**NATIONAL QUALITY FORUM**

**+ + + + +**

**SCIENTIFIC METHODS PANEL**

**SPRING 2020 MEETING**

**+ + + + +**

**THURSDAY**

**APRIL 2, 2020**

**+ + + + +**

**The Panel met via teleconference, at 9:00 a.m., Dave Cella and Dave Nerenz, Co-Chairs, presiding.**

**PRESENT:**

**DAVE CELLA, PhD, Co-Chair**
**DAVE NERENZ, PhD, Co-Chair**
**J. MATT AUSTIN, PhD**
**BIJAN BORAH, MSc, PhD**
**JOHN BOTT, MBA, MSSW**
**DANIEL DEUTSCHER, PT, PhD**
**LACY FABIAN, PhD**
**MARYBETH FARQUHAR, PhD, MSN, RN**
**JEFFREY GEPPERT, EdM, JD**
**LAURENT GLANCE, MD**
**JOSEPH HYDER, MD**
**SHERRIE KAPLAN, PhD, MPH**
**JOSEPH KUNISCH, PhD, RN-BC, CPHQ**
**PAUL KURLANSKY, MD**
**ZHENQIU LIN, PhD**
**JACK NEEDLEMAN, PhD**
**EUGENE NUCCIO, PhD**
**SEAN O'BRIEN, PhD**
**JENNIFER PERLOFF, PhD**
**PATRICK ROMANO, MD, MPH**
**SAM SIMON, PhD**

**ALEX SOX-HARRIS, PhD, MS**

**MICHAEL STOTO, PhD**

**CHRISTIE TEIGLAND, PhD**

**RONALD WALTERS, MD, MBA, MHA, MS**

**TERRI WARHOLAK, PhD, RPh, CPHQ, FAPhA**

**ERIC WEINHANDL, PhD, MS**

**SUSAN WHITE, PhD, RHIA, CHDA**

**NQF STAFF:**
**ASHLIE WILBON, MS, MPH, FNP-C**
**SAM STOLPE, PharmD, MPH**

**MIKE DiVECCHIA, PMP**

**HANNAH INGBER, MPH**

**CAITLIN FLOUTON, MS**

**ALSO PRESENT:**

**SUPARNA BAGCHI, CDC**

**NAOMI BARDACH, UCSF**

**JENEITA BELL, CDC**

**ANDREA BENIN, CDC**

**LISA BERGERSEN, Boston Children's Hospital**

**CLAUDIA DAHLERUS, University of Michigan**

**ELIZABETH DRYE, Yale CORE**

**JONATHAN EDWARDS, CDC**

**JACK KALBFLEISCH, University of Michigan**

**JOE MESSANA, University of Michigan**

**CRAIG PARZYNSKI, Yale CORE**

**DORIS PETER, Yale CORE**

**JONATHAN SEGAL, University of Michigan**

## CONTENTS

P-R-O-C-E-E-D-I-N-G-S

9:01 a.m.

MS. WILBON: Good morning, everyone. Welcome back to day two. And I see that we've got just about everyone back. So, thanks again for joining us.

We're looking forward to jumping right in. We're going to do our best to keep things brief and on time today, so that we can make sure that we get the tasks at hand completed.

Just a quick agenda review and update from yesterday, based on where we landed. We did run out of time and were not able to get to the second measure that was slated for review yesterday. We have slotted that measure in the measures we'll review after the break this morning.

In order to accommodate that, we did have to shorten the morning break, which is scheduled for 11:00 a.m. We will still take that break, but instead of a 30-minute break, we're going to shorten that to 15 minutes. So, I

realize it's not long, but we want to make sure that we can get done the tasks at hand and make sure that all the measures have adequate time for review. So, our 11:00 a.m. break will now be 15 minutes. It will be from 11:00 a.m. to 11:15. We'll come back at 11:15 and wrap up measure review for the remaining three measures.

Our goal is still to try to adjourn by 1:00 p.m. And we'll do our best, myself and the two Chairs, to keep us on task. We'll just ask developers as well as SMP members to try to keep your remarks brief and concise. Obviously, we'll make sure there is adequate time for discussion, but we want to make sure we're being as efficient as possible as well.

A couple of kind of housekeeping items about the webinar. For those of you that are already on the webinar, I think you've figured it out, but there is a separate link to get into the webinar for day two. So, if you're on the phone, if you haven't gotten to the webinar yet, it probably says you have to click on the day two

link.

Also, keep in mind, for those of you who dialed into our speaker line, your lines are open. So, please make sure your lines have been muted if you're not speaking, so we don't get the background music or background noise and feedback.

If you are dialed into the other line, which the public and some developers may have that line, we will choose you when to speak and you can hit *1 when muted. If you need operator assistance, please hit *0. We can add reminders as well to the staff about that.

If you're having issues speaking and you're dialed into the speaker line, if you're a Methods Panel member and for some reason you're not able to get unmuted, or you should be able to speak and are not able to, the best workaround for that is just to dial back in. We do have an operator that is assisting us and who is with us for the duration of the day who should be able to assist you in getting back into the call in a

timely manner.

I think those are the main reminders, and I just wanted to open it up to Dave and Dave to see if you have any opening remarks before we jump in.

CHAIR CELLA: This is Dave C. No, just to say I thought yesterday was a good meeting. We had lots of great discussion. I'm sorry that we didn't at the end keep on schedule, but we'll catch up today.

Dave N.?

CHAIR NERENZ: Yes, Dave N.

Thanks, everybody, for all the hard work yesterday and the staff for putting things together. It was really, really good.

Please, please, folks, today stay on point. We have a lot to try to get in. We have to squeeze these discussions in and stay on time.

As I look at what ground we have to cover, if for a given measure, say reliability passed but validity is the question, please limit the discussion, then, to validity. Let's not re-

discuss or re-legislate things that have already been settled.

Thanks.

MS. WILBON: Thank you both.

What we're going to do, one other thing for the Methods Panel, you should have received an email by now, an email from our team with voting instructions again, so that they're at the top of your mailbox, as well as the revised agenda for today.

Please also keep in mind we will be voting by Subgroup. As we get to your Subgroup, we will check in with those who we are expecting to be voting to make sure you're logged in and that we can achieve a quorum for voting. So, we'll do that when we get to it. We won't do roll at this time.

We're going to dive right into measure evaluation. I did want to check in, particularly with developers for our first measure that's up for this morning, 3556. And that's the National Healthcare Safety Network Nursing Home-onset CDI

Outcome Measure from the CDC. The developer is Dr. Jeneita Bell.

Dr. Bell, I think you may have the other line. If you would hit *1 to speak and let us know if you're there or enter chat in the chat box, and you can let us know if you're able to speak.

MS. BELL: Hello. Can you hear me?

CHAIR CELLA: Yes.

MS. WILBON: Yes.

MS. BELL: Okay. Great. I was actually provided a number that gave me access to the speaker line.

MS. WILBON: Okay. Great. Okay. Perfect.

MS. BELL: Yes, so I'm here and I'm joined by two subject matter experts on my team who will be here to assist with the conversation. Thank you.

MS. WILBON: Okay. Great. Thank you so much.

So, just a quick process overview

again. I'll start out with a brief introduction of the measure. I'll hand it over to the lead discussants who are John Bott and Larry Glance for this measure. We will, then, open it up to the other Subgroup members to comment. And then, we'll hand it over to the developers to provide a response to any questions raised. We'll then open it up to the full panel for any comments. And then, we'll bring it up for a final vote. Okay.

I'll just start out with a brief introduction here. I will direct you to page 10 of the discussion guide where the measure is summarized. There's also links there to the measure information form and the testing attachment, if you need to reference that.

I will just ask, if you're on the phone and on the computer, if you could turn the volume down on your computer, so you don't get the feedback, and then, if you could mute your line, so we're not getting feedback.

So, again, this is a new measure

submitted for consideration. It's the standardized infection ratio of a nursing home facility-onset incident, a CDI; laboratory-confirmed events among residents in the facility. Nursing home-onset CDI is defined as laboratory-confirmed cases that develop four days after admission. It's assessed at the facility level. It is risk-adjusted.

For this measure, they submitted data elements validity testing. And so, the focus, particularly for this measure, is, again, based on our criteria, data element validity testing can be submitted. If it is submitted, therefore, they don't need to submit additional reliability testing. So, the focus should be on the data element validity testing and whether or not that was adequate. The vote for validity at that point would, then, stand for the reliability vote.

And so, I'll focus on the data element validity question here. I'll just give a really brief overview, and then, hand it over to the

lead discussants.

The developers performed sensitivity, specificity, PPV, and NPPV populations. We do a comparison of validators' and facilities' determination of the presence of a reportable CDI testing, it was based on three states encompassing 14 nursing homes. The results are summarized here on the discussion guide. I won't read those aloud.

But again, I'll hand it over to Larry and John at this point to give a summary of some of the concerns identified by the reviewers, and we'll go from there.

Larry and John?

MEMBER GLANCE: I'm happy to start off, if you'd like.

So, as you said, validity was conducted at the data element level. The results of the validity testing, sensitivity and specificity, I thought were problematic. Because we're looking at the validity of the outcome itself as opposed to the elements for risk

adjustments. And although there are no strict criteria for what represents acceptable sensitivity/specificity, the values here seem unacceptably low, given the fact that we're looking at data elements for the outcome itself.

Differences in measure performance between nursing homes may reflect differences in the accuracy of the data conversion rather than true differences in nursing home performance.

And finally, the results in data validity are based on a convenient sample of nursing homes which may not be representative. So, I thought data validity was a real problem.

The other main threat to validity was the risk adjustment model itself. The model is based on data from 2700 facilities, but includes no patient-level risk factors, and therefore, can't account for any differences in case mix between facilities.

And what the measure developer says, quote, "Social risk factors make up for the differential incidence of CDIs resident-level,

but we were unable to assess these factors because of the methods of data collection."

The other problem with validity, again, a major threat to validity, was in their risk-adjusted model they included facility characteristics, things like percent skilled nursing and a number of the patients were admitted for C. diff treatment. Now, admittedly, that's going to be an important risk factor. And the problem is that, if you remove the portion of patients that were admitted for C. diff treatment in the risk-adjusted model, you will adjust away, potentially, differences in performance. So, if you have more of these patients, your own patients are at higher risk for infection and, in theory, you should be taking steps to try to mitigate that. And instead, you sort of get credit for having a sicker patient case mix, and you're not asked to make any adjustments.

So, overall, I thought that data validity was a major issue. I thought that the risk adjustment model was poor. In terms of

guidance to the measure developers, I think they need to include the validity of the outcome data elements before this measure can be used for public reporting. I think they need to consider including a more representative sample of nursing homes around the country in the data validity testing. I think they need to include patient-level risk factors such as age, sex, comorbidities, ADLs in the model itself. And finally, I think they need to exclude facility-level risk factors in the model.

MS. WILBON: Thank you, Larry.

John, would you like to add anything to that?

MEMBER BOTT: Well, the notes I had were really all covered very articulately by Larry. So, I just really absolutely second everything Larry said. Well said and I completely agree.

Thanks. And thanks for that great summary, Larry.

MEMBER GLANCE: Thank you.

MS. WILBON: Okay. I wanted to see if there are any other Subgroup members that wanted to identify any additional concerns maybe that Larry didn't touch on for the developer to respond to.

MEMBER AUSTIN: Yes, this is Matt Austin. Good morning.

I mean, to sort of follow up on the conversation from yesterday in terms of critical data elements, one of the gaps I identified was that they did not seem to provide any data around the data element validity for the values in the denominator, things like residents' age, et cetera. And so, I think that was a gap in their analysis.

MS. WILBON: Anyone else from Subgroup 1 have anything to add?

Okay. Dr. Bell, I'll hand it over to you.

I did just want to make one comment that the measure did not pass reliability and validity. We wouldn't have added the measure to

the agenda, but, given the timing of the COVID crisis, many of our developers have been busy responding to the crisis. And so, people were not able to submit a written response. So, we pulled the measure for discussion in order for Dr. Bell and her team to be able to respond verbally. So, there is no written response to any of these questions that we have. Obviously, we'll allow time for them to respond verbally.

So, Dr. Bell, thank you for joining us, and we'll open the floor to you and your team to respond to the panel's concerns.

MS. BELL: Hello. Good morning, and thank you so much. I appreciate the opportunity to engage in this discussion because you may know this by now, that I was actually out in the field helping with the coronavirus response. Our team actually pulled away today to be able to participate in this discussion.

I have with me at least three other subject matter experts, two of which who led the risk adjustment work. And I also have our

subject matter expert who leads data validation within our branch, which is consistent across the branch for other components within our surveillance system.

But, first, I want to say I appreciate and understand the comments that were provided by the panel members. I hear that there's two things primarily. One, there's some issues that you are all concerned about concerning the validity of the measure, the data elements that are associated with the measure itself, and the representativeness of the sample. And also, there's some concerns about the factors that were included in the risk adjustment model.

So, if I may, I'm going to ask for my colleagues, Jonathan Edwards and Elizabeth Mungai, to see if they can speak to some of the concerns regarding the risk adjustment. And then, I'll pivot over to Dr. Suparna Bagchi to see if she has comments relating to validity.

MR. EDWARDS: Hi. This is Jonathan Edwards. Good morning, everyone, and thank you

for the summarized input.

The thing I want to start with is that we definitely hear the comment about the need for patient-level data quite often in the data we collected in NHSN. And I'll just editorialize here just real briefly and say that we are trying to move in directions where we can capture data electronically at the patient or patient admission level for new type data collection, not necessarily related to this measure or this population yet.

So, that's something we want to move in the direction of, but we have to pay attention to the data collection burden. So, given that, the way the data are collected in the NHSN are they are ecological or summarized data. And while we have information on the infection events themselves that are at the detail level, the data are still summarized. And so, we do not have patient-level data on the entirety of the population, only those that have the events.

And so, what we have to rely on are,

basically, factors that might be collected in a manual survey to be able to be used as surrogates for differences in acuity. And we have had many discussions and have had many occasions where there are our own internal discussions about how well those particular factors from an annual survey may serve to distinguish differences in acuity of patients.

And it is not a new point that some may not be willing to accept some of the factors. And it also is not a new point to have folks that were reviewing our models to say, well, we want patient-level data and we want patient-level factors, and I'm not satisfied with particular annual survey surrogates.

That said, what we do, and what we did -- Elizabeth Mungai and myself and then the team -- is we used the data reported into NHSN and we used this measure to be able to understand where are there differences in the outcomes. And so, to the extent that we understand where there are differences in the outcomes, we, then,

capture those differences.

And again, it's a fair point to say that we can't adjust for certain factors. I believe that the best that we can do is to use these factors that are collected in an annual survey to say: are the outcomes different or is the outcome different across those levels of those factors?

So, in the end, the measure is based on a regression model that is the best characterization of differences in the outcome, being the CDI incidents.

And I'll just stop there and see if my colleagues have anything further they want to add in there.

MS. BELL: This is Jeneita.

I would just add, you know, Jonathan mentioned that we rely on facility-level factors because very limited information is available for a patient level. At the most, you can collect age and sex, but there is no comorbidity information or information about patient case mix

to include into the model.

Should we proceed with our remaining comments? I don't know how dynamic the conversation is expected to be. Sorry.

MS. WILBON: Yes. Hi, Dr. Bell. This is Ashlie.

Go ahead and have everyone from your team respond, and then, we'll have the follow-up questions from the Methods Panel follow that.

MS. BELL: Okay. Jonathan, do you have anything else to add?

MR. EDWARDS: No, no. I'm happy to respond to further questions, but I think I've said everything that we can say.

MS. BELL: Yes.

MR. EDWARDS: Again, our preference is that we would have better risk adjustment. We always want to seek that out. We always have to bear in mind the data collection burden.

