NATIONAL QUALITY FORUM

+ + + + +

SCIENTIFIC METHODS PANEL

+ + + + +

TUESDAY

OCTOBER 29, 2019

+ + + + +

The Scientific Methods Panel met at the National Quality Forum, 5th Floor Conference Room, 1099 14th Street, N.W., Washington, D.C., at 9:00 a.m., Dave Cella and David Nerenz, Co-Chairs, presiding.

PRESENT:

DAVID CELLA, PhD, Northwestern University; Co-Chair
DAVID NERENZ, PhD, Center for Health Policy and Health Services Research, Henry Ford Health System; Co-Chair
J. MATT AUSTIN, PhD, Armstrong Institute for Patient Safety and Quality at Johns Hopkins Medicine
BIJAN BORAH, MSc, PhD, Mayo Clinic
JOHN BOTT, MBA, MSSW, Healthcare Ratings, Consumer Reports
DANIEL DEUTSCHER, PT, PhD, Maccabi Healthcare Services
LACY FABIAN, PhD, The MITRE Corporation
MARYBETH FARQUHAR, PhD, MSN, RN, American Urological Association
JEFFREY GEPPERT, EdM, JD, Battelle Memorial Institute
LAURENT GLANCE, MD, University of Rochester School of Medicine and Dentistry
JOSEPH HYDER, MD, Mayo Clinic
SHERRIE KAPLAN, PhD, MPH, UC Irvine School of Medicine
JOSEPH KUNISCH, PhD, RN-BC, CPHQ, Memorial Hermann Health System
ZHENQIU LIN, PhD, Yale-New Haven Hospital
JACK NEEDLEMAN, PhD, UCLA
EUGENE NUCCIO, PhD, University of Colorado, Anschutz Medical Campus
SEAN O'BRIEN, PhD, Duke University Medical Center
JENNIFER PERLOFF, PhD, Institute of Healthcare Systems, Brandeis University
PATRICK ROMANO, MD, MPH, UC Davis
SAM SIMON, PhD, Mathematica Policy Research
ALEX SOX-HARRIS, PhD, MS, Stanford University
CHRISTIE TEIGLAND, PhD, Avalere Health
TERRI WARHOLAK, PhD, RPh, CPHQ, FAPha, University of Arizona, College of Pharmacy

NQF STAFF:

MICHAEL ABRAMS, Senior Director
KAREN JOHNSON, Senior Director
ANDREW LYZENGA, Senior Director
NICOLETTE MEHAS, Director
ELISA MUNTHALI, Senior Vice President

YETUNDE OGUNGBEMI, Project Manager

SAM STOLPE, Senior Director

ASHLIE WILBON, Senior Director

ALSO PRESENT:

KATHERINE ALLEN-BRIDSON, CDC*

SUPARNA BAGCHI, CDC*

MARY BARTON, NCQA

SUSANNAH BERNHEIM, Yale-CORE*

SUSAN CALI, CDC*

TARA MILLSON, CDC*

DANIEL POLLOCK, CDC*

ILANA RICHMAN, Yale-CORE*

LINDSEY ROTH, NCQA*

*** present by teleconference**

CONTENTS

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

8:58 a.m.

CO-CHAIR CELLA: Good morning. So, first, we're just going to review the agenda, which is, you know, the morning up until 11 is the continuation of yesterday, looking at the remaining submissions. And then at 11 -- so I'll do that part, and then Dave will, Dave N. will facilitate the process review and discussion of methodological issues. And then we'll get updates from NQF, and then we'll close by 2:30. So that's the day.

Karen, would you like to introduce Jen?

MS. JOHNSON: We're whispering about you, Jen. Good morning, everybody. So welcome. We're going to ask you a couple of things. One is just to tell us who you are and where you're from and that good stuff but then also our disclosures of interest. So I don't know if you want to do the spiel or have you done this enough --

MEMBER PERLOFF: I've done it before, and I think you know it, yes.

CO-CHAIR CELLA: It's not your first rodeo.

MEMBER PERLOFF: Great. Jen Perloff. I'm a health services researcher at Brandeis University. I apologize. Yesterday, I was at an annual R01 team meeting on nurse practitioner cost and quality, so I was thinking about quality measures, just not the right quality measures.

So excited to be here today. And I have no conflicts to disclose. I'm not a measure developer, just a consumer.

MS. JOHNSON: Okay. Some reminders and housekeeping. If you are on the phone and you're not talking, please mute your lines. We appreciate that. Also, once you do want to talk, if you're on the phone, just remind us who you are, say your name so we know who you are.

CO-CHAIR CELLA: We have one more measure for Subgroup 3 to discuss. Reminding you the subgroup members are John Bott, Marybeth

Farquhar, Dave Nerenz, Eugene Nuccio, and Ron Walters. It's also Paul Kurlansky, but Paul is not here.

So we're going to start with 3492, acute care use due to opioid overdose for Yale CORE/CMS submission, and Michael is going to lead us off with the NQF presentation summary.

MR. ABRAMS: Very good. Good morning, everyone. So let's see. Why don't we, can we get the slide up there? Let me get out my study guide here.

CO-CHAIR CELLA: So while you're looking for the -- consensus was not reached for reliability, four high, one low, one incomplete. And then validity, one high, three medium, three incomplete. Did I say one before? Two incomplete for reliability.

MR. ABRAMS: Yes. So if you take a look at your study guide on page 32 it should be, you'll find a header for this measure, which is, again, Measure Number 3492, acute care use due to opioid use disorder. But the actual measure is

described because of a little header switch error that's actually described subsequently on page 35. Okay. So I'm going to be going through the discussion guide materials there beginning on page 35.

Again, this is acute care use due to opioid overdose. It's a new measure. It was actually brought forward to this panel in a previous cycle and was, it failed passage here for a couple of reasons, one related to validity. We'll talk more about that later. The other was because there was no ICD-10 testing previously.

Just to remind you briefly about what the measure is about, the numerator is incident, what you might refer to as outcome events defined as an opioid overdose that results in an emergency department's use event, okay? Later during this meeting we're going to talk about whether or not that constitutes what might be considered a quality measure or not, but we don't want you to worry about that now. It is, for the purposes of our review at this moment, to be

considered as an outcome measure as described by the developer.

The denominator for this is rather a specific population, Medicare Fee-for-Service beneficiaries, engaged in Part A, the inpatient benefit, or Part B, the physician outpatient benefit. Individuals aged 18 or older residing -- and the measure is reported based on geography, county or state level. No exclusions, per se, were described for this measure. It's claims based, Medicare claims principally.

And when you all rated it previously for this current cycle it was consensus not reached for reliability. Let me remind you what the reliability looked like. It was an ADAMS R score across 25 states that they had data for, and the ADAMS R for the signal-to-noise score ranged from 0.92 to 0.99 across those 25 states. Across Maryland's 24 jurisdictions, counties and Baltimore city, the ADAMS R scores ranged from 0.6 to 0.99.

They also did a split sample analysis

with the Pearson's r, at least I'm assuming that's what they're reporting, and the correlation coefficients from that analysis were 0.94 and 0.87.

For validity also consensus was not reached. They did validity using a small panel of physicians from Yale from their institution, a sort of convenience expert panel, if you will, and used, you know, Likert scheme of voting to assess face validity.

Then they also did comparisons with AHRQ and CDC overdose data. The CDC measure is deaths per 100,000, so, you know, instead of, you remember the measure is about ED events, okay? So as a validator, they looked at mortality events from the CDC and they did that combining both NEDS and this data, NEDS being ED discharges and this data being ED-to-inpatient transfers. And then they looked at AHRQ data, as well, which was opioid-related hospitalizations.

Okay. So that's perhaps a little bit more related to the ED measure, but, again, this

was hospitalizations directly. And compared those two separate indicators of, you know, opioid burden, if you will, to the current measure. For the CDC comparison, the correlation coefficient was 0.74, simple correlation coefficient. For the AHRQ measure, as it turns out the correlation was again 0.74. I don't know what the P value is for having two correlations exactly like that. Somewhat of a coincidence, but that those are the numbers they reported.

And then they looked at meaningful difference using, you know, confidence interval overlap, did a GLM model with Poisson distribution where they took into account year effects, as well, and from that type of analysis found that there were 12 states below the mean performance level, three at the mean performance level, and ten above. So this was evidence for their suggesting that there were meaningful differences worth addressing in this measure.

So I'm going to pause there and ask that our discussant then make a comment.

CO-CHAIR CELLA: Marybeth is up for that. Thanks, Michael.

MEMBER FARQUHAR: I want to thank the measure developer for bringing this forward again. It's pretty hard to do population measures at this point. I do have some questions with regard to it. And thank you for sending the additional information. That was really helpful. I was one of the people that couldn't make head or tail out of it after I had seen a few of the other measures. So I have a question as to why Medicare Advantage was not included in the denominator. I'm not sure how come that's not included when they just did Fee-for-Service and not do the whole population.

So that was one thing I had a question on. The other thing, they have no exclusions, and I have a concern with regard to hospice and palliative care individuals who sometimes need more aggressive treatment than some of the others and they really should exclude them from this population or from the denominator.

What else did I have? I do have concern about the face validity and the five Yale physicians, which I think introduces bias. We have no idea about what their background is, whether it's oncology or general physician or what. And who knows if they're practicing. We didn't get any of that information with regard to determining whether that was appropriate or not, so, to me, I don't know that that was appropriate for the face validity piece.

