Web recording: http://eventcenter.commpartners.com/se/Meetings/Playback.aspx?meeting.id=398544

NATIONAL QUALITY FORUM

Moderator: Perinatal -
April 14, 2016
12:00 p.m. ET

OPERATOR: This is Conference #: 66719618.

Suzanne Theberge: Good afternoon everyone and welcome to the second workgroup call for the Perinatal and Reproductive Health Committee. Thank you so much for joining us today.

This is Suzanne Theberge. I'm the Senior Project Manager on the team and I am joined by the rest of my colleagues on the team, Reva, Kaitlynn and Nadine.

And before we get started, I'm going to do just do a couple of quick housekeeping notes, and then we'll see who's on the phone, and then we'll go into the content.

So, as with our prior calls, we ask that you put yourself on mute if you're not speaking to reduce interference. And that if you wish to speak on the phone, you do need to be dialed in to the conference call, not just on the webinar. But if you have the webinar and the phone running at the same time, please turn off the sound on your computer speakers as we'll get a lot of interference if that's on. And we also request that you not put us on hold because then we'll get your hold music.

If you have any questions during the call, committee members, feel free to speak up. You can also raise your hand on the webinar. But really, just feel free to jump in with comments and questions.

And with that, I think we can see committee members, if you could let us know if you're on the phone.

Sheila Owens-Collins: I'm in group four.

Suzanne Theberge: I'm sorry. Who was that? Who was in group four?

Sheila Owens-Collins: Sheila Owens-Collins.

Suzanne Theberge: OK, hi. Thanks for joining us. Yes, we did ask that, you know, other committee members are welcome to listen in. But do we have folks on this workgroup?

Matt Austin, are you on the line?

Matt Austin: I am on, yes. Good afternoon.

Suzanne Theberge: Great. Deborah Kilday?

Deborah Kilday: Yes, good afternoon.

Suzanne Theberge: Hi. Kristi Nelson?

Kristi Nelson: Yes, I'm here also.

Suzanne Theberge: Great. Juliet Nevins?

Juliet Nevins: I am she.

Suzanne Theberge: OK. Cindy Pellegrini? Cynthia said she was going to be calling in a bit late, so hopefully she'll join us later.

Diana Ramos? Rajan Wadhawan?

Rajan Wadhawan: Yes, I am here. Just a disclaimer, I am traveling to India and I don't have internet access unfortunately right now.

Suzanne Theberge: OK.

Rajan Wadhawan: I hope it will be OK if I'm just going to be on the call.

Suzanne Theberge: Oh absolutely. Thanks for letting us know.

And Carolyn Westhoff, are you here?

Carolyn Westhoff: I am here.

Suzanne Theberge: Great. Great. Do we have any other committee members listening in?

Ashley Hirai: Yes, hi. This is Ashley …

Carol Sakala: Yes, Carol Sakala is here.

Ashley Hirai: And Ashley Hirai, hi.

Suzanne Theberge: Great. Thank you. Welcome.

So, we have asked our developers colleagues to join as well, so I know that some of them are on the phone. If you have questions about their measures, hopefully they can assist and hopefully be available for that.

So, before we get started on that content, does anybody have any questions on anything be fore we get started, committee members?

Matt Austin: So, this is Matt. Where do we find the results for the surveys that we filled out?

Suzanne Theberge: Those have now been incorporated into the measure worksheet. So if you click on the measure worksheet and you scroll through and see, and actually Kaitlynn, if you could pull up 716 since that's the first one. We can show that they will be in the pink boxes.

Matt Austin: Great, thank you.

Suzanne Theberge: At the end of each criteria.

All right, so I think the first measure that we have under the session is, as I said, measure 716, Unexpected Complications and Term Newborn. This is a maintenance measure. It's a hospital level outcome measure, which is reported as the percent of incidence with unexplained newborn complications among full term newborns with no preexisting conditions. This was a, as I said, it's a maintenance measure, so it's previously endorsed under the title, Healthy Term Newborn. And the developer had changed that.

So, what we're going to do first is take a look at the importance, starting with the evidence. And at this point, I would like to ask the folks who have then the lead discussants for this measure, if they would like to make any opening remarks. And so that's Diana Ramos, Juliet Nevins, Carolyn Westhoff and Cindy Pellegrini. Would any of you like to start off the discussion?

All right, so we'll be – what we're looking at here is this is an outcome measure. The developer has provided updated evidence. And what we'd like to know is what the committee thinks about the new evidence. It seems like it's basically in the same evidences prior, going in the same direction, but it's updated. It's a little bit stronger.

Does the committee have thoughts about this they wish to discuss?

Carolyn Westhoff, I know you're here on the phone so I will call on you.

Carolyn Westhoff: Yes. Well, I mean my main question, reading all of this was that, in a sense, that there examples of actions that could influence these outcomes, their examples were limited to both Group B Strep testing and treatment during labor and the use of – appropriate use of vacuum.

And so, I thought that was maybe a little limited. I mean, as an OB, I was saying, "Really? Is that a substantial – would that account for a substantial proportion of these complications?" But in fact, I think they provided a lot of evidence and there are other interventions as well. They gave a long list of practice guidelines that could potentially have an impact.

I, you know, I couldn't find, given that this is a maintenance measure, that they actually analyzed whether application of these guidelines would in any way explain the variation they've identified. But there's – but they seem – you know, there are plenty of interventions available to improve this, and there's plenty of variability among the hospitals, and so it all seemed quite plausible to move forward. I think that, you know, the new name is fine.

Suzanne Theberge: Thank you. Juliet Nevins, do you have anything to add?

Juliet Nevins: Yes. I was actually talking a lot, and then I realized I was still on mute. But I just wanted to add that, you know, they did stipulate or did they just say that this was the balancing measure. And what's interesting about what they're trying to measure is that there are lots of different interventions around these types of income. So, I think this is fine in terms of the evidence and the additional information that they have offered to support this measure.

Suzanne Theberge: OK. So, the next piece of the importance criterion that we look at is there a gap. And that would include whether there is opportunity for improvement as well as issues around disparities. What were your thoughts on this performance that's been reported on the measure thus far?

Juliet Nevins: So, I commented that they gave us results from 2011 and 2012, but I just wondered if there was – I mean I guess if they were available, they would have provided that. But, you know, since there are so many initiatives out there to reduce the types of outcomes that they were looking at, I just wondered if there was any more modern or more recent data was available. I do think though, however, that this is the kind of measure that could be used to help elucidate disparities that definitely exist, right? This other member mentioned variability in terms of these outcomes, and I think this is the type of measure that we could use to kind of delve more into that.

Suzanne Theberge: So, I think (Elliot's Nane) is on the line. I do see him at least on the webinar. Do you have any comments about whether there might be updated evidence, (Elliot)? Operator, if you can make sure that he has an open line?

(Elliot Nane): I had to get off my mute as well. Good morning everyone. I guess it's still morning in California. We are running this measure every six months on the 500,000 births in California that we have annually. So, there is data. The data that was submitted here was initially run and submitted for an earlier attempt at a earlier ad hoc updated this measure because we did change it (inverted this), to make it more understandable.

There is interesting disparities data that we can provide, and I can send that in before the May meeting. The – in fact, I think there is some of this that you're showing out on the screen here with – yes, exactly, those African-American women are considerably statistically higher than non Hispanic (whites). And we can update this. And actually, this is easy to run. But this is the stuff that we had run in our interim submission.

In terms of the question that Dr. Westhoff asked about potential impact, we think this has a lot of potential areas for improvement. The ones that were the most documented were GBS and in vacuums that were the cleanest ones to submit. But there is evidence that people can reduce birth injuries which would fall under this with safety programs. So, that's been reported not with this measure but using other similar measures that are, we think, a little too simple for use, such as the AHRQ measure for birth injuries, and so that would fall into this as well.

But this real pick up, the breadth of unexpected outcomes, that include respiratory outcomes that you see with early gestational age C-sections, the 37, 38 weeks develop more respiratory morbidities, so would fall – that's an impact that would – that is reflected in this measure. Sepsis both from – related to chorioamnionitis and newborn sepsis, they are our opportunities to improve that are reflected in this measure, as well as birth injuries themselves. So, you know, as we strike toward lower cesarean birth rates, we want to be sure that we don't have more difficult outcomes for babies, and therefore that's an example of how this can be a balance in measure.

Suzanne Theberge: Great.

(Elliot Nane): Great, thanks.

Suzanne Theberge: Thank you. So yes, we can speak after the call about getting that additional data, the newer data for the committee, and we'll share that with folks prior to the in-person meeting.

Do any committee members have any other comments about either the evidence or the gap before we move on to the scientific acceptability?

OK. So, the next piece that we'll look at is the reliability of the measure. This measure is based on administrative claims and birth certificate data. As we've just discussed, it's been inverted since they are previously endorsed version.

Do folks have thoughts on the changes to the specifications, the data elements, the logic in there, calculation, algorithm?

Juliet, do you have any comments on that?