And I would just add, in NHSN, in terms of using these measures, we want the measure to be based on the data that are readily

available. And so, another point that may seem a little bit astray, but I think it's still relevant, is that sometimes people can share ideas of, oh, well, why don't you get these data or why don't you have use of these other data sources? Well, in the NHSN, we need to have the data available at the time. And so, the completeness and the data availability and the timeliness, certainly, there are factors there. And so, we have to weigh the burden of data collection also together with the completeness and timeliness of data.

That's all I'll say for now. Thanks.

MS. BENIN: Jeneita, it's Andrea. I can also add to Jonathan's discussion about how we're often needing to weigh the pros and cons of how we approach the metrics by, I think, emphasizing also that this particular disease we think is of enormous importance in these facilities. And without being able to get a start on a metric to quantify that and understand it better, it really inhibits our ability to

understand the landscape and move forward with the prevention activities.

And I think the recent events around infection control in these facilities highlights some of the general urgency around being able to get at some of the feasible approaches to understanding these types of facility-acquired infections.

And so, just to underscore what Jonathan is saying about why some of the approaches that we have chosen at this point are what we have at our fingertips.

MS. BELL: Thanks, Andrea.

And continue on with our comments, I'll pivot over to the validation portion of this discussion. And I heard the comments that there is concern about the representativeness of the samples and there's concern about the essence of the denominator data and the validity or validation testing.

The methods that we use were developed in-house and consistent with what we use for

other quality measures that are already NQF-approved. We did understand that there may be some concern regarding the three state validations that we presented. And we debated whether or not to include all three, but, for the sake of transparency, we decided to include Wisconsin and Minnesota, in addition to Nevada.

Nevada was our pilot state. They worked with us once we finalized the methodology and they have requirements for all their long-term care facilities to report to NHSN. So, we knew that invariably these facilities have some opportunity and some training and education about how to report to the surveillance system and do it with some level of consistency, because of the requirement that's there.

But I want to allow Suparna Bagchi the opportunity to give some explanation about the validation methods. Hopefully, that will provide some deeper understanding as to why there may be some discrepancy here.

Suparna, I want to see if you have any

comment.

MS. BAGCHI: Sure. Good morning, everyone.

Thank you, Jeneita.

So, as Jeneita mentioned, we have developed an in-house validation methodology for the long-term care CDI validations conducted by the states. And this was developed in collaboration with the Nevada Department of Public Health, primarily because Nevada is one of the states which has a mandatory requirement for reporting. So, Nevada was the State that pilot-tested our methodology around 2018.

But around the same time, prior to our development of this methodology, around the same time, Massachusetts was already conducting long-term care validations. So, they had the criteria for the facility selection was different than that that's proposed by NHSN. And Wisconsin followed the suit pretty soon.

So, the general methodology that NHSN follows is like it's trying to maximize the

resources we suggest for states with greater than 50 long-term care facilities. They would select facilities with at least greater than a hundred bed size.

And the idea was to maximize the possibility of the validators being able to identify an adequate number of charts which would provide us with a fairly good sample size to have some precision and accuracy and the precision of the estimation.

So, the methodology of Wisconsin and Massachusetts was different, and that could attribute to some of the reasons why the sample sizes are significantly smaller than Nevada. However, we did, when we started putting together the package, we wanted to be transparent that, even though we definitely worked with Nevada, and we are concerned about the methodology and the results, we still wanted to go ahead and provide the results.

And we cautioned the leaders about the results that we have identified, the Wisconsin

and Massachusetts validation. And we definitely understand that the smaller number of charts that were sort of viewed could lead to accuracy estimations with lesser precision.

As far as the denominator validation is concerned, definitely, that has been on NHSN's list. However, because of the lack of resources, our current focus has always been on the numerator validation. At some point, we definitely hope that we would be able to conduct denominator validation, too.

That concludes my summary --

MS. BELL: Yes, and I think along those lines, it might be helpful to understand that there is no specific funding dedicated to any of this. So, all the states and all the nursing homes and staff that have contributed to this validation have been voluntary. And we are grateful for their participation, but it also limits our ability to control all the factors that come into bringing this all together.

CHAIR CELLA: Hi. This is Dave Cella.

MS. BELL: I think that concludes our remarks.

CHAIR CELLA: Yes. Thank you. This is Dave Cella. I'm one of the Co-Chairs, but also on this Subgroup.

Just to move things along, does the Subgroup or anyone on the Committee have any questions in response to this response from the developers?

Thank you for that, by the way.

MS. WILBON: I think there's several hands raised. I didn't catch the order.

CHAIR NERENZ: Christie, Larry, Patrick, at least as they appear on the list I have.

CHAIR CELLA: Yes. Okay. Go ahead, Christie.

MEMBER GLANCE: I have a really quick question to the measure developers. You said that you are very concerned about the data collection element piece and that was why it was difficult to do more patient-level risk

adjustment. My question is, why not use the data that is available that is already being collected by CMS for Nursing Home Compare and use that data for patient-level risk adjustment?

MR. EDWARDS: This is Jonathan.

I would just add in here that the timeliness and completeness of data is definitely an issue we have encountered on a number of them where we have attempted the idea of can we bring in data from other data sources. One of the challenges we faced -- and I definitely understand the idea that, theoretically, you ought to be able to just have data on all facilities and be able to easily move that in. I don't know all the particulars of that myself specifically, but one of the issues is timeliness of the data and, then, the other one is the completeness.

And when I say "completeness," one of the things I'm bringing up is the ability to actually merge and match data from other data sources with those that report into NHSN. And

so, one of the things that I'll report on is, in general, not with the Nursing Home Compare data, but in other data sources we've looked at for this population of health care delivery, is to basically have to confront a 80 percent or lower match rate.

So, there are problems and technical challenges in trying to align with how data might be reported from another data source with how it is reported in NHSN. Largely what we try to promote is a brick-and-mortar structure, identification of a facility. And I think that we would need the data to be timely in NHSN.

And I'll just stop there, if other colleagues have something to add.

CHAIR NERENZ: Thank you. We're trying to move through on schedule.

Patrick, do you want to go from here? Patrick, and then, Christie, if your hand was up and you want it back up. I don't see it now. Go ahead, Patrick.

MEMBER ROMANO: Yes, thank you.

I think a big stumbling block for me is the actual data that you've recorded on validity testing. And I'm wondering if you could help us interpret this. I did look at the ancillary documents that you provided.

So, it appears, if I understand correctly, that the sampling for your validation study is based on positive laboratory reports of C. difficile. You can correct me if I'm wrong. And then, your auditors do a further evaluation of that.

Now your report from Massachusetts, of course, is very, very low sensitivity of 26 percent. It's higher in Nevada of 90 percent. So, maybe you could help us understand that for a second. I understand that only Nevada has a mandatory reporting via NHSN, but I assume that Massachusetts also has some state-level reporting or something going on that led them to participate in this pilot in the first place. So, maybe you could explain that variation in sensitivity between 26 percent and 90 percent.

And then, I'm also concerned about the variation in specificity rate as there seems to be a fair amount of over reporting going on, and you explain that. But normally, specificity of even 80 percent in Nevada would be considered unacceptable because that really, depending on the prevalence of the events, that could lead to a very large number of false positives.

So, maybe if you could just help us understand a little bit the sampling for the validation study and the wide variation and, overall, what you're doing to respond to that?

MS. BAGCHI: Sure. This is Suparna then.

So, as I mentioned earlier, yes, we understand the concerns about the differences in the findings between the states. So, Massachusetts was not a part of the pilot testing. So, they conducted their validation prior to the validation guidance being provided.

I have received the results from the state validation. They identified 10 facilities

for validation that had complete data in the six months prior to their validation timeframe, which probably was, given the concerns about the data completeness, already narrowed down the sampling pool. And from the facilities that were identified to have the complete data in the timeframe of six months prior to the validation, possibly could have less from smaller facilities with the lesser number of medical records available for validation.

So, a response from the Massachusetts Department of Health, from the facilities that they selected, the number of charts that they were able to identify ranged from like zero to 19 across these facilities. And that supplies the reason why they have such a small sample size.

But given the smaller sample size of the charts that were reviewed, any assessment of the accuracy for the 26 percent sensitivity, it makes us question the precision. So, it is, as you've mentioned, it is not a really good presentation of the long-term care settings

within even the State of Massachusetts. So, it would not be correct to compare the results of Massachusetts with even Nevada --

MEMBER ROMANO: Thank you.

MS. BAGCHI: -- in terms of any accuracy estimates.

CHAIR CELLA: Okay. Christie Teigland had her hand up. She got disconnected. I think I don't see any other hands up.

MEMBER TEIGLAND: I'm here. Can you hear me?

CHAIR CELLA: Yes, if you have a quick question? We're trying to move very quickly.

MEMBER TEIGLAND: A really quick question. Yes, I just was curious if you attempted at all to link to the Minimum Data Set where you could get a lot of those characteristics that you would need to appropriately risk adjust the measures with the demographic characteristics about the patients.

MR. EDWARDS: So, the answer is -- this is Jonathan Edwards -- no, we did not link

to the Minimum Data Set. I think to go back to the concerns that I was mentioning prior about the timeliness and the completeness of data, and when I mention "completeness," I'm really primarily thinking of the ability to have complete data to match and merge in. If we have a match rate of anything less than 100 percent, then that would mean that we would get skewed in the facilities based on the lack of matching within that data set. And that does not even bring in the fact that there is a timing issue.

Thank you.

CHAIR CELLA: Thank you. I want to thank the developers and make sure there are no final comments from the Subgroup or the developers.

Otherwise, I think it's time to move to a vote.

MS. WILBON: So, we will be voting on validity only. This will be based on the discussion that we've had around their data element validity testing. The voting results

from the validity testing will also serve as the reliability vote because of our pass criteria regarding data element validity testing.

So, those Subgroup 1 members, we're just going to do a quick roll call here to see who's on.

Daniel Deutscher, are you there?

MEMBER DEUTSCHER: Yes, I'm here.

MS. WILBON: Okay. All right. Dave, I know you're here.

Matt, I heard you.

John, I heard you.

Joe Hyder, are you there? Okay. Patrick, I heard you.

Sherrie, are you there? I think I saw Sherrie.

CHAIR CELLA: She's been on the chat.

MS. WILBON: Okay.

CHAIR CELLA: She asked for the ballot to be sent to her.

MS. WILBON: Okay. Terri, are you there?

MEMBER WARHOLAK: I am.

MS. WILBON: All right. Mike Stoto?

MEMBER STOTO: Yes, I'm here.

MS. WILBON: Okay. And, Larry, I know you're there.

So, we'll have one, two, three, four, five, six, seven, eight, nine voting.

Okay, Hannah.

For those of you in Subgroup 1, if you will try to make sure you clicked on the voting link in the email, and the survey should be up for you, and we'll display the voting results when everyone has voted.

MEMBER STOTO: Are you sending an email now with voting?

MS. WILBON: It got sent maybe about 20 minutes ago.

MS. INGBER: At roughly 9:02.

MS. WILBON: Thank you.

It should be toward the top of your email.

MEMBER STOTO: Okay.

CHAIR NERENZ: But it doesn't say "voting link" in your subject line. It starts with the agenda. The voting link is further down. You just have to scroll down.

MEMBER STOTO: Okay. Thank you.

PARTICIPANT: The link that she sent yesterday also works. That's the link I used.

MS. WILBON: Yes. We were just trying to put it at the top of the email. Whichever you get to first is fine. It's the same one. Hannah, could you just give us an update? Are you seeing votes coming in? Do we have a sense of who has voted at this point, how many --

MS. INGBER: Yes, I have eight votes in and I'm waiting for one more.

MEMBER STOTO: This is Mike. I'm having trouble finding the link. I'm not sure I'm looking --

MS. WILBON: That's okay.

MEMBER STOTO: Where is it in that email?

MS. WILBON: All the way at the

bottom. I can send it to you again.

MEMBER STOTO: All the way at the bottom? Okay. No, I'll go. Voting. I got it.

MS. WILBON: Thank you.

MEMBER STOTO: Okay. I think I've voted now.

MS. INGBER: Okay. Yes, I see your vote.

One minute while we adjudicate the results. I'm going to share the results on my screen.

Okay. You should be able to see the results for validity on Measure 3556.

We have eight votes for low and one vote for insufficient. Therefore, the measure does not pass on validity.

MS. WILBON: So, I just do want to thank the CDC team for joining us and for, obviously, all the work you're doing in the current crisis. We do hope that you are able to take the feedback from the Methods Panel and consider some other approaches to potentially

improving the measure. We, again, want to thank you for your time and engagement in the process.

MS. BELL: Thanks for your consideration.

MS. WILBON: And with that, let's go ahead and keep things moving. We'll be moving on now to Subgroup 2 and the evaluation of Measure 2496, Standardized Readmission Ratio for dialysis facilities.

And if you want to check to see if the developers are on the line? I believe that is Casey Parrotte, Joe Messana, Jesse Roach, Joel Andress, Wilfred Agbenyikey, and Jennifer Sardone.

Is anyone from the team on the phone before we get started?

MS. INGBER: I see Casey's name in the webinar, but I don't hear you. I'm not sure if you guys got the speaker line, but if you hit *1, the operator should be able to connect you and open your line.

MR. MESSANA: This is Joe Messana from

UM-KECC. Am I audible?

MS. WILBON: Yes, we can hear you.

MR. MESSANA: Okay. So, I'm not the principal discussant from our group. I believe Dr. Jack Kalbfleisch and Dr. Claudia Dahlerus were and they received the same speaker line information that I did. So, hopefully, they're on and just being shy.

MS. WILBON: Okay. Maybe unmuting your phone?

MR. KALBFLEISCH: Yes, this is Jack Kalbfleisch. Yes, I'm on the line. I'm trying to get over my shyness here.

(Laughter.)

MS. WILBON: Okay. Okay. Great.

CHAIR CELLA: Okay. This is just a reminder, everyone, we barely made it under the wire on time with the previous discussion. So, to calibrate, please try to make your comments concise. A friendly reminder.

MS. WILBON: Okay. Thanks. We'll dive in here.

So, we're now on Measure 2496. This is a maintenance measure. Its most recent endorsement was in 2016. It's a Standardized Readmission Ratio for dialysis facilities. It looks at the ratio of the number of index discharges from acute care hospitals to that facility that resulted in an unplanned readmission to an acute care hospital within 4 to 30 days of discharge to the expected number of readmissions given the discharging hospitals and the characteristics of the patients, and based on a national norm.

It's based on claims and registry data. It's measured at the facility level. It is risk-adjusted.

For this measure, there was consensus not reached for reliability and not passing a vote for validity.

I'll just do a really high-level overview of what they did for reliability and validity, and then, hand it over to our lead discussant. I'm not sure that Susan is on. So,

it will be Gene who will, hopefully, be leading us in identifying some of the concerns.

For reliability, they did four-level reliability using the IUR and the PIUR that we spent a great deal of time discussing, and the notes are there in the discussion guide. Again, we're on page 12 of the discussion guide.

For validity, they presented testing for data elements and measured four-level testing. They did present some data element validity testing, but this does not meet our requirement. So, we are going to ask the panel to focus on the measure for empirical validity testing, which they compared this measure to four other measures using Pearson's correlation.

So, with that, I'm going to hand it over to Gene Nuccio to give us a summary of some of the concerns raised by the Subgroup for this measure.

MEMBER NUCCIO: Yes. Thank you.

There are several concerns about the use of IUR and PIUR. We discussed much of that

yesterday. Let me just do a quick summary.

By the way, the developers provided a very lengthy response, beginning on page 76 in the detailed discussion guide. We thank them very much for that. That was informative.

One of the concerns was that in their previous presentation the IUR values for the measure were on average 0.55. I couldn't find a PIUR number for the previous one. However, in the new presentation with more recent data, the IUR value dropped to 0.35 and the PIUR value dropped to 0.61. Both of these are below that .7 value that we discussed yesterday with Adams. So, that was a concern.

