NATIONAL QUALITY FORUM

Moderator: Sheila Crawford
May 14, 2013
3:00 p.m. ET

Angela Franklin: Hello, and welcome to the NQF Behavioral Health first workgroup call. This is phase two of the project. This call today will be focused on depression, alcohol functioning and medication adherence. My name is Angela Franklin, I'm the senior inspector for the project. I don't have my usual voice today, so please pardon the hoarseness.

And I have with me our senior project manager, Elisa Munthali, and we also have Lauralei Dorian who is our project manager, and Jessica Weber who is our project analyst is also in the room with us here in NQF.

Our coach here, Harold Pincus is on the line. And I will turn it over to Harold in a few minutes. I just want to just quickly go over for the benefit of the steering committees the process for today, and also for the benefit of the developers who are also on the call today.

As we have sent out in our e-mails, the process for today will be that there's a lead discussant or one or more lead discussants for each measure. And we'd like for that lead person to give us their analysis of the measure and how it meets the criteria as well as briefly describe the measure first. And we'll be – we'll also be summarizing the complied sense of the workgroup on how they voted the measure overall.

We also have the developers on the call, so if there are any questions about the measures, the developers may way in. However, please keep in mind the discussion is primarily for the steering committee to deliberate on each measure. These deliberations are preliminary and we will get into more detail

and recommendations at the – in-persons during committee. With that said, are there any questions?

Female: Also just note that this call is open to members of the public so we'll be taking public comments towards the end of the call. And it's also being recorded and transcribed so we can send the conference transcription of the call once they receive it in a few days.

Angela Franklin: OK. Well, Harold, I heard you early on the call.

Harold Pincus: Yes, I'm here. And I just see agenda we're following is starting with major depressive disorder diagnostic evaluation?

Angela Franklin: This is the – yes, go ahead.

Harold Pincus: Yes.

Female: We have 1923 that's severity-adjusted effect size measure up first which I believe you were the lead discussant for.

Angela Franklin: That's correct.

Female: If that's (inaudible), if that works with …

Harold Pincus: OK, let's (inaudible). The thing that we sent out I guess on Monday have that – the other ones first, OK.

Female: OK.

Harold Pincus: But – so actually, I need to get the …

Angela Franklin: Oh, OK.

Harold Pincus: Information (inaudible) because …

Angela Franklin: If you want …

Female: We can start with 0103. (Mark) is ready to go.

Angela Franklin: Did you want more time, Harold? And …

Harold Pincus: Yes, it would be good because I have to pull it up from – from the (web).

Angela Franklin: OK. We can get started then with – is it 0103? And then (Mark) is the lead – (Mark), you're there, correct?

(Mark): Yes, I'm here. I don't have the full wording of it here. I can summarize it if that's what I'm supposed to. I don't know. This is the first time I'm going through this. So …

Angela Franklin: Oh, OK.

(Mark): … you have to bear with me.

Angela Franklin: I can bring it up on the screen as well.

(Mark): OK, why don't you do that?

This is the measure on major depression and it is combination of the measure of identifying major depression and in those where a major depression is identified assessing for severity of the depression with mild, moderate, severe. Do you need more detail on the initial description or is that sufficient?

Angela Franklin: That's sufficient. We're explaining the measure. That's good.

(Mark): OK. And I felt that it had – was high on the importance since they gave the evidence of high frequency of depression and the impact of depression. Among the five of us rating it, we were split between high and moderate with two of us having it high and three of us having it in the moderate range. The rationale – again, I rated this high because it seemed to me and from what we had that the importance of not just the depression but also determining the severity of it were – was an important process that the clinicians need to go through.

And, again, we print – we just had four out of the five of us rating that it's moderate and one as high. The evidence was again – let me move that. There were some that did not feel it was a health outcome and that was actually three

of the five. I had felt that it really was the – it's not a – was a measure of I guess, it's not the – it depends on what is included in the definition of terms of the health outcome. It certainly was an – to me, a health outcome of being able to identify the issue – the condition as first important step in its – in its treatment. So I felt it provided some of that under the evidence. There was rated lower with all of them either in the moderate, one in the lower range, and two – that felt it was insufficient information that was reported.

The rationale of – there was at least one who has it with the data linking, felt it was lacking as part of it. And the – one issue I had in terms of evidence is that while there is some description in DSM-IV for the – what is mild, moderate, severe that that was still more to – it could be needed to be better clarified and just indicating the severity without some measure of accuracy. I think we can set somewhat under the evidence decision. There were three, there were no, and two yes. And again, under importance.

There were two of us who felt it was important than three not, and going through the rest of it. As far as feasibility, there – three of us felt it was high with one moderate and one insufficient in the area and usability was again, only two of us were high and the other were (inaudible). So I think there were some differences of opinion as to the quality that they – that it had and under suitability, there were three of us that were yes and two of us no.

Angela Franklin: OK. Thank you (Mark). And at this time, I'd like to circle back to importance to measure and report, and remind committee that it's a must pass. It's one of our must pass criteria. And I'd like to throw the floor open for discussions particularly around the split between the – whether the committee members thought it passed the criteria in terms of impact and evidence. And we also have a developer on the call if we'd like to direct any questions to the developers.

Harold Pincus: So, let me say a couple of things just from my perspective. I think the – one of the questions is – and it's hard to sometimes evaluate this in some way to think about for the overall way in which these are evaluated. But the important issue, it becomes a little bit unclear whether the issue that the measure is supposed to address is important versus whether the measure, if

implemented would actually be an important measure. And I think that that – I think that made also account for some of the differences in terms of – in terms of some of the ratings. But …

Female: Well, this is – this is (inaudible). I think the other issue here is in many of the measures we – considering the path and in this batch. It may be part of the causal pathways to improvement that in and of itself may or may not be deficient to affect patient-oriented outcomes. In other words, simply measuring something doesn't usually in all of our trials have much of a lasting effect on outcomes. But, I would be willing to say that measuring severity is almost an essential part of being able to monitor the treatment and progress of a disease state like depression.

(Mark): Yes, and this is (Mark). And I thought that it was an essential step that without it, you really can't get much farther along in the process.

Harold Pincus: My perspective was that I agree that it's an essential first step but there's no evidence that measuring it one time has any impact. And as I understand it, this measure does not require a systematic assessment. And let me just post that to the measure developer. Exactly how is the level of severity characterized in it?

Angela Franklin: Is PCPI on the line?

(Sam Kearney): Yes, hi. This is (Sam Kearney) with the PCPI. So I think in terms of – I think the question from Dr. Pincus was that specifically related to severity and how that's determined or how the measure recommends that be determined?

So, you know, we have some guidance in the measure in the definitions from the DSM-IV, you know, that that judgment be based on the number of criteria symptoms, the severity of the symptoms, and the degree of functional disability and distress. It would allow for some clinical judgment and decision-making.

With the math, you know, we don't get overly prescripted related to, you know, mild must meet this criteria matter. It must meet these criteria. And I

think that's consistent with the DSM-IV. I think we also might have EPA representative …

Jack McIntyre: Yes. This is Jack McIntyre and just to elaborate on that a little bit, in the measure, it does define severity, as was mentioned according to the DSM-IV. So for example, it says that the mild episodes are characterized by five or six depressive symptoms with a mild disability and the capacity to function normally but with substantial unusual effort. Then it goes on and talks about severe and then moderate being between mild and severe. So it does spell it out to some degree but it's certainly true that it's not specifically characterized to the extent of a – on numerical rating.

Now, the – in the measure, it does talk about using tools such as PHQ-9 to assist in that regard and the PHQ-9 does have the numerical breakout in terms of what's considered mild, moderate, or severe depending on the, you know, the number out of 27, that is, would be the mild, moderate, or severe.

Harold Pincus: Is there a reason why you didn't require that there'd be a, you know, actually using the PHQ-9?

