severity of illness and they talked about that. I guess we can address whether we think that's adequate or not.

The other comment I wanted to make is I wonder whether it would be useful to report at this point the percent of patients who they could reach at six months and then among those who they could reach the reduction in the PHQ-9 scores. So you're, perhaps, encouraging clinics to do follow-ups but you're not (contaminating) that PHQ-9 score remission measure with the fact that they can only follow-up about 20 percent of patients.

Sheila Crawford: Is that a question that you'd like to ask the developer or director (inaudible)?

Tami Mark: Yes. I think, I might want to – unless other people have comments, maybe we could ask the developer to respond to some of these points.

Sheila Crawford: I'd like to call the workgroup. Is there an additional comment before we go to the developer?

Female: One question, who was the developer for this one?

Sheila Crawford: Minnesota Community, sorry, Minnesota Community.

Female: OK, so it's not a CMS?

Sheila Crawford: No, it is not. It's Minnesota Community Measurement.

And we do have the developer on the call.

Sheila Crawford: (Colette), are you available to respond to these questions?

(Colette): Yes, I am and I will try to address several of the issues that have been brought forward.

We have a set of depression measures that we are working on and fully understand the lost to follow-up before these measures were commissioned and started. We understood that patients were being lost to follow-up in huge numbers. Well, part of the reason of having this type of measure is to encourage that follow-up and reaching out to those patients of any population that I can think of that needs assistance and help in that reaching out and that's one of the things that this measure does support.

We do, let's see – I want to talk about the issue between the behavioral sites on the primary care sites. We have about 500 clinics that have been reporting this data to us for several years. It's – lots of an issue about them being able to report the data. We have many (ethics) client who have the PHQ-9 built into the system so the information is coming out of the system electronically. Also, the companion measures for this measure, the remission at six and 12 months, and the utilization of PHQ-9, those are specified as eMeasures and two of them are in meaningful use right now.

The second thing I wanted to talk about was the difference between behavioral health and primary care, and again a valid point. And we have some patient population characteristics that help with that differentiation between the

patient base. So in a behavioral health setting, those patients need to have major depression as their primary diagnosis. So the intent of that is to not include patients with more severe psychiatric conditions. So, major depression does need to be primary in the primary physician. In the primary care setting, depression is the – the diagnosis occurs in any physician for the visits. Again …

Sheila Crawford: Can you just clarify that, I thought you're saying in the primary care setting.

Female: Correct. And so, the denominator criteria states for a primary care setting you're searching for major depression or dysthymia and any diagnosis code position. So as you're coming in with a sprained ankle and the doctor's assessing you for depression, you are going to be pulled into the population. But in the behavioral health setting, major depression needs to be the primary diagnosis, the reason that you're being seen in that setting.

Female: May I ask some follow-up questions about that, please (inaudible). As EMRs and coders pull in code for billing purposes, often, old codes that aren't primary or even secondary reasons for the visit get pulled into that?

And so, in the primary care setting, if the depression can be in any place in that list of typically 5 to 10 codes that get pulled in. I'm not sure about, that makes me question even further.

Caroline Carney Doebbeling: I'd be happy to address that question.

Female: Great.

Caroline Carney Doebbeling: Since this is actually – if you think about it, it's a measure that's specified at the encounter level, but we are looking for active diagnosis at that encounter and you also need to have an elevated PHQ-9 to start. So, it's like a prospective measure when the patient is given that diagnosis so they have an active diagnosis for that encounter and their PHQ-9 is elevated, that starts the clock ticking for that patient.

And then, data-wise it's going to look six months out plus or minus 30 days to look for an additional PHQ-9 score. How we do this in Minnesota, if groups

are pretty much dumping their PHQ-9 forward, often we have portal programming that's determining that timing. And in our work with CMS and the contractors for meaningful use, we've successfully worked through how that would work in an EMR system for calculating the numerator and denominator as well.

Female: So, the index is the date that the PHQ-9 was measured?

Caroline Carney Doebbeling: Right. And we don't have the reverse, we don't simply just say, "Oh, any elevated PHQ-9 comes into the denominator", we really need that confirming diagnosis frequently as we're out auditing clinics. We see patients that have a couple of PHQ-9s that are high, but they do – they are not yet formally diagnosed as having major depression. So we do have those two things happening in tandem before they come into the denominator.

All right, I'm sorry I can't remember all the questions that everyone was having, but I would be happy to keep addressing any issues that you do have.

Dolores Kelleher: Well, this is – this is (Jodie Kelleher). I was just wondering because you didn't specify it, why you have eliminated anyone who also has access to personality disorder?

Caroline Carney Doebbeling: That's a great question. I think that was early on in the development and let me try to explain. Initially, as we were working through designing these measures, we thought that it was simply enough to be searching for ICD-9 code for major depression.

And kind of early on and this is years ago, ran into the situation especially with bipolar conditions maybe less so with personality. But the measure development workgroup decided that both two things needed to come out of the population. And I can speak more to bipolar than personality disorder despite best clinical practices of say, you're treating the patient with major depression or you believe they have major depression and perhaps spent later in their care, a month or two later they have their first manic episode and you realized you're not dealing with major depression.

Clinically, best practice is then to switch your coding practice to start coding those visits for bipolar, but what we found in actual practice especially in the behavioral setting is the providers continue to code major depression and bipolar in the same visit and they keep flipping those around depending on if the patient is more depressed this visit or not. So, we have to add those exclusions up front.

Dolores Kelleher: And I understand that. I understand the bipolar and other competing Axis 1 that might sort of confound, but they are not – unfair on the sense that they are – why you, you know, this is (inaudible) me why I would exclude or include an Axis 2 …

Caroline Carney Doebbeling: OK.

Dolores Kelleher: In the presence of an Axis 1 major depressive disorder.

Caroline Carney Doebbeling: I can appreciate that, thank you.

Female: You kind of a little bit addressed my comment about why you included the lost in follow-up, but I guess I'd like to hear more explicitly about the thoughts on having that as a separate measure, what percent they were able to follow? And then of those what was the percent that showed a significant decline in the PHQ-9? And also is there any data on how consumers have – how useful consumers have found that information on the – it's been on your website for two or three years now, so they can use it to select providers. Are they using it? Are they finding it useful? Do they have any concerns about it?

Female: Oh, sure. I'll address both of those questions.

The first one in terms of lost to follow-up. Analytically, we have looked at that data frequently behind the scenes, so we're looking at the patients that are assessed at six and 12 months and there's still lots of opportunity for improvement. Our rationale for not just looking at only those patients that we connect with is because we believe that we would never be changing care. We would – patients would continually be lost to follow-up. So, we also have additional measures …

Female: But what if you reported the lost to follow-up on the website as a separate measure?

Female: Actually, that's our plan. So we've been (expecting) these follow-up PHQ-9 rates for all the groups and that's – we've seen some incremental success at that from low 20s to now about 27 percent. We have a plan for getting that on our website in the next year. So, we are tracking that follow-up rate. The groups have always internally had those companion measures.

We were seeking guidance from NQF if they want us to bring those forward as well and their process measures that they recommended that we did not.

Female: Was that in the prior …?

Female: Just in this, just in this submission as we were bringing – oh, I should back up and explain. So, we have some other measures that are currently endorsed that look at remission at six and 12 months and we consider that to be the gold standards. Your PHQ-9 of less than four, you're achieving great symptom control. The reason why we were bringing this measure forward was the request from our community, our providers, this is clinically a difficult measure and we want to make progress on this. And they felt that if we publicly reported the response rate measures as well, that it would be encouraging for providers and also for consumers because it's a little bit depressing to look at these low numbers.