I do have an issue with the geography of the consumer. Now, because we have such a mobile society, people often go to other places to get care. For instance, I have a friend who went to Hopkins, although she lived in Virginia, to get her primary care. So to ding the people back in Virginia for something that Hopkins may have done is not appropriate in my terms. You know, we live in the Washington, D.C. area and we're so close to the Maryland, West Virginia, D.C. border, we want to make sure that we appropriately, you know, I believe that holding

these providers accountable for these overdoses in the ERs. So in order to, for instance, if I live in Loudoun County, if I go to Fairfax County to get my care done, that means that, if I had gone to a hospital for an overdose in Loudoun, they would take me to Loudoun and then, because I lived in Loudoun, they would ding Fairfax County for the overdose. And to me, that's just not a fair comparison if that's what they're planning on doing, and that's what I gathered. So if I'm wrong, please correct me on that.

CO-CHAIR CELLA: Marybeth, I believe that there's someone on the phone with the Yale CORE group. Are you on the phone?

MS. RICHMAN: Yes, this is Ilana Richman for Yale CORE. I can --

CO-CHAIR CELLA: We'll come back to you in a minute. I just wanted to confirm you're here. Okay, thank you. Keep going, Marybeth. I'm sorry. I just wanted to --

MEMBER FARQUHAR: No, that's okay. I'm almost done. I just had a couple of

highlights after we went through that information. What else did I have? Again, I think, you know, at the state level, it might be fine to do that, but at the county level I think that's probably a gross injustice to the providers in that particular area. So those are the things that I kind of pulled out and wanted to highlight. I don't know if anybody else on our team wants to jump in here.

CO-CHAIR CELLA: Let's go ahead with Dave and then maybe give Yale an interim chance to respond and then if more subgroup discussion is needed.

CO-CHAIR NERENZ: Thanks. Yes, I had a lot of concerns about this, and they were mainly conceptual concerns having even to do with the rules of NQF endorsement, not so much the technical specs in the measure. I was the one who objected originally since this isn't even a quality measure. In fact, when I first saw it in our packet as a subgroup, I sent it out to Karen saying, you know, why are we even looking at

this? This doesn't fit our template.

I specifically didn't know what to do with the whole area of validity because most of our approaches and concepts about validity have to do with some underlying concept of quality. You look at process outcome relationships and you validate outcome measures by the extent that you're quite certain of the processes and you validate process measures by the extent to which they influence the outcomes. So there's a whole frame that simply doesn't apply here.

In the original materials presented to us, there was only a very brief one-sentence mention of quality that said, like, higher scores versus quality, but there's no theory, there's no data, there's no development of any sort of theory of quality for this measure.

Now, what I'm told, though, the sponsors will go ahead and review it, and, as Michael just said, we're not supposed to somehow table it or decline it because it's not a quality measure. I won't accept that it can be an

outcome measure without being a quality measure because outcome is a subset of the larger domain of quality. It's just a type of quality measure.

I don't know what this is an outcome of. If you live in a state or county that has a high rate of this, probably the driving factor is there's a lot of opioid use. So maybe that's an outcome of the criminal justice system, it's an outcome of the education and economy of the local area.

So I'm really stuck to know what to do with this, particularly in the domain of validity. Now, if we accept that it is just simply on its face was it says it is with no inference about quality and, in fact, no inference about outcome, I mean, I guess we can go ahead and raise some validity of it, simply is this a measure of ED use for opioid overdose. That's it. No inference beyond that. That's it. I guess you can assess the validity of that, and that is, as far as I can tell, what the face validity process is, at least as I read the

materials. The folks at Yale were not asked to say is this a good quality measure or is it a good outcome measure. They were simply asked is this a good measure of the face concept of ED use.

So I'm greatly troubled by this one, and I, in some ways, don't even know what to do with it.

CO-CHAIR CELLA: You want to make one of these points --

MEMBER NUCCIO: Yes.

CO-CHAIR CELLA: Yes. So one more moment, Yale. Gene Nuccio, a comment.

MEMBER NUCCIO: I just wanted to sort of reiterate my concern about the population that's being measured here, Medicare Fee-for-Service. And that's a point that Marybeth already brought up.

It struck me that this was clearly a population kind of measure, and so I didn't know why you didn't use population data. It would seem to me that hospitals in a county report the

number of visits to their emergency department, and they also might collect data on the number of opioid-related cases. And from that, you could also identify what the age of the patient is, and it would seem to me a pretty straightforward measure of opioid use and overdose, that they went in the hospital. Why Fee-for-Service for 18 year olds? That's a fairly restricted population until you get to 65.

So I had a fundamental problem with why this denominator and why this population of the measure. Also, the point that Marybeth made about only five people on the TAP from Connecticut seemed to be a little bit restrictive in terms of how we mostly form our TAPs.

So I just want to support both of those. I came at it again understanding David's point about is this a quality measure. If it's a population measure, you use population data, not restrictive healthcare data.

CO-CHAIR CELLA: Okay. We're going to pause there and give Yale a chance to respond.

MEMBER NEEDLEMAN: Before we do that, I just want to follow up on Eugene's point about population-based data. There actually is a population-based data set here, not in all states but many states collect ED data and hospital discharge data and those are available publicly. They actually do a correlation of this measure with the rates from the ED and inpatient databases, and the correlation is 0.74. In other circumstances, that would be fine. But it just feels to me that it's a low correlation. Apropos of what Eugene is saying, if you want to use one population measure, use a population measure. A correlation of 0.74 feels awfully low comparing a population measure to a Medicare Part A/B population, Fee-for-Service population only.

So the question is are you getting good enough measurement of the population here when we know what the actual population rates are, and I'm saying I'm concerned about that.

CO-CHAIR CELLA: Okay. Thank you. So let's ask for a response from Yale to the issues

that have been raised so far.

MS. RICHMAN: Okay, sure. This is Ilana Richman from Yale CORE. So thanks for those comments. Let me begin with the issue of the population. There were a few issues raised why Medicare Fee-for-Service or another broader population, why not include Medicare Advantage? If it's a population measure, why not use the national inpatient and national emergency department sample, for example.

So there are a few reasons why we chose Medicare Fee-for-Service. One is that we're working with CMMI on this measure, and part of their model is around engaging Medicare Fee-for-Service patients. So there was a sort of reason to focus on Medicare Fee-for-Service initially because of the contextual arrangement between the State of Maryland and CMMI.

Two, there are practical reasons for developing the measure in Medicare Fee-for-Service, namely it's the only population that I can think of, with the exception of Medicare

Advantage perhaps for which there are kind of limited data, but, anyway, it's the only population where we know not only numerator events but denominator enrollment. So, for example, in the national inpatient sample and national emergency department sample we used for validation, one limitation of that data is that we don't actually know the enrollment period for all the beneficiaries. So in that data, in the AHRQ data, the denominator seemed to be the state adult population, which is probably a reasonable proxy but it is not exactly the same as the number of patients who might be couched in the numerator. So it was sort of a pragmatic reason for using Medicare Fee-for-Service.

I will say that we plan to do additional testing in an all-payer population, knowing the limitations. But for initial testing, we felt we could best count the numerator in denominator in Medicare Fee-for-Service.

In terms of why the correlation with

national inpatient and national emergency department sample is 0.72 or 0.74, I mean, there are two reasons why that is the case. One is that the populations are different. Obviously, the age distribution between Medicare Fee-for-Service and an all-payer sample, Medicare Fee-for-Service skew older. Although we included patients 18 and older, most patients are 65 and older. That's one reason.

And then the second is that the measures are actually different. So the measure that we compare it to in AHRQ data is a much broader definition of emergency department and hospitalizations for opioid-related conditions. So, for example, our measure excludes intentional overdose, but the AHRQ measure that we compare it to includes it. It includes also things like an ED visit for opioid use that's related to assault. It includes withdrawal. It includes ED visits for opioid use disorder and not just overdose. So it includes a much broader set of conditions, and so we'd expect not to have

perfect correlation. But we felt it was useful to compare the measure to these. We're really trying to ask, you know, are we getting at this broader, this broader kind of phenomenon of the opioid epidemic and opioid overdose.

We saw that even in a different population and even with a different measure outcome, we saw a relatively high correlation, and that was true also for opioid, for death from opioid overdose. Again, a different population, a different outcome, we still saw a relatively strong correlation, so we felt like that was helpful.

So that was our rationale for choosing Medicare Fee-for-Service. We felt we could really kind of count the measure numerator and denominator with high fidelity, and it was also in line with what we had discussed with CMMI in whom this measure is being developed in collaboration. And then, yes, the correlation coefficient is not a one, but I think that's, to some extent, to be expected.

There was also a question about exclusions from the measure. We did end up in this measure discussed whether to exclude patients in hospice and palliative care, and we did not for two reasons. One is that the number of patients who would contribute are probably relatively small. That's not necessarily a reason to not exclude them, but, more broadly speaking, I mean, if we have a, if there is a patient on hospice or who is receiving palliative care who has an opioid overdose, to me, that is still considered an adverse outcome. I mean, that's not preferable even for patients who are receiving those benefits. So we would want to include them. That's still an opioid overdose, and we consider that undesirable. So we decided ultimately not to exclude those patients.

Let's see. There were some questions about the validity and the face validity. So I will say that, you know, in thinking about the validity of this measure, our primary objective was to think about, one, is the measure capturing

what we think it is capturing, and here we built on the existing literature which has been shown in a number of papers that ICD codes for opioid overdose are highly specific, even though, you know, population where opioid overdoses are rare events, the positivity value is still high. So if you see it, it's really there.

Second, we compared the measure, as discussed, to two other measures: hospitalizations and ED use for opioid-related conditions, that's the AHRQ measure, and death from opioid overdose of CDC data and saw relatively high correlation coefficients. And in face validity, you know, we felt like was reasonable and valuable to discuss. It was really not kind of the foundation on which our argument for validity is built. We did it as sort of an adjunct to the argument that we were building, which is really based on empirical validity testing.