Juliet Nevins: You know, I think that adding the birth certificate data, for me, was very crucial because administer – and I think I wrote this in my comment that, you know, administrative claims don't always necessarily line up with what happened on L&D and what's in the clinical documentation. So with, you know, with more use of electronic medical records and just electronic documentation, I think that will definitely increase the accuracy of the information that where which we're retrieving and using for the aggregate data.

Suzanne Theberge: OK, and they have – oh, go ahead.

Carolyn Westhoff: Yes, this is Carolyn and I just had a question about the two submeasures that was expert advice to do that split and what – I mean, this was a wonderful long submission. And if there's any preliminary data on severe versus moderate, I didn't see that.

(Elliot Nane): We do have extensive data on severe and moderate and actually on additional submeasures like respiratory versus infection …

Female: Yes.

(Elliot Nane): … these birth injuries, which are very useful in terms of interpreting the measures for quality improvement activities. It's one thing to have a rate but the other thing is to understand where your issues are, so then you can build out for improvement.

Carolyn Westhoff: No, I agree. What I'm really asking is did you put any of those data into this application? But it may be that I just didn't find it.

(Off-Mic)

Carolyn Westhoff: Sorry.

(Elliot Nane): I don't believe it's in the application per se.

Carolyn Westhoff: OK.

Suzanne Theberge: (Elliot), if you want to add, if you, you know, have time to run that and add that in to the additional materials that you're sending, we can certainly share that with the committee as well.

(Elliot Nane): Sure.

Suzanne Theberge: OK. So, the measure was tested using signal to noise testing for the reliability. Do folks have any thoughts? Was the test sample adequate? So, it seem like the measure is sufficiently reliable? Any comments on the testing?

Hearing none, the next piece is the validity testing. They were updated both empirical and face validity testing results submitted. And those results, any comments on those results?

Female: No.

Suzanne Theberge: Does it seem – go ahead.

Matt Austin: So, this is Matt. A question and I'm sort of trying to read on the flag here a little bit too. Was the testing for reliability and validity done with the re-specified measure, or was done with the current, or the older version of the measure?

(Elliot Nane): With the re-specified measure.

Matt Austin: OK. Thank you, (Elliot).

Reva Winkler: This is Reva. I just wanted to mention that we really are going to be looking at the in-person meeting for your thoughts around, you know, does this testing demonstrate to you that the measure results are reliable and valid, and therefore appropriate to be used to make comparisons and conclusions about quality? So, think of it in those terms when you look at the testing results, are you convinced that this measure produces reliable and valid results? Thanks.

Suzanne Theberge: So, the last piece of reliability and validity that we look at is first, the validity, and that includes exclusions and risk adjustment. There are five exclusion categories for this measure with – and the developer has included frequency for them. Are they consistent with the evidence? Should there be additional exclusions? Or, are there exclusions included that should not be?

Kristi Nelson: This is Kristi. Are they going to do ICD-10 codes? I only see 9 listed as …

(Off-Mic)

(Elliot Nane): Yes, we have completed the transition to ICD-10, and we have actually tested that. We have 100,000 patients (versed) with ICD-10 data. And we've tweaked it so we actually get the same rates on a longitudinal way with ICD-10 that we get with ICD-9 for the main measures and the submeasures.

Kristi Nelson: Thank you.

(Off-Mic)

Suzanne Theberge: Go ahead. I'm sorry.

(Elliot Nane): We can send those codes in as well, the ICD-10 codes.

Suzanne Theberge: They should be posted on SharePoint. We will have to check after the call and confirm that, but we should have included that code table in the materials. It wouldn't be in this particular measure worksheet document. It would a

separate – most likely a separate Excel tables. That's generally how we post those.

So the – oh, go head.

Matt Austin: This is Matt. I may jump back – real quick back to validity and reliability testing, and I'm getting sort of maybe too focused.

Suzanne Theberge: Of course.

Matt Austin: Was that also done with the inclusion of ICD-10 or was that tested with ICD-9?

(Elliot Nane): Those were tested with ICD-9. We only got ICD-10 data within the last month or so. So, we started in March and the hospitals reported to us more recently, but the numbers actually almost identical, so we're comfortable with that.

Matt Austin: OK, that's good to know. Thank you. And when you say almost identical, (Elliot), meaning in terms of the results for each hospital?

(Elliot Nane): Yes, in terms of timelines. So, we would look to see each quarter, whether it had changed, and then we look to see whether they changed ICD-9 to ICD-10 and they were consistent.

Matt Austin: OK.

(Elliot Nane): A quarter before and a quarter after, which was very reassuring.

Matt Austin: Yes, no, that is great, thank you. Just a …

(Crosstalk)

Suzanne Theberge: So, the next – oh, go ahead.

So, the next piece that we want to look at is the risk adjustment. As we went over in the orientation and the Q&A calls, NQF is now looking at risk adjustment for SDS factors. This measure is not risk adjusted and is not – and

what we would like is your opinion on whether the rationale for not adjusting this measure for SDS factors, whether they're – or it's not a contextual basis for that. What are you thoughts on that?

Carolyn Westhoff: Hi, this is Carolyn. I have one dumb question. What is SDS?

Suzanne Theberge: Oh, I'm sorry. NQF speak, sociodemographic factors

Carolyn Westhoff: Oh OK, good, so like SDS.

Suzanne Theberge: Yes.

Carolyn Westhoff: OK, fine. So now I understand what that is. And I have one question about it. When I looked at the breakdown, and I can't find the page right now, maybe – but the hospital, it was interesting that at page three actually, very early on, the hospitals with the regional ICU were – so this isn't really SDS. It's like level of care, but they had slightly higher rates and, you know, they were sort of at the 75th or 80th percentile. And I wondered if that might mean that there's some risk stratification going on about which patients end up in those hospitals, that beyond what the exclusions, and if that's something that requires attention?

(Elliot Nane): I'm happy to respond to that. That is the area that has caught our attention as well. And we've been running the extensive analyses on that. Actually, within each of those NICU level categories which we're taking to reflect maternal levels of care indirectly, there's a big range. So, you'll see some very low hospitals and some higher hospitals.

Our strategy is as we move to use this more publicly. We've been giving feedback to hospitals now for two years in California. It's all 250 hospitals and (inaudible) been using it.

Our strategy is to stratify it further by NICUs. We compare NICU levels to NICU levels, or these level three and level four there. When – and there will be some issues going on exactly as you identified, that no matter how we do the exclusions, we're not getting – there may be more clustering of other

higher risk factors in level three centers, but there is still is a range variation there which is kind of interesting. So, that is an area that we are pursuing.

Juliet Nevins: This is Juliet and I also had another thought on this. Am I on mute or am I am being heard? Hello?

Suzanne Theberge: We can hear you. Yes, we can hear you.

Juliet Nevins: OK, great. I had another thought with respect to the SDS factor because I believe that Austin, you know, that is a variable that affects both the moderate and severe complications, given I think what is generally understood as the population with the higher rate of obesity, diabetes, and pre as well as gestational hypertensive disease, right? So, when you look at that and you look at all the attended management that comes with it, the stat C-sections, the difficult deliveries of the baby, there are macrosomic, I just wondered if, you know, doing the risk adjustment would help to flush out some of those disparities that are based on the type of patient who's walking out in the door.

(Elliot Nane): This is a complex issue as to whether you would risk adjust for race, for whether you risk adjust for those diabetes and hypertension conditions. We don't want to do away with the opportunities to see disparities on care if they exist, but we want to be fair about it. I mean, the biggest disparities with African-Americans, for example, is with prematurity and small for dates babies, which are excluded under this. So, those are known issues so that everybody would be expecting.

Hypertensive moms, hypertension, we really shouldn't be expecting some of these complications on, and should, you know, good prenatal care, good delivery care should alleviate some of those risk factors in terms of their baby outcomes, accepting, you know, the severe ones like small for dates and prematurity. So, it's – I'm on two sorts on this one, to be honest.

Juliet Nevins: But, you know, the reason I asked this question is because, you know, I worked for – I'm a Medical Director, right, but I still work as a hospitalist in a city hospital in Brooklyn. And the hospital is located in an area where we get

five distinctly different racial and ethnic groups, distinct, right, Arabic, Russian, Chinese and Latin American.

And, you know, even with, you know, we have all kinds of initiatives going in terms of when to do C-section or doing all kinds of mandatory training on reading, you know, the scripts, we're doing, you know, vigorous identification of patients who need GBS testing. I mean, all of those things, you know, are in place.

But I – the reason I brought this question up is because I feel that the outcome has a lot to do with not necessarily the race of the patient, right, but the overall health of the wellbeing that's walking in the door. And a lot of those patients are not getting prenatal care where they just, you know, come off, get off the plan from wherever, from Yemen, right? So, we've never seen them before and these are people with all sorts of, you know, morbidity if you will.

So, that's why I'm asking, not necessarily because I want to flush out disparities. That's a secondary outcome, if you will. But I am just curious as to whether this kind of data would help us kind of narrow down the – I mean I think we already know what those are in terms of the most significant underlying health factor that affects the outcome of healthy, you know, what should be a healthy full term neonate, and that's why I'm asking the question.