Again, if you indicated a lot of it is due to lost to follow-up but there have been tons of efforts. I can't tell you how widespread the PHQ-9 use is in our state because of these measures. And we're branching out into other areas of screening.

Your second question, we did have some recent consumer focus feedback about these measures, should we also publicly report the response rate measures. And the consumers agreed that that would be a good thing to do and not adding more confusion to the picture on our website.

Female: So they thought it was good to report the response rates or not?

Female: They did. Yes, so those were added to our public reporting about a year ago.

And also the main purpose of this, it helps show that progress towards remissions. So, we're not the gold standard, we're viewing it as a complimentary intermediate outcome towards reaching remission.

Female: And did consumers find the existing measure useful and did they use it to make decisions?

Female: I know that they found that the information was useful, I'm not so sure about the decision part. We could – I can follow up with the person that conducted the focus.

Female: OK, thanks.

Female: Sure.

Oh, I just wanted to share to this (inaudible) again. I just wanted to share that I know the overall rates are low, a lot of our measures historically in our community have also started out very low. We would have wish to see this one progress more quickly than it has. But I do want to point out that there – if there is variance within the measure and there are clinics that are achieving higher rates than the statewide average.

Angela Franklin: Thank you. So this is (Angela) again. We – as we were having our question and answer session, we got into the scientific acceptability questions quite a bit. From the workgroup, are there additional questions about – and discussion about scientific acceptability because that's what we had that's been told?

Bonnie Zima: This is Bonnie Zima. I was probably the one that was most vocal in my concerns about missing data. And frankly, I haven't heard a strong enough argument for how the missing data problem would be addressed with this measure.

Sheila Crawford: Do you want to ask the developer to respond or is there additional discussion?

(Colette), can you answer it?

(Colette): Sure. I just wanted to point out that the patients who are not assessed are included in the denominator. So there is an assumption there that they are not in remission if they're not assessed. So, we have that full population and we're counting that in the measure. You know, it's a success rate where we would like to see it. No. Not yet. But again, this is a patient reported outcome measure and it is involving remaining connected with that patient six months after their initial PHQ-9 score.

Oh, and can I add another comment about something else?

In terms of measure burden or that this would be a simple change in the numerator statement that goes along with the companion measures that are already developed as eMeasures.

Angela Franklin: OK. Thank you. Are there questions from the workgroup, discussion by the work group members? OK. We're moving on the usability and that is there a discussion from the workgroup additional discussion about that? We could cover that, OK?

Feasibility? OK. OK. Hearing no further comments there must be some summary comments that anyone wants to make. We can move on the next measure which is related. And that's 1885 which should the depression response at 12 months progress tours remission. Again, these developers and the set of community measurement and I believe (Jodie) is the reviewer.

Dolores Kelleher: I am. And we'll leave discussion. And so, this is very similar to the last measure. It's on adult patients 18 years and older with major depression or dysthymia and an initial PHQ-9 score of greater than nine, demonstrate a response to treatment at 12 months defined as the PH-9 score reduced by 50 percent or greater for any initial score. And this applies to those patients newly diagnosed and those that existing depression identified during the defined measurement period.

This current PHQ-9 score indicates a need for treatment. So, on – that's the description on importance to measure and report. There were three yes and one no. Evidence, there was four yes. Health outcomes, there were only three

people that responded if the health outcome is the rationale supported and there was only three responses and they were all yes. Scientific acceptability of the measure properties, there was two yes and one no. Reliability was there was a two rated as moderate and two as low, validity, two as moderate, one as low. And then with usability there was one high, two moderate and one insufficient. Feasibility, one high, two moderate, one low. And the preliminary assessment of the workgroup was split again as in depression response at six months two, two.

The comments on this were similar or it's not almost identical to the comments made for the previous depression response are six months measure. Again, I had a question about Axis 2 which is not, you know, was answered in the earlier discussion. The other comment was again same as six months measure. There was a very high no response rate at 12 months whether due to no EMR feed or no follow up activity occurring at 12 months and it makes it very difficult to (inaudible) if depression response is seen or not versus no entry for 12 months. With only 18.6 percent counting scores submitted what have been implemented to change the low PHQ-9 score entry. So again, similar to the previous measure with the same (inaudible).

Female: Hopefully up for other comments from the workgroup?

Female: (Benita), this is more of a question for the Minnesotans. Forgive me if you did respond to this question earlier. And so the six months that – that was my question is, so with having to do this – done this for I think for three or four years and you saw that the 12 months response or score submission was 18.6 percent, what has been done to try to improve that and is that 18.6 now on aggregate and you've seen it's only improved from 10 to 12 to 14 or does it kind of remain there or what's the progress in that?

(Colette): This is (Colette). I'm just assuming you wanted me to respond to Minnesota?

Female: Yes.

(Colette): There has been some progress, some incremental progress in both terms of the outcome measures about one percent per year which, you know, is not fabulous but at least it's going up. And then and the ability to obtain those

follow up PHQ-9 scores starting in the low 20s and now at 27 percent. We've done a variety of different things. This is just FYI. We have a couple of pay-for-performance programs with the employers that are really interested in the measure and are doing everything they can to promote and reward positive outcomes. And let's see, oh also we have a depression toolkit that we have on our website that were providing for providers and recommendations there's been collaboratives.

We've also been working with XCD Institute for clinical systems improvement. This started as a project in 2008 with Diamond. We have 52 Diamond Clinics that's depression improvement across Minnesota offering a new direction. And those are actually placing case managers within the clinics to do some of this reaching out and follow up. And then those learning's have been shared with other clinics across the state. Many outside clinics have systems in place now that they're trying to reach out to patients and obtain that follow ups. It's acceptable to mail the PHQ-9 to the patient and to receive that information back. So we are seeing a lot of – of those kinds of things.

Female: OK. And second question on the follow up to that. So once we implement or endorse the quality measure and let's say some other per individual organization wants to use that. How was this helpful for you to identify which providers needed more guidance or actually improving depression scores versus just improving submission of scores?

(Colette): You said you had about a one percent increase in the actual measure. So, was that enough to differentiate which providers were doing better with actual depression treatments at 12 months and follow up or this measure have been used more to try to target where you're having downfalls in submitting the actual PHQ-9 and that's what it's being used for more.

Female: I would – guess I would say it would be the former. There has been reward and incentive for improvement in those rate silver times for not only hitting a target but are you better. And then if you look at the variation of the clinics and the website, some of the clinics are actually hitting close to 40 percent even of the 12 month mark. So there are clinics that are being more

successful than what's showing as statewide average. And please understand that statewide average is about 500 clinics. So, you have varying degrees of adoption implementation workflows in their practices. So the average can be a little bit deceiving.

Female: And that 40 percent was actual improvement, the 50 percent reduction in scores or as submissions.

Female: OK, thank you.

Female: Yes.

Female: I have in the data, you know, kind of where people were falling. It's little bit like before the 12 month measure that we're talking about right now, the range of course is zero but the high end is 39 percent. And then there is some stratification in terms of where clinics are at and the majority are, you know, falling between 5 and 15 percent. And again, I fully admit that is not where we want to see but we still believe that it's a valuable measure.

Sheila Crawford: Do you have the statistic just based on the patients that they were able to follow up?

(Colette): This is (Colette). I don't have those handy. I can start searching on the computer if you guys want to keep going. I'll look for that analysis, I don't have that at my fingertips. Would you like me to do that?

Sheila Crawford: If we want to do that we can continue our discussion and circle back or we can send it out after the call?

Female: Yes, I think that's fine if we send it out after the call so we could have it in time for the meeting.

Sheila Crawford: Yes.

(Colette): Happy to do that.

Sheila Crawford: All right.