And, yes, we used a local expert panel which would not necessarily call it tech. It

wasn't intended to be kind of a comprehensive national expert panel, but we convened a local expert panel as a way to get some feedback on the measure. The composition of the panel, it consisted of -- I'm sorry. So it consisted of emergency department physicians and general internists, all of whom have a research interest in opioid use. And I would say, by research interest, I mean many of these people are national experts who have experience measuring opioid use and overdose in a variety of context, including in the emergency department. So they were not sort of, they had focused expertise in this area.

But, again, you know, the face validity, we thought of this as an adjunct to our error validity testing, not really the kind of foundational premise on which our argument for validity is built.

And then, let's see, questions about geography. So the use case of the measure, the intended use is in the total cost of care model

which is a payment model in which the State of Maryland and entities within Maryland, geographic entities, are responsible for the total cost of care for Medicare enrollees. And patients in this model are attributed geographically, whether it's small area geographies or on the state level.

And so it really made sense to define this measure on the geographic level. Agree, of course, people move around, but not only, you know, is this payment model geographically defined but also, you know, public health infrastructure is geographically defined either on the state level or often on the county level. So it's true people cross counties to receive services, but often services, particularly in the opioid treatment world, are kind of structured on a county level. So counties invest in opioid treatment and provide resources to their local jurisdictions. So we felt like that was reasonable, a reasonable way to define populations. So, again, folks are attributed to

geography based on their place of residence, not where they end up having opioid overdose.

And then, lastly, this question of is this a quality measure. So I will say this. Opioid overdose is a complex phenomenon of health outcomes. I mean, you could say the same thing for an MI or for a hospitalization or for any kind of health outcome. These are complex events that do have to do with social structures and economics and all of the rest.

I will also add that there are clear evidence-based interventions that can reduce the risk of opioid overdose. The primary one is the provision of medication-assisted therapy, which we know can save lives from opioid overdose and reduce the risk of overdose or lethal overdose. On a broader level, we also know that reducing the availability of opioids in a community reduces the risk of opioid misuse. So there are kind of two primary ways in which healthcare entities and groups of healthcare providers can engage in reducing the risk of opioid overdose in

a community.

I think, so, yes, there are lots of other things that contribute to opioid overdose, but there are also specific things that we can do within the realm of healthcare that can reduce the risk of opioid overdose. We think, in that sense, it does reflect something about the body of care that is provided in a community. Yes, there are lots of other things, and we think that the way in which the measure is used to reflect those things, so, for example, we have tested the measure for use year over year in which an entity, a state or county, we'd compare it to itself and can try to show improvement over a previous year.

We acknowledge that it's a complex phenomenon, but we think there are things that folks can be doing on a local level. And we've also tested our measure to reflect that potential use.

CO-CHAIR CELLA: Okay. Thank you, Dr. Richman, for tracking and responding to, I think,

all of the issues raised. There are some continued comments or questions from the subgroup only, so we're still on the subgroup discussion. Gene, are you next? Okay.

MEMBER NUCCIO: Thank you for responding. I neglected to mention something else that concerned me about the measure. In your forms that you submitted, 2B3.2 regarding risk adjustment, you say that you are not risk adjusting this measure, and this gets back to how are you going to use the measure? In general, we typically look at measures being used to compare across entities, as opposed to simply improving within an entity. So a year-to-year contrast in a single county would be a within entity versus how county A in a state is doing with county B in that same state would be another way that we typically use these outcome measures.

You state in your risk adjustment rationale that "although the risk of opioid overdose varies according to patient demographic characteristics," and then you cite the Kaiser

Foundation study from 2018, "this variation reflects social rather than biological difference." I find it strange that you did not make any use of the sociodemographic variables in terms of risk adjustment, recognizing that there's already been evidence to show that sociodemographic variables make a difference.

So if you could respond to that issue. We're not asking, I think I made a comment that we're not looking at race, we're looking at sociodemographic variables.

CO-CHAIR CELLA: Go ahead, Dr. Richman.

MS. RICHMAN: Sure. So, yes, I think risk adjustment is a challenging and important question here because, on one hand, I think we all would agree that different, people in different communities have different rates of opioid overdose and opioid use, and that is a complex phenomenon that is absolutely attributable to the quality of healthcare in that community.

However, if we were to, say, adjust for poverty or some other marker of socioeconomic status or sociodemographics, really two things I think would happen. One is that we could obscure important difference of the very thing that we're trying to measure, and two, you know, as I think it's come up in another discussion of social risk, it's akin to saying that we are willing to accept a higher level of opioid overdose in the communities that are suffering the greatest.

So I think, rather than risk adjusting -- so we acknowledge that those factors are important, but I would say that, rather than risk adjusting, really thought should be given to how the measure is used. So we have tested the measure for use comparing one test entity to another within a year, and that's shown that we can identify meaningful difference.

And then how that information is used is important. So is it used to penalize communities? Is it used to identify opportunities for improvement? That, I think, is

the kind of key question is how that information is used, not necessarily whether or not it's risk adjusted.

And then an adjunct, yes, we do think there's a case to be made that, once differences are identified, could communities be encouraged to improve year over year, although I realize that's not kind of the primary way in which quality measures are often specified.

CO-CHAIR CELLA: Okay. Dave.

CO-CHAIR NERENZ: Okay. This is just a comment and basically a response to the response in the area of validity. If we are to consider this an outcome measure, my question then is outcome of what? And there is no specific conceptual model theory offered to us, nor is there any empirical data provided to us and materials to review either in the original or in the follow up linking this outcome to some defined prior something.

Now, you know, we've heard some worthwhile and probably valid speculation about,

well, the opioid rate could reflect this, could reflect that, but we have no data. So in terms of the rules of the game of how we judge validity, you know, we just heard that they're not resting on face validity. They consider that as kind of an adjunct, it's not the main support. But then we have no empirical information on validity that links this outcome to some act, some process, some something. So I still have a problem.

MS. BERNHEIM: This is Susannah from the Yale team. Can I just jump in with a quick question, actually, and a comment for the NQF staff? I just need a little bit of clarification that I think will help this discussion.

You know, when we present things to the scientific committee, there are certain pieces of information we're providing. There is a more comprehensive amount of information about the conceptual model for the measures that go then when it comes to the full committee, and so I want to understand what this committee needs

and weighs in on because, David, you're asking critical questions about something that is challenging which is a population health measure, right? We're not used to thinking about the way a community's actions, the health system and the community, relate to outcomes of a population. It's not the typical thing. As Ilana said, there's actually very good evidence that there are actions that communities can take and that healthcare providers then could take to reduce the likelihood, regardless of the rate of opioid use in the community, that reduce the likelihood that people are ending up with overdoses in the emergency room. But this measure is really intended to both incentivize and measure communities acting in ways that reduce those outcomes. They are clearly outcomes, and they're clearly outcomes that can be influenced by the quality of care provided in a community.

That piece of our work is less well represented when it comes to this committee. You know, we are mostly bringing, you know,

reliability testing and validity testing at this stage. And so I'm looking for the NQF staff to help us understand what this committee needs so we can make sure the right things are in front of this group and to bound the conversations to the NQF criteria related to this measure, which you're having an important, critical, challenging conversation on because it's a hard measure.

CO-CHAIR CELLA: Jeff, you wanted to add something.

MEMBER GEPPERT: So I made that exact same argument to NCQA yesterday, and their response was that their job was to create a valid sort of health status measure and they were completely agnostic as to how that health status was accomplished. And so we were okay with that response, and I don't know why the situation is different here. You know, the solution could be housing, it could be food security, it could be healthcare, it could be behavioral health.

MS. BERNHEIM: It's not agnostic, it's that there's evidence that there's a wide range

of activities that are better quality of care for a community that can influence outcome.

CO-CHAIR CELLA: Unless the subgroup objects, we have, de facto, moved into general discussion for all. So I think Larry and then Sherrie was up and then Karen. So Larry, Sherrie, and Karen, and then John.

MEMBER GLANCE: Great. So the first issue is one of risk adjustment and, most importantly, we're not accounting for the social factors that would drive the outcome in question, which is opioid overdose. And I would submit that, you know, if you're looking at the typical safe hospital quality measure, say for readmissions, so there are patient factors that the hospital has control over. So if you have surgery and you have fewer complications, you're less likely to be readmitted. So the hospitals can control that piece of it.

On the other hand, if you have a poor social support network when you go home and no one is there to take care of you, to some extent,

the hospital really, and you end up going home maybe to a poor neighborhood and et cetera, et cetera, the hospital has much less control over that. So you can make an argument, a fairly good argument, that, for hospital readmission measures, you ought to be controlling for social risk.

In the case of opioid overdose, it's a little different. So the driving factors are primarily social in nature. It's unemployment, it's poverty, it's, you know, food deserts. It's all those different things, right? So if you were to risk adjust for all those drivers, essentially you would be giving a free lunch, so to speak, to states that don't do a very good job controlling for those economic drivers of opioid overdose.

So I would argue that you would not want to risk adjust in any way because I would want to, you know, I think if you're going to look at how different communities do in terms of opioid overdose, you want to take into account

whether or not these communities successfully adjust the social drivers of opioid overdose. You don't want to risk adjust. So that's the first point I wanted to make. And I know this is a little atypical for me because I'm always arguing for risk adjustment. It was a little difficult for me to make this argument.

And the second piece is in terms of David's point that there's not a lot of data out there that we can impact opioid overdose. And I don't know this literature, so I can't necessarily comment on this, but I would say that's okay, you know. So there's not a lot of data out there. That doesn't mean that we can't develop data, and, to me, it has a lot of face validity, the idea that, you know, this is one of the biggest challenges to our society is finding a way of cutting down on opioid overdoses. This is something that I think we ought to be looking at.