The developers did provide several reasons regarding why they think that this drop was expected; notably, dealing with the fact that these are publicly-reported values now and, also, part of a value-based purchasing program, which would suggest that people are getting better. And I suspect the assumption is that there's less variation going on. Nonetheless, the values for

reliability are rather low.

There was another set of questions regarding bootstrapping and a bootstrapping methodology. And this had to do with the size of the agencies where the agencies that are delivering dialysis service were small and didn't have sufficient information. And so, they used a bootstrapping method.

In their response, they talk about patient years, which I believe is patient-years at risk. And just out of curiosity, you know, why can't you just use age, given that people would develop these sorts of needs for dialysis perhaps at a later age, and prior to that there was no need for this? So, age of the patient.

But if the developer could explain their response to issue No. 3, that bootstrapping deals with the size of the agency, but they reported the information based on the age of the patient. And then they say in their notes that they noticed that a lot of other developers are doing this.

But again, let me just do the validity real quickly here. The Pearson correlations are notably low, down in the .1 level, and this is not an r-squared. It's just an r. So, that was a concern there.

Also, one of the interesting things, what they did -- I think a positive thing -- is that they allowed for the Medicare Advantage patients to be calculated by looking at claims data for inpatient claims because you can aggregate both the Medicare fee-for-service and the Medicare Advantage on the inpatient claims, but you can't do it for outpatient. So, this is a change from the previous measure, and perhaps that's accounting for some of that difference.

Also, they did make a positive change in looking at excluding patients who were only in service on dialysis from zero to three days. Given the lack of stability, they excluded those patients from the data set, and they should be applauded for that.

The final thing that I want to mention

-- two things. One was that they did make use of the sociodemographic variables, and they are to be congratulated on the use of sociodemographics.

The one last thing that was of some concern was the power of the model. The C-statistic for the logistic was only 0.6359, which is rather low. And so some of the discrimination characteristics of the lower-than-expected, as-expected, and worse-than-expected, basically, the numbers of patients that performed poorly, as I'm understanding the last table that they provided in their response. For the group flagged as better-than-expected, that was 136 patients. They had the same number of facilities -- the same number of facilities showed up in the as-expected in terms of worse-than-expected. The number of facilities in the as-expected group was 204, and in their flagging the number of facilities was 245. So, again, there seems to be a lack of discrimination, perhaps due to the risk model.

I'll stop there.

MS. WILBON: Thanks, Gene.

This is Ashlie.

I just wanted to point out one thing. Our discussion yesterday about thresholds, that was more of a forward-thinking discussion. And I just want to make sure that we're not applying any kind of new rule, based on discussions yesterday, to the measures we're evaluating today. So, I think the .7 threshold is desirable for many, but it's not kind of a new rule or anything that we're imposing. But, certainly, in terms of consistency from the reviews that were done by the Methods Panel, the cycle that the IUR reported is lower than other measures that passed.

I see we have a hand raised from Jeff. Did you want to make a comment?

At this point, I guess we should open it up to other Subgroup members.

Jeff, go ahead.

MEMBER GEPPERT: Sure. Just quickly, I think in a lot of respects their methods that

are used are not position exemplary and the results are presented with a lot of integrity, as Eugene mentioned.

So, for a mature measure like this, I think a more compelling demonstration of validity is warranted. Demonstrating validity with correlations among more measures seems like a circular argument, in that each measure is essentially used to demonstrate the validity of the others. And a more compelling demonstration might be to give evidence that the quality construct itself is causally related to the likelihood of experiencing these outcomes of interest. For example, if there's something that better dialysis centers do or have that worse dialysis centers do not do or do not have, you know, that is causally-related. So, that would be simply a more compelling demonstration for a very mature quality measure.

Thank you.

MS. WILBON: Hi, Christie. Go ahead. I see your hand.

MEMBER TEIGLAND: Hi. I'll try not to disconnect myself this time.

So, I had a concern, and to make sure I understood this correctly, that the model was developed using a year's worth of data, and I believe just from Medicare fee-for-service patients. And this model is being applied to Medicare Advantage, which we know is a little bit different. And I know you just have adjustments for that.

But, then, the model, the way it's implemented is it gives you the only diagnoses on discharge claims. So, having extensive experience working with claims data, I know that discharge claims only contain very few diagnoses. Obviously, the primary diagnosis for those is the hospitalization stay and maybe a secondary diagnosis that's also related to the stay. But in no way do they document all the chronic conditions. So, that means if you have less patients who were discharged for some type of inpatient stay that didn't involve their heart

failure or their diabetes, or their whatever else is in the model, they're going to not get properly risk-adjusted.

So, I'm concerned that you developed the model using a year's worth of data which it does take significant time to document all the conditions that patients have. But the way it's implemented, it's only using a subset of that, which could have resulted in some of those very small validity findings.

CHAIR CELLA: I think we should move to developer discussion if there are no other -- oh, wait. Bijan and Jack. Sorry. Go ahead, Bijan.

MEMBER BORAH: Hi. Good morning.

So, I think this is something that has been pointed out earlier. So, there was a significant drop in IUR from the prior submission to the last submission, from .55 to I think .35. The only difference was that we are using more outdated data. And I think as was pointed out earlier, we also had Medicare Advantage data. I

guess it could be a good question as to what is the rationale or what was the thinking behind what is dropped.

And my other sort of question was, the initial analysis they have used, taken here, as opposed to the number taken, and I guess in terms of how Medicare patients are penalized. I mean, particularly the unique age of the patients. I think it would be really helpful to understand how including patient-years versus simply the number of patients, I mean, how would the results change? It would be insightful if they would present the results both ways, including patient-years as well as simply patients.

Thank you.

CHAIR CELLA: Jack? Yes. Okay. Thank you for lots and lots of comments.

MEMBER NEEDLEMAN: This is Jack. Sorry, it took me a while to get to my mute button.

I just have like three comments, some of them responding to the response of the

developers. I, again, want to echo we're getting excellent documentation of these measures, and nothing is nefarious about the documentation.

With respect to the IUR versus PIUR, I think the IUR is low, lower than we should expect. And given that centers will be evaluated across the distribution of experience, not simply the outliers, the IUR for me is the relevant measure, unless we're being told we're only looking for outliers in this measure, which we have not been told.

Second, the developers make a specific comment in their response about the IUR going down because we're getting better risk adjustment. But I'd simply point out that, if we're getting better risk adjustment, that means that we're getting a more accurate measure of the expected. And when we do that, the range of distribution across the facilities is narrowing. So, it's harder to find differences across the providers, but the conclusion that a lower IUR is an indication we're doing a better job of

discriminating across providers is simply wrong.

The third thing is, there's a comment on the analysis comparing to other measures about, while the correlations have gone down, the directions are right, which is terrific, but they are very precise. The estimates are very precise. The statistical significance is -- the p-values in the statistical significance are very, very low.

All that means is you've got precise estimates. And what really matters is the magnitude of the correlations, not the p-values. You can't point to a p-value and say, because the p-value is so good, it's showing that these measures are correlated.

The question is, what's the level of correlation? And you noted the correlations are low.

I'm done.

CHAIR CELLA: Okay. Now to the developers. Thank you for your patience. I know that's a lot of questions.

MR. KALBFLEISCH: Okay. So sorry, there are two Jacks around obviously.

CHAIR CELLA: Yeah, sorry about that.

MR. KALBFLEISCH: This is Jack Kalbfleisch. I'm a professor of biostatistics and statistics at the University of Michigan.

I'm going to talk about -- mostly about the reliability issues and a few other things. And I think that Dr. Dahlerus will talk more with regard to the -- with regard to validity. And we'll also kind of pick up a number of other comments as well.

And one of the first ones, one that's frequently asserted a change in the IUR this time from the last time which, of course, was of considerable concern to us too, to see that change so large.

And we really don't know why that happened. There's been a lot of changes since the last time. And I think a number of you have already mentioned in the discussion there's sort of been changes in the data, changes in measuring

instruments, and to some extent, changes in the response that we use now. We don't use the first three days which had been used before in the IUR measure.

So there are a number of changes in there. One which I think is substantial is that there has been involvement of a measure in equipment and in the five-star ratings and also in Dialysis Facility Compare presentations. So there's more attention paid to this measure by facilities than there was in the past. And that's a good thing, of course.

I think we also more completely account for comorbidities in that we changed the way we measure comorbidities from CMS hierarchical condition categories to the AHRQ CCS diagnostic categories.

There are more measurements of comorbidities than there used to be. And we've expanded the number of measures that would be used. So all of these have some impact on reliability.

I certainly agree that reliability goes down doesn't mean necessarily that you're doing a better job of accounting for facility differences in a model. Reliability can go down for two reasons.

One is because the facility -- the adjustment in facility, the differences are smaller, that you get a smaller variance corresponding between facilities. Or the other is that you get more variation among patients within a facility. And either of those can account for change in reliability -- in the IUR rather.

I guess one sort of general concern, I think, is with IUR itself in that IUR is inter-unit reliability. It was called that for some reason. I don't know. It's really just a correlation.

It's an inter-class correlation between the -- with respect to the measure rather than the individual, so that the individual gives you the ICC, the inter-class correlation; the IUR

is with respect to the measure. And so it's a measure of signal to noise or correlation.

It's given the name reliability, and I think one gets a transfer then to everyday language -- from everyday language to the technical language that one needs to worry about. I think the original technical report by Adams, which led to the 0.7 which again the committee is considering, it prefaces the comments in that technical report that his arguments are based on the assumption that inter-facility differences are entirely due to the quality of care.

But I think this is seldom the case that there typically are unmeasured confounders like genetic differences or diet, socioeconomic factors that we really can't measure very well and that affect the responses and that also differ in their distribution between facilities.

So IUR should be interpreted with some care, I think. And even with relatively large IURs, one should be cautious about making comparisons of facilities, especially in the

center part of the distribution. There's always the possibility on unmeasured confounders affecting the variability.

Also as you pointed out, a low IUR doesn't mean that the measure is not useful for profiling because there can be a relatively smaller subset of providers that are extreme and of interest. This is partly because the IUR is to be based on normal of function. And it doesn't really recognize what's going on the tail very well.

So you can have an ultimately smaller group be extreme and certainly of interest because they're actually likely to get a faulty improvement program in. But the IUR doesn't pick that up as something the measure is capable of.

And so that's the motivation of the PIUR. And it's really based on the rather simple idea that a measure can be viewed as reliable if the probability will be reidentifying the same facilities in the same category and perhaps an extreme category -- those that are verified

outside the range in the upper 2.5 percent, 5 percent, or 10 percent -- reidentifying those facilities in a new sample under the same conditions. If that probability is relatively high with that, it could be taken as a measure of reliability.

I think in the introduction actually to the form, on reporting on the measure to the NQF panel, it actually describes reliability in terms like that, basically the ability to repeat the classification.

So unlike the IUR, if the IUR is not based on normal assumptions, if the probability of reidentification is estimated based on data splitting, and it doesn't really involve normal assumptions at all. And I think that's something of an advantage. But it does concentrate on specific things like the tail area.

The standing committee made, I think, a comment that there's an indication of the measure are useful for identifying extremes is quite true. But on the other hand, I would argue

that that's an important aspect for information too.

You could standardize the PIUR in various ways. Although we use the properties for the IUR under a normal model to calibrate the PIUR so that we basically calibrate the PIUR by picking a value of IUR that we give the same probability of reidentification. So the same probability of reflagging individuals.

So that's done -- in the IUR, it's done in a very normal assumption. And it's really just way of trying to get something which is trying to get at a way of calibrating the PIUR that relates to measures of some of the claims.

The PIUR would give has a reference to a tail area of 2.5 percent which you can get similar results with other fill areas as well at 5 percent or 10 percent. But it's focusing basically on tails of the distribution.

I think the higher PIUR, and it's still I think 0.7 is a very high bar. But the high PIUR I think indicates that it is very

useful in identifying facilities that are relatively extreme compared to the central part of the distribution.

It's less useful, I think, for ranking facilities closer to the center of the distribution. Although I think it should be recognized that ranking facilities in the center of the distribution is a difficult task for most measures, even moderate or even high IUR, especially if you account for the likely existence of unmeasured confounders.

So I think basically the PIUR indicates that the measure is quite useful for identifying extreme values.

MS. WILBON: Hi, Dr. Kalbfleisch. This is Ashlie from NQF. I apologize for interrupting, but we want to try to make sure that we have enough time for you guys to finish your response for validity as well as allow the methods panel to ask any additional questions. And we still have one more measure review during this time frame.

So if I can just encourage you guys to be succinct in your remarks, and then we can try to stay on time.

MR. KALBFLEISCH: Okay. I think that was basically what I wanted to say about IUR, I think at this stage. So I could pass it on to Dr. Dahlerus to talk about validity or take questions. I don't know.

MS. WILBON: Yes, we'll hear from your colleague about validity. Thank you.

MS. DAHLERUS: Hi, this is Claudia Dahlerus from University of Michigan. Can you hear me? I want to make sure that I'm connected.

CHAIR CELLA: Yes, we can hear you.

MS. DAHLERUS: Wonderful. Thank you. Good morning. Okay. I will be concise in our responses regarding the validity testing we did for the SRR. So we do recognize that the correlation coefficients were a little lower than in the prior submission from 2014.

What I would reiterate that the hypothesized association for the direction of the

correlations are very consistent and in the expected direction for all of the other outcome and intermediate outcome measures that we validated the SRR against.

I did want to highlight that there has been some changes to the underlying data source since the original submission six year ago which used actually 2009 data. And here, we're using more current data that includes ICD-10 diagnostic codes versus ICD-9. So that's one substantial change.

Given that, we felt it was reassuring to see a consistency in the direction of the correlation coefficients and that the declines weren't huge, so we sort of stand by what we report in terms of our validity results.

Also the measures that we validated the SRR against have also undergone some definition changes. This would include all of the measures, so the hospitalization and mortality measures in the prior round did not include adjustment for prevalent comorbidities

and now they do. And the SMR is also underlying the population change where it was restricted down to the Medicare-only population.

So this likely may have influenced potentially some change in the correlation coefficient. But at least, again, we maintain that the direction of those associations were very consistent and still very statistically significant.

There are two intermediate outcome measures that really reflect the delivery of care by the dialysis facility. And so I think this may address one is the panel member's concerns about using sort of measures to validate against reflecting what the dialysis center can do to help manage outcomes in their patients.

And so the fistula measure and the catheter measure both reflect care that is delivered directly by the dialysis center where they have some control over outcomes that may help them manage hospitalization readmission in their patients.

Both of these measures were correlated with the SRR in the expected direction. The measures that we used however have undergone changes since the original submission. So that would account for some of the changes that we're seeing in the correlation coefficient.

So the other change that was made was how we handle Medicare Advantage patients. They were included in the measure before. But they were not accounted for in a way that makes sure we are capturing all available comorbidities that are used for a risk adjustment. So we've modified that in the current SRR model.

So given sort of the range of changes and, both to measure specifications as well as to underlying data source, we do feel very comfortable with the correlation coefficients and the empirical validity testing results.

And we think that this really demonstrates quite good stability from the previous submission in 2014 to this current submission and think that the results represent

both stability and robustness in light of those changes that were made.

I will stop there because I know we're short for time, unless there were other specific issues in validity that you would like addressed regarding Medicare Advantage -- the inclusion of Medicare Advantage patients.

CHAIR CELLA: Thank you very much, and thank you for being concise. Any other questions or comments from the committee, from the subgroup?

MR. MESSANA: If I may respectfully ask, this is Joe Messana. I would add one last, final comment. The -- one of the discussants at the beginning of this section talked about or questioned the use of inpatient claims only.

I think it's important to know that both for Fee-for-service and Medicare Advantage patients, the average number of claims-based diagnoses identified from inpatient claims as in the 12 to 14 range in this population and didn't differ between Fee-for-service and Medicare

Advantage, differed by single-digit, small percentage.

So we feel very comfortable with the decision to use inpatient-only claims as a reasonable compromise between the possibility of excluding Medicare Advantage patients entirely from the measure which would be, we think, an overreaction. And we're comfortable with the use of inpatient claims as a reasonable source for identification of claims-based comorbidities here. Thank you.