Suzanne Theberge: So, I hate to cut off discussions because this is important and also very interesting, but we also have a number of other measures to get to, and we still have some more to discuss on this. I think this is an issue that we want to be sure to flag for the full committee in the in-person meeting and have some more conversation about there and, you know, and continue to discuss it. But I also want to make sure that we have time to discuss the other remaining measures on the call.

So, with that said, I just want to jump on to the last couple of questions under scientific acceptability, that as meaningful difference and missing data, and see what the committee's thoughts are on whether this measure can demonstrate meaningful difference and quality, and whether there are any

concerns about missing data. We do have a comment about missing data and adding data from electronic medical records.

Would any – and I'm not sure who made that comment. Would that person would like to expand on that? And the comment that I have is that adding data from electronic medical records may enhance the validity, what is billed and what is documented is not always congruent, but EMRs are not universal yet. And there are many different platforms, so this could also lead to feasibility issues, basically the essence of the comment.

Any further thoughts on that as an issue?

So, hearing none, I would like to move on to feasibility unless anybody has any other burning questions on either reliability or validity. So, for feasibility, one thing that we know in maintenance measures is that we wouldn't possibly expect more information on implementation issues. And we did see some of that from this developer.

One thing that they noted was the measure originally is patient discharge data. But they had some difficulty identifying term infants and they – once that data was going linked with birth certificate data, we are able to get better – more accurate and complete results. And the other item that the developer noted was code, in fact, is just very – some facilities over code, some facilities under code. And they have made some adjustments to account for this.

So our question for the committee is, you know, do you think that these adjustments to the measure have addressed these feasibility issues? Do you think that, generally, this measure is usable for you to use?

Cindy Pellegrini: Good afternoon. I'm sorry to interrupt. This is Cindy Pellegrini. I just want to let you know, I just dialed in. And could you tell me which measure we're on?

Suzanne Theberge: Of course, yes. We are talking about 716, which was the unexpected complications in term newborn.

Cindy Pellegrini: Great, thank you.

Suzanne Theberge: So, any thoughts on feasibility for that measure?

OK, hearing none. The last criteria that we're going to look at for this measure is usability and use. And this is a maintenance measure. We really want to focus on that, more than we would with the new measures since we would expect – you know, we would hope that the measures being used, they might want to see how that is playing out in practice.

So, you know, I think we've already discussed usability to some extent. Its come up earlier. But what does the community think? You know, we discussed how this is – the developers encouraging this be used as a balancing measure. It has been used and improved on based on user feedback. And the developers are expecting to begin public reporting with the measure this year for California with 2015 data.

Any thoughts on usability, whether the benefits of the measure outweigh any potential unintended consequences, whether the measure has been vetted? Any thoughts?

Matt Austin: So, this is Matt. I'm looking at the feedback item about the MAP revealing the measure, and the technical concerns. Were they specific with what those technical concerns were?

Suzanne Theberge: I don't have that data in front of me. And I don't know if any of my colleagues know, off the top of my head. That's certainly something that we can pull up prior to the in-person meeting.

Matt Austin: Yes, maybe interesting to understand what their concerns were, and whether those have been addressed or – and maybe this committee feels differently than what the (MAP's) out. It might be …

Reva Winkler: Hi Matt, this is Reva. Yes, number one, the MAP saw – looked at this measure, first version, when it was a healthy term newborn probably about three or fours years ago. And I think one of the issues with that was people were just not used to looking at a measure of good. And the conversation was around, "OK, that's nice. But I'm more interested on what's – where the

problems are." And I think it might be also instructive to ask (Elliot), why did they decide to change the construct of measure to inverted because, perhaps, it was something similar along those lines.

(Elliot Nane): It was that kind of feedback, exactly. And it's sort of an interesting conceptual view of a quality measure if you're looking to something that's 97 percent. That seems like an awfully good grade. But if you're thinking of a 3 percent problem rate, that's a bigger issue when you're talking babies. And it was – so 3 to 5 versus 95 to 97. It is an interesting frame. And we got a lot more traction of people looking at it from a quality improvement standpoint when we inverted the frame and framed it as unexpected newborn complications.

So, it's the same concept. Everything is similar, you know. You know, we did tweak it as we went along and actually tweaked some of the codes. Learning how codes were actually used code by code in the real world, and that's the other significant adjustments is the inversion and then adjustments to the real world use of some of these ICD codes.

Suzanne Theberge: Great. OK. Well, in the interest of time, I think we need to move on to the next measure and …

Cindy Pellegrini: Could I ask one quick question just for the record?

Suzanne Theberge: Of course.

Cindy Pellegrini: Sorry. This is Cindy. And we don't have to discuss this but I just like it to be out there. When I was reviewing the reliability testing, it generally looked very impressive. But in looking at the chart of the individual hospital results, reliability did below the threshold in a number of cases. Most of which have low numbers of birth that were – but there were more than 200 births, so like between 200 and 500 or maybe 600. And so, I'm just wondering. Again, (Elliot) doesn't have to answer now. But for the future, right now, the recommendation was that the measure should be limited to hospitals with more than 200 births. Should that be a little bit higher like maybe 500?

Suzanne Theberge: OK. Well, I think now that – to bring this to committee's attention or, you know, you raise it again. Once we get to the in-person meeting, we'll definitely want to look at that further.

(Elliot Nane): Happy to comment then.

Suzanne Theberge: Great. Thank you. Well, thank you very much for your time, (Elliot). And now, we would like to move on to the next measure which is 1382, percentage of low birth weight births. This is another maintenance measure, another outcome measure. And looking at the percentage of births with birth weight less than 2,500 grams.

The developer states that there has been no changes to the evidence since the measure was last updated. And so, I'd like to turn it over to the committee for their comments, their thoughts, whether you think that there should be additional new data – evidence, or are you comfortable with the data that we have on the evidence, and whether you are comfortable with the idea that there are actions within the healthcare systems that can affect the results of the performance of this measure?

Our lead discussants for this measure are Carolyn Westhoff, Kristi Nelson and Cindy Pellegrini. Would one of you like to start off?

Carolyn Westhoff: This is Carolyn. I'm enthusiastic about the measure and having it, you know, achieve maintenance. The one thing I was interested in, because there was no new evidence presented, but given in the O.B. world, there is a lot of attention to protocols that involved delaying elective induction and scheduled C-sections. I didn't try looking it up myself, but I was actually a little surprised that, I mean, because those interventions would adhere to – have an impact on this measure.

Suzanne Theberge: Do we have the developer on the line for this?

CDC, do we have anybody on the line from CDC?

All right, I'm not sure if we have somebody. If you are here, please let us know. Either raise your hand on the webinar or please try and get your line

umuted. I think, you know, we'll flag that question and make sure that we share that with the developer.

Carolyn Westhoff: Yes. I don't know if there's anybody on the committee who, in fact, will be more familiar with the literature on this.

Rajan Wadhawan: This is Rajan Wadhawan. I doubt that would have any significant effect on low birth weight infants. The elective C-section rate, really, if you look 36, 37 weeks gestation baby, unless your SGA, you're over 2.5 kilos. I'm not familiar with the literature, but my thought is that probably it won't have any effect.

Carolyn Westhoff: Well, no, the problem is that people misestimate in fact. But yes, I don't know what evidence there is. I think one of the reasons for the big push to delay the elective induction and delay scheduled C-sections is because there is substantial misclassification of gestational age and birth weight.

Rajan Wadhawan: Sure. All I was trying say was that if you're above 34 weeks, the mean, gestational mean birth weight for that maybe would be over 2.5 kilos. Although it may affect the birth weight, it may not affect the proportion of kids under 2,500 grams. That's what I was trying say.

Carolyn Westhoff: Yes.

Suzanne Theberge: OK.

Sheila Owens-Collins: Yes, I'm (inaudible) and I agree with that.

Suzanne Theberge: OK, thanks.

Sheila Owens-Collins: Sheila Owens-Collins, yes.

Cindy Pellegrini: Thank you. So, this is Cindy.

Suzanne Theberge: Oh, go ahead.

Cindy Pellegrini: Sure. As the other lead discussant, this was actually the only measure that I didn't end up having any questions about. That it seems like it was a very well

established measure. There are no changes being proposed. So, I'm comfortable with approving it again.

Suzanne Theberge: And just the committee, moving on to the committee, think that there's still a gap. There sounds like there's an opportunity for improvement or there are disparities issues here that still need to be addressed.

Cindy Pellegrini: Absolutely.

Female: Agree.

Carolyn Westhoff: Yes, I agree.

Suzanne Theberge: All right. I think there were no real comments …

(Off-Mic)

Suzanne Theberge: Oh, go ahead.

(Elliot Nane): If I could ask one quick question. Is this meant to be a population measure or a hospital level measure?

Reva Winkler: Yes. Hi, this is Reva. This is a population level measure. NQF does endorse a population level measures because, again, it provides a different perspective on the overall quality of care that's provided. So – but this one is definitely a population level measure.

(Elliot Nane): Perfect, and I think it's highly appropriate as that.