And then, finally, the validity issue. I think if we, and I don't always follow the

rules, as Karen will, I freely admit to, but I think this is an example where they have addressed empirical validity by looking at construct validity, by comparing this measure to some other accepted measures. So I think that they've kind of crossed that threshold in that respect. Thanks.

CO-CHAIR CELLA: Jerry, thank you. Sherrie.

MEMBER KAPLAN: Susannah, this is Sherrie Kaplan. I was thinking to myself about why this is a quality measure, following up on David's point. And we do use things like ER visits for asthma, and the quality link is things like, well, for ER visits or hospitalizations for diabetes, you failed the primary care imperative. You know, you failed to treat those people adequately in primary care. So if that's the argument or even if it's not, if it's community-based stuff, you could actually sort of, if that was a model for looking at is this a failure of primary care or a failure to reach people to

services that actually could have mediated their opioid use or abuse, that's one argument you could make. I was trying to think out loud of why this would be a qualification.

But second thing is is that why wouldn't you then sort of link things with, the CDC lit in 2017, a bunch of states, I think it was 20 for naloxone use, for other kinds of opioid mediation stuff, could you not reach, as long as you have the CDC data, could you not then look at those communities that have benefitted from that and lots of new stuff compared with other counties or states that didn't?

CO-CHAIR CELLA: Karen?

MS. JOHNSON: Thank you. These are a little tricky. We do get, very rarely but we do get population-based measures, so we do have those. And we are going to tee up the discussion of, you know, quality measures and other things after we kind of finish this discussion.

But a couple of things that I did want to say, number one, I think Susannah is right.

One case that she asked about was I think part of what Dave would love to see, that linkage with the interventions. That is something that we asked for under the evidence subcriterion, which you guys aren't seeing that information yet. So I think there's probably more to come, it's just under a different criteria. So you're feeling the lack now, and that's just how the criteria are set up.

But I did go to look at our criteria. We do have somewhat, sometimes things are a tad bit different if you're talking about a population health measure versus the usual health outcome measure. So I was looking at what we have for validity testing, so let me read you what we have right now for validity testing. It demonstrates that the measure data elements are correct and/or that the measure score correctly reflects the effect of interventions to improve population health, adequately identifying difference in effectiveness.

So I don't know if that's helpful or

if that actually is less than helpful, but that's kind of how we are looking at the pop health validity testing.

CO-CHAIR NERENZ: Well, just to follow directly on that, the last phrase you mentioned I think sounds like the typical thing that we look at when we evaluate validity is that there's some statistical linkage between some defined interventions and the outcome and we are given statistical evidence on the percent of variance and the outcome that's explained by variation in the processes. And I just don't see that here, and I'm just looking for guidance about, if we don't have it, what do we do?

CO-CHAIR CELLA: Bijan and then John and then Jack.

MEMBER BORAH: Yes, I think this is a question for the measure developers. It's about correlation part. I think that the thing was it's right to do correlation between data is instead of focusing on the entire database, I think that both the database actually provide

insurance information, could they not have sort of enough correlation with Medicare population between those databases and then what they have? So that could have, again, if we are looking at population effects, that could have brought out that correlation. So focusing on the sub-cohort that is Medicare, we need those two databases, NEDS and NIS data.

And then another question, to bring back David's point about what's the intervention here? What is it that we are looking at? Again, I think that, essentially, we're looking at status quo, what is the effect of just the health services within the community, and so, therefore, we're looking at events in terms of what we get to see in all periods. So that's, you know, we don't have any, but we have the status quo, and that's what we are looking at. That's how I approach the argument for this particular measure.

CO-CHAIR CELLA: So there are two questions there for Yale. Dr. Richman, you want

to respond?

MS. RICHMAN: I'm sorry. I'm in clinic right now, so I got distracted. But somebody was asking me a question. Can you repeat the question for me?

MEMBER BORAH: Do you try to do the correlation with HCAHP and NIS data? I think you did cover it, and I asked NIS data and --

MS. RICHMAN: Yes.

MEMBER BORAH: -- they both have Medicare component into it. Like, did you try to do it? Are those subcohorts for Medicare data?

MS. RICHMAN: Yes. So, no, we used publicly-available aggregated NEDS and NIS data for our validation, yes, rather than look at specifically at Medicare, although, again, I think, you know, part of our rationale for using this broader population was to say that we could look within the Medicare population, but, actually, we felt like it was more powerful to look in a different population or a slightly different measure to say, like, okay, even in a

broader population, with a broader measure, we see that these two measures track together to a reasonable extent.

CO-CHAIR CELLA: And, Bijan, was your second question more rhetorical, or do you want a response for your --

MEMBER BORAH: No, the second one responds to David's point. I feel that it's more into looking at what's the effects of status quo health services within the community on opioids. And I feel that's how I look at this measure.

CO-CHAIR CELLA: Okay. I think we can leave that without a response. John and then Jack.

MEMBER BOTT: Yes. I mainly note the following in regard to, perhaps this is a bigger issue to talk about in the future but to use this example as an example of the issue of face validity. It seems like the measures I've reviewed, we've all reviewed, boy, when it comes to face validity, we just get soup to nuts for what we're seeing from when a measure developer

submits information on face validity. It seems that's likely suggesting that we could provide more guidance on face validity.

And as some folks noted, you know, as an example, I think we should provide some guidance for the team regarding the team members, the types of disciplines that should be involved. You should not be involving folks from your own institutions. That seems like a slam-dunk kind of rule. And if you would have articulated that, I'm sure Yale would have followed that.

And as Dave noted, perhaps some guidance on if you're going to go down the road of face validity, here's the types of validity issues you should address so we get the types of information that would be helpful to us in ascertaining validity. And maybe we could provide some guidance on the process by which that face validity occurs.

I didn't really see, and this is yet another example, how did they go about conducting that? We should ask for more specifics on the

results that we get that occur from face validity. I think their initial response, there was kind of a lack of response of what was the upshot of that face validity? They discussed it and what methods did you use, but when it came to what are the results it was silent.

But they did provide some initial information on page 84, and it was still a rather high level. The measure got a four on face validity, but really what was the question, what was voted on? I'm just really thinking we really need to firm up our guidance on face validity because we continue to get these kind of nebulous responses back from face validity, and the developer says, well, it's the initial round with the measure so we can do face validity, but oftentimes we're left kind of cold with the response we get, as here, I'm a bit disappointed with the lack of any understanding of true validity of the measure because of this use of face validity.

CO-CHAIR CELLA: Let's bring that back

for discussion after we go through the measure reviews because it's on the agenda for that sort of issue. So Jack and then Joe and then Patrick.

MEMBER NEEDLEMAN: So let me begin by saying I have some sympathy for, was it Susannah who said, you know, what are the terms here because I think some of the discussion has stretched into what the substantive committee would be deciding about the relative value of this measure against other measures that are doing the same thing, the usability of the measure. Those are not in our wheelhouse here, so I have some sympathy for the developers.

So I'm trying to figure out in my own head how I bring this back down to the issues that this committee should legitimately be discussing. And at the moment, I've got three, which is probably a lack of imagination on my part. One is it's presented as a population-based measure, as a measure of how effective the healthcare system and social service system and other systems in a state have been in dealing

with opioids, and the first question I would ask is, given the breadth of the interventions that are being assessed and that it's a population-based measure, is a Medicare-only measure valid on its face? Not is it good because, if I were on the standing committee, I'd probably say the population-based measures from the ED databases or the CDC measures are better measures and this is redundant, but that's not the question we're being asked. The question is is a Medicare-only measure in this space valid on its face? I'm inclined to give a pass on that, but I think that's one of the key questions that has sort of emerged from our discussions.

The second is the risk adjustment. If the goal is to figure out how well the systems are performing and compare performance across counties, across states, then the question is how hard is the job that they're facing there and how do you take that into account? So the counties in Southeastern Ohio have a much harder problem because the OxyContin folks dumped an awful lot

of product there. So the question is would you make any adjustments to the size of their problem in assessing how well the system is performing?

I don't think the answer is you have to have risk adjustment and if you don't -- I would like to see both a risk adjusted and unadjusted measure here to capture different dimensions of the product. So I'm not comfortable that there's no risk adjustment here. I'd like to see two versions of the measures.

The third question, and this one I'm going to throw at the developers, is about the reliability of the data. And we've got the standard language in the package here about Medicare data is used for billing, so we code for it. The main issues that Medicare is concerned about is over-billing.

And so the question I would ask is, given the social stigma associated with any inappropriate use of opioids, I believe that, where you've got something counted, it's there. So I believe the measure is specific. But I'm

not sure how sensitive it is, and I'd like to know what analysis has been done of the data that's the underlying data here. We have the coding data for the discharge abstracts that are in the AHRQ and other data which you say are correlated with this measure of undercoding, and has anybody looked at that and is that a concern we should have about the reliability of the data that undergirds this measure?

CO-CHAIR CELLA: Before you respond, this is just the timekeeper noting that this is really a very rich discussion, lots of people are very engaged and have a lot to say and ask about. We have still three more people that want to make comments, so let's try to make our points as succinctly as we can.

Please go ahead and answer that last question, if you can, Yale.

MS. RICHMAN: Sure. So there have been a number of studies that published literature that have compared diagnostic codes for opioid overdose to chart review using a

standard case definition, and we know that there is some undercoding. Claims-based measures, in general, are highly specific but not completely sensitive.

Now, whether or not complete sensitivity matters is dependent, I think, on the measure use. So you know, if we were a public health agency trying to count every opioid overdose in a jurisdiction, yes, we'd want complete sensitivity. But if the goal is to try to understand the comparative performance of entities or a change in entities over time, then I think complete sensitivity is less important.