CHAIR CELLA: Thanks. Jack Needleman, your hand is up.

MEMBER NEEDLEMAN: Yeah, hi. Thank you. That's actually very relevant information and very helpful in thinking about how well the risk adjustment is performing from the hospital-only data.

But was any analysis done for the Medicare Fee-for-service population comparing the risk adjustment estimated expected patient and then by facility using just the inpatient data

and also using the look-back on the outpatient claims as well?

CHAIR CELLA: Response?

MS. DAHLERUS: So do you mean just including -- so excluding Medicare Advantage patients and just --

MEMBER NEEDLEMAN: Basically yeah. So you're making this change in method and you have an old method and you have new method. And you're applying the Fee-for-service patients, and I get that. But did you do the comparison of what would be of value from the Fee-for-service patients using the old method? What do we get using the new method? And how comparable are the results?

MS. DAHLERUS: Okay. So I'd have to check with our analyst. I don't think we did an analysis that's just restricted to Medicare Fee-for-service patients. We did compare the predictability of using -- of outpatient versus inpatient claims with respect to how well they predict the outcome.

And as expected, use of inpatient claims were far more predictive than outpatient claims. And because we do get inpatient comorbidities for Medicare Advantage patients, we felt very comfortable again with the decision to include them.

But in terms of the availability of comorbidities for Medicare Advantage patients, it's quite similar to the Fee-for-service population. So we don't think that we are missing a lot by excluding the outpatient claim. We did not -- but I don't think that we compared Fee-for-service versus Fee-for-service plus Medicare Advantage.

MR. MESSANA: Claudia, I'll add a general observation that may help in response or may provide some information in response. So in other measure evaluation, we have compared the impact of inpatient only versus inpatient and outpatient in the Fee-for-service population for mortality measure development. And we saw a very small effect -- a very small difference in the

modeling and in the C-statistic using inpatient-only claims.

And I think that's generally consistent with the approach that many in the nephrology field use, including United States Renal Data System, who basically apply a premium to inpatient claims.

The way they score inpatient -- excuse me, score claims-based diagnoses is that if it's present on an inpatient claim, they consider it identification of a comorbidity. For outpatient claims, they have a different threshold that has to be -- it has to be present on more than one outpatient claim separated in time and/or venue to be considered evidence of a comorbidity.

So they take the general approach of putting less emphasis on outpatient claims in this population, if that helps the answer.

CHAIR CELLA: Thank you.

MR. KALBFLEISCH: It may also be relevant to note that Hospital Compare also uses just inpatient claims has made that change as

well. So they also looked at that question.

CHAIR CELLA: Thank you, again. I think it's time to move to a vote unless there's any final urgent comments.

MS. WILBON: Hi, Dave. It's Ashlie. I did just want to just -- before we vote on validity, just a point of consistency. The approached used for demonstrating validity in terms of a correlation with other measures using experiments was used on for this measure as well as for a couple of others submitted from U of M: 1453, 0369.

Both had kind of similarly low correlations but seemed to be all statistically significant and also in the direction that the developers hypothesized.

So I just want to make sure, from my understanding and perhaps for others from a consistency perspective, what may be different about this measure and the parts that were reported versus the other two that passed. And I just want to make sure we're being consistent.

If there's something different, then certainly let's make sure that's clear.

CHAIR CELLA: Can somebody address that? Were all three of these with Subgroup 2?

MEMBER NEEDLEMAN: This is Jack. I believe so. And we haven't used -- we have accepted those correlations, that methodology for establishing empirical validity before, not only in this cycle but in prior cycles.

And the issue is whether the correlations are high enough given that they're measuring different things. And I actually found the correlations high enough. My complaint was not with the correlations, per se. It was with the leaning on statistical significance rather than on the magnitude of the correlation.

MS. WILBON: So sorry, this Ash. I do want to say that Measures 1463 and 0369 were reviewed by Subgroup 3. So there was kind of a different set of panel members reviewing those. But again, for the sake of consistency, I just want to make sure that we're consistent across

all of the subgroups in that way.

CHAIR CELLA: Well, I think on the one hand, Jack is clarifying that he's focused not on the methodology but on the use of statistical significance over magnitude or scientific correlation.

I don't know that anyone can speak to whether that's the basis for Subgroup 3 passing the other two submissions, but it sounds like it's independent. Can anyone else comment on that?

I mean, Subgroup 2 members didn't dig into Subgroup 3 measures. And Subgroup 3 members didn't dig into Subgroup 2 measures. So you're raising something that might not be that easy for anyone to speak to.

MS. WILBON: Right. I wonder if some members of Subgroup 3 which we're going to be moving to them shortly. So hopefully, there's a few on the phone might be able to just talk about their consideration of the correlations that were submitted for the other -- those two measures:

1463 and 0369.

Hopefully, I'm not putting anyone on the spot. But if we could, if there is anyone that could speak to that and we could just -- I just want to make sure that before we move forward with a vote, that we are just being consistent so it doesn't -- it's not something that we have to come back and rectify. ZQ, I see your hand raised.

MEMBER LIU: Yeah, I think it was in the past cycle and also in this cycle including the measure in this subgroup, I have seen other measures and correlations to validate reason for validity test. I think our focus should not be on, again, the scientific association, right? It's based on your hypothesis whether you think there should be what kind of association.

So it's not just, like, you really need to have some really strong association. Sometimes it's a negative association. Sometimes there's no association. As long as you expect -- a reason to expect that, I think that's fine.

MS. WILBON: Thanks. That's actually really helpful. I think that will help and certainly for us in writing up summaries that we will be able to explain that.

MR. KALBFLEISCH: And if I could just comment on that briefly. We certainly take the point that it's the value of the coefficient that matters, not the significance. If we gave a different impression, we didn't mean to.

CHAIR CELLA: And Gene, would you like to comment? You have your hand up.

MEMBER NUCCIO: Yes, sorry. Just I mean, the values, with all due respect to Jack, and for standardized mortality rate, the R-value is 0.1. For long-term catheter, 0.04. And for fistula, it's -0.06.

So I mean, there's only one that's above 0.1. And I did not find that as demonstrating a meaningful difference or potentially creating how these things are meaningfully related, the statistical significance to p-value notwithstanding. And I

did consider those and didn't find them persuasive.

CHAIR CELLA: Ashlie, do you feel like you have enough information to differentiate this from the two that are in Subgroup 3?

MS. WILBON: I think so. I think so.

CHAIR CELLA: So then we should vote.

MS. WILBON: Yes. So Subgroup 2 members, I do just want to do a quick check. I think I've heard or seen most everyone on the webinar at this point. I've heard Bijan. I've heard Christie. I've heard Gene. I've heard Jack. I've heard Jeff. Jen, I've seen you, but I didn't hear you. Are you there?

MEMBER PERLOFF: Yes, I'm here.

MS. WILBON: Okay. John, I think I've seen you --

(Simultaneous speaking.)

MEMBER BOTT: I'm here.

MS. WILBON: Okay, perfect. ZQ just commented. I think the only one missing from Subgroup 2 is Susan, unless she joined us since I

last checked.

Okay. So we'll have eight Subgroup 2 members voting. If you could find the voting link in your email that was sent this morning. And click there, you should see we're going to be voting on both reliability and validity.

MS. INGBER: Right. So voting is now open for reliability on 2496.

Okay. I'm just going to share my screen real quick. Okay. So as you can see, for 2496, we have zero votes for high, five votes for moderate, and three votes for low, and zero votes for insufficient. The measure therefore passes on reliability.

MS. WILBON: Okay. Thanks, Hannah. So the next vote we'll be making will be on validity, and Hannah will give you the process forward.

MS. INGBER: Okay. So voting is now open on validity for Measure 2496. Your options are high, moderate, low, and insufficient.

Oops. I'm sorry. I was on mute.

Okay. I'm going to share my screen again to show the results. Okay. So you can see here for 2496, for validity, there was zero for high, three for moderate, five for low, and zero for insufficient. Therefore, the measure does not pass on validity.

MS. WILBON: Okay. Thank you to the University of Michigan development team and for joining us and for engaging the discussion today. And I hope that the feedback from the panel was helpful and that you will continue to engage in the process and that we will be in touch.

The next measure up for discussion I think may also be some of the same team members. It is Measure 3566, Standardized Ratio of Emergency Department Encounters Within 30 Days of Hospital Discharge for Dialysis Facilities.

Can someone from University of Michigan let us know? Is it the same team, or will be a different set of colleagues?

MS. DAHLERUS: Hi, this is Claudia Dahlerus from University of Michigan. It'll be a

subset of the same team. So it will be myself and I believe that Dr. Jonathan Segal is also on the line --

MS. WILBON: Okay.

MS. DAHLERUS: -- and also Dr. Kalbfleisch.

MS. WILBON: Okay. I think Dr. Segal, I don't know if we spoke to him. Do you want to just test to make sure we can hear you?

MR. SEGAL: Yes, this is Jonathan Segal. Can you hear me?

MS. WILBON: We can; great. Thank you. So I wanted to thank you guys for joining and just give a brief overview of this measure, 3566. It's a new measure that was submitted by the University of Michigan team. It is a measure initially that we made an error in the tally of the reliability vote.

So initially, we communicated that it had passed reliability. And we identified the error as we were preparing for the meeting and finalizing materials. So the developers did not

have an opportunity to submit a written response. So we did ask them to join in order to provide a verbal response to the panel's concerns.

And I think we may find some efficiencies because they used very similar methodology for reliability that we just discussed with the IUR and the PIUR. But we'll get into that shortly. But again, just wanted to thank you all for joining us and for being accommodating.

So I'll just briefly review the description of the measure, and we'll hand it over to the lead discussants, Eric and Marybeth, to discuss some of the concerns with reliability before we open it up for discussion and conclude with a vote.

So again, this is a new measure. It's the Standardized Ratio for Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge for Dialysis Facilities. It's a ratio observed to affected events. And it is based on claims and registry data measured at the facility

level. It is risk-adjusted.

Again, the concerns for this measure were with the reliability. The developer used the IUR with bootstrapping as well as presented a PIUR value which is listed here in the discussion guide. And I think there were a couple of concerns here, both with the results and some of the specifications. But I'll hand it over to Eric and Marybeth to review some of the concerns in more detail.

MEMBER WEINHANDL: All right. This is Eric. Can you hear me?

MS. WILBON: Yes, we can.

MEMBER WEINHANDL: All right. Excellent. So this is the Standardized Ratio of Emergency Department Encounters 30 Days after Hospital Discharge. I will say that my initial reaction to this measure was extremely positive. I think that this is the measure that fills an obvious gap that exists in readmission metrics from dialysis facility landscape.

Just to give a little bit of a

background, as the group just discussed, the measure around hospital readmission during the 30 days after discharge, what there has been in the dialysis population over the last approximately five, seven years is a pretty steady increase in emergency department encounters to the point where they're occurring essentially as frequently as hospital admissions are.

So during the 30-day readmission or post-discharge period, a patient could present for any number of acute care needs. They may go to a hospital, be admitted as an inpatient. Or they may go to an emergency department and be discharged to home.

So I think that thinking about it from that perspective helps to clarify what the rationale or need is for this metric. And I think it's obvious to me.

And so that dovetails into the exact specifications of this measure. Very much of this measure is very analogous and perfectly harmonious to the standardized readmission ratio

that we just discussed.

Insofar as patients are discharged from the hospital, the first three days of follow-up during the post-discharge are not tracked by the metric, but instead emergency department encounters are included from days 40 to 30 plus discharge.

Now because these outcomes, the emergency department encounters, are truly not just emergency department and then immediate transfer into a hospital bed, the outcomes that are tracked are only those are taken from outpatient claims insofar as the patients is in the emergency room or observation status and then is discharged back home.

Because of that difference with the standardized readmission ratio, that does necessitate that the measure is limited to the Medicare Fee-for-service population. And that may be something for the subgroup and for the panel to consider insofar as it is a subset of the denominator that's tracked in the

standardized readmission ratio. Those would just be for service.

And to give you a qualitative sense of what we're talking about, I would say that approximately 80 percent of the dialysis population is either in Fee-for-service or Medicare Advantage. And of those 80 percent, about three in four are Fee-for-service. So we're dealing with about 75 percent if not more.

The measure is risk-adjusted -- a wide variety of factors, demographic factors, comorbid factors. As was discussed with reliability and validity, reliability was the domain in which this measure did not have.

And I'll specifically note that the inter-unit reliability to 0.451. The profile inter-unit reliability was 0.570. That is the number. It's the 0.570, even for detecting outliers but probably caught the attention of many reviewers within the subgroup.

To the extent that there are forces moving around the dialysis population,

particularly with respect to potential Medicare Advantage enrollment in coming years, it's reasonable to think that the number of index discharges per dialysis facility may actually decrease so long as this measure is restricted to the Fee-for-service population.

And so I would encourage the group to consider whether these profile inter-unit reliability values are likely to remain stable or potentially decrease with some erosion of sample size in the future. Not a guarantee but a potentiality.

And as far as validity is concerned, the measure did pass. I won't speak a lot to it, but I do want to pay attention in the discussion guide to some of the values that are in Table 2 where the standardized measure was correlated with respect to other measures.

I actually found this to be quite heartening in many regards and quite positive. You'll notice that there is a modest but notable correlation with the standardized mortality

ratio. As one would expect, the higher mortality during the post-discharge period would be associated with any demand for acute care.

I'd also point out that that correlation between this measure and the standardized transfusion ratio which is the second one in Table 2 is quite a positive thing.

The dialysis population is approximately 20 percent of transfusions occur in the emergency department setting. And you would expect patients to generally be relatively more anemic, potentially blood loss, during the post-discharge period.

So to see that difference in the standardized transfusion ratio between better than expected and worse than expected facilities, that's an extremely positive sign with respect to the validity -- face validity of this measure.

And then I'll also point out the second to last row, which is an interesting thing in its own right, the standardized readmission ratio takes the value of 1.00 on average in those

facility with a better or as expected. And it also takes the value 1.00 in those worse than expected.

To the extent that these two outcomes, the standardized readmission ratio taking hospital readmissions, this measure is taking emergency department readmissions so to speak, emergency department visits in the post-discharge period but the patient goes home instead. One might hypothesize that these measures ought to be sort of orthogonal and that they should reflect differences in disease severity.

So actually, to see the absence of correlation for me -- it may not be the case for every other person in the subgroup -- but for me, the absence of correlation was actually a strong feature of the measure as it demonstrated that this really does fill an information gap that the readmission ratio is currently not filling.

So I think that when it comes down to it, the validity for me was quite strong. Obviously, most of the grades were medium. But

reliability and the nature of the profile inter-unit reliability being 0.57 is definitely a concern for the panel to consider.

MS. WILBON: Were there other subgroup members -- Subgroup 3 members who wanted to comment on the reliability in particular. I know we had a very similar discussion with the prior measure. But if there's anything you or any comments that you would like --

(Simultaneous speaking.)

OPERATOR: Pardon me. Pardon me. Hello? Hello? Can you hear me?

MS. WILBON: Hi, yeah.

OPERATOR: Hi. My name is Gigi. I'm the lead operator for today's call.

MS. WILBON: Yes?

OPERATOR: Yes. Okay. Can you hear me, all?

MS. WILBON: Yes, we're on a call. Hello?

CHAIR NERENZ: Let's go ahead.

OPERATOR: Wait one moment. Never

mind. Sorry. Disregard. Disregard. Never mind.

MS. WILBON: Hello?

CHAIR CELLA: Go ahead, Ashlie. Go ahead, Ashlie.

MS. WILBON: Hi. Sorry about that. I'm not sure what happened.

CHAIR CELLA: She made a mistake.