Suzanne Theberge: OK, moving on the reliability and validity. Again, this is a maintenance measure. The developer did not submit additional testing data other than what was submitted previously. They performed data element validity testing. And since we don't have any new testing data, does the committee agree that the criteria has been met?

Female: Yes.

Female: Yes, I would agree.

Female: Yes.

Suzanne Theberge: OK, and again, with validity, any thoughts on the validity of the measures?

Carolyn Westhoff: The references that are presented in the application were, you know, solid a decade ago, and they remain solid to in my opinion.

Suzanne Theberge: OK, and so …

(Off-Mic)

Suzanne Theberge: Oh, go ahead.

Matt Austin: Maybe a question for the lead discussants, those leading the discussion, did they look at – I mean, obviously, their main data point is collecting the birth weight. But when we look at say, for instance, like the race of the baby, did they address that as well in terms of ensuring that that is sort of reliable?

Carolyn Westhoff: I didn't understand your question.

Matt Austin: I guess my question is, you know, obviously, the main data point they're collecting is birth weight, but then that birth weight gets stratified by different sociodemographic variables, one being race. So how good are the race data that they collect? Did they address that at all?

Carolyn Westhoff: I don't think they addressed it at all. It hadn't occurred to me to – I think it's a good – I mean, it's …

Reva Winkler: This is Reva. This is a measure that's based on – for vital statistics and there are standard definitions. We can look up what those are for vital statistics data in terms of collecting race and ethnicity just to be complete.

Matt Austin: OK, or if you can even just point me to something that might better educate me, that'd be helpful. Thank you.

Suzanne Theberge: Sure. We can – yes, we do that prior to the in-person meeting. So, the – if there's no further questions on reliability or validity, you know, we can move on to feasibility. This is, again, it's based on birth certificate data. Is this – does this look feasible? Are there data elements and data collections? Are they ready to be used?

Female: I would agree.

Carolyn Westhoff: Yes.

Suzanne Theberge: OK.

Female: I just have a quick question. I know that in Maryland, the birth weight is on a birth certificate but not the gestational age. Would that be a problem?

Suzanne Theberge: I am not sure. Reva, can you speak to that question? Or, is that something we would want to spend – sent to the developer? Reva, are you on mute?

Reva Winkler: Sorry. Yes. I got distracted for a second. What was the question?

Female: In Maryland, the birth weight is on the birth certificate but not the gestational age. And so at which, you know, I'm trying to help them with that and that's a big disadvantage. Is that a problem with this measure, if you're using birth certificate data only? Because in the State of Maryland, you won't have the gestational age, you just have the birth weight.

Reva Winkler: I mean, let's take a look at the specifications because my understanding is birth weight is the only data element. I don't believe they are including gestational age.

Female: Yes, I didn't see gestational age anywhere on the data.

Female: OK. All right. Thank you.

Suzanne Theberge: OK. And the last criteria is usability and use. The measure is currently publicly reported and CDC does report improvement results from 8.26 percent in 2006 and it's now dropped to 8 percent in 2014. So, they are demonstrating some improvements. And I think the questions that we would want to look at

here for this maintenance measure is, why are the measure information is useful and meaningful for patients, to payers, to policymakers? That if the measure could be used to further the goal of high quality healthcare? Are there any unintended consequences? Any thoughts on this?

Female: I think it's a good starting point, you know, for your population. And that, you know, decide the (inaudible). I'm just kind of – like if it's the first the birth or a second subsequent preterm, you know, birth. You know, there's other things that you could do versus prenatal care, like, (17P) and stuff like that. But I think it's a good starting point to see what your rate is and then see how you population is (affected).

Suzanne Theberge: OK. I think we can move on to the next measure unless folks have any final comments on this.

OK, hearing none. I think we can move on to measure number 0304, and that is a late sepsis measure. Do we have the folks from Vermont Oxford Network on the line?

Female: Yes.

Suzanne Theberge: Great. All right, so this, again, is another maintenance measure. It is an outcome measure, looking at the standardized morbidity ratio, and observed minus expected measure for nosocomial bacterial infection after day three of life and very low birth weight rate infants.

Our lead discussants for this measure are Rajan Wadhawan, Matt Austin and Deborah Kilday. The developer has provider updated evidence for this measure since the prior review in 2012. And I would like to ask our lead discussants, what your thoughts are on the updated guidelines and the new information that's been presented on the evidence.

Matt Austin: So, this is Matt. I guess I'll go first. I thought the updated evidence seem to enforce the notion that there are specific quality improvement interventions and activities that hospitals could put in place to improve their performance on this measure.

Rajan Wadhawan: This is Rajan Wadhawan. I agree with that as well. I think that it's certainly additional evidence that's there's a possibility of impacting this outcome and multiple ways of doing that.

Suzanne Theberge: OK.

Deborah Kilday: And this is Deb Kilday. I would support that as well. It just reinforces some of the earlier evidence.

Suzanne Theberge: So, in this case, you know, and just the maintenance measure, we would, you know, not – we don't really need to discuss it further if it's previously going to – if it's going to continue to pass the evidence.

I think we can move onto – excuse me. We can move on to the gap and care on the opportunity for improvement. We do have some tables in the measure worksheet that demonstrates the performance how in the gap and care over the last several years, from 2006 to 2014, and disparities data as well for that time period. And so what we would like to ask the committee, again, here is whether there is still a gap in care, and whether the measure does provide information about disparities in care.

Rajan Wadhawan: Hi, this is Rajan Wadhawan. There is substantial variation as has been reported between units in the sepsis and meningitis. And so, there are clearly significant variability in care and there's opportunities for improvement.

Matt Austin: Yes, and this is Matt.

Rajan Wadhawan: Based on this data for …

(Crosstalk)

Rajan Wadhawan: Sorry.

Matt Austin: Yes. I mean, there still seems to be some level of disparities across different subgroups, although that seems to be shrinking overtime, which is I think a good thing. You know, if you look at black non Hispanics there, the rates

went from 0.234, almost cut in half, to 0.126, and so much more – or closer to each other than we were seeing back in 2006.

(Off-Mic)

Suzanne Theberge: And the survey comments that we received from the committee members do seem to agree that there is still a gap in care and there remains room for improvement here.

Reva Winkler: Yes. Suzanne, this is Reva.

Suzanne Theberge: Yes.

Reva Winkler: I just wanted to ask just for context. This is all data from the Vermont Oxford Network. And if the developer could just tell us, you know, what the – how many hospitals participate and provide data to the network, and whether that's increasing so we can get a sense of how these numbers represent, you know, what portion of the population.

Suzanne Theberge: I think we have someone from Vermont Oxford on the line. Are you …

(Crosstalk)

(Erika): And I thought that I unmuted myself and I didn't, so I'm sorry about that. I was merrily talking away.

So, the network actually does continue to grow. There are currently, in 2014, we had 938 hospitals that contributed data. About 2/3 or actually – I'm sorry, 3/4 of those are in the United States. And we have a wide range of different types of NICU use ranging from those that are the quaternary level centers that perform cardiac surgery requiring bypass to your typical NICU that does surgery on infants to those that may not do surgery. So we, overall, we estimate that we collect data on over 85 percent of the very low birth weight infants that are born and admitted to a NICU in the U.S., and we continue to grow.

Rajan Wadhawan: I just have an additional question for the Vermont Network people and this is a question about the database. I can't recall. This measure was based on birth weight on 401 grams to 1,500 grams. Although somewhere in the submission, it states that Vermont Oxford Network collects data on 501 to 1,500 grams on the VLBW registry. I can't recall based on the database, which one of these is true. I think it's 401 to 1,500, but we should probably clarify that.

(Erika): Right. The eligibility for VLBW is 401 to 1,500 grams or 22 to 29 weeks gestational age. The risk adjusted measures are calculated on infants, 501 to 1,500 grams, and this is a risk adjusted measure.

Rajan Wadhawan: OK. Yes, thank you.

Suzanne Theberge: So it actually sounds like, the committee is interested in discussing the specifications and some of the reliability and validity, so I think we can move on. I think, was the question raised on the specifications, has that been clarified? We …

Rajan Wadhawan: Well, you know, that just brings me to the question, if the risk adjustment methodology is developed on kids, 501 to 1,500 grams, can it be reliably applied to different birth weight category which is 401 to 1,500?

(Erika): It likely could be. We have not – just historically, from the time the network started in 1990, the risk models were run on infants, 501 to 1,500 grams. And unfortunately, that continues. We'd never extend (at) the model to 401.

Suzanne Theberge: So, that might be something we want to flag for further discussion at the in-person meeting. Are there other thoughts on the specifications?

Matt Austin: Yes, so this is Matt. This is maybe my own lack of clinical knowledge. But when I look at criteria number two, specifically item number two, one or more signs of generalized infection. I'm just wondering is that – is there more sort of vagueness in that than maybe that that would be desired? And I don't know how that gets – I mean maybe there is sort of a standard definition to everyone's mind of what generalized infection, how that would manifest itself. But that was an area where I was like, "Huh". I'm wondering if that might – could be tighter in terms of the definition.

Suzanne Theberge: What do our neonatologist committee members think?