And I will say that one of the things that emerged from our discussion with the expert panel is that, you know, obviously, in general, there's always a tradeoff between sensitive and specificity, but there's also some clinical uncertainty about what constitutes opioid overdose. So a patient could come into the emergency department unconscious and there could be a variety of reasons for that.

So in the face of that kind of tradeoff between sensitivity and specificity, as well as the kind of underlying clinical uncertainty, the general feeling, and I would agree with this, too, is we prefer a measure that is highly specific, even at the risk of some tradeoff in sensitivity.

So, yes, these issues have been looked at. Yes, they're imperfectly sensitive. But we believe they're specific and we believe they can be used to compare entities over time, particularly since the goal is not to do kind of a census of all of the opioid overdoses in a community but, rather, to have a signal that we can use to compare in between entities.

CO-CHAIR CELLA: Thank you. Joe, are you still up? No? And, Patrick, you also --

MEMBER ROMANO: It's covered.

CO-CHAIR CELLA: Okay. Then Gene.

MEMBER NUCCIO: Hi. Eugene Nuccio. Just to go back to the risk adjustment argument, I did some quick surfing on the net and

discovered from the U.S. Census Bureau quick facts that the workforce weight in Mississippi is 57.2 percent and in West Virginia 57.5 percent, so, looking at those, it would seem like, you know, there's maybe a slight difference. When you look at the deaths due to opioid from the National Institute of Drug Abuse, Mississippi's weight was 21.7 and West Virginia's was 49.6. So I would suggest that there are some ability to make use of at least some sociodemographic variables that then will allow you to more carefully tease out what makes a difference in terms of interventions, getting back to David's point about, if we have an outcome, we need to know how to change that outcome. And by controlling for at least some of the sociodemographics, I suspect that we could do that a little better than just using the raw rates, as is proposed.

MS. RICHMAN: I will say that we did some preliminary testing on a state level looking at the relationship between, for example, the

proportion of residents in poverty and opioid overdose rates, and the relationship is complicated. It's not perfectly linear. For example, Maryland has one of the highest rates of opioid overdose in the country, and it's also one of the richest states in the country. So, you know, part of that is sort of an ecological fallacy. There's a poor part of the state that it has a definitely concentrated population that uses and misuses opioids and also has a population of people who are wealthy, and those are two not necessarily the same population but just to illustrate the fact that, although you think about these things as tracking together, it is not actually perfectly sort of correlated.

CO-CHAIR CELLA: Okay. Trying to bring us to a close without foreclosing. Sean has a comment, and then I'm going to ask if anyone else has any comments or questions and then one last opportunity for Yale. Go ahead, Sean.

MEMBER O'BRIEN: This is about risk

adjustment. I didn't know if you had the idea of developing both adjusted and unadjusted versions. I think they address two different questions, but both questions are relevant and meaningful. But, fundamentally, I don't have a problem with the unadjusted version, just going back to what question had asked and answered, but most measures, what I think about risk adjustment attempting to do is trying to replicate a hypothetical randomized experiment where patients could be assigned to different providers. So when you don't do that, in the case of comparing two states, A and B, if state A had a better outcome for this measure, I wouldn't necessarily assume that if you switch their populations that state A would still have it better. It could be that state B was investing more resources and using their resources more effectively than state A, but it's still relevant to know that they have a worse problem and perhaps need to invest more resources. And that answers the questions that I think I'm more interested in from a public health

perspective. I think it's addressing something that's probably helpful.

CO-CHAIR CELLA: Okay. Thanks. Any other comments from anyone? Okay. So thank you, Dr. Richman. You represented the measure well, and I think that we're finished with the discussion. It's time for a vote. And Dr. Walters is not here, nor is Kurlansky, so there will be four subgroup voters and the rest of us on the shadow vote.

MS. OGUNBEMI: Okay. We are now voting on the reliability of Measure 3492. Your voting is open.

(Voting.)

MS. OGUNBEMI: Okay. Voting is closed. High is one vote at 25 percent, moderate is two votes at 50 percent, low is one vote at 25 percent, and insufficient is zero votes and zero percent. So the measure passes reliability, and we'll go now next to validity.

Voting is now open for validity of Measure 3492. Your options are high, moderate,

low, and insufficient.

(Voting.)

MS. OGUNBEMI: Voting is closed for validity of Measure 3492. Results are zero votes high at zero percent; moderate, one vote with 25 percent; low, two votes, 50 percent; and insufficient, one vote at 25 percent. Measure 3492 fails validity.

CO-CHAIR CELLA: Okay. Thank you. Well, Subgroup 3, that concludes your job. We now move to Subgroup 4 and the members are Matt Austin, Bijan Borah, Lacy Fabian, Joe Kunisch, Sean O'Brien, Patrick Romano, Sam Simon. And the first measure to discuss is 3528, CDC and VON harmonized outcome measure for late onset sepsis and meningitis in very low birthweight neonates, CDC. And Michael again is going to lead us off.

MR. ABRAMS: Thank you. This --

CO-CHAIR CELLA: One second, Michael. One second.

MS. WILBON: Are the developers on the line? Is anyone from CDC there?

MR. POLLOCK: Yes, yes, we are.

MS. WILBON: Okay. Thank you. I just wanted to check before we got started. Thanks.

MR. ABRAMS: Good. So this measure is looking at sepsis and meningitis infections in very low birthweight neonates, and the status of it previously was consensus not reached on reliability. Let me make sure that that's the status. Actually a failing grade on reliability and a passing grade on validity, so we'll be focusing on reliability in this.

Our setup for you is one of the unusual situations where we're talking about element-level validity standing in or substituting for reliability, okay? That's going to be the main focus of our discussion.

But let me go through the details of the measure with you, and this appears on page 32 of the discussion guide. This time, this is the little swapping error here. It's under the heading for the opioid measure, but it is at the bottom of page 32 for this particular sepsis

measure. And it's looking at late onset sepsis or meningitis infections in very low birthweight infants. And there's some preface in there about the prevalence that's in the introductory material there and the descriptive material on why this is an issue of importance. I won't review that explicitly. But the numerator of this measure is the number of sepsis infections or, separately, meningitis infection events. And this is a survival measure where there's a comparison of those without, who don't experience, infants who don't experience those events.

The denominator, the focused population, are neonates in a specific age range, days of life 4 to 121. Also, a certain birthweight range, in the very low birthweight range, 400 to 1500 grams, and a gestational age of 22 to 29 weeks. The other thing you should know about the denominator of this measure is that it's those in facilities that do not include level 2, 3, or 4 nurseries.

So we're talking about very specific intensive care, neonatal intensive care units is the population of focus. It's described as an outcome measure. The negative outcome in this case being, of course, the infection. The data comes from abstracted health records, level of analysis at the facility level. And there is risk adjustment that is conducted in order to create this measure, to create a standardized infection ratio, and this briefly summarizes a ratio of observed to predicted infection rates across a number of different stratified groups. And it's not calculated when the prediction infection rates are below 0.2 in a certain facility, so when they're sort of exceeding the rate in a certain facility there's an exclusion applied.

The specifications list something like 3,000 - 3,100 specific codes for a variety of different infection types. Included in the definition is something about qualified anti-microbial days, although it's a little bit

unclear why that particular definition is proffered. It seemed a bit of a non sequitur in the specifications.

The observed versus expected rates are derived from a Bayesian statistical model, which is to say not purely proportional but more nuanced and based on other information that the developers brought to bear. And the data comes from 716 Vermont hospitals. Over 40,000 neonates were assessed between 2010 and 2016 to look at this analysis.

Now, with regard to the reliability, and, again, this is going to be the central focus of your re-voting, it previously did not pass, but, arguably, the reason it didn't pass is because it wasn't reliability that was actually conducted. Instead, there was a sort of validity assessment that was done where the measure rates were compared to what the developers referred to as an online calculator. And this online calculator was another way to get at the events of interest, the infection events of interest,

with sort of an automated approach, and the developers can correct me if I'm misrepresenting that. But that was compared then to the measure as their way to demonstrate reliability. Those of you on the subcommittee reviewed that and said, huh, this feels like validity, it feels like an external standard is being applied, and, thus, you did not pass it as a reliability marker.

Again, this is a situation where, because it was element level, it was done with specific neonate measures or neonate events that this is a situation where we're talking about element-level validity, perhaps supplanting the need for reliability.

The results from that, by the way, are presented at the bottom of page 33 in a simple two-by-two table giving precision at 100 percent, recall, and the formulas are there, of 96 percent, and a Cohen's kappa at 96 percent.

So part of the question certainly is how these 320 cases were selected in order to do

this element-level validity. That, again, is part of what you will discuss is whether or not this can supplant the need for reliability.

The validity results were quite similar.

CO-CHAIR CELLA: Sorry. You mentioned bottom of page 33, but those are the previous, right? Am I wrong?

MR. ABRAMS: Bottom of page 33, there's a two-by-two table that has the comparisons of the --

CO-CHAIR CELLA: Oh, of their submission?

MR. ABRAMS: Yes --

CO-CHAIR CELLA: From the discussion guide?

MR. ABRAMS: No, no --

(Simultaneous speaking.)

CO-CHAIR CELLA: There's a table in the discussion guide, but I'm on page 33 from the previous measure. But I'm -- I would look at the discussion guide, so --

MR. ABRAMS: Okay, yes. On the bottom -- in the -- page 33 in the discussion guide. I pulled it out from the -- from the submission. You'll find this reliability table that I'm speaking of.

(Simultaneous speaking.)

MR. ABRAMS: Yes, the headers are wrong on the measure, so it is --

(Simultaneous speaking.)

MR. ABRAMS: It is from the bottom of page 33, but wrongly labeled. Maybe they got swapped on our item. Okay? Everybody clear on where those numbers are? Those reliability numbers.