Jack McIntyre: Yes. The notion was that the – for the initial diagnosis that the PHQ-9 by itself is inadequate both in terms of some of the exclusions criteria and the need for more comprehensive assessment. In terms of the follow up, it was – now, this doesn't pertain to this measure but it was clear that the PHQ-9 is very helpful in terms of follow up but for the initial diagnosis that would solve that – a more comprehensive assessment was needed rather than just the PHQ-9. But the PHQ-9 has a tool to assist. It was certainly written into the measure.

Harold Pincus: Yes, I mean, my view is that, you know, again, I don't want to dominate this but my view is that, you know, the problem with that with the measure is number one, there's a fairly, you know, low levels of reliability for depression assessment using, you know, DSM-IV criteria in general. And there's – it's even lower once applying that, you know, the model moderate, severe. But I guess for the purposes of using this as part of a measurement-based case strategy and, you know, it doesn't give you a common baseline.

So, you know, so I'm saying not so much to PHQ-9 as a comprehensive measure but it's giving you a common baseline, but then subsequent follow up would make more sense in that point of view.

(CROSSTALK)

Male: The fact that you have to capture the information through a chart review is fairly expensive and time-consuming.

Female: That's right.

Harold Pincus: Yes. There's something more.

(CROSSTALK)

Jack McIntyre: My …

Harold Pincus: Can I finish please?

Jack McIntyre: Oh, go ahead.

Harold Pincus: Yes. And so, it didn't seem that the, you know, the effort was worth the result.

Jack McIntyre: Well, I was just going to comment. Was the team like that the amount of extraction required from doing it from the chart makes it a lot harder as a process?

Harold Pincus: And, you know, from an important standpoint, the fact that this is just measuring severity on an index – index diagnosis, it's sort of like OK but that really in and of itself is we're going to be strict definitely to any change in outcome. And certainly, not any change in patient-oriented outcome.

Now, I – you know, I could be convinced that that's part of a pathway that ultimately could but it's a particularly weak measure. Put it that way.

Male: Are there other comments that you would like to make about this measure?

(Carlynn): Well this is (Carlynn). I just wonder how adept primary care physicians are at determining the severity level of depression if we're looking at that proof as well with this?

Male: Well, what I would say is that this (inaudible) done a lot of work with primary care if you give them a tool which is relatively easily implemented like a PHQ-9. They're intimately unable to use that. It does require changes sometimes in office practice and flow, or provision in electronic health record as people go to the EHR. But there's nothing that's magical about this that's beyond the typical primary care office. And indeed, there are a number of trials that's been based using such tools. So, I think we've demonstrated convincingly that these are tools that can be used in usual practice.

(Carlynn): And I guess that's – well, I'm sorry, I didn't complete my thought. Without a tool, it just gets a little more nebulous when you're trying to decide if they're severe. For some physicians, that's kind of the complaints I've heard..

Harold Pincus: Right, I would agree with that. Are there questions or comments you would like to make about this measure?

Jack McIntyre: Well, this is Jack McIntyre again. You know, I think the point that was raised seems to me to be a crucial one in terms of the first step. You need to establish the diagnosis in a way that is, you know, that there are some consistency of results now. As Harold said, you know, the degree of consistency certainly can be debated but I think without a diagnosis, it seems to me that further measurement or question about, you know, what the – and how much it leads to improved outcome is, you know, it's very difficult to ascertain.

Harold Pincus: Yes, I – all right.

Male: Well, this is really a two part. I mean, you know, it's the having the criteria and documenting them, and then it's the depression severity measurement.

Jack McIntyre: Yes.

Kendra Hanley: This is Kendra Hanley from the PCPI. I wanted to address a comment that I heard from a couple of the steering committee members that it sounded like there was – thinking that this was only allowable for chart abstractions. I wanted to make sure that the committee did see the electronic specification that we submitted with our measure submission where we have outlined the measure, the data elements, and the corresponding value set for this to be incorporated into an electronic health record.

Male: So if – do we have to vote on this …

Harold Pincus: Yes, that's – that's the question I have. Should we be voting on this or is it something that we present at the full steering committee in terms of people's perspectives?

Female: Oh, go ahead.

Elisa Munthali: Hi, this is Elisa Munthali. We just wanted to have this call so that we can go through the sub-criteria for all of the measures. And all of the nuances and the – the major discussion pieces that you've pulled out. This is a sort of discussion that we'd like you to bring up during the steering committee meeting, the in-person meeting on June 5th and 6th. And at that point, the steering committee will make a recommendation on the major criterion and overall suitability for endorsement.

Female: And also note that if any of the conversation on the calls today, you feel warrants a change in your initial vote, you will have until next Friday to go back into the SurveyMonkey tool and change your vote if you wish to do so.

Harold Pincus: So let me make sure I understand the process. Are we going to be voting on these as a group when we get together in-person or is the vote we've entered on SurveyMonkey (it) …

Angela Franklin: I'm sorry, this is Angela. The vote that you've entered in on the SurveyMonkey will be for review at this full committee meeting in June. And the final – not the final vote, but the vote of the steering – full steering committee will take place at that in-person meeting. And those are the votes which carry a weight of recommended or not recommended for endorsement.

Male: OK, that's what I assumed but I just wanted to clarify it. Thank you.

Harold Pincus: OK, any other comments on 0103? Should we go to – what the – again, my agenda seems to be somewhat different from the one that you're using.

Angela Franklin: Well, we do have a request from Optum.

Harold Pincus: Oh, we do?

Angela Franklin: Is OptumHealth on the phone?

So well maybe we'll leave that one for later then.

Harold Pincus: OK.

Angela Franklin: Maybe we'll move onto 0104, adult major depressive disorder, suicide risk assessment. And I believe that's (Carlynn).

(Carlynn): Yes, this is (Carlynn). This measure is to measure that with a new diagnosis or recurrent episode of major depressive disorder that a suicide risk would be completed during the visit in which that diagnosis was identified. So, that is the purpose of this. And under importance to measure (inaudible) to two, and itself that it was important to measure and that there was evidence that, you know, there is a logical reason to measure suicide in patients with major-depressive disorder.

Let see here. Under the liability, there was little evidence regarding the – there was also under evidence that didn't seem like just the assessment of suicide led to an improved outcome. So there was some question around that as well. And none of the data really showed how usable it was or what would happen if you did assess for suicide although it makes sense that you should do that with someone with depression. But it didn't show that by just simply measuring the suicide what would the outcome be if you didn't do anything else.

And this also was difficult to pull out of an electronic medical record because it tended to be in a dictated note versus, again, (inaudible) use. So, that's what I have.

Male: Other comments on this?

Male: You know, I think again, this seems like many of these measures were keys. I mean, you know, you're not going to know if someone is suicidal and I have a chance to intervene if you don't find out. Yet, there is a real positive data that connects that with the ultimate outcome of completed suicide or more appropriate intervention. But, you know, I guess I would tend to at the end of the day, vote that yes, this is suitable for endorsement despite not being the methodologically rigorous, you know, causal pathways, if you would.

Male: Other comments?

Male: I tend to agree, (Jeff), with your perspective but I'd like to raise a couple of other issues. One is, you know, my reading on this, and actually applies to a few of these, is that only – currently, this is already endorsed and it's part of PQRS. But less than one percent of providers who were eligible to participate actually report on this. Does that indicate that it is the fact that there's going to be such a total lack of update that this is not a useful measure?

And I'm just curious about how, you know, that kind of metric about measures that is the, you know, the proportion of people who are eligible sort of up – using the measure, reporting on the measure is, you know, to what extent that should represent a concern?

Male: I mean, in some ways, you know, I think we're dealing a bit with what people need and what they want. You know, at least my experience would suggest that many clinicians, probably most primary care clinicians aren't particularly comfortable talking about suicide. And that it isn't necessarily part of their routine despite it being I think pretty clear consensus that that's something we should be doing.

So you know if you ask people, "Should we do this?" I expect most people will tell, "Yes, sure." If you really look carefully at what people do or what they want to voluntarily commit themselves to, I'd say, not so much.