Female: So additional discussion on any of the criteria for 1885. I believe we've covered important but scientific acceptability or another criteria?

Sheila Crawford: OK. We can cover a lot of the same issues.

OK, hearing none, I think we can move on to our next measure which is measure 105, Anti Depressant Medication Management. NCQA is the developer and of the need of the (pendula) is our lead discussant for that measure.

Female: All right. So this measure, measure 0105 is the percentage of 18 year old and older with the diagnosis of major depression and were nearly treated with antidepressant medication who remained on the treatment at two parts, the effective acute phase as the first three months and effective continuation phase is six months. Those measure has been in place since 2009 and it's up for renewal and it's by the NCQA.

We're going through the measures and first important to measure in report, we have three yeses and one no for that and there's a lot of detail in there but I think a lot has to do with the ICT 9 that's in there that's allowed and direct clearly defined major depressive or does that cause leakage and allow more different forms of depression I think is the question there. And the evidence, three yeses and one no. Health out – and (inaudible) outcomes, one yes and two non-applicable. Quantity of data of evidence, three highs, one low. Quality, three highs, one low. Inconsistency is three high and one low. Do you want to talk about that or do you want me to run through all of the points right now?

Female: You can run through all the points and then we'll come back to important.

Female: OK. So scientific acceptability was four yeses, no no. Usability was three highs and one low. Feasibility, three high, one medium, and preliminary assessment criteria suitability for endorsement was three yeses and one no.

Female: If we want to go back to the top now?

Female: Yes, let's go back to the top and the importance and would invite comments from the workgroup.

Female: Well, the developers might want to comment on the – someone has the issue about you saying 311 but it's my understanding that especially in the primary care arena but in general that is often used as a review default code when someone finds significant depression. Oh, I don't know how you could leave it out but I'm, you know, I guess it's open for a discussion.

Tami Mark: Yes, but this is Tami Mark. I thought maybe I could just clarify my comment because I made that comment about the 311 and my concern about the importance. My concern stems from the fact that I don't see lack of adherence and underuse of antidepressants as the public health problem which needs outcome measure such as this. The latest statistics show that at least 11 percent of Americans of 12 years of age and over take an antidepressant in any given year.

We also know that adherence of role is very high. More than 60 percent of Americans have taken antidepressants for more than two years or longer from an – citing this from a CDC report. So I think the data cited there that we have a problem with the antidepressant adherence is from a 2002 assignment article that's very old. So I think that in terms of the importance and improve, I do think that for people who are severely depressed that they're probably are people for whom adherence is an issue and has serious consequences but as the population public health measure I don't think it's important.

I also have concerns about using the measure for populations that aren't severely – don't have severe depression because if you look at the meta-analysis that they cite, the foreign year meta-analysis in (JAMA) in 2010, it actually says that the benefit of antidepressant medication may be minimal or non-existing on average for patients with mild and moderate symptoms. So we're encouraging providers to provide antidepressants and to keep patients adherent to antidepressants when there's no scientific justification for doing that in that population of – you know, I hear you that some of those patients may actually have severe depression but many of them or most of them may not. So, those were my concerns.

Female: I'm going to address the one quick concern about the low use of antidepressants in the CDC from 2002. I believe there's been just so much movement in trying to even get depression diagnosed in back into early 2000 and we didn't have that initiative as much. However, we do have a lot more patients on antidepressants than that low number these days that was here in Southeast Michigan and the second part, it is stated in the APA, the American Psychiatric Association, and looking at the Texas algorithm for depression, I believe it's in their queue of the importance of taking the medications not only for six months but really now the data is really (gleaning) towards even longer for 12 months. This measure is only going out for 3 to 6 and the problem we have is a huge drop off that occurs after the first three months and a lot of that is because people just don't feel like it's working because they're been put on a starting dose and they don't get a titration off. So that is the concern I have with this measure because you can kind of gain this system, it's an (entry) QA so it becomes a heat of measure.

So far a health plan that's trying to get top, you know, 90 up percentile, they want to make sure the scores look good. So if you get a one 90 day sale, your score is good, however, more than likely that patient – it didn't get a dose titration and they're going to stop taking their medication. So we actually see that as a bigger problem, not in here and still is an issue that we see here, I don't know that's unique for our area.

Female: But do you see that, I mean these are – is there good population that begin the logic data that lack of adherence is a problem. I mean what I'm seeing in this national survey is it says that 60 percent of people are taking it for two years or longer and 14 percent have taken it for 10 years or longer. So I'm not seeing a lot of, you know, national data suggesting that adherence is a big population public health issue.

Female: I have to look up and see actual data but I believe the data is much – the volume is much more than the 8 percent. I don't know if NCQA has that data already and they can make a remark but I can definitely research that and get that back to you.

Female: Do we have someone from NCQA on the line?

(Jerry): Yes, we have quite a few people here but (Mary Baron) will respond to that.

Female: Thanks (Jerry).

(Mary Baron): So I'm just looking actually to enhance CDC survey which says that 11 percent of Americans are taking anti-depressant medication, so to that question about the 8 percent. I think the, you know, the question of how to link up survey data and a patient's response to a survey where you do imagine that there's going to be some social desirability bias to say to the surveyor, oh yes I've been taking my medicine, you know, could potentially be a different window onto the problem than actual medical record.

I don't just – I could not argue at all with the point made by the work group member earlier about the fact that a fill of a prescription, you know, that doesn't necessarily mean that the patient is taking every dose and doesn't speak to the question of dose intensification when that's the appropriate thing to do. But I do think that we have with this measure a clarity about the – about the diagnosis because it does require major depression to be in the measure and then clarity about the – at least opportunity for medication treatment because right, if you haven't filled your prescriptions then you certainly aren't taking them.

So in a way it's a bit of a fuzzy filter on our camera but I think it gives us a better picture than we could actually get with the survey.

Female: Can you just clarify that comment about that you – is it always what would be the major depression because I thought you included 3-11 in the denominator which is not, major depression includes a whole bunch of other things.

Male: Right, we do include 3-11 in the denominator and we describe it as major depression and 3-11 has come up within previous re-evaluations of this measure and like what was said earlier by one of the work group members it is – that lack of a catch-all, although we, you know, we do make sure that there is a diagnosis requirement with this measure as well. Yes, the medication requirement.

(Mary Baron): And I think actually the number of the …

Caroline Carney Doebbeling: This is Caroline and I would like to chime in that I also have an issue with using the 3-11 because it's – I kept all for many, many things. There maybe a valid reason why people are not staying on medication for that period of time or may have erroneously have been started on medications when perhaps they weren't indicated especially in the cases of something like an (inaudible) disorder. So I am just speaking to support a stronger look at whether 3-11 should be included or not.

Female: Does NCQA know how many of the people that commence with this and those that are using the 3-11 because that could help us understand what large – is it a large proportion of the people that get into the denominator?

Male: When a field test was done in 2007, we saw that between 31 and 41 percent the diagnosis codes, we're using 3-11 and just to clarify for our measure, we require that a code come with an encounter and so if that code 3-11 is included within in-patient stay, you would be in the denominator but otherwise, for any outpatients, they're either that we require two visits and a code for you to go in the measure. So that's part of our effort to make sure that this is a reliable diagnosis of major depression associated with each of those anti-depressant medications.

Caroline Carney Doebbeling: And I will bring up the same point that I brought up in a discussion of a prior measure which is coding. There are issues with coding where codes from prior visits are pulled forward to a current visit for billing purposes only and that condition may have never been a direct or all during a visit. So that's possible that with the 3-11 or in any case, any of the outpatient encounters that two of them can occur. I will speak from very practical experience in doing – (inaudible) at our health plan which is with diabetes, often diabetes is erroneously code in the emergency room setting based on a high blood sugar for whatever reason, (DED) codes it and those stay in the measure or stay in the coding set and hang with that number over and over and over and so we have to manually remove that member from the denominator to say no there was never a proven diagnosis.