CO-CHAIR CELLA: Yes, now I am.

MR. ABRAMS: Very good.

CO-CHAIR CELLA: Okay, thank you.

MR. ABRAMS: The validity analysis looked quite similar. But they said instead of using whatever this online calculator was, they used manually extracted data as a gold standard and they looked -- looked at similar numbers,

300. So they have separate validity. And then they have validity that we're asking you to think about in terms of reliability -- is it -- because element-level reliability. So you've already passed on validity element-level based on manual chart abstractions. The question before you is what do you think of this reliability presentation using the online calculator and a similar number of measures at the element level? Just to wrap up the other details of the measure related to validity -- risk adjustment was a maximum likelihood estimator deployed using what they said were both univariate and multivariate comparisons, ultimately including the effective birth weight, gestational age, gender and whether or not the baby was transferred to a new facility as a neonate, presumably that reflecting some urgency to get the -- the child to a more intensive setting.

Meaningful differences were somewhat evident. I've given you at the bottom of page 34 of the discussion guide there a time plot which

gives on the y-axis hospital rates in terms of percents of infections that occurred with time. And you can see when you get to 2007 the gap between the tenth and the 90th percentile -- that widest interval is -- is actually, you know, somewhere in the neighborhood of five percent or something like that, in that fairly wide interval. So in any case, that's their demonstration of the gap. Missing data, they comment about, is quite rare and not a problem. So with that I will pause and open it up for --

CO-CHAIR CELLA: Sam?

MEMBER SIMON: Great. Thank you for the summary, Michael. So this year we have another case where because data element validity was evaluated, the measure gets passed, and I do look forward to kind of coming back and discussing those in general someday. Put that aside. There was a problem with feasibility, but I've -- one of the problems really was actually trying to discern the measures -- the specifications of the measures. They're -- and I

say measures, because from my reading, we're talking about four different measures here. There's accrued, cumulative, infection rate -- the monthly accrued infection rate. There's an overall accrued survival rate. And then there is the standardized infection rate -- that ratio. So the specifications, you know, when you look at the hierarchy of things you are evaluating for a liability, is the measure precisely specified, I think. It really felt, to me that I -- I couldn't even, with the details given, make heads or tails of this measure appropriately. So that -- that was one very fundamental issue.

The second, though -- another issue I want to raise is sort of more of a process issue. It's not a measure-development issue. But from NQF's perspective, given that there are -- the analytic activity looked at the designation of sort of this late-onset sepsis and meningitis, and there -- there was a reasonable validity check of that comparing the -- these two approaches. But I guess the question is, these

-- the four different measures that are applied in this form seem to warrant separate applications. I don't know that it's appropriate that we sort of cobble them together into one -- one evaluation. So that's more of a process issue.

The other -- one other issue, I think, that I wanted to raise with subcommittee and the -- and the larger group is just a general lack of information. And in addition to the specifications, you know, there is a standardized infection ratio, but we have no model-fit statistics. We don't know how -- how well a measure is calibrated, so I -- I question even the validity of this measure. I know we're being asked to just look at reliability, but those -- those are some of the very -- at a very high level, some real concerns I have with this measure.

CO-CHAIR CELLA: Patrick?

MEMBER ROMANO: Yeah, this is Patrick Romano. Yeah, I just want to emphasize that the

-- the very first step of the guidance for evaluating reliability is, are submitted specifications precise, unambiguous and complete so that they can be consistently implemented? And if the answer to that question is no, it must be rated as low. And clearly, in this case, the answer to that question is no. There are four different denominators that are given here. It's really unclear which denominator is being proposed as the performance measure and how the -- how the fundamental construction of the indicator works is not explained. So it fails on the very first step. So the data that's being provided about data element validity is helpful, but it doesn't obviate the low score on reliability.

CO-CHAIR CELLA: I think -- well, go ahead, Joe, and then we'll give CDC a chance to respond.

MEMBER KUNISCH: Okay. So first let me start out by saying thank you to the measure developer for providing that supplemental

material that actually clarified quite a few things for me. And full disclaimer, I actually met with our infection prevention team at our organization to kind of, you know, talk about their process on, you know, IRR and -- and how they use that and if they actually use the calculator. And that -- you know, in my review, that was one of the things. There was no initial information on that calculator -- what were the data elements used? And it is -- to this date I still can't find that calculator and my infection prevention people were unfamiliar with it also. So one of the questions -- maybe speak to that. And is it an actual calculator, or is it one of the worksheets?

And then on the reliability testing, if I understand it correctly the way you did it is the abstractors -- probably your infection preventionists -- took the data elements on the worksheet, then the epidemiologists reviewed those before entering them into now this calculator, which is the gold standard, and then

you did that comparison between the -- basically the epidemiologists and the calculator.

I am always a person that likes to see what I call real-world testing. Meaning the abstractors are really going to be the ones doing this work if this becomes a measure. So I am wondering why you didn't even report the agreement between the abstractors and the epidemiologists. That would have maybe made a stronger case for reliability -- or actually, at our organization it's the infection preventionists doing reliability between each other in agreement. They don't even use the online calculator for that. And I'll leave it at that.

CO-CHAIR CELLA: Bijan, do you want to say something before --

(Simultaneous speaking.)

MEMBER BORAH: Yes, I think there is some of the questions that we had initially. And they did provide pretty detailed response. And I think some of the questions -- and I still agree

with Joe that -- I agree -- even last night I tried to find out this -- this calculator and I could not. And I have really appreciate the developer, if they -- they can tell us as to where -- or how it is being used. And that -- and that part I completely agree. It's very much less in detail, but I think now that they provided like quite a lot of details, I think that question, at least to me, some of that concern is sort of mitigated.

CO-CHAIR CELLA: So could the folks from the CDC please introduce yourselves and respond to what's been raised so far?

MR. POLLOCK: Sure. And this is Dan Pollock. I am the Surveillance Branch Chief in CDC's Division of Healthcare Quality Promotion. And I am joined by several colleagues. I will let them introduce themselves.

MS. ALLEN-BRIDSON: This is Kathy Allen-Bridson. I am the team lead for the NHS and Protocol and Validation Team and an infection preventionist.

MS. BAGCHI: Hello, this is Suparna Bagchi the validation (telephonic interference).

CO-CHAIR CELLA: Okay. Anyone else?

MS. MILLSON: This is Tara Millson. I am a member of the Protocol and Validation Team and an infection preventionist.

CO-CHAIR CELLA: Say your name again?

MS. MILLSON: Tara Millson.

CO-CHAIR CELLA: Thank you.

MEASURE DEVELOPER: Good morning, (telephonic interference). I am also with the Protocol and Validation Team at NHS (telephonic interference).

MS. CALI: I am Susan Cali, and I am also an infection preventionist for the Protocol and Validation Team.

CO-CHAIR CELLA: Okay, Dr. Pollock, who is going to lead the response?

MR. POLLOCK: I will lead the response. Thank you very much for the opportunity. Let me preface by saying that we recognize that our original submission created

some confusion and we recognize that we needed to be clearer and we've attempted to clarify on the reliability issues that were brought to our attention. So let me preface the response to the specific issues that were raised with a description of our plans for implementing the whole data supply chain that will serve up data for this proposed measure. And our design is to have a purely electronic process whereby data at the reporting facilities would be extracted from electronic data sources that are in routine use throughout the United States. And these are the electronic health record system itself, the laboratory information system, the medication administration system and the admission, discharge, transfer system. So the measure is designed so that with those systems up and running, implementers can extract the requisite data to consider the infant as a candidate both for the numerator and the denominator of the measure. This will obviate the need for infection preventionists or others to abstract

data from charts. It helps us move in the direction of what we consider to be electronic measures. And in our definition this -- this is an electronic measure.

So the goal is to have a standard way to implement a set of rules applied against the available electronic data. And the calculator is that set of standard rules developed by CDC for purposes first of evaluating the measure itself, but then as we move forward, the calculator and the logic embedded in it will serve as a reference implementation -- a set of specifications where vendor implementers could build into their implementations so that the way in which the data are processed on the sending sites will mimic completely what the calculator does right now. It's not a -- a calculator that we have yet made publically available because we are going to be launching the whole process of capturing LOS meningitis data next year in NHS, and at that point, we're glad to make it public. And we're glad to share it right now with

interested parties. But because this will be a new event in NHSN, it doesn't replace anything that we're currently doing, it's going to be a -- a new addition for us.

That said, do we recognize that there are always concerns about the reliability of using electronic data sources compared with the traditional or conventional approach of manual processing by experts of the data in records? So the way that we designed our reliability testing was to, in effect, look at two raters using the same data. The data that are available via an abstractor reviewing records was served up to an epidemiologist with expertise in neonatal late onset sepsis and meningitis. And those data were reviewed manually, cognitively by that individual with a case determination rendered.

The same data were processed through the calculator. And again, the calculator is designed both to test the reliability of the measure as well as eventually to become a technical linchpin to the actual implementation

of the measure. And when we evaluated those two sets of determinations, we summarized that in a variety of ways, including a kappa statistic, which is a test of inter-rater reliability, with a very high kappa value of 0.96. So that's the rationale for the approach that we took to evaluating reliability.

Now when we -- I can stop there and ask if there are any questions, or move into the specific issues. What would you prefer?

CO-CHAIR CELLA: Why don't -- why don't we see -- thank you. Why don't we see if there are any questions following up on that? All right, go ahead, Jack.

MEMBER NEEDLEMAN: Jack Needleman. I just want a -- a quick clarification. You've got an existing measure based upon chart abstraction. You'd like to move it to a more automated way of -- of capturing the numerator, and that's the calculator. So you're -- this is a test if whether the calculator does is -- so if you've done is try to test whether the calculator is

doing as well as the manual abstraction was doing under the old measure?