Parinda Khatri: Yes, I think – this is Parinda. And, you know, we are – we have a primary care organization and work a lot, you know, with our primary care providers. And what we hear a lot is, don't ask a question you don't want to be answered to, you know? So what – I think what happens is most primary care practices are not equipped to manage the yes response.

So if someone says, "Yes," then that is – then what do you do? And typically, they don't have the behavioral health support. They don't have their resources or even kind of a clear decision-making tree about what needs to happen with that patient. And so I think in terms of implementation and feasibility that has been one reason that utilization of this measure has been so low.

Male: So actually, what you're saying is sort of an argument to my mind, almost in favor of this kind of measure because if they're not prepared to deal with it, then that's something that they need to be prepared to do. And …

Harold Pincus: It's really about organizational capacity.

Parinda Khatri: Yes, exactly.

Male: In my mind and, you know, if you have the Pandora's box (inaudible) effect and people feel like it's going to be punishing them because there is no referral source so you have to take 45 minutes to find somebody who's available for the actively suicidal patient or there are not psych emergency rooms or whatever resource available, people will avoid that. But, I mean, you know …

Male: (Inaudible) patient-oriented stands to take.

Male: Right, exactly.

Parinda Khatri: Right. The question is, you know, are we guided more by feasibility versus, you know, our resource capacity as you mentioned? Or are we that really

guided by, you know, is this a critical clinical quality measure that we need to establish as a standard of care, and then once that is established, then the systems will be organized as they need to because now it's important.

(Mark): Yes, I would agree that if it – if it's important which is, I certainly feel like it is, then by making it more prominent, it brings out the issues and is going to put more pressure on trying to get better services as part of it. Because you can identify them if you don't also have some resource on how to deal with it.

Harold Pincus: Mark, I would tend to agree with you. And what – the one thing that worries me a little bit about this, is somebody – that somebody mentioned earlier is that it would be a much better measure if it – if it was – if it was linked to doing something about it, not just assessing.

(Mark): Yes, I would agree.

Parinda Khatri: Right, right.

Male: And I guess I wonder also, I can't recall and maybe NQF staff know this off the top of their heads, are there other competing measures that are better? In other words, a – you know, would that vitality and appropriate referral would be a lot better than just simply assessing. Are there other NQF-endorsed measures that are of – I think, stronger?

Angela Franklin: We're taking a look right now but not in this project. There are – and it can go to …

Harold Pincus: I think beyond this project. I mean, I think that one of the things that we're assessing I think is – and ideally, we'd like to have as much sort of alignment as possible if there's a batch of things that are measuring similar stuff. You know, ideally, we'd want to endorse testing class and lecture some good reason why some variance of the measure might be better in certain circumstances than others.

Parinda Khatri: Yes, I think in general in community mental health, you know, when we look at quality metrics, it isn't just the screening. So, it's the screening but also the follow up plan. Is there an appropriate follow up plan in place? And so, one

would think that for that primary care realm, you'd want something parallel to that.

Harold Pincus: Yes. And this is not just for primary care but – now, I do know that recently, I think the U.S. Public Health Services Task Force or preventive care did not endorse suicide to community general population but they did not – I don't know if they had taken a stand about among those with major depression. It would obviously make sense from a clinical point of view to endorse, you know, assessment of suicide risk. But as a measure, it would be much better if we were more than just, did you assess, but did you assess and then develop a plan, and ideally, if you developed a plan, did you implement it?

Female: Yes, I think the U.S. Preventive Services Task Force didn't suggest screening. In other words, all comers?

Male: Right yes, they …

Female: But did look at special populations where there was …

Harold Pincus: More specific populations. I think that's not their mandate. Yes.

Female: So, in any case, I agree with Harold in the sense that at least that seems to me these expert groups should be assessing among measures, and at least recognizing where one has a greater potential or impact on patient-oriented outcome.

(Sam Kearney): This is (Sam Kearney) with the PCPI. Can I jump in for a second to respond to a couple other recent comments? Thanks. So, I know that there were some question about what other measures there might be available in the state, and I just wanted to let you know from our perspective at least related to other endorsed measures that deal with suicide, there is one for bipolar disorder and then there is one for child and adolescent major depressive disorder. So, this is the only one that I'm aware that deals with suicide risk assessment in the MDD population.

And then the other point if – if you don't mind, that I would also like to make, I know there was a comment some time ago about the Physician Quality

Reporting System program and whether or not the – what low adoption might mean. And I just wanted to point out, you know, the PQRS as it's commonly known, is currently a voluntary reporting program and the rate of adoption across all eligible professionals are fairly low. Its about – in 2010, it was about 24 percent of eligible professionals which some say it's actually a significant jump from the earlier years of the program where adoption was around 15 percent or maybe a little bit more than that. So the fact that only a few, you know, maybe – I'm not quite sure of the numbers but …

Male: It's OK.

(Sam Kearney): … a lower percentage of eligible professionals referred on these measures. I think it's more a function of the program as opposed to the measure itself.

Male: Although in your case, it's less that one percent.

(Sam Kearney): Yes, I have to – I have to look back at the data to confirm that. But, yes, I wouldn't necessarily read too much into the rates of adoption. I think we've seen that they vary quite a bit and that, you know, it depends on how the various professionals choose to participate in the program.

You know, one of the – the recent trends that we've seen is a growth towards reporting on measures group and these measures exist as individual measures. So that could be part of the reason that the measures aren't reported on us frequently.

And also just to mention, this measure is going to be in the meaningful youth program starting in 2014.

Male: Other comments on the measure?

(Angela): This is (Angela). I just wanted to confirm. I was – we were looking up the (inaudible) – the measures that might be related to this one, and I think I just want to confirm that – Sam had already noticed that there's a measure in the portfolio 0111 that's for percentage of patients with the bipolar disorder with evidence of initial assessments that include an appraisal for the risk of suicide.

Plus, there is another measure, 1365 in the portfolio that looks at the percentage of patients aged 18 years and older with a new diagnosis of a recurrent episode of major depressive disorder who had a suicide risk assessment completed and that's the PCPI measurement as well.

Male: Isn't that – isn't that to say, and almost the same as this one? Except that it's recurrent …

(Angela): Oh, I'm sorry, I'm sorry, no. There's – there's – I'm sorry. 1365 goes to the percentage of patient visits for those patients aged six to 17 with a diagnosis of major depressive disorder with an assessment of suicide risk. So, the population is younger.

Male: So these are kids?

Jack McIntyre: This is Jack McIntyre. This would be the sort of the companion measure in terms of both, you know, bipolar and also for children. This is for adults and major depressive disorder.

What we did in our workgroup in developing this measure had felt like a sense of discussion about if we should also include something in terms of, you know, next – next steps. And we thought that that was – as we went through what the next steps might be with some variable in terms of the setting. And folks have pointed out before the resources available that we're really stuck in terms of how to define it in terms of what the next step should be but felt that, you know, if we're going to make any movement in this arena that we need to first have folks ask the question have said.

Sometimes also, I'm willing to ask the question because I'm unsure what to do next but what we thought that that was not a good, you know, patient standard reason for not asking the question and it does highlight then where the gaps are in terms of what additional services are needed in order to then deal with the answer that you received.

Harold Pincus: And then I – you know, I don't know if I agree with you but I think that you really wanted, you know, provide incentive than expectations for people to ask a question. But I really think, you know, there need to be some

expectation that people take action. If there's, you know, if people had no idea about what the action to be taken is that it really raises a lot of questions (inaudible).

Male: Hello?

Harold Pincus: You know, it really raises questions about, you know, whether it's ultimately going to make a difference in outcomes because part of that, you know, the criteria for evaluating evidence is – is there approximate linkage with outcomes? And simply asking about it, well important is it's still pretty distal.

Jack McIntyre: Yes. Yes, I agree with that Harold but it seems like it's a necessary step that …

Harold Pincus: Yes. I mean, I just would encourage the developers to really work on, you know, sort of like a companion piece to it that, you know, that has some kind of expectation for action. You know, I guess having watched this now for, I don't know, 15, 20 years, anyway, it seems like we should be getting to the point where we expect not just measurement but we should expect the next step. Other comments?