So basing it only on codes I do think that's problematic.

Female: Are there additional comments from the work group on this issue?

Male: Can NCQA – I'll just provide some additional information.

Female: Yes, thank you. Thank you.

Male: I think if you look at the – when you look at our performance results for this measure, I don't have it open but I'm just kind of going from memory, we'll report this in several different types of plans, Medicaid commercial and Medicare. And we have probably, you know, 7 or 800 plans reporting this measure. And if you look at the average denominator size for this, these are not large numbers when we compare them to other measure sets that we have. They range anywhere from I think in the mid-200s maybe up to about 900 and some of these plans as you will – can imagine are really quite large, you know, the (inaudible) as well points et cetera.

So, I do think that, you know, measurement and the population health states and in the health plan environment is a little bit distinct from – at the physician level and we understand there is some noise as (Mary) characterized, there's some fuzziness to our lens, but we feel that it's a telling story, the performance rates on this are, you know, not in the stratosphere and even the 90th percentile are not in the stratosphere, so I do think that it is telling a story that it's important for moving the agenda.

Female: Question, what – kind of forgetting, but what is the median rate that's not in the stratosphere but …

Female: The mean for the commercial in the acute phase was the latest in 2011 was 65.3.

Female: What did you say?

Female: 65.3 percent for at least, yes, for the effective acute phase for the commercial population in 2011.

Male: And continuation was around 50, so is that right (Mary)?

Male: Yes.

Female: Yes.

Male: So I mean, so you're saying declining performance for the very thing you're trying to capture? So I mean I think that – there's a certain logic for that.

Female: So that's 65 percent were adherent at three months, am I getting that right?

Female: Correct. That's how I interpreted it, yes the average or the mean 65.3 were adherent at 12 weeks.

Female: For 12 weeks and that's – I'm getting started – getting this fixed up with another measure. That's 80 percent – are you using the 80 percent tradition ratio?

Male: No, it is not a 80 percent MPR. We're looking at 84 days specifically and will allow certain amount of days for wash out or changes in your medication and for the acute rate that adds up to 30 days of an allowable gap in the medication.

Female: So basically 12 weeks so it would be 114 days and they look to see that you had at least 84 days.

Male: Correct, 5 percent which is around 75 percent.

Female: So it's like equivalent of the 75 percent position ratio if I'm thinking about the other measure that we reviewed on the last call. Is there any reason for selecting that number of days or?

Male: The exact number of days was decided on the initial development of this measure which was I think in 1998 so it's been a while, you know, since that task and obviously so I don't know exactly how the exact – that number was developed and when we responded to those, we did say that we used a – the strength to look up (inaudible) a guideline from 1993. At that time, this measure was developed with the Robert Wood Johnson Foundation. And so I think we're looking at allowing 10 days of gap for every 30-day prescription.

So when we're looking at 84 days or put some 90 days, we're looking at almost 30 days work doesn't allow a gap which I believe where they came up with that number.

Female: So that makes sense back in 1998 to have done that. I understand that but now with all the data that's come out and how everyone is moving towards it more of a standardized I guess is the best way to say it, of what is it in that acceptable medication possession ratio or proportion for days, that's really become 80 percent that something NCQA would consider moving towards what, is becoming more standard?

Male: I think one of the – so, you know, the answer is of course, yes, we're willing to entertain those, you know, because we're obviously we're harmonizing, we have a hundred measures in the NQF space alone. So we're harmonizing, it seems everyday of the year. So that's you know something we would look at. I do think that our health plan to – or the people we are holding accountable generally liked, you know, we've been trending this measure a long time.

I've concluded in our accreditation scoring, was just 50 percent of an accreditation score for a health plan CPM. They – like many people, they don't want to change the exchange rate between, you know, if you will the dollar and the euro. So they're used to that, people can kind of calibrate the difference between 75 percent and 80 percent if that's the dominant either MPR or PDC whichever version you want to use. But I think we are attentive to the fact that, you know, we work closely with PQA, Pharmacy Quality Alliance and finding helpful and we are collaborating with them more on the areas where pharmacy measures are included.

So I think that, you know, I would imagine that in the next couple of years or so, we may see a shift towards that. We have lots of measures – would have different variations on a theme of this are actually measures for example or COPD measures, so let's say it's a little bit on sometimes art. It's not exactly science that 81 – 80 is optimal and sometimes it's dependent on the treatment and the condition.

Female: Right, I guess that's why I'm asking and I mean in this condition, is it acceptable for every 30 days there's 10 days in this therapy and is that really – is there data that say that's acceptable, second this whole thing of, you know, such enlarged PBM, as being allowed to just do 190 days, so that they promote you got your 84 days if you move it to 80 percent for a proportion per day, the possession ratio, you would have to have 91 days. And that's truly what speaks to the measure of did they get titration or did they really get seen and I mean there's all of that right the sixth week measure that they had to be seen by a doctor that was all removed, that used to be there with those measure as well that got removed.

And I think that's something that should be at least considered instead of just saying – I'm with the health plan and I understand but all they would have to do is just change the calculation, there's nothing else because everybody else would have to move to that. So they're at the 90th percentile and 75th percentile would move right along with everybody else.

Male: Yes, that's true.

Female: I agree with that statement.

Female: I also would like NCQA to address the comment that is made later that does go back to being able to do this measure correctly and that is the use of generic antidepressants many of which are on the $3 or $4 payment list that are not captured in claims data by a health plan. So I would like to hear your comments about the potential underreporting of people who truly are adherent but their claims are not appearing in the data sets.

Male: We're certainly cognizant of this CPM – panel has often reference it, to be you know, that actually to speak quite honest the quality of the published literature are really getting after this – are not very extensive and (inaudible) asset we're looking to. This measure does require the – the prescription and the claim, so therefore, someone was stealing all their prescription outside the system, paying cash for not seeking reimbursement most health plans offer a reimbursement when you're paying cash and, you know, they did that for flu and then they just expanded that and again, it may not be universal but you

would be missing those people in the denominator because they're simply not showing up.

It wouldn't change the – the quality, the measure. It simply means that those people who choose to go outside the system are going outside the system. It's almost like people who choose to go outside the system for STI, you know, screening because they just simply don't want to health plans – they don't want the parents to know, you know, whatever. They're – they're not –

Female: So it's – if a member is identified by the code and claims, because there her claim will appear for being depressed. Is it that they'll never appear because they – I'm a little bit confused about that.

Male: Even though you're first looking for diagnosis …

Male: Yes, just – yes I think, (Jeremy's) got a response.

Male: And – to complete (success) to get into the – the denominator, you still need to find an index prescription date and that's where you – we look at the medication. So, we're really not looking for new diagnosis. The intent of this measure is to find people who are nearly treated. So, to get into the denominator, you need that first prescribing event. And from there …

Female: I understand but then what the measure is really looking at are those people who were diagnosed and who had their prescription refilled and paid for by the health plan. Not people who were diagnosed as depressed and whether or not they were adherent, just only looking a set of those (inaudible).

Male: Correct.

Female: So, it's not as universally applicable as perhaps, it's not to be.

Male: I think, we're all operating within a somewhat fractured health care system that allows people choice about what they – where they go and where they see. It would be very hard to imagine a way to capture those people who choose to go outside their pharmacy benefit because of, you know, probably,

the 5 or 10-dollar difference of pay. So, I understand it's a phenomenon. We just would be challenged to understand how best the approach are.