MR. POLLOCK: Yes, because --

(Simultaneous speaking.)

MR. NEEDLEMAN: I got what you were doing here.

MR. POLLOCK: Yes, you've got the gist of it.

MR. NEEDLEMAN: Okay.

CO-CHAIR CELLA: Okay, thanks. I don't see any other hands raised or cards raised here, Dr. Pollock, so maybe you can continue on with the more detailed response.

MR. POLLOCK: Sure. So issue number one that we addressed was a difficulty in assessing reliability because the measure specs cover three different measures. Well, as we said in the response, the specs for the proposed measure include numerator and denominator details. There are two types of neonatal infections, namely, late onset sepsis and meningitis, for which measure data will be

analyzed and summarized using several different outcome statistics and -- and these have been referenced already in the conversation this morning. Cumulative admission risk, crude monthly risk, survival probability and standardized infection ratio.

In our response, we go through each of what those measure statistics amounts to. And the cumulative admission risk will be the lead outcome that reflects the risk of acquiring LOS or meningitis for any eligible neonate during their admission to an eligible neonatal unit. So we see that as really being the prime outcome statistics with these others available to complement what the cumulative admission risk would summarize.

So we have used the standardized infection ratio in other of our healthcare-associated infection measures. And we anticipate with sufficient data we'll be reporting out for those facilities that have enough denominator exposure and SIR. But these

other measures can be calculated as well, even without the SIR. So we wanted to have an array of different ways in which the measure data submitted -- the numerator and denominator data -- can be summarized. And again, we've done that with our other HAI measures where, in some instances we use both rates in the measure as well as standardized infection ratios, which are risk adjusted on summary measures, as well as a third adjustment type of the SIR -- a Bayesian process that reliability adjusts, taking volume of exposure into account. So for us, you know, we -- we've had a pattern of submitting measure proposals to NQF that have numerator and denominator specs that are accompanied by different ways that those numerator and denominator data can be summarized and used for -- for measure purposes. So I will stop there with respect to issue one and -- want to see if you have any questions or comments? Or if you want me to proceed to the subsequent issues.

CO-CHAIR CELLA: Keep going, please.

Thank you.

MR. POLLOCK: Issue 2 was additional background as needed on the data source and data collection methods -- more details. Used on the online calculator. I guess I -- I addressed some of that in my prefatory comments. In the response that we provided we -- we summarize it by saying our method of reliability testing is designed to demonstrate that the measure data elements are repeatable, producing the same results a high proportion of time when assessed from the same population and the same time period. And that's a definition or a concept -- a conceptual view of reliability that's drawn straight from the NQF Measure Information Form 282.

We've outlined -- specified the particular data elements that are -- that figure into the whole process of identifying both numerator and denominator data. And these are placed in a table on page 2 of our response. I won't read each of them, but I would say across

the board that these are intentionally structured, machine-processable, readily available data with the goal, again, of having this measure be an electronic measure in which there will be a set of algorithms applied against the available data. And those algorithms are, by the way, reflected in the flow charts that are also a part of our response. And that was Issue number 3, LOS and NET are not operationally defined, so we provided these flow charts for the denominator and the NET determination as part of the package we returned.

Issue 4 is a clear explanation of the methods used for reliability testing. So again we used samples of records at three different facilities and had a trained record abstractor produce the -- the summary data that were then reviewed by a trained epidemiologist with expertise. And neonatal laid out the sepsis and meningitis and that same data was packaged and rendered in a way where it could be processed by the online calculator, producing the results that

we've summarized in -- in the table that lines up the manual abstraction data that were used by the trained epidemiologist against the calculator results. And here again, the precision is 100 percent, recall 96 percent and the kappa statistic 0.96. So we think these are very strong indications of reliability in the process of identifying the cases that we've been discussing.

CO-CHAIR CELLA: Thank you --

(Simultaneous speaking.)

MR. POLLOCK: Issue 5 --

CO-CHAIR CELLA: Sorry.

MR. POLLOCK: Sorry?

CO-CHAIR CELLA: I am sorry, I thought you were done.

MR. POLLOCK: Issue 5 is reliability cutting across organizations. What organizations were -- we actually involved in testing and we've included a table, page 4 of our response, that describes the three hospitals and there are clearly some differences across the various

dimensions that we've reported, including the EHR systems in use in those hospitals. So we -- we have confidence given that we kept it -- this measure in -- in -- several different information environments, that it is going to be feasible to have a more wide use of the whole measure data submission process.

And Issue 6 is the extent to which -- the rationale for using previous measure NQF 304 to justify validity and reliability. And NQF 304 was used only indirectly. The validity demonstrated for that measure, for the infection definitions, is relevant but not really the focus of the reliability conversation this morning. And the reliability testing that we've done is independent of that -- that testing that was done for 304. So that's -- that's an overview of the issues that were raised and our efforts to respond to them.

CO-CHAIR CELLA: Okay, thank you very much.

MR. POLLOCK: Questions?

CO-CHAIR CELLA: Thank you very much. Sean?

MEMBER O'BRIEN: Yes, this is a comment slash question for NQF as well -- potentially the measure developer. And just kind of -- trying to figure out where I am going in the re-voting. And I answered low based on the NQF instructions to rate low if you believe this specification is not precise, ambiguous and incomplete. And just to give an example of that -- and even combined with information provided by their response -- just the concept of a survival probability is mentioned in the measure information form. But the amount of information that's present about it basically says the numerator, denominator statements. The numerator is the number of eligible neonates without an LOS or neonates within an eligible location, the denominator is the total number of eligible neonates. And that's -- that's all the information about the survival probability. In their response -- they mention that basically

their survival measure can be used to calculate Kaplan-Meier plots -- the -- staff can better understand patients length of stay -- so it -- I completely trust that they're able to do that, but I don't know what about it is that they're collecting and measuring and presenting that allows them to construct those type of plots. So it's just not completely there. I guess my question is, on some level I don't have a -- I assume that they have very detailed documentation in place for everything they're doing.

So on some level, I am not sure I have to know all the ins and outs as long as it's clear that they have the documentation and it's being implemented in a way that's repeatable and reproducible and -- and clear. Perhaps it's good enough. But if on some level -- on some literal level, I think I'm still stuck voting low unless either the developer or NQF staff can clarify or address my concern.

MR. POLLOCK: Appreciate that. You know, the -- these are all statistics that are

calculable with the data that we will have in hand. They're spelled out both in the table at the bottom of page 2 -- the data elements that are included in the algorithm, as well as indicated in -- in the flow charts that are -- at the -- at the end of the response. So these are calculable outcomes with the data that we have specified. And the statistics are statistics that are -- are -- are used in other types of quality measures and -- and our intent is to follow the statistical calculations accordingly in the way we produce these measure data.

CO-CHAIR CELLA: Joe and then Sam and then Matt.

MEMBER KUNISCH: So -- this is Joe Kunisch. I am going to push back a little more on that reliability because I guess it wasn't clear to me, the plan was to electronically extract this with no human review in between. So you know, one, does that make it an official e-measure because there's different testing criteria under the NQF? So I will kind of leave

that question over to them. But in the way that you did it, basically, you scrubbed the data before you entered it into the algorithm. And we've done numerous electronic clinical quality measure testing, and if we did that, we would probably get perfect results too because we're already taking out the dirty data, per se -- or the cases that wouldn't be because of some data element that was erroneously entered in. So I am wondering, you know, what that -- the way we typically do it is we extract all the data elements from the EHR database. Then you would use that data, run it through your algorithm, randomly select some cases for, say, the epidemiologist to review. Then you want to see if your epidemiologist and your calculator are then in agreement. And you're getting a better picture, again, of real world data, which you probably are aware, is a lot of the time it's not really good.

So, you know, I -- I'd rather see testing like that, being it's going to be

electronic, basically clinical quality measure. Because again, if I am scrubbing the data ahead of time before putting it in, I expect to get near-perfect results.

MR. POLLOCK: Right, so -- appreciate that. The -- you using the word scrub, we can use the word extraction because a date of birth -- a date of NICU admission, transfer, discharge -- whether the birth is in-born, the hospital or -- the baby arrives from elsewhere, these are structured data that can be manually extracted from a record, or electronically extracted from a record. We opted to have a process of manual extraction, but we are confident from other -- other measures that we have in use, that the implementers are going to be able to execute that in an electronic way -- to extract those data. So the difference between what you're describing as scrubbing and what we would call an extraction step is one is being done manually and the other is being done electronically. We're -- we're confident, based on our prior experience, in the

capabilities of EHR vendors and infection surveillance vendors with whom we work, that there are -- their processes -- their extract, transform, load processes -- ETL processes are going to work. And we can, of course -- and will, with implementers -- validate that they have appropriately implemented the protocol steps, as we've done with our other measures that are electronic. We see that as a -- an operational implementation issue, not a measure reliability test.

CO-CHAIR CELLA: Okay, thank you. A couple of more -- we'll have time for efficient points. Go ahead, Sam and then Matt.

MEMBER SIMON: Yes, this will just be really quick. This is just a clarification to the subcommittee and the broader group as well. Just wanted to clarify something that Michael said. There is a chart in the submitted information that looks at meaningful differences. But I want to clarify that that is not this measure. It is actually NQF 304. It's an

analogous measure, but it's not actually -- we don't have any measure information. We have the validity information. We don't have any measure of the whole summaries for statistics of how we -- we measured those --

(Simultaneous speaking.)

MR. ABRAMS: That's right. Thank you for --

(Simultaneous speaking.)

CO-CHAIR CELLA: Matt?