MS. WILBON: Oh, okay. I wasn't sure if I hit a button something. Okay. So we'd already discussed reliability with the previous measure. And so I just wanted to open up to the subgroup to see if there's any additional comments regarding reliability that might be new or different for the developers to respond to before we give them an opportunity to comment on that.

Again, I'm not sure if there's anything different. But I do want to give the developers an opportunity to respond if there are any additional comments about that.

CHAIR NERENZ: Yeah, Dave Nerenz here.

And I appreciate the great similarity between the issues in front of us here and what was discussed in the last one since we turned to the developers.

So I would appreciate it if they would focus on anything that might be different in the realm of reliability between this measure and the one we just discussed. Otherwise, I think we had a very clear and very thorough presentation just a few minutes ago.

MR. KALBFLEISCH: This is Jack Kalbfleisch. It seems to me that the issues are quite similar here with the IUR and PIUR. I think that the level of the IUR and PIUR certainly within the bounds of things depicted and have been approved before, although don't meet the very high threshold that the Commission talked about recently.

MS. WILBON: Thanks. I did want to just check in with other Subgroup 3 members or anyone else from the panel who may have comments. And if there are no further comments given the

conversation we had for the previous measure, we can call the vote.

I do just want to do a quick check-in before we do that with Subgroup 3 members to make sure we have a quorum for the voting. Alex, I see you. Can you hear us?

MEMBER SOX-HARRIS: Yes.

MS. WILBON: Dave, I know you're there. Eric, I know you're there. Joe Kunisch, are you there?

MEMBER KUNISCH: I'm on.

MS. WILBON: Okay, great. Lacy?

MEMBER FABIAN: Yes, I'm here.

MS. WILBON: Okay. Marybeth?

MEMBER FARQUHAR: Yes, I'm here.

MS. WILBON: Hi. Paul? Paul, are you there? Okay. Sam?

MEMBER SIMON: Yes, I'm here.

MS. WILBON: Okay. And Sean, I think I saw or heard you.

MEMBER O'BRIEN: I'm here.

MS. WILBON: Okay. We'll have eight

people voting. Hannah, Subgroup 3 members, if you could locate the email from this morning that was sent from our team with the voting link in there. If you could locate that, Hannah will be pulling up the vote for reliability for Measure 3566.

MS. INGBER: Thanks, Ashlie. The vote for 3566, rating for reliability is now open.

We're just waiting for one more vote.

Okay. I'm going to share my screen to show the results. Okay. You can see here for Measure 3566, the overall rating for reliability, we received zero votes for high, three votes for moderate, five votes for low, and zero votes for insufficient. Therefore, the measure does not pass on reliability.

MS. WILBON: So this is Ashlie. I just want to, again, call a point of consistency. So we just reviewed another similar measure with an IUR and PIUR that was lower. So I think it is a little problematic in terms of consistency for us to make the case for the voting results.

Can someone speak to that? Or potentially if we might revote, I just want to make sure we're being consistent, and it doesn't appear that we are kind of in sequential order.

Also, considering that it is a different subgroup. But subgroups, we're part of a whole. And the votes represent the whole panel. So I just want to make sure that we are being consistent and we have a discussion about that.

CHAIR NERENZ: Ashlie, Dave Nerenz here. I fully appreciate that concern and problem. One possible task, if both of these measures can make their way onto the standing committee, it seems to be the reason we had this discussion for the past hour or so and the reason that things came out the way they did is this is a really, really close call on both these measures in terms of whether the reliability is acceptable or not. Part of it is just simply the numeric values of these two statistics on each one.

But also part of it is the idea that when we endorse a measure, when ultimately NQF does, it's not for specific use like identifying outliers. It's for a whole range of uses, including building in the star ratings and things like that.

It's, in my mind, very, very problematic. I'm right teeter-totter on this on whether, at least in our hands, this or the other one should go forward or not. And it's perhaps that ambiguity or that close call nature could be passed on to the standing committee.

Eventually, we end up these pass-fail distinctions. But at least in my mind, both of these are just teetering right on the edge. It's really hard.

MS. WILBON: Thanks for that, Dave. Any other comments from the subgroup? Or I will say that with our current criteria that issues around threshold values strictly for reliability is grounds for the standing committee to reconsider the measure and revote. So the

measure could and likely would be reconsidered by the standing committee.

But I do think it's an important point of consistency for the panel as well because certainly again while we do divide the group, the panel and the subgroup vote on that, we do want to maintain some cohesiveness as a panel and what the votes and recommendations are for the panel. So if there are any other comments about that, I think it would be really helpful.

MEMBER SOX-HARRIS: So this is Alex. I would just add that this highlights the importance of our discussion yesterday and the planned future work to try to tighten our understanding and consensus on standards.

Prior to that discussion, there had been a very loose consensus. Lots of different interpretations about what constitutes that reliability coefficient. So I think what we're seeing here is a reflection of that state of affairs.

MS. WILBON: Hi, Joe. I see your

hand.

MEMBER KUNISCH: Yeah. Just I kind of second Dave and the comments that it has to be consistent because experience coming from the same measure developer. If you pass one and they're very similar in results. And then they go back to say, okay, how do we get this one passed through the second round?

And it was really failed on -- just dependent on who reviewed it. I struggled with this and the low scores too, but I did from the very beginning when we started this in our very first round of not really having clear direction on thresholds. Because when I first started, anything that I thought was too low a threshold, I was failing right away. And then had a lot of offline discussions with Karen at the time.

And I still struggle with it because, yeah, I see some of these results and I just think they're very low. But again, unless we have something to say this is the cut point, you have to give a pass.

So if you get a pass on one measure with similar results, it just wouldn't be good to not pass the second measure again because they have to go back and prove it. And it may depend on who's going to review it in the second round.

CHAIR CELLA: This is Dave Cella. Just to chime in, I want to return to what Dave Nerenz suggested which is to bring this to the standing committee and educate them about how this really is right on the brink.

It would've been interesting if we did the shadow voting this cycle to see if there would've been a consistent up/down if the whole committee weighed in because to some extent there is this issue of different reviewers in subgroups. And when you're at 5-3, 3-5, that's all it really takes to see it go one way one time and another way the other.

So I would endorse kind of moving on from here. The vote is vote. Pass it along to the standing committee and let them deal with the issue, looking at both submissions or all three

of them.

MS. WILBON: I see a hand from Sherrie. But I'd like to just push back a little bit because I think it presents an issue I think for the methods panel, right? Because the vote that moves forward are from the methods panel. So it doesn't say subgroup. It says methods panel.

And so I think the signal that it sends from -- an inconsistent vote coming from the methods panel I think is probably one that we should try to rectify. I'd like to just put out a suggestion that we revote and see how folks feel about that because it is a little disconcerting.

And I just want to -- and certainly I've heard responses from several folks about that. But I just want to kind of put it out there and we do have a mechanism in place. But I think the signal that it sends, I think -- which we talked about before with the signal that it sends when we can't vote on risk adjustment or

social risk factors.

Again, I think the vote coming out of the methods panel as a whole I think makes a difference. So I'll just stop there. There's a couple hands from Eric and Sherrie. I think Sherrie was first.

MEMBER KAPLAN: Ashlie, can you hear me?

CHAIR CELLA: Yep.

MEMBER KAPLAN: So I agree and sitting on another steering committee. And so the signal would be very confusing and it would really, I think, cause confusion in the steering committee. I mean, if you send it back with this kind of inconsistency, it'll be very disconcerting to those sitting on the committee.

And I do think being valid but not reliable is another signal we've got -- issue we have to address. Because if you're accurately -- if you're inconsistently accurate, it's a very curious kind of -- from a measurement standpoint, it's a very curious situation to be in.

So I think that the issue, I think you're absolutely right. I think we should probably revote.

MS. WILBON: Eric and then Alex.

MEMBER WEINHANDL: Yes. So this is Eric. I think that revoting is reasonable. However, I think that there's a bit of a cognitive bias to the extent that Subgroup 3 is being asking to revote in light of the results of Group 2's evaluation of readmission.

So I wonder if we should be doing a bit of a crossover, and this my perspective, where Subgroup 3 has a chance to vote on the reliability of the readmission metric and Subgroup 2 has a chance to vote on the reliability of this metric.

MS. WILBON: Okay. That's a discussion we'll put on the table. I think it was Alex, Jack, and then Lacy, I think.

MEMBER SOX-HARRIS: I was going to suggest a version of what was just suggested, either having the entire panel vote or have both

subgroups vote on both measures instead of just having Subgroup 3 revote or just Subgroup 2 revote.

MS. WILBON: Okay, thanks. Jack and then Lacy.

MEMBER NEEDLEMAN: Yeah, I suspect what we're seeing in -- the two votes are not that far apart. They just work in opposite directions. And I think what we're seeing here is real disagreement among the two groups of eight about what level of reliability is acceptable.

I suspect if we had both groups voting on both measures, we'd wind up 4-4, consensus not reached. And I think the fact that consensus is not being reached with a pass on reliability for one of the measures and not passed on reliability on the other measure just reflects the fact that, as somebody said, the reliability number here is different. It makes some people uncomfortable and less people less uncomfortable.

So I have no suggestion other than I

think if we revote, we're going to wind up with the same results.

MS. WILBON: Okay. Lacy and then Alex.

MEMBER FABIAN: My thought was just to second that we either revote as a whole group on the couple of measures or we -- the subcommittees vote on -- revote on the measures because otherwise really the only option is to default to the prior vote of the other subcommittee to address the issue which presents the bias.

MS. WILBON: Okay, thanks. Alex, did you have a -- I saw your hand go down. I wasn't sure if you had any comment.

MEMBER SOX-HARRIS: No, I put my hand down.

MS. WILBON: Okay. Dave and Dave, do you have any thoughts on this?

CHAIR NERENZ: Yeah, just a couple of things. One thing, just watching the time and also just thinking about kind of make these kind of decisions on the fly. The one that's easy to

do is just have a Subgroup 3 revote right now.

Go ahead then and take the break. And depending on how that comes out, we may want to pick out the different strategy that we could do, like, at the end of the day today or something. I assume people will still have the issues reasonably fresh in mind. But I don't think we have the opportunity here to try two or three different variations and sort of make them up on the fly as we go on.

MS. WILBON: Sure. Fair enough. Dave C., do you any thoughts?

CHAIR CELLA: Well, I think the options that we have, also looking at the time, I'm actually least comfortable with just having a Group 3 revote and going with that. Having a Group 3 revote to come to a discussion would work better for me.

But I do think that in a sense it's like asking people in Group 3 to reconsider and go with the direction of Group 2 which is really just an order effect on something that's -- I

guess I don't necessarily agree that it would be terribly difficult for a standing committee to understand that this was just where we are at this point in time in terms of this issue, unless you want to try to resolve it later in the day if we can move through the other reviews.

MS. WILBON: Okay. Let's do that. Also, there's a lot of options on the table. And given that we're already into our break time, I'm inclined to table this. We'll talk offline and come back with another plan to see how we can resolve this before the end of the day.

So thanks to all. We'll break now. Oh, sorry. Let me do a brief public comment. Thank you to the developers again, all of you for going overtime. But is there anyone who'd like to make a comment, press *1 for the operator to open your line and we'll take that now. Or you can enter a chat into the chat box.

Okay. Seeing none, we'll go ahead and break. We'll return from break at 11:15. I realize the slide says 30 minutes, but it's

actually 15. We're squeezing in an additional measure on the second half of the discussion. So a short break. We'll see you back in 10 minutes. Thanks, all.

(Whereupon, the above-entitled matter went off the record at 11:05 a.m. and resumed at 11:18 a.m.)

MS. WILBON: Hi, everyone; welcome back from a very short, brief, break. We are here to convene and put our heads together, and we're going to share our path forward here. Essentially, since the panel seems to have not reached consensus on reliability for these two measures, we'd like to put these forward to the Standing Committee; put the vote for reliability for both 2496 and 3566 as consensus not reached.

That kind of eliminates kind of re-voting and getting the same inconsistencies. And so essentially what the Standing Committee would need when they look at these measures for reliability for both would be consensus not reached from the SMP. We would not share

individual votes per se, for high, moderate, and low.

What we're going to do in order to solidify that decision is, after you guys do respond to a brief survey that is going to be opened up via the survey link that you all have access in the last two minutes, if you could open that up, I think Hannah is going to have it open for you. And we're just asking whether or not you agree that these measures will be put forward as consensus not reached in the Standing Committee; just a yes or no response.

So I do just want to, before we do that, if there's questions from the Message Panel members or others before we cast votes. Can you guys hear me?

Are we in the full --

MEMBER AUSTIN: Yes, this is Member Austin, so I can --

MS. WILBON: Oh, okay. There was radio silence; I wasn't sure if I was actually speaking to everyone --

MEMBER AUSTIN: That sounds like a fine solution, Ashlie.

MS. WILBON: Okay, perfect; absolutely perfect.

CHAIR CELLA: This is Dave Cella. I'm back Ashlie, hi.

MS. WILBON: Hi, hi. Okay, great. So if everyone could click on the link for voting, there will be a survey there for you to respond to. We just want some acknowledgment from the group, and this will be a full-panel vote, not just Subgroup 2 and 3 members. It will be all Message Panel members will vote on whether or not you agree that these measures will be put forward, consensus not reached for reliability to the Standing Committee.

MS. WILBON: You should have 25 votes, Hannah.

MS. INGBER: Right. I'm reporting 20.

CHAIR CELLA: If you have 13 in one direction, you could probably close it, if the number of voting members is 25.

MS. WILBON: It still looks like we've got unanimous agreement there; I think Hannah will show it briefly. But thanks to all for that quick resolution, and these measures will be put forward to the Standing Committee as this is not reached for reliability.

Thanks again, and we'll dive back into evaluating the remainder of the measures on the docket for this afternoon. For this morning, we're starting with Measure 2539, Facility 7-Day Risk-Standardized Risk Hospital Visit Rate after Outpatient Colonoscopy, presented -- this involved the Yale CORE team and CMS. They just want to check in to make sure the Yale CORE team is on the line.

MS. PETER: Hi, this is Doris Peter. Yes, I think we're all on the line. Elizabeth, are you there?

MS. DRYE: Hi, yes, thanks.

MS. WILBON: Hi, great. We can hear you. So we will get started here; let me find my place. Okay. So this is a maintenance measure.

Again, I'll just give a brief overview here, and I'll hand it over to the lead discussion, Alex and Sean, also noting that DQ is recused from this measure, and we'll file through here.

So maintenance measure, last endorsement was in 2014 at the Facility level Risk Standardized Rate of acute, unplanned hospital visits within seven days of a colonoscopy for a hospital outpatient department or ambulatory service center among Medicare Fee-for-Service stations 65 years and older.

The claims-based facility-level measure is risk reducted. The measure did pass reliability, so we will not spend time there. The focus of discussion will be on the one with validity where consensus was not reached. Because this is a maintenance measure, the developer is asked to submit empirical validity testing at this point. They did re-share the base validity assessment they had done before and also provided a rationale for why they were unable to do empirical validity testing and

provided a description of their analysis and consideration of other measures.

I did just want to note also that you should be looking at the version of the discussion guide that was attached to the viewing meeting invitation, whereby the five measures that they considered us potential comparators are listed in the discussion guide under Action Items.

With that, I would hand it over to Alex and Sean to walk us through the methods panel concerns, and we'll focus here on validity.

MEMBER SOX-HARRIS: Great, thank you. This is Alex. So I wanted to commend the developers for the methodology and results and their face validity analysis, because it's really well done and, as far as face facility goes, pretty compelling, although as Ashlie just mentioned, the NQF criteria for maintenance measures are that there's a requirement to submit empirical validity testing at the time of maintenance or a rationale.

So the rationale provided for not doing empirical validity testing was mostly focused on the inability to find a relevant measure to correlate with that also had adequate sample size within entities, because different entities do single procedures and therefore wouldn't be able to calculate both measures of the same site and so forth.

So I'm going to, for the purposes of this discussion, cede that point and say, Okay, there's not another measure to correlate with. So that particular form of empirical validity testing is not feasible.