Female: OK. Thanks. Next on our agenda, we have measure number 1880, and the lead discussion for that measure is Parinda Khatri. Parinda, are you ready?

Parinda Khatri: Yes.

Harold Pincus: Adherence to mood stabilizers indicative with the bipolar one.

Female: That's correct.

Parinda Khatri: Yes. OK, so as she mentioned, this adherence to mood stabilizers for individuals with bipolar one and this measure calculates the percentage of adults who have essentially been prescribed a mood stabilizer medication. They have a diagnosis of bipolar disorder. And then they have a measure of adherence. So, what is the percentage of folks who have had at least two claims for the mood stabilize, and the initial in terms – the initial evaluation in terms in importance to measure seemed to be fairly positive from the group

that most people agreed that this is an important measure to report and monitor, and there clearly seems to be a performance gap that has pre-significant impact on patient care and quality of life. OK.

I know that there was one note when I reviewed the note about using kind of old guidelines, APA 2002 guidelines, so people were concerned a little bit about the quality and the kind of currentness of the data. Although I did look up on APA and there – that really the most recent they have for guidelines for bipolar disorder is just 2005 which is a guideline watch update, and in general we're, you know, whereas they had more expanded information. It was generally consistent with the 2002 guidelines.

Number two, the scientific acceptability of measure properties. Again, the group seemed to be fairly consistent voting yes. There were, you know, generally high to moderate marks for reliability and then in general moderate overview for validity. I did have a question for the developers. When I looked at the specific codes that were used for the assessment, they seemed to be based on the old 2012 psychiatric, the CPT codes, and there was kind of a notation later about the ICD9 conversion crosswalk to ICD10 but I didn't see anything about the 2013 CPT codes? Did I miss something?

Kyle Campbell: Yes, so this is Kyle Campbell from FMQAI, measure developer. These measures undergo an annual update so this would represent a measure that would be calculated, you know, as a 2012 measure because it's calculated retrospectively. The 2013 update would be the next annual update for the measure.

Parinda Khatri: Right. So for the annual update, are there – are you going to update the codes? Because when you look at the code, not the 2013 CPT codes, they're not included.

Kyle Campbell: Yes, absolutely. So, for the annual update process, all the codes get reviewed and any updates that are met, that's taken cared of during the annual maintenance and then submitted to NQF.

Parinda Khatri: OK, great. Thank you very much.

Angela Franklin: And Kyle – Kyle, this is Angela. Just to be clear, that updating process is on your end, are you discussing the NQF updating, annual updates?

Kyle Campbell: Yes. So that's on our end but we submit our annual updates to NQF on a regular basis to the fourth quarter of the year.

Angela Franklin: Great. I just want to establish that link.

Kyle Campbell: OK, thank you.

Parinda Khatri: OK. In terms of usability, this did seem to in general, be viewed as being a usable, helpful measure for quality improvement in public reporting although, you know, I think there was some concern that the actual kind of specific components were not really clear in terms of how it would be used at how that – how you would go from having this data and how the impact would – would be a cause in terms of patient, you know, functioning and quality of care. But in general, I think most people agreed that it's at least is a start in those early stages of addressing this issue in terms of quality improvement.

In terms of feasibility, this was again in general viewed to be fairly feasible primarily because of the data is based on claim so you're not having to do chart review or chart obstruction. So, overall, people tended to be in favor of use of this – yet with some general concerns. So, I just wanted to, you know, I guess we open it up and do we want to start with the importance to measure and report?

Harold Pincus: Oh, let's just see. Let's have an (inaudible) there to the general comments that people have with regard to this measure.

So one question I had, is there a benchmark for this measure? What is the expected sort of level that would – that's considered an acceptable level?

Kyle Campbell: So this is Kyle Campbell again from FMQAI. So, as you know, the measure and setup, it's dichotomous and then it calculates the proportion of the patients that are considered adherent at a rate of a proportion of days covered of 0.8. You know, I don't think really the way we were thinking of in terms of benchmarks to be calculated as that physician groups or accountable care

organizations or plans would be compared to their mean and then those statistically significantly different from the mean, either higher in performance or lower in performance, it's really how we would envision, you know, the measure being benchmarked. But in terms of – in terms of the adherence rate for a benchmark, it's 0.8. I don't know if that answers your question but …

Harold Pincus: Do we have the actual data from, you know, surveys of organizations to know or providers to – or patients to see what that – where it is? I mean …

Kate Watkins: The – hi, this is Kate Watkins and yes, we do. The rate of adherence range between 16 and 76 percent and there are a number of different studies …

Harold Pincus: So, in other words, nobody gets to the benchmark?

Kate Watkins: Nobody gets higher than 76 percent. There is no absolute benchmark in the sense of how it's rated. It's considered as comparison to the other – the other organizations or physician groups that are being measured.

Harold Pincus: Kate, just to clarify a bit, when you say from what percent to 76 percent that's – that the proportion of patients who have a medication – position ratio above 0.8?

Kate Watkins: That's correct.

Harold Pincus: Right. OK, all right. So that's – and just a question. So, you know, to think about it, because we're really talking about two different thresholds.

Male: Exactly.

Harold Pincus: So, could you say something about the choice of 0.8 as the threshold? What's the evidence about selecting that as the basis?

Kyle Campbell: So this is Kyle Campbell again from FMQAI.. So the majority of studies that have looked at chronic medication adherence have used 0.8 as a threshold and all of the studies that we site in the form in terms of linking this to improvement and outcomes used 0.8 as the threshold. And so that's why, you know, that's why we selected 0.8 as the code.

Harold Pincus: But, I mean, they had it – they had it all the studies come up with 0.8. What was the right – was there again, some sort of a receiver operating curve kind of analysis done? What was the …

Kyle Campbell: That, I would have to get back to you on. I know just in general in terms of the medication adherence literature, 0.8 has been the level that's been set but I don't know how it was originally set. I don't have to get back to you at the face to face meeting on that.

Male: Yes.

Female: Just curious. That, you know, was that empirically derived or just someone just pulled it out off the (inaudible) 80 percent, that sounds like reasonable.

Kyle Campbell: Yes. And there is – there is some variation in it. So, like for patients with HIV for example, HIV medication adherent, it would set a higher threshold but for the majority of chronic medication, you know, for example, statins and (inaudible) and other types of chronic readministered medication, 0.8 is pretty much been considered the standard in which all the outcomes have been measured again.

Female: OK, thank you.

Kyle Campbell: But I'll bring more information (inaudible).

Harold Pincus: Other comments, questions? In case one, I guess your point is that in terms of among the performance of different organizations, you said that that varied in terms of the population range from – was what it?

Female: 15 to 76 percent.

Harold Pincus: 15 to 76 percent.

Female: Right.

Harold Pincus: And, is there a sense about what, you know, what's considered an optimal kind of a performance?

Female: I mean, I think the optimum performance is probably approaching a 100 percent because – I mean, these are people for whom that's been – has since we limit it to people who have queue filled prescriptions. These are people for whom the physician and the patient have decided that a mood stabilizer is indicated. And, you know, and that this particular mood stabilizer at least doesn't have immediate side effects that are problematic.

So that presumption is that they should be on long-term treatment and that it should approach a 100 percent (inaudible). We're not giving you a cut-off for as, you know, we're not saying you have to be at a particular percentage as being used in comparison with the other organizations you're being compared to.

Harold Pincus: And then one wouldn't expect it to vary by the nature of the population that much?

Female: Correct.

Male: Same number?

Female: It varies by your ethnicity but that is considered to be a disparity.

Harold Pincus: You know, one would think that perhaps low (FTS) or folks who are homeless, or what have you, certainly might have lower adherence.

Female: That's – but you would also hope – I mean, that the intense would be that you would focus on that population, in that population you would want to be as adherent as any other population.