Rajan Wadhawan: This is Rajan Wadhawan. I'm a neonatologist. In my mind, this was the hardest piece is to when you get a one blood culture that is positive for coagulate-negative staph, trying to determine whether this is a true infection or not. But if you look at all these generalized signs as you mentioned, all of these can be associated with infection, but they don't necessarily have to be (associated) with infection. They can be (fine) otherwise as well.

And I don't know what else we could do to improve the reliability of calling this coagulase-negative staph as a real infection versus a contaminant. But this is probably what we use as the best case scenario to define and decide.

And then also third piece to it, that you have to have an intention to treat. Even if you have these signs and they are attributed, go to something else and the physician decided not to treat and the kid survived and did fine, that really truly wasn't a coagulase-negative staph infection.

(Erika): Right, it probably could be clear that the coagulase-negative staph, part of the numerator, an infant has to have all three positive blood or CSF culture and one or more signs of generalized infection and treatment with five or more days of antibiotics.

Rajan Wadhawan: Yes. That's why – that's how I (inaudible). Like correct, right, you have to have all three, if you are claiming (staph) infection. Yes.

(Erika): Yes.

Rajan Wadhawan: And I think that's the right thing to do.

Male: OK. I'm a neonatologist. And I have more times than not, been convinced that there's infection but the blood culture was not positive. So, you know, I understand that, you know, you're sort of concentrating on specifically staph coagulase-negative, but is there any consideration for that, that they have two out of the three, they could have, you know, two of the signs of sepsis but not a positive blood culture?

(Erika): No, the definition of …

Male: OK, that would confuse – that would confuse the picture too much …

(Erika): Yes, exactly, exactly.

Male: Yes. OK. All right, it's fine.

(Erika): And this measure also does – it's not just coag-negative staph. It's any – it's a bacteria or pathogen in blood or CSF, but it does – we do need a positive culture.

Male: OK, got it. Thank you.

Suzanne Theberge: Any further thoughts on the specifications before we move on to the testing? So – go ahead.

OK. So, the developer did provide some updated reliability testing at the measure score level. However, NQF staff noted that the reliability testing was not performed with the data source and level of analysis indicated for this measure. It was a split half analysis. And they provided aggregate reliability. And there was a note that the developers concluded the correlation coefficient for lower than expected, which contest that the definition may not be applied in the same manner across all infants at all hospitals. And they did notice the coefficients increased as the number of infants at the hospitals increased.

So, what are the committee's thoughts here? Do you feel that the test sample is adequate and generalized for widespread implementation? Is the measure going to be reliable – and sufficiently reliable to demonstrate differences in performance? What are the thoughts here?

Matt Austin: So this is Matt, maybe a clarification for our VON colleague. I guess it was unclear to me how many hospitals they used for the samples because it sounded like within any hospital or a center, they used 100 as a sample, but over how many centers?

(Erika): Yes. So, what we did was we actually started with a number hospitals that closed out for the year, and we ended up needing – we dropped hospitals that had fewer than 10 infants. And I don't know how many ended up getting dropped. It probably would not be that many, but I can – I'll double check and I can bring that to the in-person meeting.

Matt Austin: So there was – OK.

(Erika): And then for each hospital, we randomly divided the infants in half, and then compared the rates in other group and looked at the correlations over the years. And I was the one that wrote that. Yes, I was surprised that the correlation – that the correlations were lower than I would have expected. So, the correlation coefficient in 2014, well, it's 0.63.

Matt Austin: And your hypothesis is that centers may not be applying the definitions in the same way?

(Erika): Sorry, go ahead.

Matt Austin: Yes. So yes, can you talk more about that? That would be helpful.

(Erika): Well, from what I know about the split-half analysis design, you would expect – the closer you get to one with the correlation coefficient, the more highly correlated or, you know, the responses are, or that the groups are. So, between group one and group – random group of infants one and random group of infants two at a given hospital. So, I would – given that we have this standardized definition, I would've expected that the correlation, you know, that in either half of the sample, infants should have been equally likely to have an infection or not, and that the correlation coefficient should have been higher.

I do – once we – the larger the hospital got, so you know, you increased the sample size, the higher the correlations got into the 0.7 and 0.8 range. So, there could be some of this low – some random noise that's contributing to this lower correlation coefficients.

And certainly, I – with the definition that – with these definitions that are kind of complicated than a little – get still a little squishy, you know, it may be that it's not being applied in a way that – in a standardized way as we would have expect it.

We have – since we have so many hospitals, we really rely on those hospitals to review our manuals of operations that they received every year to ask us questions about the data definitions if they're aren't ensure. And we haven't talked about auditing at the hospital level, but it is a fairly challenging thing to consider.

Matt Austin: Yes, absolutely. OK. Thank you.

Suzanne Theberge: OK. So, I think in the interest of time, we should move on. The next piece is validity. This measure would assess for face validity at the prior review. We don't have any updated validity testing. Any – do folks have any quick thoughts on the validity testing or are you comfortable with moving on?

Rajan Wadhawan: I'm comfortable with moving on. This is Rajan.

Suzanne Theberge: OK, great. So, the next piece of the trust to validity, there are number of exclusions. Are there any concerns there? Are the exclusions consistent with the evidence?

(Off-Mic)

Rajan Wadhawan: Yes, there appear to be, I don't have any concerns personally.

Matt Austin: Yes.

Suzanne Theberge: OK.

Deborah Kilday: I didn't have anything, no.

Suzanne Theberge: OK.

Matt Austin: They seem to be either maintaining the percentage or maybe decreasing slightly. So, that's probably indicative of them being appropriate.

Suzanne Theberge: Great. OK. So, risk adjustment, statistical model, risk adjustment with no SDS factors, do folks have thoughts here? We did get some comments on this. Do you agree with the developer's rationale, there's no conceptual basis for risk adjustment on SDS factors?

Rajan Wadhawan: Yes. That I do agree with.

(Crosstalk)

(Erika): So, this is – sorry, this is (Erika) at VON. We used to address for race. And I think that that's what the star indicates for next to birth year 2012 is we based on that conversations from the last NQF round, we dropped (phrase). It's the only SDS, "SDS" measure that we collect, because others are too complicated for hospitals to collect. And the model actually performs much better now without it for what that's worth.

Suzanne Theberge: OK. Does folks have any further thoughts?

Matt Austin: Hey (Erika), can you repeat that again? You said the model performs much better with it to without it?

(Erika): Without it.

Matt Austin: Without it.

(Erika): It would be – race is not – race is an insufficient, race alone is an insufficient sociodemographic measure. And it's also its maternal race. And obviously, these are infants, so it's determined by either asking the mother. Or, it's just really not a sufficient and probably unreliable measure ultimately for a network or a registry like ours.

Matt Austin: OK.

Suzanne Theberge: OK. And finally, any thoughts on whether this measure does show meaningful difference and performance?

OK, hearing none. Let's move on to feasibility. This is a measure that's being from clinical registry data, and there are no fees to use this measure. However, members of the Vermont Oxford Network do pay an annual membership fee to be in the network.

So, do you folks have thoughts about that? Any feasibility concerns?

Cindy Pellegrini: I had a question here which is tied actually to a related question usability, but it wasn't clear to me, the extent to which the data being gathered here through the VON is more granular or different from that, which might be available through other resources, whether it's EHR, customary EHRs, or claims, or things like that. Could the VON representative address that?

(Erika): Sure. Our – I guess that in – our take would be that this data are coming from the clinical record and not from billing, claims, administrative data. So from that sense, hopefully you're getting something that's, hopefully more accurate, hopefully more granular than you would in another realm. However, I would imagine, I have not looked – our measure is not ICD code based and I haven't looked closely at the new ICD-10. It may be possible to achieve a measure like this using ICD-10. I'm sure it is. It has to be. But in general …

(Off-Mic)

Rajan Wadhawan: Yes, I think, you know, one …

Cindy Pellegrini: … extracting this from clinical record, specifically to report to VON, right?

(Erika): Right.

Cindy Pellegrini: Thank you.

Rajan Wadhawan: I think we have to get this level of granularity no matter what you did, because if you look at the definition that is in this is the – if you have coagulase-negative staph, you have to have two other factors for it to be called infection. And this (prospect) to reflect the data. I think was probably the best we can get. Knowing that it still has its challenges, it is probably that, still, that we can get from an infection standpoint.

Suzanne Theberge: OK. So, last criteria to look at here is usability and use. Again, this is a maintenance measure and we are looking to for an increased emphasis on usability and use at this point. NQF has endorsed this measure since 2007, and the NQF criteria for usability and use is looking for performance results are used in at least one accountability application within three years after initial endorsement, and are publicly reported within six years after initial endorsement, or that the data on performance results are available. However, at this time, we don't have any public reporting or experience, ultimately, of use and then accountability program.

Do folks have thoughts on that? You could just likely …

(Off-Mic)

Cindy Pellegrini: So, I think it's a question, Suzanne, for you and Reva. I don't have any quibbles with the value of this measure as being an important issue around instant health, but I know – or as I understand it, it's NQF policy that you don't endorse measures that are only going to be used internally for specific programs. And I'm wondering if this cleared that bar or not because based on what we just heard, it doesn't seem like anyone outside VON could easily use this measure. So, this is really going to be a VON specific measure that would be difficult or impossible to use in any kind of publicly reported program.