Female: And in regard to that – I mean, I understand that is – that is the day and age we live. And in our Medicare population, we actually did analyze for a one-year period but it does also have a (inaudible) so having fulfilled differently a lot of cash, out of you know, cash they could save but it's almost 12 percent at that point. So, the reason I put that comment in the four dollars. I know you can't really do much about that because you got to make it easier, I mean it's harder to submit if you have to manually get those electronic files it would be good.

But have you thought about maybe breaking up the measure of by (self-economic) by region, understanding whether the zip code, I'm not sure exactly how but I do know that there are people that do look at measures differently for different populations because this measure is how many health plans are viewed compared to others. And so, for local plans that are in more of a disparity area versus others, you know, or just look lower but it doesn't mean that they actually – the other group is doing better. It just – they probably have more patients using the free programs because their – their patients just can't afford the co-pay.

So have you thought of that? I know that you have changed of how it's captured but can you change how you knew the population?

Male: That's an interesting question. You know, and I'll let some others lay on this as well. So at one point a lot of people believed we're getting member-level data here when people report these measures. We're getting numerators and denominators. So, we don't have, you know, we don't have that level of detail. We certainly prepare state regional – HHS regional reporting on all of our measures through Quality Compass and we prepare national, sometimes to reach new data, in their state of health care quality.

We also, you know, report this by product line and while product line is not a very sensitive – to breaking out different routes. There – there is some utility in looking at, you know, Medicaid plans versus commercial versus Medicare

and we don't adjust any of our health plan measures by, you know, for instance, by SES or other approaches to risk adjustment.

We don't want to adjust the way the health plans' responsibility and out markets responsibility for providing good care whether that's breast cancer screening, cervical cancer screening or taking care of children. So you just – we appreciate that. That's just our general orientation but, you know, at the measurement level, you know, this issue of – cause the question at hand really is generics, low questionnaires paid by cash. That's –that's an interesting problem.

And, you know, hopefully with more literature, we may be able to figure out ways to adjust for that. And – and, you know, provide a kind of corollary estimate of – of performance.

Female: So we go back to reliability and scientific (inaudible) question. The second item

Female: Yes, let's go. Move on to second – to our second item and open it up for work group discussion and then we can direct question to developer as needed.

Female: So, the point that somebody entered as a comment, I also was questioning the part on the tricyclic internist to get to that page, I'm sorry (inaudible). But, it's a (feel) threshold that the TCAs only account for about 2.25 percent of antidepressants prescribed. And because of TCA, they are still commonly used for neurologic pain and other disorders. Is that a drug class that still needs to be included in the numerator and denominator to qualify?

Male: You know, I've been, so I think I understand what you're suggesting is that – that that particular drug because of its broad applicability to other diagnosis may be misleading in some way.

Female: Misleading but then also many of the patients definitely don't take their drugs for three to six months. So – but it's not being used for depression most – many times.

Female: I think that the issue that the list of medications are meant to cover all of the medications that would be used and could be used for the treatment of major depressions and having myself in, you know, in the involvement of treatment of depression and within primary care, I would say that tricyclics are still in use and there are people for whom, you know, the old drugs are working very well. And so, the – you know, I think that to toss it out because of the fact, you know, I think we feel confident that the requirement of the inpatient claim or the two outpatient claims is sufficient to specify that we're talking about major depression here and then to associate that with the continuation of an appropriate medication.

Female: Are there any additional question?

Bonnie Zima: So this is Bonnie Zima and I was – I was the author of a lot of those comments. So, I was wondering if NCQA could – could address some of the issues that are written down there.

Male: I'm sorry. Can you maybe question (inaudible) because we're having some internet (inaudible).

Bonnie Zima: Oh, OK. OK. And – and I think, you know, one of the things I wondered about was – was the – and I'm sorry, you know what, actually there's a typo in – in the – the comments here.

Is the beta binomial model an acceptable approach to – it should read, determine reliability. I did look at (Dr. Adams') technical report. It appears that NCQA contracted was (RAND) for the technical report and in reviewing the technical report – it – it seems like it's describing – it choose more to look at variation by physician.

Male: I think you maybe talking about a different measure. I'm not familiar with a (RAND) report on our NCQA measure. That's on …

Bonnie Zima: No. It's an – it's a tutorial on the …

Male: Oh, I'm sorry.

Bonnie Zima: On the reliability of provider profiling? And – so I would just sort of – again sort of wondering, is the beta binomial model in an acceptable approach for reliability that the tutorial comment is a lot more on – how it could be useful with variation by physician. It also comments that it depends on how different the providers are from another and that – and that as performance – as performance of providers improve, the reliability actually decreases because of the way this is calculated?

Female: I think, I'm not sure what tutorial – I don't think we refer to a tutorial. But let me just say that what – my understand – in what I understand is that NCQA uses the beta-binomial model to look at very – at the applicability and in usefulness of a measure construct in comparing health plans to each other. And so there are a few different inputs that are going to be influential in the beta-binomial. One is the variation in health plan performance from the highest to the lowest and the other is the variability within each plans performance.

And when the estimates – when the point estimates for many plans are clustered tightly together and are exceeded by each of their – within plan variability, that ceases to be a useful tool for measurement and comparison between plans. That is the basic idea of what our beta-binomial approach goes and looks at and we have done work on it here at NCQA specific to health plans. And I'm not necessarily familiar with the physician profiling tutorial that you're talking about.

Female: Well, it actually was I think a background material that we got from NQF …

Female: Well …

Female: … which helps.

Male: If it's about physician profiling then I don't know if there's a direct link between physician profiling and health plan profiling …

Female: Yes.

Male: They're equipped with different dynamics there.

Female: Yes. Exactly.

Male: And again, it is tough, you know, we've been through measurements and, you know, steering committees where there was only one health plan measure and 32 physician level measures and it's sometimes hard to and, you know, kind of understand that, you know, they're kind of like the operating theater we're in compared with the more well-understood if you will, operating theater that clinicians reside. So, it is a little different and maybe we could talk with NQF (inaudible) line about providing maybe more appropriate background materials for when health plan measures are being assessed. So I'd be happy to …

Female: Circulated are – (TIMES) for measure evaluation. Are you referring to the citation perhaps in that guidance?

Female: Possibly.

Female: OK.

Female: But basically the – it's a very good technical report by (Dr. Adams) and I think it was maybe commented on by some of the other measure developers and maybe different measure where that citation was given. But it's really – it looks like it's really a technique to look at provider profiling and not necessarily health plans. So that's probably where there's (confusement). So actually if there is additional information about the beta-binomial model and how that robustly establishes reliability, you know, I would be interested in that.

Male: Yes, I …

Female: I think the other concern is it just sort of raises a bigger question of the NCQA approach, as now NQF reviewers are being asked to be more stringent about reviewing the scientific acceptability. It appears that at least with this measure, only safe validity of the measure is established by an expert panel.

Male: I'm sorry. We just cropped up from reliability to validity. Did you want us to provide some comment on the reliability? Because again, the random report …

Female: Sure.

Male: …specific to physician profiling it was actually we just folded it up. It was requested by NCQA some time ago as I recall to better understand physician profiling went large. It was not intended to be a tool for use or the measure, you know, the measurement of health plan performance.

Female: OK.

Male: That's one thing. I think, you know, if I can just characterize (Ellen Burstyn) quite often providing commentary to at least 13 or 14 steering committees that we participated in last year. She says, "OK guys, don't get lost here. This is a signal to noise ratio, a high number reflex. It is a telling a – that this is painting a reliable picture about health plan performance and variation they're in" and that's kind of that's that. And, you know, we provide kind of summary description of how this works. And, you know, we can certainly bring in statisticians here on staff who can explain it but mostly, I would have a hard time following it.