Female: OK, bye-bye. Thank you.

Female: Certainly is the goal but I mean comparing that population to another that's got ready access to mental health services might be apples and oranges, related but not the two here.

Harold Pincus: Right. And then – but, you know, you can imagine many things, may impinge upon it but I guess it's the challenge and the expectations that clinicians and

payer groups and, you know, those that are accountable would be doing something more intensively to deal with that. You know …

Female: Yes.

Male: We're those populations.

Female: Yes.

Harold Pincus: I guess that, you know, the one criteria I have been looking it – you know, you wouldn't expect it to be a 100 percent because there's only going to be some people who, you know, at some point during the measurement period, they, you know, they and their doctor, they know they had prescriptions and decided that they no longer need this. and their doctor, you know, they had two prescriptions in the side of that. You know, I – they no longer need this.

Female: So …

Harold Pincus: And there would be some people who drop off.

Female: Yes. This could have met you – there could be some people who drop off because of patient preference.

Harold Pincus: Right.

Female: That emerges later on in the course of treatment.

Harold Pincus: So, are there any other questions or comments about this measure? OK.

Lauralie Dorian: So, this is Lauralie from NQF. Before we move on to the next measure in 1923, (Natalie), if you're there, I believe we have a group of people who can help us and put in the – on the speaker line there from OptumHealth care. So it will be Rebecca Cate and the rest of her team.

Rebecca Cate: Hi Lauralie, yes, this is Rebecca Cate, and we are all on now. Sorry about that.

Female: Oh, perfect. Wonderful.

Operator: All lines are open, Lauralie.

Lauralie Dorian: OK, great.

Harold Pincus: OK, so I'll – so that's the one that I had to lead on. This is a fairly unusual measure because it is stated that there's no traditional numerator and denominator for this measure. It's basically a measure that's derived from several different psychological tests that have been tested for reliability and validity. And then it is applied to where individuals, patients are received a baseline – receive this instrument at baselines as I understand that it's given out by the provider and into a heterogeneously diagnosed group of individuals aren't receiving behavioral health services.

And then it's given out a second time either by – as I answer by mail or by a computer from the pair or by handed out to the patient by the provider and then it's reassessed over time. And then various statistical elements are done to potentially create in effect sides over the course of time that suggested for both severity and the length of time that existed between the two measures.

What's a little bit unclear to me was exactly how that is reported and used. So in terms of the specific – in terms of a criteria. So, one issue is in terms of importance and I guess, in my point of view, I thought it was important that there'd be some kind of measure of – in a systematic way of individuals. So I think it's important to encourage measurements of clinical severity as, you know, as a guide for longer term management.

Number two is in terms of – although there was no evidence about disparities data about this type of measurement. The linkage between outcomes is it was a little bit unclear. And actually, I had a little bit of difficulty understanding what was expected in terms of this rationale because this measure is not paired with any measure of process to assess whether they're – it's something that's done in between the two measures that's expected to be done that would result in – that would – to which the improvement or lack of improvement would be attributable.

In terms of scientific acceptability, the items for the measure are derived from, you know, series of psychological test that has been reasonably well tested for reliability and validity. But it's – but I – in turn is the measure as used. I couldn't actually find – there was apparently the actual design of the reliability studies that was used to actually report on the reliability. And there were some elements that were – of concern to me.

I didn't really understand why they're – there was no traditional numerator or denominator. It could have seems to me that you could easily provide a numerator, denominator of, you know, the proportion of patients for whom the affect size on the measure was greater than 50 percent. And I'm not sure why that, you know, is not – that it wasn't clear to me how in fact the results of this measure are reported.

There were other concerns that were raised. The fact that almost 40 percent of the sample was being the ineligible for the measure because at baseline they were not – they were below the threshold– the clinical significance which raised questions about, then why were they in treatment? And as I said before, it was not clear how the measure is actually reported either to providers or outsiders. There was something special, that thinking got some kind quality star mechanism but exactly how this measure translator into a quality star was not – was not clear.

Also the fact that in the description of the evidence that they were unable to report anything with regard to disparities due to sample size but in fact they have samples that includes, you know, tens of thousands of people. Yes – or, questions about how the proportion of the provider's caseload that actually reports on the measures factored in. So there were number of sort of, you know, of questions that I have about the – about this that would need some clarification.

Female: So, we do have – Harold, we do have the developer on the line. Do you want to have them way in now or do you want to have discussions?

Harold Pincus: Why don't we – why don't we open up for discussion and then have – have them way in?

Female: OK.

Harold Pincus: I – and just in terms of usability, I guess the – they have before. It's unclear exactly how the results are actually used and presented. So in terms of what people – what decision actually get made on the base of this from how it's used wasn't entirely clear. It did – the developer did say that there were some evidence that actually reporting on results of their patient population to providers actually demonstrated some improvement over time. And so that – that was a positive thing but exactly again, how that process work wasn't entirely clear.

And in terms of feasibility, you know, they're doing it so it obviously it's feasible but I still worry about the extent which it ultimately sort of represent a valid approach to how to sort of asses performance over time when so many people are excluded and also the extent to which the data really represent, you know, true effects of the clinical interventions since the fact that the time between – between assessments could be as low as two weeks, as long as, you know, six months or more.

Female: You know, one of the things I did not see and perhaps it's in here and I missed it where it's been done otherwise is to look at the reliability when applied at a individual or provider level of this and what are meaningful clinical differences. It – I'm worried that given the many transformations of the data, they're required to come up with a severity adjusted effect size that we could be doing one of two very opposite things. One, is to wipe out meaningful differences, or two, to look at differences among providers that are either clinically insignificant or artifacts of what is a fairly complicated statistical process, and what adequate number of patients does one need to have on a provider level to make inter-provider differences that are both real and clinically important.

Harold Pincus: Other comments, questions before we (inaudible) the measure developer to speak?

Male: I just one other – I had was whether all of the methodologies for doing this including the formulary for the severity adjustment, are they in the public domain and would they be available for other people to use?

Female: And I think the other issue is, are the reported measures open to scrutiny by providers who are being measured? Frankly, this seems like a very complicated – it's not black box, certainly very complicated inflows, lots of stuff that out-pops a number. I'm very concerned about the usefulness, the usability in the real world.

Harold Pincus: Are there comments or questions from, you know, people on the workgroup?

Comments from the measure developer?

Rebecca Cate: Oh, hi, this is Rebecca Cate and I was the measure steward but we do have a few of the Optum developers on the line so I want to open us up to them. But we have Francisca Azocar who's Vice President of Research and Evaluation, (Brent Boltrom) who's Director of Analytics, Joyce McCulloch who is Vice President for Behavioral Informatics, and Bruce Bobbitt who is Vice President for Quality Management and Improvement. So I want to open some of these questions and comments up to them.

Operator: And all lines are open.

Joyce McCulloch: And so, this is Joyce McCulloch and, you know, you certainly have offered a lot of feedback and I'm wondering how you want us to respond because I'm not sure really at what point we should respond to some of your comments or how you would like us to orchestrate the respond. Are there specific questions you'd like us to address?

Harold Pincus: Well, I – there are a number of items that I raised and, you know, I – whichever way you care to respond is fine.

Joyce McCulloch: OK. Let me – let me do speak to the – one of the last questions that was just raised around at being in the public domain? Yes, the item that – the outcome instrument that we uses in the public domain, the wellness assessment. And right now – and we would be more than happy to share our methodology, our code for how to actually apply the metric in the methodology, should the measure be enforced. So, yes, it would be made out in the public domain.

And in terms of it not being a traditional numerator and denominator, and I'll be honest, we actually struggled as we worked with this. We believe that this measure is an important measure of outcome. We believe that being able to measure patient outcomes through the course of psychotherapy and to be able to then use a measure that adjust for patient differences within a provider practice and uses an effect side methodology to measure the treatment effectiveness is an important – is an important measure that's within the field.