My concern and suggestion was that another kind of empirical validity testing be done that is feasible, which is to do some testing on the validity of the outcome; in this case, ED visits or admissions to the hospital seven days after colonoscopy.

So my validity question is simply, out the additions that are being counted, how many of them are plausibly related to the colonoscopy?

And that's a simple, element-level validity question that I proposed in my initial review that could be done.

So the developers came back with the response that, in fact, they had done work related to that, which is, on the one hand, great, but also undercuts the rationale for not doing an empirical validity testing when, in fact, it had been at least partially already completed.

The evidence that was presented was twofold, one is a reference to research that had been done in a single site where 68 percent of the ED visits after colonoscopy were plausibly related to the procedure itself. That's the 68 percent ascertainment of what I think the measure is trying to get at.

The issue with that is, that's just one site, so I'm really curious about the distribution of the proportion of admissions that are related to colonoscopy.

The other point of evidence that was

offered was a paper that was published, and if you go to the very end of the discussion guide, there's a Table 2, copied and pasted from a paper which I also looked up and reviewed. This is the top 10 most frequent diagnoses accompanying an unplanned hospital visit within seven days of an outpatient colonoscopy.

So this is exactly the kind of information I would have accepted as empirical validity testing of the outcome. My current issue is that this only represents about 30 percent of the unplanned hospital visits, and I don't have the information on what the other 70 percent -- you know, what were the reasons for the other 70 percent of the visits.

I looked through the paper and some of the supplemental materials briefly, but I couldn't easily put my finger on that.

So this issue I'm left with -- And I think the reviewers may have an answer to this question or may be able to get it, but I would be satisfied with what's the distribution of the

proportion of visits that are plausibly related to colonoscopy? And that's -- I guess it is simple current validity testing outcome that could have been provided. Therefore, I'm not persuaded by the rationale that empiric testing is not feasible.

That's one issue. The other issue that was related to the validity ratings for this measure had to do with the inclusion, the analysis, and decisions to include SES variables in the risk-adjustment model. And there were some specific questions raised by the subcommittee on that, and also some responses from the developers. I'll leave it to my co-discussant and other subcommittee members to get into the details on that issue, and I'll stop there.

MS. WILBON: Okay, thanks, Alex. Sean, did you have anything to add to that before we open it up to the other sub-group members?

MEMBER O'BRIEN: I'd say go ahead and open up to the other members, and I may weigh in

on their responses.

MS. WILBON: Thanks.

MEMBER O'BRIEN: Very comprehensive.

MS. WILBON: Other Sub-group 3 members have comments on this measure?

CHAIR NERENZ: Yes, Dave Nerenz here. Just on the issue of time and to stay focused, I would be perfectly happy if we just didn't get into the social risk factor discussion because, at least in my mind, it didn't bear on my vote on either the reliability or validity. I've raised concerns in that area, but it's not something that we're voting on. So if it was up to me, I'd leave that discussion aside and focus purely on what leads to the focus on validity.

MS. WILBON: Thanks. Other sub-group members comments? Otherwise, we'll allow the developers to provide their response at this point.

Okay. Yale CORE team, would you like to respond to the concerns laid out?

MS. DRYE: Sure. Hi, this is

Elizabeth Drye from CORE. I'm one of the senior presenters. I'm one of the key members on the phone, who worked on the center for a long time, so we appreciate the committee's careful consideration, and I apologize for connection issues, can you guys hear me okay? This should end in a second.

MS. WILBON: Yes.

MS. DRYE: I'm working from unpredictable settings. So I think the main question, the residual concern I'm hearing on empirical validity really has to do with the validity of the outcome as reflecting the quality of care. And so as the committee saw, we tried to provide more detailed information.

We had not run those analyses on the current data set, but we have them from prior -- you know, very similar data sets; they're just earlier years of that cohort, using the same hospital and ambulatory surgeries in the claims. So they should be completely consistent with what validities in our current application. We didn't

rerun them because they're actually kind of burdensome to do that especially now with limited access to data, given the shelter in place requirements.

So the reason we presented this type -- you know, most frequent is because to give you a good picture, there's a long tail those which, I think we could probably get those up and share, but in the bigger picture, so if he wanted that, we could try to hold those prior spreadsheets out; I think we would be fine with that.

But in the big picture of thinking about this outcome as a signal of quality. You know, when we're thinking about it and its validity, we looked at these agreements for return, we run them by our clinical experts, in this case largely gastroenterologists and even physicians and others we think about, could they be related or not?

And there's not a bright line between the difference, we've talked about this many times of related versus -- You can't say, Is this

related or not, for sure. Somebody comes back with abdominal pain, you know, maybe it's one thing, maybe it's not. Probably it is. If they came back with symptoms, is it related to the course of care, their prep for the procedure, the anesthesia, the post-anesthesia care, the restarting of their medications.

There's hammering of the post-procedure period, how good the instructions were. We don't know because some of them are related or not related.

So we have not tried to -- we couldn't give you exactly what percent are related or non-related because it's just not knowable. And I just -- I hear the concern, but I'm also just not sure where to go next because I think those of us who engaged with clinicians on this -- I'm a pediatrician, not a gastroenterologist or adult medicine doctor, but just have vetted it in conversations, looking at data, looking at variations of it, we're comfortable with it. I'm not sure what else we can give you at this time.

We're really -- Supporting research show that, you know, careful review of medical records, those tend to be single site studies, that there are presentable reasons for these that have been submitted within seven days.

So I'm not sure what else I can give you that's totally going to get you to the comfort level of, yes, this is a valid outcome. I'm a little bit limited. I think our group is a little bit limited in what we can do in the very short term, but we're willing to bring back more, if it's helpful and it's feasible from out end.

Because we do have more occasions that we haven't shared data from our prior work. And I'm going to stop there for Craig Parzynski, our senior analyst on this, and he knows what he can access and not in the current COVID environment. So, Craig, did you want to add anything that we might share?

MR. PARZYNSKI: Yes. You know, I think in just the time that we have turned around, we weren't able to do a lot more besides

what we've done in the past. I think if we had a little bit more time we can do a lot of the same thing in our newer data, but we just, given time constraints, weren't able to turn that around quickly enough, given the recent changes in the country and the world.

MEMBER SOX-HARRIS: This is Alex Sox-Harris. So just, of course, speaking for myself, I understand the issues of not being able to completely determine which diagnoses are related to the colonoscopy or not. But I think it's possible to extract some signal, like if there are leg breaks. There are certain things, a proportion of things that are clearly not related to colonoscopy. It might be another way to think about it.

But I think you can do it, and I also think it would be informative and helpful for the interpretation of a measure to know roughly the proportion of things that are being counted that should be counted and how that varies from site to site. I think that aspect of it is important

because it does bear directly on the validity of comparisons.

MS. DRYE: So I don't think we can draw -- this is Elizabeth again from Yale -- It's really hard. We can share that with you for you to make your own conclusions, but it's really hard for us to draw a line about what's related and non-related. That is why EMS has decided to go with admission measures, and in this case that is the 70 is a rate that includes admission. But our decline is all cause on client.

You know, in the early years, looking at readmission, you might look at engagement and remember it, and there were algorithms for related versus unrelated. There were people who really tried to parse that out; there still are; 3M is really focused that way.

But our goal, our hypothesis and the conceptual model for this is that it's not particular outcomes that can be prevented or unprevented, but we can lower the risk of many types of outcomes with better care. So we can

lower the risk of urinary retention, and I mean they would return to the hospital for urinary retention care, which would lower the risk of syncope; lower the risk abdominal pain.

Of course, you can lower the risk because these are fairly low within the procedural realm of influence of that person's study, doing procedure of the hemorrhagic complications in more severe complications, certainly as infections.

So the goal is not in any one sense, kind of, you know, well this is pneumonia, that is urinary tract infection was, you know, must have been and prevented or not, it's the overall. We're charting the total and incentivizing a reduction in the risk across the board.

So the one thing I don't want to comment is that we should come back and say that these are related or unrelated. We can definitely share more information if we have a reasonable time frame for doing that, and that's helpful. But I don't think that we can -- I

think it would be hard to pivot and try to draw a line on what's related or unrelated at this juncture. It's not consistent with our conceptual approach, I guess. That's how I would describe it.

I definitely want to be -- I don't want to say no. I want to get you what you need, but I'm just thinking how it wouldn't be, I don't think possible for our group to produce that distinction for you.

MEMBER SOX-HARRIS: Yes. I'll just say one more thing, and then hopefully some of my other subgroup members can weigh in on this.

But I completely understand what you're saying, and I'm not asking for a bright line. I'm just asking for some empirical work on the validity of the outcome which is what's required from NQF. I think that's possible and would be informative.

CHAIR CELLA: And there's Eric with a hand up.

MEMBER WEINHANDL: Yes, thank you.

Yes, so I have a question or two for the developer and his comments. I agree with you; I've worked extensively with claims data, and I'm generally hesitant to make the related versus unrelated distinction on the basis of retrospective data. So that is not something that I would personally add or be compelled one way or the other by.

I do have two concerns I have with prospective trade validity, and please correct me if I'm wrong. But I believe in all the materials I reviewed, but the subrgroup reviewed, everything we did see, including in that discussion guide within the context of ICD-9 codes, and maybe it's just because of lag of data, but it would be nice to be able to see what some of the principals as far as diagnoses are on the hospitalizations in the ICD-10 era; provide some level of comfort with what we're looking at. So that's one thing.

And then I guess the other thing that I wondered about with respect to validity was

just that as far as I could see, the conclusion criteria did not include the possibility that the patient recently did discharge from the hospital, such that the seven-day period during which hospitalizations are being tracked could potentially be a readmission.

And so the colonoscopy is potentially occurring in the middle of the period of hospitalization and readmission. That's more of a question, and it's whether the developer has had any consideration to that, whether that interferes with matters of certification, or if it loses its interpretability.

MS. DRYE: Yes. Okay. On the first point, yes, I think they could provide us a more and longer list of reasons for return in ICD-10. Craig, do you think that's doable? Again, sometimes we get a one- or two-day turn around; and that won't be doable because we have really limited ability to have analysts running this data right now because of all the restrictions in Connecticut. But we are all actually doing that

work.

But, Craig, what do you think might be doable between the ICD-10 code and now?

MR. PARZYNSKI: Yes, we can definitely run that. I think, in terms of how long it would take, it would probably be in the realm of two weeks, just given the restrictions we have.

But it is something that's definitely feasible, and you know, and as others have mentioned, there would be a lot of speculation as to what's related or not, but certainly is useful. You know, something that definitely appears or not. But definitely feasible and probably about two weeks, I would say.

MS. DRYE: Okay. And then on the second part about whether the colonoscopy could be occurring post-admission, I feel like we might have looked at that before. I'm not sure, but that might be something that we can look at as well, but a conceptual basis for the cohort is that these are elective procedures. The patient can't be in the middle of a hospitalization.

But could it have been in the last week? I think we can probably look at that. I'm not sure if it would really change anything. The limit is drawn to private, you know, patients who are acutely ill. For example, with diverticulitis. But if it's feasible we'll get that, and I can even ask Craig to confirm. I think it will be doable.

MR. PARZYNSKI: If I may add something, I think we might be restricted in that one, then, just because the data comes from, you know, the production contractor, and he might not have some of that data necessary to complete that. But that's something we can look at. I think I'm about 90 percent sure we can look at that.

MEMBER SOX-HARRIS: Okay. If it turns out that two percent or three percent of the patients with a denominator, but if it turns out that two or three of them were hospitalized in the last month before the colonoscopy, then I would say well, my consideration is largely due

to, you know, unfortunate --

I think what caught my attention was just the fact that of looking through the technical report, I was surprised that the risk-standardized rate of hospital patients were as high as they were for just a seven-day follow-up period. That led me to wonder, without data of course, whether or not there were hospital patients that were recurring before the colonoscopy.

MS. PETER: Hi, this is Doris. Let me just say that the rates are per thousand colonoscopies, so maybe that point is not obvious.

MEMBER SOX-HARRIS: Yes, and I took note of that. It's closer, maybe a little bit one way or the other. So that's why I asked the question.

CHAIR CELLA: So we've got Joe with his hand up, and just for time's check, we should probably be wrapping up this discussion in the next five minutes or so, so as to move on to

including these other two measures.

MEMBER KUNISCH: Yes. Hi, this is Joe Kunisch. You know, I did support what the other subgroup panel members are requesting, and I applaud your effort to try to turn this around and do it. But, you know, this might be more of a question to ask NQF.

Are we being consistent? There are a lot of readmission measures out there that have been NQF endorsed, and I just don't want to say, Let's again hold another measurer/developer on specific measure to a higher standard to then those other readmission measures that have already been passed. That started with a note to the same thing; you know, readmission is always 30 days readmissions. So was it actually related to that event when they were discharged?

So, again, just trying to be consistent with it in the overall scheme.

MS. WILBON: Sure. Hi, Joe, this is Ashlie. I think -- so it's hard to say across all readmission measures. I'm not as familiar

with the portfolio to say what type of empirical validity actually has been done.

I think certainly our requirement about empirical validity testing at maintenance is consistent, and I think certainly maybe even the types of validity testing that have come up as recommendations from the panel, I think certainly could be considered by the developer.

I don't personally have a sense of whether or not those requests are more or less than other recommendations that have been put forward, so I don't have a great response to that at this point without diving deeper. But hopefully, the requirement for empirical testing is standard, and I think ultimately the decision for the subgroup is to determine whether or not, based on this discussion, it would have been submitted by the developer based on what's feasible; that the rationale is acceptable or not.

CHAIR CELLA: This is Alex. Just to be clear, I'm not insisting that that particular

kind of validity testing be done. It's just that that's one example, and therefore it's possible to do something in that realm, and that standard should be applied consistently across measures.

MS. DRYE: Hi, it's Elizabeth Drye. I wanted to raise a question along those lines, and I'm thinking about it in real time. I'm not sure that I have a conclusion, but we've moved, and I think it's okay to thinking about the outcome, and while we're thinking about the outcome about the indicator of quality, and we talked about true analyses, redoing the signature return analogy ICD-10 codes.

And to the second one, it's giving me a little bit of a -- we could go back and look at where patients recently in the hospital, I think we want to say, before we do that analysis, why would that invalidate the outcome in our view?

When I think about it, and if I'm doing my feet so I may not be thinking right. These are outpatient colonoscopies we've tried to exclude with a lot of experience, the ones that

probably reflect really -- We tried to pull out admissions and not count if they are about solitaire for example, a colonoscopy to the extent colon cancer stays and then return them to the hospital.

So those are out, and I think if we saw -- if we're going to run analysis, I want to know how we would interpret it, that if two percent of patients were in the hospital in the last 30 days, I think we need to think, does that invalidate the outcome as a valid indicator of quality; either patients selected for same-day outpatient procedures who really I don't think would be expected to return to the hospital for unplanned events. But I think that we would be implying that those unplanned events were somewhat not related to Colonoscopy, they were related to prior admissions.

I just want to make sure that we're not going down a tangential path that maybe if we think about -- we just really need to know why it would unravel the validity of five percent, 10

percent. I really don't think it's true. But if it were, why would it unravel the validity that takes gastroenterologists and surgeons saying you need patients in a center, then they get admitted for unplanned reasons, and the next seven days there's not likely a single cause, there's no admission rate -- the use of hospital spikes within a few days after operation, then it drops off again.

But with all of our senses not likely being related, in many cases the procedure unravels if they've been admitted 25 days ago. I'm not sure that's true, and I think we should go into analysis knowing a clear sense of how they reflect, conceptually, what we're trying to measure.

MS. WILBON: Are there other subgroup or other members of the panel who have other thoughts for the developer, or is there a sense that we could bring this to a vote?

Okay. Hearing no comments, I think that means we will bring it to a vote then. So

we will be having our Subgroup 3 members voting again, and Hannah is willing to make sure that that is up for you.