Suzanne Theberge: Well, I think – and Reva, Reva will probably have more to add. I mean, you know, as you know, the usability and use is not a must have criterion, although it is important. So, I think, you know, the committee could rate it low but also still choose to recommend the measure if you felt that it was, you know, it had – to met the other criteria adequately, you know. But again, we do hope that we do look for things that are being used and that folks can look too for information.

So Reva, do you have anything to add about that?

Reva Winkler: No. I mean, I think you've basically covered the bases about use and usability. I mean NQF, originally from its origins, has been promoting, you

know, public reporting and certainly use of measures that we endorse for accountability purposes. There are lots and lots of measures out there than can be used for quality improvement but we're looking to identify those measures that can be used in the accountability realm, and that includes public reporting. And so, I think this is something that the committee needs to discuss. And again, as Suzanne said, you know, not passing usability criterion is not fatal to the measure being endorsed, but it's certainly is something that needs to be identified and considered.

(Crosstalk)

(Erika): And I would say that – I'm sorry. This is (Erika) from VON. I would say that this question was actually really hard for me to answer because I know of hospitals that are publicly reporting this measure to – on their website in their annual reports to constituents, but Vermont Oxford Network does not publicly report it.

So, on a hospital by hospital basis, there are hospitals that are using it. And if we ever get to a point where we have CMS measures in the neonatal population, we would hope that something like this would be considered in that realm as well, and we're actively working on that. Sorry.

Suzanne Theberge: All right. I think this is something, again, that we will flag for the committee to discuss, and that that is useful information that you may wish to mention, (Erika), at the in-person meeting. So – but in the interest of time, again, I want to move on, make sure we have enough time to discuss everything. We still have two measures left.

The next measure is 483 and that's also a Vermont Oxford Network measure. And I am just going to the right place in my document there. This is, again, another maintenance measure. It's a process measure. Looking at the proportion of infants 22 to 29 weeks gestation, who were screened for retinopathy of prematurity. Our lead discussants for this measure are Juliet Nevins, Deborah Kilday and Kristi Nelson.

So, you know, I think the area that we want to start with, again, is of course the evidence. And the developer does state that they're having no changes to the evidence since the last review in, I think, 2012.

Are there thoughts on – you know, it's based on clinical practice guidelines, studies or systemic review. Do you have any thoughts on that?

There is, I believe, a change in the guideline but it looks like things are kind of moving in the same direction. Any thoughts on evidence?

Deborah Kilday: This is Deb and it appears – I'm sorry. It appears as though it's going, again, directionally the same as the previous evidence.

Juliet Nevins: I agree.

Suzanne Theberge: OK. In terms of gaps, do you still – there has been some improvements over time, but do you – are you still seeing that there's a gap, that there is room for improvement here, or that there are – any thoughts on that?

Juliet Nevins: They're up to 74 percent and that's over a 9-year span. So, I think that given the importance of this testing, I think there are certainly room for improvement.

Suzanne Theberge: OK. And the developers, thus, report that there's no difference by race. So, there is – it seems like, perhaps, there's not race disparities but, perhaps, low birth weight babies. That might be an area where there's a gap in performance. And I think …

(Crosstalk)

Sheila Owens-Collins: Well I think – this is Sheila Owens-Collins. I haven't read through it, but the other potential gap is the availability of pediatric ophthalmologist, which the last time I looked at that, they were the ones that could do these exams. And so in rural areas, that could potentially be a problem. I was in a rural area and we purchased a (ret cam) to bridge that gap. But in terms of, you know, areas of improvement and potential problems, the qualification of the person who does the exam may be (one).

Suzanne Theberge: OK. I think that's a good point and one that, you know, if the developer has any information on that, that might be useful to bring to the in-person meeting if you have any kind of geographic data.

Sheila Owens-Collins: OK. All right, I'll do that. Thanks.

Suzanne Theberge: OK, great. Yes, we'll follow up with you later, (Erika).

I think the next, unless there aren't further thoughts on evidence or gap, I think we can move on to reliability and validity. Again, this is part of Vermont Oxford Network measures, so it is a registry data and it's a process measure. So, any thoughts here? There were some updated reliability testing provided. Again, this would have analysis testing for the measure score level.

Folks have any thoughts …

(Crosstalk)

Juliet Nevins: I've got a question. With respect to the information we agreed from the registry, and certainly that would be reliable given that it's coming for a registry. But in terms of which neonates are being – or babies at this point, right, 22 to 29, which are being tested. The accuracy of that information, is that – would that be coming from the registry as well in terms of who is being tested and who is not, who actually have the exam by the pediatric or ophthalmologist?

(Erika): We don't ask who does the exam and we don't ask how it was done. We just ask whether they have an exam. So …

Juliet Nevins: And is that something that's gotten from the registry that the exam was completed?

(Erika): Yes.

Juliet Nevins: OK. OK, so both information is received, so the gestational age of the child and whether or not test was done are both through the registry.

(Erika): Right.

Juliet Nevins: OK.

Rajan Wadhawan: And you also ask the question as to when the test was done because that is one of the criteria? Whether it was done at the right time or not?

(Erika): We use the gestational age at birth, and we – whether the infant was in the hospital during – at the recommended postmenstrual age. Yes.

Suzanne Theberge: We did get a question here on the specifications in the committee pre-call survey about if the facility is not a VON center, would you need to exclude transfer patients? (Erika), do you have thoughts on that?

(Erika): Well, that if an infant is not in the hospital, that they're recommended for postmenstrual age because he or she was transferred out, then we don't ask the hospital that did the transfer of hospital A to contact the hospital and find out if they had the exam, unless the baby was re-admitted to hospital A. So, that's how we end up losing some of the infants.

Female: What about the ones that were out born and transferred in, they're still hospitalized, (inaudible) screening time?

(Erika): They are recorded. Those get counted. So, they were born during the specified time and they're in your hospital.

Suzanne Theberge: OK, moving on to validity unless there are any further questions on the reliability?

This measure, and similar to the last one, with also face validity testing was performed. And are there any thoughts on that? Do you folks have any concerns about the validity of the measure?

Juliet Nevins: No.

Deborah Kilday: No.

Suzanne Theberge: OK, and I think we actually have just covered this in terms of the exclusions. They do exclude for transfers, and there is some data in there in the worksheet about that.

Any concerns on that exclusion?

Rajan Wadhawan: Just one of the question about exclusion, so same, (submitted) to what I asked before. The exclusion for an infant having an exam but have done at their own time, is that just based on the yes or no question, whether exam was at the right time or not? Or does ask for whether the exam was done at six weeks postmenstrual age for babies under 28 weeks and so forth?

(Erika): No, it's based on the birth weight – all right, I'm sorry, the gestational age at birth, and whether the infant was in the hospital at the time that they have the exam, so.

Rajan Wadhawan: Sure, but that's just one criteria of whether they have exam or not. But the exclusion also says that the infants would be excluded if they had an exam but at their own time.

(Erika): Right, so, if they were not in …

Rajan Wadhawan: So, how would that be determined?

(Erika): Right, that's I'm just trying to come up. I'm not sure that that exclusion actually would apply. And I apologize for that because they – if they're not in the hospital, we wouldn't know if they weren't at the hospital at the right time. We wouldn't know if they had the exam.

(Crosstalk)

Rajan Wadhawan: Sure. You know you could still have a problem though because you may have an infant who was born at 22 weeks and he was supposed to get his first eye exam at 34 – well, at 32 weeks (corrective) gestation but he didn't get until 35 weeks, that is still a problem. And – but that those infants should not be excluded. And based on this, it seems like that those infants might be excluded because the infant had an exam, but had it at 35 weeks instead of 32,

they would be excluded. So, this may need some clarification so to what this exclusion really means.

(Erika): Right. I need to go back and clarify that.

Suzanne Theberge: So, if you can get us that information prior to the meeting. (Erika), we can share that with folks along with the other updates that we'll be sending.

(Erika): OK.

Suzanne Theberge: All right, this measure is not risk adjusted. It's a process measure. And (inaudible) we can move on meaningful difference. There is a table of results provided. Any thoughts on whether this identifies meaningful differences about quality?

Juliet Nevins: I mean, I think it certainly does.

Deborah Kilday: I would agree.

Rajan Wadhawan: Yes, absolutely. I agree as well.

Suzanne Theberge: OK. So, I think the last two criteria are feasibility and usability and use. I think, you know, they're going to be somewhat similar in the registry related issues. We did – I have a comment, identifying the feasibility issue for this particular measure. The question about how to ensure that we are getting the correct data on which babies are tested or not tested is still not clear to me. And also, there's a comment, it may require some manual extraction to obtain the retinal exam.

Does the committee members who made those comments wish to expand on that all?

Deborah Kilday: I just know in our facility that the information – this information often has to be manually abstracted, so.