But it's not – it was not easily narrowed down to a numerator and denominator. And I may need to have my technical person speak in here but again because we're using hierarchical in your modeling, what we end up was just looking at the individual treatment episodes. But then, we're using a modeling technique that allows us to nest that within a provider and look therefore at how a provider's performance measures against benchmark. And in this case, the benchmark being an effect size of greater than 0.5 which is represent – typically in the health services, a moderate effect size.

Harold Pincus: So, why couldn't you say – if I could give a patient, of a provider, with an effect size of 0.5 or greater? I mean, why would (inaudible) numerator-denominator?

Joyce McCulloch: The effect size is being – and (Brent), if you could help me with this one, that the effect size is being measured at the level of a provider not overall and it's – again, I need my modeler to speak up to this to how it comes – but it doesn't come out as cleanly as a numerator-denominator.

(Brent Boltrom): Yes, and part of the issue is that this is a clinician level measure, so that the issue is rolling up the episode, the individual patient, the numerator and denominator to the clinician level.

Harold Pincus: We're given the clinician, a percentage of the clinician's patients who achieve the effect size on the measure of 0.5 or greater. That seems pretty simple but there's something I'm missing.

(Brent Boltrom): I mean, that's something I guess we could consider because that would make it more interpretable to people.

Harold Pincus: Well, I mean, if you don't know, how do you present it to people because it's not that?

(Brent Boltrom): We present …

Joyce McCulloch: What we present …

(Brent Boltrom): Go ahead.

Joyce McCulloch: What we present to clinicians and we are currently using this and sharing it with our clinicians in our network. What we present to the clinician is their overall severity adjusted effect size, and then we also present to them their lower and their upper confidence limit. And we make it clear that the lower confidence limit needs to be at or above 0.5.

And when that occurs, we are able to, you know, that's a measure of meeting the fact that they have met the efficiency – I'm sorry, the quality measure. If they have too much variability and their interval – their confidence interval spans, that then, we're unable to give them the designation. And should their upper confidence level be below 0.5, then that would mean not being effective.

Now we're yet to have a clinician achieve that, so we've not had to run into that. But we're looking for the confidence level around the severity adjusted

effect size which is measured at the clinician's level, not at the individual episode level. And I think that's the point, Dr. Pincus, is at least the effect size is measured at the clinician level.

It's not at – and so, we're not looking at what percentage of their episodes, we're really looking at as the results of the modeling (inaudible) episodes. What's the clinician's effect size?

Harold Pincus: And how do you attribute clinicians in the – at least in some settings, clinicians will share patients or those exclusions?

Joyce McCulloch: The treatment episode is actually initiated by the clinician. So, it's the clinician who administers and the wellness assessment to the patient of course of treatment. That is the clinician who's going to be attributed the episode.

Harold Pincus: Why – what if somebody is providing psychotherapy and somebody else is providing medication?

Joyce McCulloch: It would be attached to the – it would still be attached to whatever clinician is measuring the outcome. So if it is the psychotherapist who is administering it or if it is the – the psychiatrist, it would be whatever clinician is administering the well …

Harold Pincus: What if they both are? I mean, what I refer a patient to Harold and he does medication management and then there's a social worker or psychologist who is doing their psychotherapy, and of course, I'm seeing them in continuity as the primary care doc. Attribution to me is a really important issue here.

Joyce McCulloch: I think – Sorry, is someone else about to speak? OK. Yes – I see – I understand what you're saying and I agree. I think the issue here is that what we're dealing with is that in the course of psychotherapy, a lot of things can happen. Obviously, that can influence outcome, including being placed on medication by either psychiatrist or a primary care physician, including receiving additional support from community services, and including the actual course of the psychotherapy itself.

So lots of things can factor into the outcome. We do our best to control for that and, within our modeling but we're not going to be able to control for all the factors. But what we're – but what we're really looking at is when we look at overall for clinician, we're looking overall at the effect of their treatment episodes and the degree of change reported by their member.

Bruce Bobbitt: Hey, Harold, it's Bruce Bobbitt. Let me just elaborate on Joyce's comment. You know, when – you know, when I take a look at (inaudible) measures which are based on population metrics, they often have discrete events that are measured. And in any natural real world situation, which is the one that we encounter, you know, the kind of work that we're in, you're right. But that would almost be logically true for any of a number of outcome measures when you're comparing something over time to tease out the exact effect to the medication as opposed to the psychotherapy.

It's very difficult when we're looking at it across a wide range of people which is our intent.

Harold Pincus: Are there – one thing I wasn't clear about also is, is this only being applied to people receiving psychotherapy?

Joyce McCulloch: Yes, a patient psychotherapy.

Harold Pincus: And are there any diagnostic exclusions?

Female: No.

Joyce McCulloch: No.

Harold Pincus: So it's a, you know, so it's being applied to, you know, equally to a panel of severe – people with severe schizophrenia as well as the people who have mild anxiety disorders?

Joyce McCulloch: To the extent that the clinician is administering the wellness assessment to those patients. Obviously, patients with strong cognitive disorders may not be

the appropriate patients to be taking a paper and pencil outcomes instrument. But that, you know, again, we're looking to the clinician to administer it and we're looking at it within the heterogenous outpatients.

Female: But with the incentive of getting a star, and if you're leaving the kind of decision to the clinician, how do you account for the bias, the selection bias? So, a clinician who knows that, OK, this person is having a good day today or is responding well, then I'll just give it – person, I'm not going to give it to them. And so, the actual – the kind of evaluation of the clinician or the star rating is really can be based just as much on selection as it is on actually quality of clinical practice.

Joyce McCulloch: And in part, I think it has to be more – more than, you know, we're administering or (assessing) – using as a supplement to your – to you – or at the provider level and given the time to be transparent about how we're using it. We're trying to look at the effects of these to the provider. And so, truly at – to the provider to (inaudible) …

Harold Pincus: It's hard to hear you. You're kind of muffled.

Joyce McCulloch: I'm sorry, what's that? And so, what we're trying to do is we're using an outcomes measure on the way of informing the clinician how well they do in terms of specific patients. But it's really, we're trying to reach (inaudible) about why we use it or – so they can use it online, they can use it in (inaudible) they can use it multiple times (inaudible).

Harold Pincus: For example, do you – does entering it to the formula – I think it's getting a little bit at the previous question. Does entering it to the formula, do you have any information about what – and build in to the measure what proportion of a clinician's caseload is actually subjected to getting the measures?

Joyce McCulloch: We don't have that built in. We could look at building that in. One of our – one of our hopes is that, you know, through the use of this measure and by really being transparent with our clinician network that we are encouraging and would want to see patient informed outcomes within our network that that

– and that that will affect their severity adjusted effect size, that we will get clinicians to administer and participate on a more consistent basis.

But we do actually have and could look at integrating what proportion of their network – of their patients they have administered the wellness on.

Harold Pincus: Do you think that that would be, you know, sort of something that would be really important to know whether they're sort of cherry picking or not? The other question I had is the point about almost 40 percent of the sample was deemed ineligible because they were below the clinical threshold. How do we – how do you understand that?

Joyce McCulloch: That's actually something that, you know, where I have been – I have been comparing notes with some of my colleagues and other entities who all sort of using outcomes instruments, similar to our wellness assessment. And what we have found consistently is that about a third of the populations who respond do fall below the clinical threshold. And that was true in the number of other agencies that I have compared notes with.

And we – we're not exactly sure what that's about. I do think that part of it may be the populations we serve …

Harold Pincus: (Inaudible) of the instrument that like if – if they're below – if they're starting off in treatment below the clinical, the meaningful levels, does that raise questions about the validity of the instrument?

Joyce McCulloch: I don't believe it does. We've vetted this instrument with a number of external psychometricians. And I think it may speak more to the type of population that is typically seen in our network which is an employed population. Many of our – many of our benefit plans begin with an EAP benefit. So these are patients who may actually be accessing services initially through their EAP benefit. So this may be relative – you know, I think what we're seeing is that we have a relatively high majority of patients seeking treatment in our network. So we're actually relatively high functioning and not exhibiting high levels of symptoms.