Again, we're only voting on validity for this measure, 2539. I do just want to do a quick check for our denominator to make sure everyone came back after the break. I heard Alex; I heard David; I heard Eric; I heard Joe. Lacy, are you there?

MEMBER FABIAN: Yes, I'm here.

MS. WILBON: Okay. Marybeth?

MEMBER FARQUHAR: Yes, I'm here.

MS. WILBON: Okay. Sam?

MEMBER SIMON: Yes, I'm here.

MS. WILBON: Okay. And Sean?

MEMBER O'BRIEN: Here.

MS. WILBON: Okay. We'll have eight people voting again. Looking to please revisit that link for voting. The issue is up here for you to vote for validity on two 2539.

MS. INGBER: Okay. We have our votes. We'll just be adjudicating the results. Okay,

thank you for your patience. I'll share my screen now to share the results.

Okay. So 2539, overall rating is that for validity we have one vote for high, four votes for moderate, one vote for low, and two votes for insufficient, the measure passing on validity.

MS. WILBON: Thanks, Hannah, and thanks to the Yale CORE team for joining us. We will keep moving on to the next measure.

CHAIR CELLA: This is Dave Cella. So we go back to Subgroup 1, which seemed to be the most disagreeable group. That's maybe because I'm being a member, 0715.

MS. WILBON: Yes. We're going back to the measure that we --

CHAIR CELLA: The Boston Children's Hospital. Are they on? It should be 0715. There you go; that's it.

MS. WILBON: I think Lisa, are you there? If you cannot speak at *1, you should let your operator know that you're there to speak.

MS. BERGERSEN: Hello?

MS. WILBON: Lisa, go ahead.

MS. BERGERSEN: Okay.

CHAIR CELLA: Yes, we can hear you. Hi, Lisa. On behalf of the committee, let me apologize again for deferring you until now. Thank you for coming back today for this discussion.

We have limited time, so we're going to move as quickly as we can to the lead discussions. Or, I guess first you're going to set up, ask me and then Patrick and Matt, we have a discussion.

MS. BERGERSEN: Yes, thanks, Dave. I'll do just a really brief overview and then hand it over Patrick and Matt to take it from there.

This is a maintenance measure. The last endorsement was in 2015. It is a ratio of observed to be expected major adverse events among patients undergoing congenital cardiac cath, risk-adjusted using the Catheterization for

Congenital Heart Disease Risk Model. The risk methods are at CHARM II.

It uses electronic health data, health records and registry data measured at the facility level it is risk adjusted. There were concerns for both about the reliability and validity, so we will be voting on both. The measure developer did submit a response

And I will hand it over to Patrick and Matt to give us a more detailed view of the subgroup.

MR. ROMANO: Thank you very much. This is Patrick Romano; can you hear me?

MS. WILBON: Yes, we can.

MR. ROMANO: Okay, great. So this is Measure 0715, Standardized adverse event ratio for congenital cardiac catheterization. This is a risk-standardized outcome measure that basically represents an observed over expected ratio for a set of adverse events that can occur after catheterization for congenital heart disease.

Matt and I have conferred, so let me kind of quickly run through. So the measure was initially sourced for both reliability and validity not passed in CNR respectively.

I think that -- Or we think really that the correct assessment on both reliability and validity would have been insufficient, based on the original submission. But a lot of additional materials were acquired by the developers, and those materials are in the packet staring at page 49.

So in summary, the reliability issue was really unclear because the investigators were reporting only on the reliability of the outcome assessment, and they've corrected that with additional information that is in their submission. So it's now clear that what they do is both denominator reliability assessment by matching chief volumes from their registry of institutional records. They report 97 to 99 percent agreement on this denominator case ascertainment.

Further, they did a random audit of 650 cases, 50 from each of the 13 participating centers, and they now report the agreement for the outcome measure, which was very high in the seven percent, as well as for the procedure type, which was 100 percent, and key risk adjuster, hemodynamic indicator, which was also 97 percent agreement.

So I think we're prepared to say that the reliability issues have been addressed, which is supplemental information. And we'll be recommending -- of course, it doesn't qualify for a high score because there's no information on score level reliability; they're only reporting on data element reliability which means it qualifies for a moderate rating now on four levels -- (unintelligible).

So I'll turn it over to Matt or any other members of the subgroup if they have a different view or have any specific questions.

CHAIR CELLA: Anything to add to add to that?

MEMBER AUSTIN: No, I don't. Patrick and I sort of chatted yesterday in an email, and that was where we mentioned.

CHAIR CELLA: Okay. You didn't comment on validity. Are you saying that the validity vote would flow from the improved measure report or reliability?

MR. ROMANO: No. That's a separate issue, right? Are we voting on reliability first?

CHAIR CELLA: I think we're talking about all of it, and then we'll vote on both after the discussion. So if you can talk about the validity?

MR. ROMANO: Okay, fine. I'll go forward and talk about validity. Now, validity is a little bit more difficult because the menu developers say on that checklist or form, they say that they're recording performance at score-level validity. However, they are not doing so.

MEMBER AUSTIN: Yeah, and this is Matt. The only thing I would add to that is, you know, the

measure developer did, in their documentation, provide back an explanation that they felt like there was no natural gold standard for them to compare against.

And I feel like in some ways it overlaps maybe with the prior measure's conversation about when there is this sort of lack of a natural measure to compare against, how do we evaluate that?

But given the guidelines that NQF has set out that's for maintenance measures, that there is the expectation of empirical score level validity testing. There seems to be insufficient information where we live at.

MEMBER ROMANO: Right, another way to address that obviously, besides having some kind of external gold standard, would be to have some process measures at the facility level to demonstrate that in this case, they were out of 13 testing facilities.

Two had better than expected or low SAERs, and one had a significantly high SAER, and

so you know, were there processes of care differences across those facilities? Was there any explanation for those differences that would provide some validation of the performance score? That would be another way to go, but the developers were not able to do that.

Obviously with 13 units in the reliability and validity studies, their ability to look at unit characteristics and so forth is very limited.

CHAIR CELLA: Thanks, Patrick and Matt. Would anyone else on the subgroup like to add anything particularly to the validity discussion? I don't see any hands up. Larry's got his hand up now. Go ahead, Larry.

MEMBER GLANCE: Yeah, hi. So I appreciate the discussion with Patrick and Matt as well. I think I'd like to address the validity issue.

So there are many different ways to assess validity, and certainly one of the main emphases in the NQF approach has been to view

empiric validity testing by using construct validity to compare a supposed measure to existing measures.

This is something that we talk about quite a bit, and the limitation of this approach is that other credible measures are not gold standards, and so comparing a new measure to supposed credible measures, I think and others may agree, sometimes has really some good validity, no pun intended.

The emphasis in our validity evaluation has typically been on the risk adjustment model itself, on whether or not it's valid. Does it show good discrimination? Does it show good calibration?

And the reason for that is that in a perfect world, if you had a perfect risk assessment model, then you would know exactly what the expected outcomes would be at the provider level or whatever group level you're evaluating. So if you knew what their expected outcomes were, then you could compare the

observed to the expected and use that ultimately as a measure of quality.

And so I think that what we tried to argue, and this is not trying to change NQF policy in any way, but what we tried to argue in the whitepaper that we wrote evaluating scientific accessibility of a specific outcome measure, that this is probably the most important way to evaluate the validity of a risk adjusted outcome measure is to focus on the risk adjustment model itself.

So I would argue that the fact that these developers did not provide evidence on what NQF refers to as empiric validity testing, meaning looking at construct validity, should not be a reason to rate this as insufficient. I would strongly suggest that this should be moderate, not low and not insufficient. Thanks.

CHAIR CELLA: Thanks, Larry.

MEMBER ROMANO: This is Patrick. I guess that maybe Ashlie can address this, but I think that we are forced to implement current NQF

policy as described in algorithm three of the measure evaluation criteria effective September 2019, and that's on page 25 of the measure evaluation criteria document.

And so I mean I understand that in some cases, knowing that the model performs well in itself may be sufficient. It begs the question of whether there is anything that providers can do to reduce the outcome rate, whether there is any preventability or actionability in the measure.

So I think that is the alternative type of evidence that we would be looking for, that there is something that high performing entities are doing differently than what low performing entities are doing.

And you know, if the C-statistic of the model was 0.95, then that may leave essentially no room for quality, but a C-statistic of 0.75 in a model with three risk adjusters -- in this case, age, procedure classification, and a scoring of hemodynamic

status -- certainly leaves a lot of room for either quality factors or unobserved confounders, and we don't know which.

CHAIR CELLA: So this is Dave C. Ashlie, Patrick is asking if maybe you could provide some guidance. I guess if you could do that under the context of helping us sort out the role of this committee in looking at validity as it relates to actionability versus the sort of more basic idea that the risk model itself is accurate and is well validated, that that could, for this committee, be sufficient for a determination of moderate validity as an example.

MS. WILBON: Sure, hi, this is Ashlie. Yeah, actually according to our criteria, it is actually consistent with what Patrick said in that risk adjustment is but one sub criteria of the validity criteria, and so we do --

We are, you know, asking for, as you guys know, validity testing of the measure's score which would tell us something about the accuracy of the measure and being able to, you

know, accurately reflect the performance of providers and compare providers with the measure. And so with that, I would agree that testing of the risk model alone does not meet our current criteria.

CHAIR CELLA: Okay, I see no other hands up. Thank you, Ashlie. I think it's time for us to allow the developers a chance to weigh in.

MS. BERGERSEN: Hi, thank you very much for this rich discussion over the past couple of days and all of the time. I'd first like to just start by addressing this measure as we endorsed it. We had discussions early on when signing this initial letter --- this was actually (unintelligible) given the significant updates. It looks similar to Chart 1.

The significant updates used in the model, specifically the strategic risk group and a new committee, and as well as a change in the outcomes, I think clinically relevant adverse events to major -- to limiting them to major life

threatening adverse events.

However, I think I can, I'll try to address some of the concerns, but it is very similar, but I'll try to address some of the concerns related to validity at the score level.

We did respond as stated that there is no similar constructs by which to compare it to, and some in our field and other fields (unintelligible) proposed volume, and you can see with the supplemental data that we provided across the SAERs, (unintelligible) the highest variability in the volume from the highest performers to the low performers, we could not follow these metrics as a surrogate for quality, which is, you know, why we pursued this measure.

We do have some experience with Chart 1, and we've been actively running this registry as a quality improvement initiative for the past decade.

When the measure was first endorsed, I wouldn't be able to say with confidence, you know, whether it was differentiating between

confounders versus quality. Obviously there is still some (unintelligible).

However, I can provide expert data to show validity. Someone mentioned some evidential terms and process measures. We have looked into outliers (unintelligible), and specifically in the data provided to you, there was one outlier institution.

And we were able over the past year to do a root cause analysis into the major adverse events, and through that look, we were able to identify some differences in their practices around financial management in the (unintelligible) lab which was actionable for the center.

Moving forward, this is, you know, that sort of root cause analysis is just that in the past year when that site reached out with questions about there being an outlier. I don't know if that addresses the concerns of this potentially, but I wanted to share that. And --

CHAIR CELLA: Okay, thank you. Are

there any other comments from either the committee or the developer? This is Dave Cella again.

I think the layout and reliability, more of a stay forward recommendation of validity. You have a dichotomous choice here as to which perspective you take in your votes. I think we can go to a vote if there is no other discussion.

MS. WILBON: Thanks, Dave. This is Ashlie. I just want to point out we will push forward the vote for subgroup one. I do just want to do a quick check for who is on the line so we make sure we have the right denominator.

The vote on reliability will go first, and then the vote on validity should be focused on whether or not you accept the developer's rationale for not including empirical validity testing for much of what we did for the prior measure.

So I think with that, let me just do a quick check of who is on the line, and if you

could also be locating the link or pulling up the voting tool, that would be great. Daniel Deutscher, are you there?

MEMBER DEUTSCHER: Yes, I'm here.

MS. WILBON: All right, Dave Cella is there. Matt is there. John Bott?

MEMBER BOTT: Yep, here.

MS. WILBON: Joe Hyder? Okay, Patrick, I know you're there. Sherrie had to step away. Terri, are you there?

MEMBER WARHOLAK: I am.

MS. WILBON: Okay, Mike Stoto?

MEMBER STOTO: Yes, I'm here.

MS. WILBON: Okay, and Larry, you're there. Okay, we've got one, two, three, four, five, six, seven, eight. Okay, Hannah, we'll have eight people voting, and please go ahead and find your voting link and submit your votes.

MS. INGBER: Thanks, Ashlie. Voting is now open on measure 0715. Your options for reliability, sorry, your options are high, moderate, low, or insufficient.

CHAIR CELLA: Just be sure to note that this is the reliability vote.

MS. INGBER: Thank you, yes.

CHAIR CELLA: I know you said that. I'm just repeating it, so --

MS. INGBER: Yeah.

CHAIR CELLA: -- it's clear.

MS. INGBER: Okay, I'm going to go ahead and show the results. Just bear with me one moment. Okay, so as you can see, voting is closed for measure 0715.

We have zero -- for the rating for reliability, we have zero votes for high, eight votes for moderate, zero votes for low, and zero votes for insufficient. Therefore, the measure passes. Thanks, everyone.

MS. WILBON: So next we'll do a vote for validity.

CHAIR CELLA: Some people are really thinking about this one.

MS. WILBON: I think we have all our votes in. Hannah is working on getting it on the

screen.

MS. INGBER: Yeah, I was more thinking about the tech. Okay, I'm ready now.

CHAIR CELLA: Okay.

MS. INGBER: Okay, ready, so you can see here that the overall rating for, sorry, for validity for 0715, we have zero votes for high, three votes for moderate, one vote for low, and four votes for insufficient. Therefore, the measure does not pass on validity.

CHAIR CELLA: Okay. Thank you.

MS. BERGERSEN: Thank you.

CHAIR CELLA: Let's move -- thank you very much for calling back in today again. Let's move on now to 3576 if we can get that on the screen.

This is pediatric asthma emergency department use. It did not pass reliability or validity, and Ashlie, do you want to set it up or do we first check and see if the developer is around?

MS. WILBON: Yes, let's do a quick

check. I think they were checking in as well to make sure they were getting on the line in time. Is the developer up there?

DR. BARDACH: Yeah, Naomi Bardach is online.

MS. WILBON: Hi, Naomi. Thanks for joining us.

DR. BARDACH: No problem. Thanks.

MS. WILBON: Okay, so we will get started with 3576. I will do a brief overview and we'll hand it over to John and Daniel to do some more in-depth review. One second here.

Okay, there we are, 3576, pediatric asthma emergency department use is a new measure. The measure estimates the rate of emergency department visits for children ages three to 21 who are being managed for identifiable asthma using very specified definitions. The measure is reported in visits per 100 child-years.

It is based on claims and measured at the health plan level. It is risk adjusted and includes some social factors. The measure did

not pass reliability or validity, and the developer did submit additional materials for consideration based on the panel's preliminary analysis.

So with that, I will hand it over to John and Daniel to review some of the panel's concerns.

(Simultaneous speaking.)

CHAIR NERENZ: Yeah, I think Terri is listed as --

MS. WILBON: Yeah.

CHAIR NERENZ: -- the first to discuss, no? Okay, Terri, do you want to go first?

MS. WILBON: Apologies, Terri and Daniel, sorry about that.

MEMBER BOTT: Yeah, I didn't think I was on this one, okay.

MEMBER WARHOLAK: Yeah, okay. So good morning and afternoon, everybody. I wanted to just give an overview of what the subgroups thought.

For reliability, there were some concerns that the ICCs were extremely low. There were also some concerns about the complexity of the specifications in calculating age in child years.

In addition, there were some concerns noted with the approach and the results for testing reliability and plan variation, in that the ICC was not performed on a randomly selected split sample. And they go onto validity, and then we'll talk a little bit about the developer's response.