Suzanne Theberge: I'm sorry. It was a bit hard to hear you. Could you repeat that?

Deborah Kilday: I just – I know in our facility that some of this information needs to be manually abstracted from the chart.

Juliet Nevins: I think that is probably true that – and that's true of most places.

Deborah Kilday: Yes.

(Erika): This does seem like another very VON specific measure. And I just kind of – I wish there's a way to make it more broadly applicable and usable.

Juliet Nevins: I think …

(Crosstalk)

Juliet Nevins: I think that this one is more easily made from the – broadly applicable from the standpoint of the hospital – it's something that they could be keep – any hospital can keep track of this fairly easily. I agree that the infection measure is very much a VON specific measure. But this one is something that hospitals could be keeping track of – should be keeping track of whether they're VON members or not.

Suzanne Theberge: OK. Are there any further thoughts on this measure before we move on to our final measure?

OK. Thank you. Hearing none, I think we can move on to the last measure of the day, and actually our first new measure that has not previously been endorsed. (Erika), thanks very much. And let's talk after the call about that additional information.

So the next and final measure is measure 2895, Thermal Condition of Low Birth Weight Neonates Admitted to Level 2 or Higher Nurseries in the First 24 Hours of Life. And this is an intermediate clinical outcome measure, looking at temperature on admission to a level 2 or higher NICU for neonates less than 2,500 grams.

Our lead discussants for this measure are Rajan Wadhawan, Matt Austin and Diana Ramos. Because this is a new measure, we do look a little bit more

carefully at the evidence since it hasn't been previously reviewed. And so, I think we'll start there. Any of our lead discussants have any thoughts on the evidence of the – for this measure?

Rajan Wadhawan: This is Rajan Wadhawan. I guess I can start on this. I think there is the conceptual model makes sense. There's a correlation between neonatal temperature and neonatal outcomes. There is enough published evidence that also supports that, may not entirely in the low birth weight category but certainly in very low birth weight category. There is – there are strong data that hypothermia remission impacts outcomes adversely. So, I think there is enough evidence that this is a (process) intermediate measure that is of value in evaluating outcomes.

Diana Ramos: OK …

Matt Austin: This is Matt. One question I have and I would need to do a little more research on this myself is, is there something about a level 2 make you higher that makes this – like, could it be appropriate in other level 1 NICUs? And why would we restrict it to level 2 and higher I guess?

(Larry): Would you like me to speak …

(Crosstalk)

(Larry): This is (Larry). Sorry.

Rajan Wadhawan: I have a question with that regard as well. But if you could please elaborate on that, maybe that'll answer my question too.

Matt Austin: Sure.

Female: Sure …

(Crosstalk)

(Larry): Sure. We were looking at infants who became sick in the first 24 hours of life, sick or were small enough that they needed to be – needed the supportive care of a special care nursery. That was how we defined the population. I

actually think this would be a useful measure in younger children as its own measure or fully stratified differently – in level 1. But there had been previous discussions when VON had submitted temperature measures, that these were essentially rare outcomes. They were a failure. Really, it was really a failure measure before we used that term when VON did that.

Here, these are failure measures but they're actually common failures, and that's why we did it that way. I think that there's a good conceptual argument, but we didn't do the work to validate it. We have the resources we had, and this is also how our expert panel guided us.

Rajan Wadhawan: My comment on that regard is that this is potentially open to gaming and influence based – not even gaming, it's just based on the influence by the practice, you know, particularly unit. The birth weight criteria for admission to a level 2 nursery varies from place to place. Some place would admit everybody under 2,000 grams. Most would admit everybody under 1,800 grams. Some may go to lower.

So, it depends on how many kids you admit and the higher birthrate category within this group could influence your numerator and denominator significantly, and may falsely show whether you're better or worse than what you really are. Because most of these kids in the 1,800 and 2,500 gram category are going to go newborn nursery and will never be counted in the numerator or the denominator. Those kids are also less likely to be hypothermic as compared to infants who are under 1,500 or 1,600 grams. So, I think that is one potential problem that we see in this.

(Larry): Yes. I think that the last point you made is, to us, the key point. That the most vulnerable population are those who make it into the level 2 or higher nursery into the special care nursery, and therefore that's where the value for this measure lies. Keep in mind, if someone goes into the level 1 nursery and crashes and is admitted within the first 24 hours, they qualify for this measure. It's admitted within the first 24 hours. It's not only admitted directly from the nursery.

Rajan Wadhawan: Sure. But I think I'm not sure if you understand my concern correctly. My concern is not the kids who crash and come into a level 2. My concern is the kids who come into level 2 just because the difference in local policy, who are healthy …

(Crosstalk)

(Larry): I understand that. One of the things we advocate is performing stratified analyses, and birth weight category would be one. So therefore, there would be the opportunity in reviewing it by the accountability agency to compare like to like. I think that including level 1 nurseries creates all sorts of other problems well beyond the solution that it offers in terms of lessening gaming. And as I said, this is – we had a national expert panel. This was their judgment as well.

Rajan Wadhawan: Was there any consideration ever given to restricting this measure to VLBW and not LBW?

(Larry): Yes, there was criterion discussion about it, and it was felt to not be a good idea, because what we actually felt is that there's a tremendous amount of occult risk, occult meaning not recognized in the larger infants.

And a point in fact, in our data, looking at three hospitals in New York City, we found the risk of hypothermia on mortality to be meaningful in that LBW group. So there, we have internal data that supports that. I'm not sure if we shared that because we weren't thinking that was going to be a focus. But if need be, we could. So, we have some of that data.

Suzanne Theberge: Yes, you could pull that together and send it to the staff. We can share it with the committee prior to the meeting.

(Larry): OK, great.

Suzanne Theberge: So, you know, again, in the interest of time, we don't have much of it left in today's call. I want to move on to performance gap and see if the committee has any thoughts there. I think we have discussed that a little bit

already, but what are your thoughts on whether there's a gap in care that warrants the national performance measure?

Matt Austin: This is Matt. In terms of the disparities, are they able to provide anymore detailed data? I mean, they sort of summarized about the findings were around disparities, but it'd be interesting to see those data like we've seen further measures, so stratified by race.

Suzanne Theberge: There is some additional information other than what's in that worksheet summary if you scroll down to or click on the link, the one before.

Matt Austin: OK, great.

Suzanne Theberge: There's a bit more information. But I said, developer has additional information, we'd be happy to provide that.

(Larry): We have some …

Carolyn Westhoff: Hi, this is Carolyn …

(Larry): … but we don't have others. What we have, we're happy to share.

Suzanne Theberge: OK. Yes, (Larry), if you could send that to us, that would be great.

So, you know, we have about 10 minutes left for measure discussion on today's call. So, I think we should move on to scientific acceptability, reliability and validity and see what our committee members think.

This measure is a little bit different, and that it's reported as a distribution. And so, that's a little different than our typical measure performance. And so, we'd like to know whether that's – what your thoughts are on that? Is that an appropriate way to report on an accountability measure or the category is appropriate? What does the committee think about the specifications of this measure?

Matt Austin: So, this is Matt. I don't have enough notes myself to comment on the categories themselves. It sounds like from previous comments that, perhaps, the national panel helped establish those in terms of actually using a

distribution, I just – in terms of actually someone who does public reporting of measures, I just don't know how one would report this out in terms of benchmarks or targets. I mean, obviously, it seems like where we'd (won) is they used to be in the about right category. I don't know about overly warm that sounds somewhat alarming, but maybe I'm misinterpreting that.

So, I think that suggestion, perhaps the measure, could be reconfigured as percentage of babies that were cool, very cool or cold, or it can be reconfigured as percentage that were about right or warm.

(Larry): Yes. This is (Larry). We spent a lot of time talking about this. And this actually, again, came out of our read of the minutes of NQF conversations when VON had submitted a temperature measure, a dichotomous temperature measure. And there was a lot of argument about where that should be split. And we didn't want – we wanted to avoid that argument, frankly. So, to us, the categorization gives – is intended to give a handle, but the real measure is the distribution.

In terms of gaming, the warm, the too warm is an attempt to identify gaming by overheating. The other categories actually have graded risk. About right is where you'd like them to be. But if you – we were tasked by CMS and by AHRQ, because this was developed by one of the CHIP presenters of excellence. We were tasked by AHRQ and CMS to develop innovative measures to enhance the capacity for measurement. So we knew we were going to be pushing NQF sensibilities.

In terms of – we think the visual representation of the distribution is actually very helpful because you could look at how fat it is, how tall it is, how long the tails are, and it puts the onus on the accountability entity to give direction to the entities it's holding accountable, where it places the priority while still capturing the real variance.

Suzanne Theberge: So, I think, I just want to sit back for a moment and just let folks know (Larry's) referring to a measure that was endorsed at one time but lost the endorsement. You know, a previous NQF perinatal project a few years ago on

temperature, and that was a Vermont Oxford Network measure. So, just so folks who are new to our work have that background information.