Joyce McCulloch: Go ahead.

Female: Potentially sort of the worried realm. It's more that they not be clinically severely ill, and where you didn't see it probably very little change in terms of the global distress scale which – score which is really would (inaudible) to look at to determine the level of change clinically.

Harold Pincus: Yes – no, I understand why, you know, you – you know, there's – you know, you've sort of essential have the ceiling effect below that but it's just raising the question about more people that are coming in for special behavioral health services and they don't seem to be of the threshold for distress as measured by this instrument …

Male: So is the instrument able to detect appropriate levels of distress or is it being too stringent?

Joyce McCulloch: Again, we've subjected – we've engaged external psychometricians who've evaluated it, and do believe that it is capturing a range of symptoms and a range of functioning. Again, I think what we're speaking to the fact that, you know, within the population we serve, I believe we have, you know, we are getting a cohort of patients and our utilization data support that. These are the cohorts, the patients who come in and endorse subclinical on the wellness assessment, actually use one to two sessions and that's it.

So these are patients who are not seeking ongoing psychotherapy. These are not patients for whom we would expect to see a great deal of change, but they are a common feature of not only our experience but of other managed care organizations that are doing similar types of outcome studies or are seeing similar patterns in their data as well.

Female: And I think it – that also speaks sort of – the patient population that we are seeing also speaks to potentially a lack of disparity that you are talking about. You know, when we have done studies on looking at our patient population, we find about 80 percent are Caucasian by working folks. And so, it – and

they come to be white-collar workers. So, the patient population I think is reflected on the lack of disparity.

Harold Pincus: Well then, I guess that by that some was – not so much that – I don't think that materials – there was a lack of disparity, and so that there's no data on disparities to the sample sizes. It couldn't make sense to me because some of the – some of the data that was reported was on tens of thousands of people.

Female: No, that it's not in – it's not an item that typically ensures well the plot.

Joyce McCulloch: Right. We just simply do not – we're not given that necessary data. So, while we have thousands of patients in our database, we do not – we do not obtain those necessary data.

Harold Pincus: Oh, because that wasn't clear. You just said that the sample size was too small.

Joyce McCulloch: Yes.

Female: Yes. You know, the – I mean, the struggle of the – using this tool is really trying to find some way of measuring a clinician's effectiveness whether we do it looking at – whether they're using it in space practices or whether they're improving their patients clinically and trying to be transparent about it and make sure that they use outcome measures because at least the research is showing that.

Those who do use outcome measures tend to have – be more effective in efficacious clinicians. I mean, and so, part of it is trying to do that in a way that's fair to the clinician by using severity adjusted effect sizes rather than just using, you know, the raw scores. And so, part of this struggle is, how do you do that with, you know, a population of clinicians that read to, you know, 110,000 clinicians with different kinds of professional degrees and levels of experience?

Harold Pincus: No, I definitely admire your efforts to try to actually instantiate, you know, some serious expectations about measuring. I think it's great. Just, you know,

my question have more to do with sort of like how it actually works and, you know, and thinking about what are some of the threats of the validity and utility of it.

Female: I guess what worries me if you're trying to look at arraying people on a spectrum but the data are self selected by the clinicians. I can't see how that would yield a valid comparison. Maybe I'm missing what you're actually doing in practice.

Female: No, I mean I think it's a drawback of asking clinicians to fill the questionnaire out. On the other hand, we can't obligate them to do it. And so, the intent is really around the more episodes of care that you're able to get into our system with some, you know, baseline and an outcome measure, the more likelihood you are to get a star and the more likelihood we are to refer patients to you because then, it becomes a partnership, a collaborative process in which we're trying to get patients to the clinician and the clinician is letting us know how they're doing.

Harold Pincus: That's what I hadn't realized. So, you're saying that the star is not based just upon their score on this but how many patients they actually enter into it?

Joyce McCulloch: No, I'm not sure that that's – no. I mean, the – for us to be able to measure the severity adjusted effect size, we need to have obviously enough patients for whom the provider has actually administered the wellness assessment that we can actually measure change.

So, by being – by making this a very central measure within our network, what we're really trying to encourage with clinicians is to say, we want to be able to designate clinicians that's effective in our network. We want to be able to give them that recognition. But to do that, we need to have them administer the wellness assessment consistently so that we have enough data points. So, the message is really administer the wellness assessment, make it a regular part of the practice, such that A, to your point, we can eliminate the potential for selection bias, and B, you know, we can eliminate the variability that is naturally going to be found with smaller samples.

So, the more data points we get, the greater likelihood that we will be able to measure a clinician and be able to grant the clinician designation of effectiveness. So that's something – and that's really our measure – our message to the clinician network is, you know, that we want them to partner with us to measure outcomes so that we in turn are going to be able to – grant them designation that they are indeed effective because the data will support that.

Harold Pincus: Other questions or comments by the workgroup members or others on the call?

Lauralei Dorian: Harold, it's Lauralei here.

Harold Pincus: Yes.

Lauralei Dorian: I just noticed in looking at the ratings that – for reliability and validity, that there are a lot – and a few insufficient ratings. And so, I'm just wondering if you or the committees have thought on additional materials that the developers might be able to submit to the entire committee for review that, that would be more helpful.

Or any …

Male: Yes?

Lauralei Dorian: Yes, that was just discussed on this call. Do you think it will be useful for all of them to go back in and add details (inaudible) the mission?

Harold Pincus: Yes, I mean some of the questions that I posted in my comments. You know, in terms of that, you know, the actually design of the reliability studies that was not clear to me, in terms of how that was done.

The other thing is that how the range of providers, case loads that's actually captured in this and how that is adjusted for. You know, the degree to which

they're able to, you know, if not, audit at least sort of, you know, validate that there isn't cherry-picking giving of the way in which the data's collected.

Male: In this interaction between number of episodes who are – individuals who are treated. And then, how that factors into the designation of its (inaudible) and it seems like, at the end of the day, there's the sort of dichotomous yes or no vision.

And I don't quite understand in a concrete terms that you know, if I give you 10 on my best ones, am I going to get it? If I give you 50, how prone or susceptible to gaining is this.

And I guess that'll (inaudible) so the range of – as in the characteristics on patients and whom this has been designated is a little bit unclear.

You know, there are some illusion made about, you know, it will typically – (inaudible) provided to people that, you know, have very severe forms of mental illness. But then if that's the case then, you know, had the clinicians know that, and you know, to what extent could this be a basis for variation of my penalized people that are seeing more severe people because they're not, they've been essentially excluded from that, from the testing.

So those would be the kind of questions that I will have.

You know, I guess a part of this, you know, obviously this is (inaudible) really for a lot about it, and really worked hard on developing it.

And the question is, if it's endorsed, and it's let out to the world, are they, you know, are the people who want to apply it can have that, you know, sort of have the inside baseball knowledge of how – what the limitations are.

Female: Well, we certainly extend the comments from the entire committee. You know, this is a measure we've been working on for over 12 years. We have conducted a lot of research on it, and both that external folks as well as internally.

And, you know, whenever we put it up for external opinion, we, you know, we'd like the challenge. We'd like to hear what others have been thinking about it and what ways it can be improved?

If we continue to use it as a part of the improvement tool to ensure that the quality of care that we provide to our members is adequate. And certainly some of the feedback that we record this day, will help us to continue to improve the measuring. So I thank you for the opportunity.

Harold Pincus: OK. Thank you. And so, we have one more measure to consider.

Male: Yes, this is number 1923, no, not 1923 we only go to this by actual papers. Yes, 2152 which is preventive care and screening – of the alcohol use screening and brief counseling. And this measures the percentage of patients age 18 and older who screen for unhealthy alcohol use which is once during the two year measurement period. There's been a systematic screening method and who received brief counseling and identified as an unhealthy alcohol user.