But that's our challenge, to make sure that we're providing a reasonably good that what you're reading there is the same description we've used for, you know, 95 measures endorsed as early as just late last year. So I mean, I do think that it's a common theme and I think the approach of sound, all be it and maybe somewhat distinctive from what you're used to saying with (inaudible) and other things.

Female: And any comments on – it's simply face validity, right not clinical validity established?

Male: We prefer to say that it's face validity not that it's just simple. It is face validity. We have multiple levels of creating a measure. A measure development process is highly iterative with essentially the moral equivalent of three or four year type panels and sequence. And sometimes developing

measures over a period of years, you know, these are not created in a day. This is one of our longest standing measures that goes back to 1998. It was developed with Robert Wood Johnson. And I can't remember if the Washington Circle was involved in this. (Jeremy) if that was the other set of behavioral ….

Male: That was (inaudible).

Male: Right . So, again, yes you're correct face validity is our approach. And again, these are long-standing, measures of about 15 years.

Female: Right. And it looks like RWJ study was done in 1999 with two health plans. And now they think it's where they only see health findings which linked it to symptom reduction. Because it looks like this – the more recent field test of concordance between performance rates and denominator percentages.

So at least in 1999, among two health plans and here it was related to symptom reduction.

Female: We're just looking it up right now.

Female: OK. I'm just working out of the materials that we're provided. And I have …

Male: You know, in general I think we can speak to measures developed in 1998-'99 the slightly different world then in terms of …

Female: Yes.

(Bob): … measure specification, clearly if we were developing a measure de novo which we do all the time here, you know, that – that we would have probably a clearer link. And I do believe that given the amount of a couple comment and scrutiny of our measures, if the measure was not – was not seen as being valid. People who are responsible for implementing the measures in real communities and in real regions and in real health plans, they actually have quite a large and strong voice in our measure development process, we would be hearing about this.

We have a pretty, a very tight feedback loop and I'm not sure, you know, it's in the details of our submission form. But any measure user has a 24-hour access to our patient – pardon me, policy clarification support system and this is the staff that responds to questions from the level of highly specific, you know, coding detail to even larger issues of measure intent. We get those all the time. Those are an integral part of our performance of our reevaluation process and what we take to our measurement advisory panels and ultimately the committee on performance measurement and then the board of directors and also public comments. And so, over the years, we literally are crafting, refining measures all the time. And so I do believe that our measure development process from a validity perspective has generally the market has told us that there is a strong confidence in that.

(Angela): Thanks Bob. This (Angela). I just want to just –looking at the time, I wanted to see if there is any maybe wrap up comments on this measure? According to the voting there – this seemed to be systems (inaudible) that are on usability and feasibility but I want to open the door for discussion in case there's some questions.

Female: Can you just comment quickly on how it's being used and also if there's a targeted rate of which we say it's good enough or is it sort of let – there's no level on which you would say they can't get better. So if I'm looking for plan and I see 60 percent is that – how do I interpret that versus a plan that has 80 percent of – should I choose the 80 percent over the 60 percent and not choose the 30 percent or, you know, and there is that – be as the same the consumers are a little bit to help understanding that.

Male: Sure, I think that what we know about – well there's just a couple of dimensions to this that they just – let's be clear how health plans operate in the real universe, health plans operate in the real universe trying to get business from people willing to pay the price. The people willing to pay the price are employers and large purchasers, the unions, you know, municipalities. You know the national business coalitions, they use our data to help frame, you know, RFPs so that they can be considered. So adding that competes with dwell point across Blue Shield competes with Kaiser whatever and one of the metrics that they used is cost and for of the metrics they use is quality.

So I think no one purchasing from the purchaser perspective would say I'm going to purchase this plan on the strength of one metric. And I don't think any consumer, so a consumer might be going out to Medicare advantage sites, looking at STAR programs and looking at results from the PQRS program and/or result from, you know, the STARS rating system would say I am going to take this plan because of their performance on X.

Even people with diabetes, you know will probably look at diabetes and say, oh that's important to me but also what's unimportant to me is preventive health, what's important to me is because I have children …

Female: Right.

Male: You know and …

Female: No I take that point but how do I look at this depression measure and know is it – how do I interpret, right? We have finally 60 percent and 50 percent and I want to really choose a plan based on how well they're doing on depression. Should I choose the 60 over the 50? Or should I wait until there's an 80?

Male: I don't know, I think you should look at the national benchmark and then depending on your kind of like, yes, elasticity of interest you would say, oh I'd certainly want to pick one who's above average. That's a lot of people might think that way, consumers, I think what we know from consumer behavior and from studies is that their interest in quality measurement and stuff is developing. It's probably really not there yet but the tools for conveying information can be challenging and, you know, so I think that this is kind of like a work in progress.

But I do think that if I was speaking a plan and this behavioral help was a critical error for me and I happened to have a plan that was in the 90th percentile which is very high level of performance, then I would probably say, gee if my choices are same (inaudible) to same dollar and this person is there and they are a provider – their provider accessibility is good and in fact it includes my clinician, then I'll pick them.

Female: Well let's say – I mean, but you can't say that there's a certain level on which you can say how this plan is doing as well as can be expected. There's always room to be getting better and …

Male: I think this is why there's so much a great advantage to having national benchmarks like the ones that we provide because we're actually providing actionable information.

Female: Well, if you're just saying – yes, the benchmark to just telling, how you're doing relative to someone else that's not telling you when your population is going to be healthy enough or treated well enough with antidepressants that you don't need to do any better. So I'm trying to say as a person looking at population health what point should I say, you know, OK and this is good.

Angela Franklin: So this is Angela, I'm sorry, I don't want to cut of the bait, I would propose that if there's additional; questions about this topic that you could pour that to Lauralei and myself and we can talk off line with the developer to provide those questions before the steering committee meeting in person I'm just – say we have two other measures and two other developers waiting to discuss their measures and we want to make sure we give everyone a fair coverage.

So I'm sorry to interrupt the debate but please be sure that you know Lauralei and myself we can get those answers that you're seeking.

Male: Thank you we'd be happy to respond.

Angela Franklin: Thanks. So our next measure is 418 Preventive Care and Screening, Screening for Clinical Depression and Follow-up Plan. And our Developer is CMS Quality Insights of Pennsylvania. Our major discussion for this is Dolores Kelleher.

Dolores Kelleher: Yes, so measure number 0418 Preventive Care and Screening, Screening for Clinical Depression and Follow-up Plan. This is a maintenance review of a process measure that was originally endorsed and last reviewed July 31st, 2008. And the process measure is as follows. The percentage of all patients 12 years and older screened for clinical depression using an age appropriate

standardized tool and with a follow up plan documented when the screen is positive.

In terms of the evaluation, the importance to measure and report – it was three yes and one no given that there's a lengthy rationale or comment section so I could read through it but I think it would make more sense if the – that the discussant be the person who is making the comment since it's fairly involved. The impact is three high, one medium performance gap, one high, three – two moderate, and one low. Evidence based on subject is three yes, one no with quantity being high three, moderate one, quality being high one, moderate three, consistency being high two, moderate one, and low one. So quite a red there.

And if the rationale there is based mostly on reviews, USPSTF recommends consensus statements psychometric properties of a measure, statistically significant effects reported but not presented. So they're saying that there was – it was noted but not presented and the overall sort recommendation is B. So again, I would think the person who made that comment would want to elaborate on that.

The scientific acceptability of measure properties, three yes, one no with reliability being viewed as high one, moderate two. Validity being no votes for high, three moderate, and one insufficient evidence. The rationale being a three-month time period Medicare part B claims data and the question mark about voluntary reported data from providers as part of the physician quality reporting system and there's a question about that, and again, from one of the reviewers.