For validity, the committee subgroups had concerns concerning the empirical validity testing that was done at the score level using a difference-in-difference model and a negative binomial regression. They thought that the approach didn't adequately demonstrate the validity of the measure, but rather the QI intervention.

There were some concerns that the measure was specified at the health plan level

and that testing was done at the facility level, and then also finally, that the R-squared value for the risk model was very low.

So I want to move on now to just give a little overview, and I really have to commend the developer. You did an enormous amount of work in your response. Daniel and I did a little sidebar here offline, and it seems to address a lot of concerns.

I think there might still be some left, but it seems to me, and Daniel mentioned this as well -- and NQF, if you could weigh in on this -- it looks, because it's so methodologically and conceptually different than the original submission, should this now be considered in the next cycle as the new submission?

Because it seems like a lot of information to digest and evaluate in such a brief period of time. So if NQF could jump in, and then Daniel, if you want to jump in as well?

(Simultaneous speaking.)

MS. WILBON: Go ahead, Daniel, sorry.

MEMBER DEUTSCHER: Yeah, okay, sorry. So I just wanted to maybe note a couple of things, one for reliability and one for validity.

So if I understand correctly, the method used originally to assess reliability used a mixed model that accounted for variables included in the risk adjustment model, and clustering within providers.

To my understanding, this is more of an appropriate first step in determining the proportion of the total variance, the performance outcome that is accounted by clustering within providers. However, this might not demonstrate the provider-specific reliability estimates from which overall provider level reliability estimate and be count related.

I think the initial approach, and I'll mention the additional analysis that was submitted in a minute, but this approach I think allows an overall estimation of whether there is a significant amount of variance in performance

that is explained by the provider level, and if it is significant enough, then maybe developers can then go to the next step which is to calculate the provider specific reliability estimate.

This next step, to my understanding, initially was not conducted, and the results reported mentioned by Terri suggested there is not enough or maybe almost no between-provider variance to enable detecting differences between provider performance or not in a reliable way.

Now, and I also commend the developers for the large amount of additional materials that were submitted, and within those materials, the developers have included a totally new set of analysis for reliability, and this time it was first based on the health plan level, so that's one big fix that was done.

The method used, again if I understood correctly, was a split sample reliability testing basically comparing two sets of provider scores from those split samples, and then an ICC was

calculated with the result being 0.7, I think, for one set of health plans, and a little bit under 0.9 for the larger set.

I wanted to note that this ICC is conceptually different than the signal-to-noise approach or ICC type, and maybe the developers could explain why the ICC, which some could also argue being relatively low, but in the fact that they assess the accuracy or stability of the provider level score, where does that provide evidence that the measure is effective at detecting reliability differences between providers' performance? So that's maybe one question the developers could address.

The second comment I wanted to add for risk adjustment, the Chair mentioned the initial results were very low, almost nonexistent R-squared, and then the additional analysis which was totally on a different level, again moved to the health plan level.

R-squares were higher, but there were large differences between plans. So for one

plan, the model has an R-square of 0.13, for another it was almost 0.6 or 0.56, I noted here. I wonder if this questions the model's predictive or external validity.

So those were the two additional comments I wanted to add, and I agree that the amount of additional information, which I think some of that information addresses the concerns and maybe some do not. I think it really deserves a new submission since it's really not only different, it's also different content wise, but also method wise.

So yeah, so I'll turn this back to you, Ashlie. Just as Terri mentioned, I'm wondering what was the NQF's policy in such a case. Thanks.

MS. WILBON: Sure, thanks, Terri and Daniel, for that. So a couple of things. I think some of it may be, there may be two different issues, maybe a process of how the submission actually gets updated to reflect the information that was submitted as a response to

the preliminary analyses, and then kind of what is considered for voting today.

So with the process changes that we made last cycle, we do now allow for the developer to submit additional information for consideration. So we would ask that the votes that we submit today in reconsideration of the measure do take into consideration the additional materials that they submitted, and that your vote should kind of weigh those additional, that additional information.

In terms of kind of the new submission, our policy right now is that we do ask that developers kind of maintain the same submission per se, so we have the same measure number, same information in terms of the testing attachment and so forth, but what we will do is make sure that the information they have submitted as additional information becomes a part of their packet.

What we want to do is kind of tell a story of the measure throughout the process. We

don't want to kind of, you know, remove information that was there before so that the standing committee really has a holistic view of the measure as it's matriculated through the process.

We will work with the developer after the committee reviews the measure to clean up the submission so that, you know, by the time it's endorsed, that the packet actually represents the measure that was actually recommended for endorsement if that is the outcome. And so hopefully that's helpful, but if not, please let me know and I'll try to clarify.

MEMBER DEUTSCHER: Yeah, thanks for that.

CHAIR CELLA: This is Dave C. Any other comments from the committee or from the developer, from Dr. Bardach?

DR. BARDACH: Sorry, this is Dr. Bardach. Did you want me to say something?

CHAIR CELLA: If you would like, yeah, please do, and then we have another hand raised,

but go ahead.

DR. BARDACH: Okay, I wasn't sure if you would include me in the request for comments.

CHAIR CELLA: If you want to comment on what was just discussed, please go ahead, and then we can go to Matt Austin.

DR. BARDACH: Okay. No problem. Yeah, I wanted to also preface my comments by saying thank you very much for the careful review and the very thoughtful method feedback.

I think it's probably helpful to give a little bit of explanation and context to what the testing was and what the new submission of testing is because I think this issue of should we submit a new application or not, it's probably helpful to clarify what we did submit for the whole committee to understand.

The original submission technical specifications as a measure are still the same. The technical specifications haven't changed at all. The new data that we gave when you guys requested additional (unintelligible), and you

gave us all the comments and you gave us an opportunity to respond.

We changed our testing for reliability and validity in response to the comments from reviewers, and there was actually -- we used the methodology that was suggested by the reviewers in the comments, and so that's what we gave to you guys to review again.

It is quite different, and let me explain why it's so different, and it was actually extremely helpful to get those comments for us to be able to understand more about our measure and how to do the testing.

The basic issue, and this is super deep in the weeds, but I think it's really important to explain, the way the data set comes out of our measure specification is that it is a patient-month level data set.

That means what we're trying to look at in that data set is each line is: how many ED visits did you have in that month for that one patient? That number is going to be actually

very hard to predict and very unreliable if you just are looking at the month number line.

So our original reliability constant actually was looking at that, and we were amazed that our ICC was still very low, and we submitted because we couldn't quite figure out what was the right way to do it.

And so the comments back from the methods panel were actually very helpful in us being able to think about the fact that actually the reliability we're looking at is, you know, the conceptual and important reliability is to think about the plan level variation across the measure, which is 100 child-years.

So the reliability testing that we did and presented in our second submission, and that you guys are talking about now, is the reliability of looking at the plan level reliability rather than the method level reliability.

So that's what that split sample testing was able to allow us to accomplish, and

for the rest of the committee I'll say out loud, it is a plan level measure. It is not a provider level measure. So we're looking at just trying to do plan level reliability and validity.

So let me just look at my notes again. I think those were probably the biggest responses to give. Oh, sorry, the one other thing, the validity, to just clarify, there's two types of validity that you guys have already sort of mentioned during the course of my listening into the last hour-and-a-half or so, but there's validity of the model itself, the risk model and then there's construct validity. The first set of results that we gave was actually more focused on construct validity or face validity.

We basically took a plotting from the collaborative with a lot of focus in the quality group collaborative on cost of patient care, and then we said for those people who participate in the quality collaborative, the clinics that participated versus the clinics that did not participate, is there a difference in this

outcome measure of asthma ED utilization in order to basically say: is this actionable for a health plan that wants to actually improve care?

Can they go to their clinics and say, "Hey, we need you to focus on asthma," and their asthma care processes, and will that actually be related to this measure? And so that was what we were trying to present in that type of validity analysis. That was what we did looking at those differences.

The response that we sent you was to actually -- the request from reviewers was for the validity of the model, and so we repeated all of our analyses and then the new data set would show you just Massachusetts data, and the new analysis was in California data to show, you know, what are the results in the two different states and how do they compare? So that's just to clarify why the analyses look so different.

CHAIR CELLA: Thank you. Matt, would you like to comment now, Matt Austin?

MEMBER AUSTIN: Yes, thank you so

much. So my concern around validity was -- and I've sort of been trained to sort of weed through all of this and sort through all of this somewhat on the fly too.

My recollection is that the validity testing was not done at a health plan level, and it was done at some sort of lower level of analysis. Did you all provide any validity testing that was done at the health plan? Because my understanding of NQF's requirements is that the testing has to be done at the level of the measure's specification.

DR. BARDACH: We, in our resubmission, we did it at the health plan level. Yeah, we looked at the health plan. We gave a whole new set of analyses that looked at the R-squared and the variation in the -- I'm sorry, hold on.

I'm trying to find where it is in our document -- the variation in performance at the health plan level, both in California and in Massachusetts. So it looks like most of that data is all in (unintelligible) number 20 in the

document.

MEMBER AUSTIN: Okay, yes, thank you for directing me to that. I'll take a further look at that. Thank you.

DR. BARDACH: No problem. Yeah, I know it was a lot.

CHAIR CELLA: Yes, on the one hand, it's good to get the opportunity to hear from this panelist as to the concerns, but then by the time we get there, then if they're long responses, it's hard to find the time to really, really sink it in, and I think that's where the question of, you know, should this be --

There's so much in the response. Should this be a new submission? But I think we are going to -- I'm not saying we're closing to vote now, but I think there will be a vote coming soon. And if I misheard, Ashlie, is that correct? We will be voting on this as a submission with the developer's response?

MS. WILBON: Yes, we should be --

CHAIR CELLA: Okay.

MS. WILBON: -- now considering what they submitted, yes.

CHAIR CELLA: Okay, so as we're all sort of, you know, looking through the response to shore up the original concerns, I'll just ask if there are any other comments from the subgroup or anyone on the panel, anyone in the overall committee, or any last words from the developer?

Okay. Well then I guess, Ashlie, you can walk us through the voting.

MS. WILBON: Sure, so then subgroup one is up again. So I won't redo roll call. I'm hoping everyone is still on. We'll be working with a denominator of --

CHAIR CELLA: Well, one second, Ashlie. Mike Stoto left, I think, unless --

MS. WILBON: Oh, no, you're right. You're right. Thank you for verifying that. So Mike --

CHAIR CELLA: And Larry is still here, but we're going to lose Larry in 10 minutes or so, so you know, we're fine, but I think we're at

seven.

MS. WILBON: Yeah, so one other note to make is that Patrick is recused from this matter, so our denominator goes down as well, so I think we'll be having six.

CHAIR CELLA: Okay.

MS. WILBON: Let me just double check my notes, six voting. We'll have, yeah, six voting on this measure for the subgroup, and Hannah -- and we'll be voting on both reliability and validity.

And again, votes should be taking into consideration the additional material submitted by the developer, and please locate the voting link, and Hannah will put forth the voting for reliability first. We'll post the results and then vote on validity.

MEMBER AUSTIN: Ashlie, this is Matt. Is six enough for a quorum?

MS. WILBON: It is. It is. We did make quorum because Patrick's recusal kind of takes the denominator down, so we're still within

quorum. Thank you for bringing that up.

MS. INGBER: Okay. We're just waiting on one more vote.

MS. WILBON: Okay. Hannah is going to show us the results.

MS. INGBER: Sorry, I was just having a little trouble getting unmuted. Okay, I will share my screen now.

Okay. As you can see, the responses for 3576 overall rating of reliability, we have zero votes for high, three votes for moderate, two votes for low, and one vote for insufficient. Therefore, a consensus is not reached on this measure for reliability.

MS. WILBON: Thanks, Hannah. So Hannah will work on getting the validity vote up for you, and then we'll go from there.

MS. INGBER: Okay, voting is now open on validity for 3576. Okay. I'm going to go ahead and share my screen again. So voting is closed on -- oops, sorry -- 3576 for an overall rating of validity.

We have two votes for moderate, three votes for low, and one vote for insufficient. Therefore, the measure fails on validity. I'm sorry, it does not pass on validity. My bad.

CHAIR CELLA: Okay. Thank you, and thank you to UCSF and Dr. Bardach for joining us. I think it's time for public comment, Ashlie. We're back on schedule.

MS. WILBON: We are. Kudos to the panel. Thanks for getting us back on schedule. Yes, we will open it up for public comment at this time, and we'll do a very brief wrap up, and we will look to close on time.

If you would like to make a public comment, please press *1 to let the operator know you'd like to speak, or you can enter a comment in the chat box.

OPERATOR: We have no comments over the phone lines.

MS. WILBON: Thank you. I also don't see any comments in the chat box, so we'll keep moving. If there are comments, please feel free

to add them in the chat box until we close.

For a brief follow up, we did have an opportunity -- we did have a placeholder here for some process improvement feedback. Given where we are right now with the timing, I am not going to spend time on that. We will have some time allocated at our next meeting to do a brief debrief, and we'll do so with the co-chairs as well.

I did also just want to keep on everyone's radar about some of the papers we discussed, and Jack will be reaching out with an email reminder for folks to respond with their interest in participating on one of the writing groups for the paper.

I think we at least identified at least three papers on reliability, risk adjustment, and our social risk adjustment, and discussing validity of cost measures, or just kind of general issues in evaluating cost measures, so we'll be in touch about that.

Also a couple of next steps for this

cycle, the measure submission deadline for measures that we just reviewed in terms of them submitting their full submission for the committees to consider, which would be the important feasibility --- use and feasibility sections, will be coming up or is in process now for the next couple of weeks, and we will be summarizing the discussions of the methods panel for these measures and providing those results to the various standing committees.

The standing committees will be meeting to do recommendations for endorsement early summer, the May, June time frame, and we expect to see that decision for endorsement for this set of measures around October.

And our next cycle begins for measure review in the fall. Our intent to submit deadline is August 3, so just a few dates to keep in mind, and also we just wanted to list here for you the next webinar date, or the remaining webinar and in-person meeting dates we have scheduled.

Certainly depending on where we are with this crisis, hopefully we'll all be able to meet in person again by October, but those dates are scheduled through the end of the year, so hopefully you have them on your calendar, and I think that's it.

I did just want to give a big thanks to our co-chairs, Dave Cella and Dave Nerenz, for keeping us on task today and for facilitating such thoughtful and engaging discussions over the last couple of days, and for all of our other methods panel members for bearing through a very long but fruitful meeting over the last couple of days via webinar, which isn't ideal, but I think everyone did a great job and we appreciate everyone staying engaged. Dave and Dave, any final words for the group?

(Simultaneous speaking.)

CHAIR CELLA: Go ahead, Dave.

CHAIR NERENZ: It's hard to go alphabetical. Just to echo the thanks back a couple other ways, the staff did a great job of

putting back on the (unintelligible) supporting this whole enterprise.

This thing could have totally blown up given how we had to adjust on short notice doing a virtual meeting, and it went amazingly well, so thanks to Ashlie and Hannah, and the whole team for the way this was put together, and for the whole panel.

This is hard, especially in these times. We're pulled in different directions. We have lots of other things on our minds, other competing priorities. It's hard to do this without the face-to-face engagement, and again, I think this went amazingly well under the circumstances. So thanks, you all, so much for your thought, dedication, and seriousness of effort, really good.

CHAIR CELLA: Well this is Dave C. I have nothing to add. I think you said it all, Dave, and I think it's pretty impressive that we're able to give you six minutes back. Thanks again to NQF for really a great setup of all of

this and helping get the technology to work and everything. Thank you.

MS. WILBON: Thanks, everyone, and I appreciate your time. And I hope you all stay healthy and well, and we will meet again. Take care, everyone.

CHAIR CELLA: Bye-bye.

CHAIR NERENZ: Yes. Thank you.

(Whereupon, the above-entitled matter went off the record at 12:54 p.m.)

**S**

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Scientific Methods Panel
Spring 2020 Meeting

Before: NQF

Date: 04-02-20

Place: teleconference

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

Neal R Gross

---

Court Reporter