And I think most people believe that there is an indeed high impact condition here and that there is a fair amount of evidence available. It was interest – reviewed the previous data from the US preventive services task force but more did – most of you today the most recent update which was – as of today on the evidence and I think it's, it's a pretty good and comprehensive review and probably about as good as it gets when we're looking at making the measure to outcomes in the causal pathway. So, I think as far as impact and evidence, there appears to me to be fairly good and strong.

With regard to the scientific acceptability, reliability and validity, it was a little bit more askew but in general the agreement, clearly when you're doing this in a chart review format, it's going to be more challenging. But as we transition more and more to standardize tools embedded in (EA charts) won't certainly concede that this will become more easily accessed and that there are going to be a fair amount of reliability in this and fair amount of phase validity. Usability …

Angela Franklin: Dr. (Inaudible), can I you know, just a sec, this is Angela.

I just like to remind everyone if you're not speaking to place your phones on …

If you're not speaking to the measure please put your phones mute. Thank you.

Hello?

Female: OK, well I'll try to carry on here.

Female: And (Natalie), if you can help us with that fix.

Male: Despite the messages from (Mars) or wherever they're coming, the usability – I think that there is pretty good evidence that this is a measure that had some, you know usability that it's currently in PQRS and it provides some meaningful data to permission in patients and that there is probably a number four, relative feasibility of this measure.
So the vote came out to be five to zero here.

You know, I guess if, you were extremely stringent in your application of each of these, you could find faults but as far as I am concerned, my sense is, this is about as good as it gets today in the behavioral health arena and I'm happy to try to answer the questions or hear what other folks thought about this one.

Male: So, other comments to our group?

Harold Pincus: I had one comment or question that – about sort of going back to the issue about measure alignment. There are a number of other measures that are similar and how does this fit in with regard to other similar measures? The fact that this one is a chart review on – and as I understand, there are also our electronics specifications for this. But are there others that are capturing similar concept that might be, that might be more or less feasible?

Angela Franklin: Harold, this is Angela and Lauralei. We do already have a measure that was endorsed in the first phase of the project and that's, that's the NCQA measure that goes to the percentage of adolescent and adult members with a new episode alcohol or other drugs dependents to receive care. And then there is a suite of joint commission measures that are intended to capture some of the spectrum from assessment through follow ups. And those are in this phase of the project and we'll be reviewing those on later calls.

Female: And just to explain to you the NQF process for harmonization. We will be looking in this project at measures that are related to one another. So, if you have the one that you're discussing today and then this – the joint commission measures and during our in-person meeting after both are – after all those measures are evaluated and they're recommended for endorsement, that's the point of which we'll have a discussion about harmonization and how those measures can be better align and then for measures that are not in this project, which Angela just mentioned, there was one that was endorsed in the first phase, you can discuss harmonization of the measure that will take place probably later on in this year during the annual update period.

So, it's a sort of off-cycle. It's a new – we've changed things around a little bit. So, we will definitely have a dedicated time slot to discuss related measures during the up coming in-person meeting.

Male: I have one question, how much should the question of what's best in class influence us?

Female: You have to evaluate, you have to evaluate each of these measures on their own merit first then you can look at the best in class discussion.

Male: It would be helpful to have that you know, because either two which is – one is what's left in class and the other is if things are equally rated, you know, are they measuring you know, is there some similarity in the way the different numerators and denominators are defined so that people aren't comparing apples and oranges.

Female: Well, we will be providing that kind of comparative information at the in-person meeting. The thinking behind having each measure stand on its own is that we want to be sure each of the components we've looked at before one of them can be knocked out.

The harmonization process best in class will happen at the in-person once we know what the merits of each measure are.

Female: And the comparative table with the measure specification, so all of those that might be related or competing will be sent out to you prior to the in-person meeting with the in-person meeting materials. And we do have strict guidance that will be sent to you as well that will help you walk through that process. OK, thank you.

(Sam Kearney): This is (Sam Kearney) with the (AMA) could I ask something real quick to the other point?

Female: Sure, go ahead.

Female: Thank you. I understand that the discussion of harmonization will come up later but just for the committees, for the committee to be aware of the measures that were described that are somewhat similar to the alcohol measure from (inaudible) which is OO4, deals with the patient population that has a diagnosis of alcohol or substance abuse. And it really focuses on initiating treatment and ensuring that patient stay-in treatment. So it's fairly different than this measure.

The other measure from the joint commission and substance use screening and treatment, there are two measures that are similar but those are focused on a different setting of care in the hospital level and I believe they also focused on a different level of measurement at facility level.

So I just wanted to provide that additional information.

Female: Thank, (Sam). And in those charts of comparison chart (inaudible) provide all that information to the steering committee.

Male: Other comments about the measure?

(Jeff), I don't know whether you've mention this but, given the fact that it's a chart review measure. What did the data show about reliability?

I would have to go back and …

(Jeff): Maybe the measures developers can tell us that.

Male: Would (inaudible) care to answer that question?

Female: I'm sorry, can you repeat the question please?

Male: Given the fact that it's a chart review measure, what were the data about reliability?

Female: You know, I think I don't have a form in front of me and like I'm having computer problems at the moment. I apologize, but I believe that the – it's demonstrated moderate with reliability.

Male: In numbers are better because there's been a lot of fuzziness about the concept of moderate.

Female: The numbers are better, OK. You can give us a minute. We can pull that out.

Male: Yes, cap is – are better.

Female: Sorry we're just pulling it up right now.

(Dan): A cap of 0.82 is on page 19.

Female: Thank you.

Female: Thank you.

Female: Thank you, (Dan).

(Dan): And I'd also point out that the specs certainly appropriate for chart, but there are electronic specifications that will certainly improve their reliability.

Male: Yes, if I remember it correctly there was some comparison of chart abstracted versus electronic.

Male: Well 0.82 is actually quite good for chart.

Male: Yes. Yes I would concur. Any other questions I can answer?

I want to thank everybody. Is there anything else that we need to do?

Female: We'll just pause for a moment to see if there's any NQF members or members of the public who are on the line who would like to make a comment?

Male: The line is open, I guess.

Female: OK. It sounds like that's a no. We thank you Harold for running this meeting. And thank you everybody for participating so well. We though we – it was a really great discussion and we're really looking forward to seeing it at the in person meeting.

I'd like to remind you I will send all of this information out in an e-mail. But if you do want to change your votes, you can do that by going to back to the (SurveyMonkey) by next Friday. And if you could just put a two after your name when you answer your name, just so that we know that it's your second time (inaudible) votes for that measure.

At the beginning of this month, you should have received that logistics e-mail from our meetings department here, so that you can setup your travel arrangement. We will be starting at 9:00 a.m. both days of the meeting, but we will have a continental breakfast at 8:30 a.m. And following all of these

work group projects are going to be summarized in exactly what was discussed and re-circulating the information to you. And for the measures for which you are lead discussant, you'll be expected to introduce those measures at the in-person meeting as well.

Male: It's going to be at the NQF offices?

Female: It is. At our headquarters here at …

Male: The hotel is going to be …

Female: What was the question? Sorry.

Male: Do you know where the hotel is?

Female: The hotel is actually right around the block from the NQF headquarters, we think.

Female: We're not sure, our meeting team will follow up very shortly if they haven't already, please e-mail us, they should have sent out an e-mail already. So if you can send an e-mail to Lauralei that would be great, if you haven't heard from them yet.

Male: Yes, I think they didn't say what the hotel was.

Female: OK. We'll follow up with them and perhaps send it to you in an e-mail.

Male: OK. Thank you.

Male: What time do you anticipate we'll be done? I'm sorry that maybe in some of the communications?

Female: No, that's OK, the first day ends at 5:00 p.m. and the second day ends at 4:30 p.m.

Male: OK, thank you.

Female: That’s June (7, 6).

Female: Great.

Female: So I think that's all from us. Thank you very much again and thank you to the developers who dialed in and we're willing to answer questions. And we are looking forward to seeing you here in Washington.

Male: OK, thank you very much for your support.

Female: Thank you.

Male: Thank you.

Female: Bye.

END