Poor record and provider was response rate for data sample raises concerns regarding integrity of data, reliability based on agreement between claims and in the independent reviewer. And my understanding is that there was some – some investigation and adjustment done on that, so that, I would like to hear more about that.

And then, there's more comments about records without valid denominator criteria removed prior to a reliability assessment leading to a denominator of

(inaudible) a hundred percent. Again, I would ask that numbers talk about it, and then, again, another question about validity established by a technical expert (Piano) "for facing content validity". So there's also question there.

Usability was three high, low one, feasibility one high, two moderate, low one. The comments there were poor record and provider view rates and voluntary data, again, questioned. And another that the rating, and I assume this is a better rating was because of the EHR in near term 2014, so preliminary assessment for the criteria being met or the measure being suitable for continuing endorsement is three yes and one no.

Open it for discussion and I would ask that those who made comments, perhaps, question elaborate on them.

Bonnie Zima: OK, so this is Bonnie Zima. I was probably the one with the most involved comments and they're really almost frankly questions back to the developer because it obviously – I had some confusion about – things about – it varied from client's data to claim sample reviewed. We had 275 records from 77 providers but 240 records reviewed. So some additional information about some of the message so that we can interpret better what the performance gap is will be helpful.

Female: We have the developer on the phone. (Inaudible).

(Tish): Hi. This is (Tish), (Inaudible) Pennsylvania. We do have a number of people on the call including our statistician, Gary Rezek. So I would like Gary to just give a little synopsis on how we actually obtain our records for testing. Gary, could you just give a synopsis of that, please.

Gary Rezek: Sure. Where we begin is with the Part B claims. This measure is supported by, you know, just putting a G code on the claim. And so we request all of the claims that have reported one of the enumerated G codes for this measure. For that quarter that we requested that the data for which we, the first quarter of 2012, there were about 10,000 total claims which reported this measure.

The sample that we used for reliability testing is the – is a random sample of the providers who submitted those 10,000 claims.

Male: So is that the 77 providers?

Gary Rezek: Yes. What we – we try to sample close to 300 claims. We also tried to not sample more than 10 claims per provider. So when we request the records, it's not overly burdensome for them to submit those.

Female: And we also, Gary, I think we tried to make it as it how many providers were we're shooting for with that 80?

Gary Rezek: Yes, we shoot between – yes, that's between 50 and 100. So we came right around 75.

Female: Thank you, Gary.

Female: I think it would've been helpful if there was a little bit more description on the methods in the application.

Gary Rezek: I think there's some supplementary materials but I don't know what you guys are saying. I know we did submit several documents.

Female: OK.

Female: And how was underserved and non-underserved defined? These performance rates by category, with (high score) probability and then it looks like just some simple stratification. But how was underserved and non-underserved defined?

Gary Rezek: Well, that's kind of terminology that's based from other CMS projects but it's essentially white and non-white.

(Tish): I don’t know if you have this – sorry, this is (Tish). The supplementary document that's provided. On page 17, we do have a definition for non-underserved and underserved population. And it's posted that the underserved category is defined by the racial and ethnic designations of African-American, Asian-American, Hispanic, and Native American.

Female: So it's the same as race and ethnicity?

(Tish): We also have that rural and urban.

Gary Rezek: Right. And race and ethnicity, we're looking at each individual race category as opposed to white and non-white.

Female: OK so – OK. So it's really race. OK.

Female: And this is NQF, just to note that all of that supplementary material is included with the measures on the SharePoint page if you want to.

Dolores Kelleher: And this is Dodi, I'd like to make a comment 'cause I don't want it to be missed. I also think that there's opportunity in terms of performance gap with the fact that you're now doing 12 and older and you're including pediatricians in the mix here because of the sort of goes back importance to measure the – even more profound, you know, profound need to be able to screen and follow up without adolescents. So I think that's something that is different from your initial measure and endorsement, correct?

Female: That is correct. We have expanded the age. I will say, however, I know we did try to get some records from pediatricians and there was a lot of concern about the testing for those measures because of the age of the client. So we did – of all the people, we requested records from, I know that we did have some pediatricians who do not are feel comfortable sending the records because of the child's age. So the testing in that case is primarily adults.

Female: Are there additional questions? And we're looking at importance to measure and report I believe.

OK, any summary questions from our discussions from the work group? OK hearing none, we'll move up onto our next measure which is 0518, Depression Assessment Conducted and CMS and Acumen are the measure developers. Bonnie Zima is our – lead discuss for this measure.

Bonnie Zima: Yes, OK. This is a process measure defined as percent of patients who were screened for depression using the standardized depression screening tool, a start or resumption of home health care. The data source is the way to see

data from Medicare certified home health agencies, and the level of analysis is at the facility or home health care agency level.

As mentioned, the stewards of CMS and it appears that they've contracted to do this work with Acumen which is a company that develops scientific support and analytic tools to help stake holders make decisions from effectiveness, appropriateness, and safety efficacy. According to their website, it is not clear, however, from the website if they have the resources or infrastructure to assess whether adherence to propose indicator improves access to care for depression, quality of care received or clinical outcomes.

It appears that this measure was endorsed in March 2009, updated 2013, and obviously, is under review. For importance to measure, the rationale for high impact was basically prevalence rates of depression among these target population as well as comorbidity related to depression like repeat hospitalizations (inaudible) and higher cost. And although the summary evidence states there's room for improvement, there wasn't any study that described that support of screening for depression and its target population relates to improved access to care, receive a quality care or improve clinical outcomes.

What's really interesting is that this quality measure is part of the (OASIS) assessment core which is actually mandated for CMS certification for all home health care facilities. So this is probably why we have such very high adherence rate at such a tiny performance gap with an average of 96 percent and a 58th percentile at 99 percent. So if Medicare or certified home health agencies are required to collect and submit this data for certification, it seems like – I anticipate that this measure's going to stop and review just because of the performance gap. It's so small.

Besides evidence, again, I didn't think data supporting the process of outcome relationship was presented on reliability. Again, it's mandated so that was – it was very hard. They too use the beta binomial method that we've already discussed. And validity appears to be solely based on (safe) validity based on a technical expert panel. And there's no meaningful difference in performance again, because it's mandated.

Female: Thank you so much. Are there other discussion from the work group?

Female: Can we back up and go through the, you know, the group…

Female: You know, we – on importance we had three yeses, one no. Evidence, three yeses, one no. Reliability four yeses, no no's, usability high one, moderate two, one low. Feasibility high two, moderate two, low zero. And it was a split vote on the endorsement.

(Anita): So (Anita) and my comment I think totally aligns with what you're saying and just looking at, you know, there's numbers that I put in there for – pulled from theirs, and they're obtaining 94 percent because it's mandated by CMS for OASIS so they're getting the time out. It's a really a surrogate marker and we don't have outcomes of the time to move this to an outcome of can home health care agency have, you know, have it measured more? How many either referrals do they make or how depression management plans do they make out of this resulting score? Or something like that, I think, would be more meaningful than just to say that they did a screening.

Female: Did we want to go the developer for a response there or additional discussion from work group?

Female: So this – I just wanted to clarify – can you just – maybe the developer could clarify this in their responses "was endorsed previously."

Keziah Cook: Yes. This is Keziah Cook from Acumen. This measure was given time limited endorsement in 2009, and then it was granted for endorsement the summer of 2011. So yes, this is a previously endorsed measure.

Female: What did it achieve in 2011? Like what was the percentage that were completing the PHQ-2 screening.