(Diana Ramos): So, thinking about this one from a consumer perspective, it seems like the distribution concept, while I understand it and I appreciate that it has utility in certain circumstances, that it would be pretty hard for a lay person or a person who wasn't, you know, really deep into measurement to interpret. And so, I was kind of leaning and thinking, is it possible to – I appreciated the discussion of the dichotomous measure and I understand the challenges of that. But the dichotomous measure, of course, does give you kind of a better sense of, is the institution doing well or doing not as well?

(Larry): I'm not sure if you're looking for a response. I would just say that I think consumers are more sophisticated. Like consumer groups in particular are sophisticated enough that the combination, the way we ask it to be presented, they could look at both the proportion who are about right and the shape of the distribution and make a pretty good judgment about how well the institution is doing, but that's my belief.

Again, we're trying – our center was trying to move away from really dumb down measures. I don't mean that in a pejorative. I just mean that where we were developmentally in a state of measurement into the future and what measurement ought to look like moving towards the future, and that's where we were aiming, that's where we were charged by, again, AHRQ and CMS to look. And that's how we came up with this.

And yes, all the details were developed including which percentiles were chosen by an expert panel using a RAND modified Delphi method.

Suzanne Theberge: OK.

Rajan Wadhawan: I wanted to comment in that regard. I think just looking at just the two tools, that means combining the category one to three can be very misleading because you could actually be hurting a lot of kids by all warming them and still look great because you don't have any – too many kids in the category one, two or three but you haven't had a lot in category five. So, it is a

balancing measure but it cannot be ignored. That's just one comment that I had implications to that.

Matt Austin: So, could a measure be then the percentage of babies that are about right, if that's ultimately what we're trying to drive towards? Is that the measure of success? And if they're too hot or too cold, that it's a failure?

(Diana Ramos): Right. That would certainly be the more consumer friendly version because that's really what – if you're pitching any of this at all towards families, that's what they want to know is are you taking proper care of babies?

(Larry): We did have families as a part of our development team. That's not where they steered us. And we had family groups, the Institute of Patient and Family Centered Care was a part of our steering committee. They actually wanted the more granular. They also felt that by giving the more granular detail, it guided improvement as well as accountability.

Whereas as was noted earlier, if the number is too low, if they're already over warming and you're not catching that, you're losing – you could be causing problems as opposed to resolving them through the kinds of things that might go on.

Matt Austin: And then I guess, to me, it might be sort of this debate about what gets publicly reported as opposed to what's collected and used for internal quality improvement purposes, and maybe the difference there.

(Diana Ramos): Right.

(Larry): And I think this was designed to be adaptable to both purposes and we could imagine plans or states choosing to report portions and not all of it or all of it as suits their needs. We were tying to develop something that was valid, reliable and useful, recognizing that use varies in context.

Suzanne Theberge: So, I know there's a lot here. And unfortunately, we are very close to the end of the call, so I want to stop here and see. We did receive a number of comments on this portion of the survey on testing, and see if anybody wanted

to mention anything about the reliability or the validity testing in the last couple of minutes that we have for discussion on this.

Thoughts from the committee on the testing, it's either element testing was performed. Any thoughts on whether all the critical data elements were tested, whether the test sample was adequate?

All right, well hearing none, let's move on to the validity. Thoughts on the validity? Are the specifications consistent with the evidence?

Matt Austin: So, this is Matt. And sort of quickly, when you're looking at this, it looks like there's data element testing using the ICD-9 codes. Are there plans to update to ICD-10?

(Larry): We have mapped ICD-10. No, grant funds have run out and we don't have the kind of money that would require to do that again. I would say that the algorithms here and the codes map well. And if anything, they're actually a little better in ICD-10, a little more specific. But we did it through a mapping algorithm and then over – reviewed it over to make sure there weren't any inadvertent – excuse me, problems therein.

Matt Austin: All right, thank you.

Suzanne Theberge: So, the – (inaudible), I know that on the validity here, in their review, NQF noted that the construct validity analysis used the temperatures from categories one and two. This is another version of the measure and I think should demonstrate the validity of the measure as specified.

And our question to you the committee is – whether the version of the measure tested the same as the measures submitted for (possibly) more endorsement. What your thoughts are about that?

Matt Austin: So, this is Matt. I would say one (inaudible), I mean I think I would agree with the principle that the validity testing should align with how the measure is specified. But based on earlier conversations, I'm wondering if we first need to have further conversations about how to specify the measure.

(Larry): One thing to make clear is some of this – much of this testing was actually done before we knew what the measure was, in part, because the CAPQuaM process, we didn't start with a measure improvement. We started with constructs and guidance from our panel and the literature and developed measures that met those.

And so, the measures that – the stuff that we presented, looking at mortality as a dichotomous and we also looked at a continuous. We found that across the spectrum, one degree of temperature, under 37 degrees, one degree of temperature conferred about the same additional risk as losing 100 grams of birth weight.

So, I mean, we've done it throughout the spectrum. We just didn't present it that way. We could reanalyze that data, I suspect, if that were necessary. But the reality is we found – one of the reasons we presented it as a continuous variable is we actually found that the risk itself was continuous and not threshold based.

Suzanne Theberge: OK. Thanks, (Larry). I'm sorry, we only have a couple of more minutes on the call and we still have to do comment and next steps. So, I'm going to stop the discussion here because we've already taken a couple of hours of those time and I want to see, in the last minute of this portion, do any of the committee members have any final comments on feasibility, usability of use of this measure?

All right, well, hearing none, you know, we will have more time to discuss this at the in-person meeting. And I think, unfortunately, we are only able to really just get a little bit of this discussion and – but we'll have more time, although not a lot more time unfortunately, but we'll have more time at the in-person meeting.

So, we will pause here and see if there are any public comments, and then we will do next steps and any questions from the committee. So operator, could you open the lines and see if there are any public comments?

Operator: Thank you. At this time, if you would like to make a comment, please press star then the number one on your telephone keypad. We'll pause for just a moment.

And there are no public comments at this time.

Suzanne Theberge: OK. And if folks who are just tuning on the webinar and have any comments, you know, type those into the chat box. And in the mean time, we'll move on to the next steps.

First, the committee members, do you have any questions about what we discussed on today's call about the NQF criteria? Anything like that?

(Crosstalk)

Suzanne Theberge: Great.

Female: In the in-person meeting, we'll be able to sort of – there were some topics that people wanted to talk about more. I assume that we'll have more time then hopefully to kind of delve in to some of the things that were of concern.

Suzanne Theberge: Yes. So, you know, what we've done is taken some notes on the additional information that was requested as well as areas that we want to be sure to highlight at the in-person meetings, things that we'll need discussion. And if we get additional information, we'll be sharing that with the committee, and then we'll also be sharing the transcripts from today's call with everyone.

So what will happen next is that we ask everyone to review all of the remaining measures in the project and just take a look and be prepared to discuss everything at the in-person meeting. And when we get there, we will be going through with the full committee, each of the criteria. And, you know, of course if a measure doesn't pass, one of them must have criterion as we have previously discussed, you know, such as importance or reliability, validity, something like that, then the discussion would stop. But otherwise, we will have the full committee discuss and vote on each of the criteria. We will ask those of you who are lead discussants to kick off the discussion at the

in-person meeting for the same measures that you've been assigned as a lead discussant, too. So, be prepared to be doing that at the in-person meeting and have that in your mind as you prepare.

Reva Winker: Suzanne? Suzanne?

Suzanne Theberge: We – yes?

Reva Winkler: It's Reva. I just wanted to mention that we will be sending you, soon I think, a suggested script that you can use to help organize your presentation for leading the discussion at the in-person meeting. Committee members from other committees have told us this is very helpful in helping them organize the conversation so that we can efficiently cover all the bases, get all the discussion topics out on the table, and yet, move through things expeditiously. So, we'll be sharing that with you soon.

Suzanne Theberge: Yes, we will. Thank you, Reva. And I also wanted to add, we'll be continuing to update the worksheets as we finish these – the last couple of workgroups tomorrow and next week. We're going to be putting in the surveys. If you have not yet completed surveys, you know, you don't at this point have to because we are using those surveys to guide these workgroup discussions. Although if you would like to, if you find that helpful as you complete your review of the measures, you know, please feel free to enter more information in there, and we can incorporate that into these worksheets for the rest of your colleagues.

If you were on today's workgroup as a member of the workgroup, you don't have to listen to the remaining two calls, but you are definitely welcome to hear how the rest of the discussion goes. And I know we had some folks from other workgroups on today's call listening in.

So, in the last two or three minutes of the call, I will just stop there and see if our committee has any questions about logistics or anything else before we close.

OK, well hearing none, please don't hesitate to e-mail us if you have any questions. If you have any trouble with SharePoint, I know we've already reset a few passwords. And if that happens, don't hesitate to contact us. The whole team is checking that e-mail box. It gets checked several times a day. So somebody will get back to you as soon as they can.

And with that, I think we'll close out. I want to thank all of our committee members and all of our developers very much for your time today. We very much appreciate your giving us a couple of hours, and we're really looking forward to seeing you all in person in just a couple of weeks. Thank you.

Matt Austin: Thank you. Take care.

Operator: Ladies and gentlemen, this does conclude today's conference call. You may now disconnect.

END