Keziah Cook: So our first – our initial submission was they found at nine months of data from 2010 which was the first year that this – the item supporting this measure were collected on the OASIS. And at that point, the performance rate was around 89 percent. So we have seen a substantial improvement in

performance since this item, you know, since this measure has been publicly reported. So it does seem to have had an impact on compliance to publicly report this item.

I also just wanted to clarify what the discussant was saying about this being a mandatory CMS requirement. The CMS requirement is that home health agencies conducts the OASIS assessment. And one of the items on the OASIS assessment asks whether depression assessment has been conducted and there's even a short PHQ-2 assessment that an agency can choose to perform directly as part of the OASIS assessment

But the mandate is to conduct an OASIS assessment. From that OASIS assessment, we learn whether the agency did also do a depression assessment.

Female: So there's no additional fees or offered from Medicare if they do that additional part for the OASIS?

Keziah Cook: No. Agencies are, you know, they are required to establish their patient's need at the start of care. So there are certainly a variety of ways that agencies are expected to establish their patient's need.

Harold Pincus: Hi. This is Harold Pincus, I just came on at the very end. But I had a question that I've bringing up on several of these calls about the extent to which this measure is similar to other measures, and so getting to the whole sort of measure coordination alignment and so forth. And it's – maybe somebody (inaudible) can speak to that (inaudible) neighbor measures and how similar this is.

Angela Franklin: Hi, Harold. This is Angela. We're looking at our chart. I think we did identify some additional measures and we can pull that up if you want to continue discussion. We'll get that back to you in one minute.

Harold Pincus: OK and the other question I had is this measure, as I understand, does not require any kind of assessment or follow up, correct?

Keziah Cook: Hi. This is Keziah again. This measure does not require follow up but we do have two additional home health measures that are reported to home health

agencies. One of which tracks how often interventions for depression are included in the patient's plan of care. And the second which tracks how often those interventions are implemented.

Harold Pincus: (Inaudible) choosing the denominator of positive assessment from this one?

Female: Approximately, there are certain cases where a patient can have a negative assessment where the assessment could show they're not at risk but depression interventions could still be included in their plan of care, you know, based on clinical judgment, of course but roughly.

Harold Pincus: You know, the reason why those kinds of – why we're not given those, you know, sort of this package of measure altogether to review?

Keziah Cook: You know, I think in large part, that's historical. When we submitted these measures in 2009, all three depression measures were submitted to NQF. And at that point, NQF felt that, you know, tracking the plan of care and the implementation applies too high of a burden on home health agencies and chose not to endorse those measures. It sounds like the situation maybe different now that we're a few years later, but that that's why, you know, we are – this measure was up for maintenance because it have been previously endorsed.

Harold Pincus: I think, so – so the others have not been endorsed?

Keziah Cook: No, they were not endorsed in 2009.

Harold Pincus: And so they weren't submitted on this – resubmitted for this call for measure?

Keziah Cook: No.

Harold Pincus: Is there reason why that's the case?

Female: Could you repeat that, I'm sorry?

Harold Pincus: Is there a reason why there wasn't a resubmission of those measures?

Female: I'm not sure.

Keziah Cook: I think the CMS have to answer that.

Harold Pincus: OK because I mean, one thing that turns out (inaudible) is that, you know, that there's limited data. And I think Bonnie alluded to that that simply screening or assessment alone actually influence the outcome. You know, unless, it's part of the – sort of a suite of measures that capture, you know, some kind of process.

(Deborah Dee): This is (Deborah Dee). It's also from the development team. And I think that one of the reasons that we didn't submit it is because it was previously rejected by NQF. But we are very much exploring the possibility and the option for incorporating the idea that there was a follow up plan incorporated into the measure. So we're developing a response document to the request from this committee about harmonization of these measures, and we are looking in to whether or not we can incorporate that – the follow up plan into the measure.

Harold Pincus: OK, thank you.

Female: Harold, this is NQF. You're asking about the other measures that might be related to …

Keziah Cook: So we do – there are number of other depression screening measures. There are two NCQA measures which are depression screening by 13 years and screening by 18 years. There is another success measure, that's the percentage of patient to have depressive symptoms. Another – oh wait. Sorry that's – yes, that's CMS. There's another NCQA, one, a maternal depression screening and two, PCPI measures, one of which is in this project, a major depressive disorder and diagnostic evaluation. And they also have a corresponding child and adolescent major depressive disorder diagnostic evaluation measure.

So the way that all of these developers have been contacted by NQF and asked about why – ways in which they can harmonize their measures. And then during the in-person meeting, we have dedicated time to discuss harmonization of these measures. So we have a whole process. We'll be sending the material shortly, just as an FYI.

Harold Pincus: OK. But I think one of the things to think about as we go through this is, you know, I know that our roles doesn't include sort of trying to get, you know, necessarily trying to choose the best in class or to pushed some kind of, you know, coordination on this. But it does I think helpfully inform our discussion, you know, the different – what other near neighbor measures there are and how they're different.

Female: Absolutely. We agree. So we'll get that information out to you, and the comparison chart so you can see that clearly across all the measures.

Harold Pincus: OK, great. Thank you. I'm going to ask to sign off now, thank you.

Female: OK. Are there other additional discussion points? I think we have covered the gaps but were there additional questions about the performance gap which is …

Bonnie Zima: No. This is Bonnie Zima. I think that developer did a good job of the clarification.

Female: Thank you. So any additional questions, comments on 518? OK, hearing none. Thank you.

At this time, we'd like to open the line – all lines so that anyone who cares to make a public comment can do so.

Female: And (Natalie), can you just make sure everybody can push over from the public line?

Operator: Yes, all lines are open.

Female: Great. Thanks.

Female: (Inaudible).

Female: OK. Well, if there are no public comments then I guess that concludes our call. We just want to thank you so much for your engaged participation today. We think it was a really great conversation. Thanks to the Steering

Committee members and to the developers for taking the time to be on this call. In terms of next steps, what we'll do is we will summarize what happened on these calls and then well send that material out to you in time for the in-person meeting. And the measures for what you are lead discussant for the work group calls, you'll also be the lead discussant for the in-person meeting so you can summarize what happened on the calls, and then voting as well.

And if, by the way, any of the conversation today, sort of made you want to change your vote, your initial vote, you can go back into the SurveyMonkey tool and resubmit any different votes and just be sure to put a 2 or a B after your name so that we know that it's a new mission.

As you know, in-person meeting is coming up on June 5th and 6th and that all be here are headquarters in Washington, D.C. and you should've received a logistics e-mail a few weeks ago about booking travel and making hotel arrangements. So if you didn't, just e-mail us and we can – we can get that sent to you. Ad the meeting will start both days at 9 a.m. but we will have a continental breakfast at 8:30 a.m. should you care to join us earlier.

I'm really looking forward to seeing everybody in Washington. Are there any questions about the upcoming meeting before we end the call?

Dolores Kelleher: This is Dodi. I believe the last thing I read was the hotel was still not determined. Is that changed?

Female: Right, we're checking on that right now actually, and we'll have the meeting's teams send you an e-mail with the updated hotel information. Hopefully we can this afternoon.

Dolores Kelleher: All right. Thank you.

(Bonita): Just one quick question. This is (Bonita). I always get this mixed with up with the D.C. airport. Which one is the one that's closest to where we have to go?

Female: Washington Reagan.

(Bonita): Reagan? Thank you.

Female: Yes, (PCA), I think that is.

Female: Yes, (PCA).

(Bonita): Thanks.

Female: Great. Well, thank you everybody for joining us and we'll see you in a few short weeks.

Dolores Kelleher: Thank you.

Female: Have a great afternoon.

Female: Thank you, folks. Bye.

END