MEMBER AUSTIN: Yes, thanks. So my comments earlier -- yesterday, I think, sort of reflected some of my challenges with this measure in terms of it sort of being multiple measures within one form. And -- when NQF makes the decision to sort of separate those out into individual measures versus keeping them all together. I think might be an example of where we want to have some further conversation about them. At least, from a reviewer's standpoint, where it might be easier to have things as individual measures.

It sounds like -- or what I come away with is that really the measure developer is focusing on maybe data element validity testing as a way to demonstrate reliability. Which does require that all critical data elements be validated. I don't know if we have a clear definition of what is a critical data element. But what struck me was it seems like the focus was really on the numerator, and matching those. In the SIR, there would be an expected number of infections, of which there are some parameters that go into that as well. So -- just for future consideration, I don't know how we want those tested, or if those need to be tested. But if we're going to rely on data element validity testing, that might be something to consider.

CO-CHAIR CELLA: Patrick?

MEMBER ROMANO: And just to amplify on that, with an electronic extraction procedure, presumably both the numerator and the denominator are important. And so there wasn't any effort to assess the reliability, or the -- the validity of

the data-element level of the denominator, which I would think would be relevant here. And again, it -- it sort of gets back to the issue that it wasn't clear initially what the preferred outcome -- I should point out that in the submission form, they actually describe four different numerators. We presume that what's being tested is the composite of any or none of those four different numerators. But that's another element that added to the confusion here. So that could be clarified, but -- anyway, yes. So this -- this question of -- of what are the key data elements? Is it enough just to validate the numerator when the measure is risk-adjusted and has a specified denominator?

CO-CHAIR CELLA: Well let's give the CDC a chance to respond briefly, please. And then Jack with final thoughts, if you're raising -- I see you raising your hand. Dr. Pollock?

MR. POLLOCK: Yes, so -- appreciate that. There are two conditions here the -- for which the outcome statistics are going to be

applied. One is late onset sepsis, the other is meningitis. Both of those are defined primarily on the basis of microbiology results. If there is a particular type of microbiology result -- a pathogen that it is sometimes described as the commensal organism, that's where the qualifying antimicrobial days, come in because we wanted an indication that the clinical team viewed that pathogen and has warranted treatment. We have denominators to find for both LOS -- and for both -- late onset sepsis and meningitis. They are the same. We've got numerators that differ on the basis of the type of specimen that's being obtained. That's the fundamental difference for us, from our perspective. The key issue in the reliability testing really has to do with the case determinations. It's absolutely correct; we did not test the reliability of the denominator determinations. But we are confident that that's going to be a straightaway process, again. That will certainly work as it's worked in -- in other types of denominators that are -- are captured

and reported electronically in NHSN.

So we focus on -- in our efforts, our -- our reliability efforts just to evaluate, again, the calculator because the -- the calculator and the algorithms embedded in it really afford case terminations, from our perspective, are the main methodologic issue with this measure. We recognize full well that the measure data -- the basic numerator and denominator data -- can and will be summarized in different ways. We see, again, that the cumulative admission risk will be the lead outcome measure that we'll -- we'll use as -- as a primary outcome measure for performance measurement purposes. So this is our -- this is what we've got.

CO-CHAIR CELLA: Yes, thank you. Jack.

MEMBER NEEDLEMAN: I've got three quick questions. First of all, I just want to clarify your answer to Joe's question about data scrubbing. I understand that you pulled the data

manually rather than having it pulled electronically. But I just want to confirm that you did not do any exclusion of cases where data was clearly wrong, or edit any of the data that was pulled manually. It was pulled exactly the way you would expect it to be pulled in the -- from the electronic data, the health record. Is that correct?

CO-CHAIR CELLA: Yes or no answer, please?

MR. POLLOCK: That is correct.

CO-CHAIR CELLA: Thank you.

MEMBER NEEDLEMAN: Okay. Because of the SIR, your risk adjuster is a critical component to this measure. It is what produces the denominator to the SIR. And the description of it is not clear. The statistics for it are not clear. The variables that were excluded by your process are not clear. So I have two questions about it. It looks almost as though you simply pulled the SIR from the earlier 304 measure. Is that correct or not?

MR. POLLOCK: We used data from a blog, but it's not from, you know, the one -- 304 measure proposal itself.

(Simultaneous speaking.)

MEMBER NEEDLEMAN: Okay. And the -- okay. The second question is whether the risk adjuster is done only on Vermont data. And if so, how valid do you think these coefficients are as you bring the measure nationally?

MR. POLLOCK: Well, we -- we expect that they're valid as we bring the measure nationally. We haven't built the measure into our system at this point. But when we do, we'll of course have additional data with which to evaluate the risk adjustment and the coefficients -- as we've done with all of our risk-adjusted SIR measures. And if we find on the basis of additional data that our risk adjustment needs to be changed to reflect what we then have available, by all means we will update as we've done with other NQF-endorsed measures -- our predictive models used in the SIR calculation.

(Simultaneous speaking.)

CO-CHAIR CELLA: Okay, Dr. Pollock -- Dr. Pollock?

MR. POLLOCK: We have implemented --

CO-CHAIR CELLA: Hello?

MR. POLLOCK: We -- we -- we needed a data source to include the data that we're -- building into NHSN to produce models and include them in our measure proposal --

(Simultaneous speaking.)

CO-CHAIR CELLA: Dr. Pollock?

MR. POLLOCK: -- which is exactly what we --

CO-CHAIR CELLA: We're not -- we're not voting on validity, so I am sorry to cut you off. But we're not voting on this issue. We're only voting on reliability. So I apologize for cutting you off, but we're really running against the clock. Any other --

MR. POLLOCK: Got you. Thank you.

CO-CHAIR CELLA: Go ahead, Bijan.

MEMBER BORAH: So I -- sort of two

quick questions. So one is following up -- can you please confirm that the two flow chart that you provided as response, they are now going to be used as -- sort of in a defining numerator and the denominator? And if that is correct -- so I am just looking at the -- sort of second flow chart. And I think that both the bubbles in there sort of would be defining the critical data elements that-- that will be going into this algorithm. So I guess my question is, if I have a hospital with NICU, how easy for me to have this data elements in one place? I mean, for implementation?

MR. POLLOCK: How easy? Well --

MEMBER BORAH: I mean, can you -- can you -- can you get all these data elements in sort of in a -- in a form? I mean, how -- how easy or difficult to get it, you know? I mean, I don't think they will be coming from one data system alone. I guess the question is, you know, what -- what is your plan in implementing or executing this, you know, in the real world?

MR. POLLOCK: Sure. Good question. Clearly the data are not coming from a single system. There would be several systems from which data would be drawn electronically. But again, our experience with electronic measures -- or as we would define electronic measure -- as substantiated that vendor implementers, or homegrown implementers, are able to tap into laboratory, antimicrobial administration, EHR, hospital systems, ADT systems -- and produce from both -- from all of those systems, files that incorporate the data and process those data against decision rules that are standardized. Very -- we're not -- we're not concerned about the feasibility here. We're confident that this is going to be readily implementable in multiple environments.

CO-CHAIR CELLA: Okay, and thank you. Any objections to moving to vote?

Go ahead, Yetunde.

MS. OGUNBEMI: We are now voting on Measure 3528, Reliability. Your options are --

CO-CHAIR CELLA: Hang on one second, sorry. Go ahead.

MS. JOHNSON: Yes, just wanted to clarify -- just a couple questions of you guys from NQF. First of all, multiple measures under one number -- we do allow that. So your questions -- and maybe it's good that we talked about this before we voted. So one of the things I heard was a question about specifications. So you have to ask yourself, do you understand this measure enough that you feel like that the different implementers would be able to implement it consistently? Okay, so that's one of the questions that you would answer. If you feel like that maybe there's enough for one of the four, or two of the four, or three of the four -- but maybe not another, let's talk about that before we vote because we may be interested in passing some pieces of it, but not all of it. That's possible. Okay?

In terms of data element validity, as you know -- and we'll talk about this a little

bit later today -- if -- if you're satisfied with the validation of the data elements, you can use that as your rating. We talked about imputing your rating for validity -- so you can do that. You asked about a definition of critical data elements. If you'll bear with me. Let's see. It should include those elements that contribute most to the computed measure score -- that is, account for identifying greatest proportion of the target condition, event or outcome being measured -- the numerator, target population denominator, the population excluded -- or the exclusions -- and, when applicable, risk factors with largest contribution to variability and outcomes. So you have to ask yourself, do you feel like what was presented reflects the critical data elements? Now, sometimes people will say, this measure -- you know, A, B, C, D -- what is really important is the numerator and if they tell me the numerator and I'm happy with that, then I'm happy. Other people will say, no, you know, I need to see -- you know, denominator,

exclusions. So that's, again, a judgment call for you. So did I hit Matt and -- did I hit all the questions?

MEMBER KUNISCH: I think -- just the -- the question about the e-measure. So, you know, to me this is extract -- pure electronic extraction. It doesn't fall under the -- you know, the typical HQMF QRDA format. If they're going to take the data directly, though, you know --I mean, should it be in QRDA format at that point, or some other accepted -- but what's -- where do you cut off where this is an eCQM, and it requires this additional testing and this doesn't?

MS. JOHNSON: Right. So we would not see this as an eCQM because it's not in the HQMF format with the -- I am forgetting the letters. The quality data model -- all that kind of stuff that you have to do for eCQMs. And also, I didn't see anything about value sets. So we're pretty specific about what we would expect to be a formal eCQM.

MEMBER KUNISCH: And -- and well, you know, just -- kind of a, I don't know -- say, word of caution, or -- when you talk about aligning programs and, you know -- CMS is doing this. CMS under the hybrid, I'll call it, free admissions is doing it under the QRDA format. So the 13 critical data elements have to be submitted in QRDA format. Not just directly submitted into their systems. So, you know, I would encourage CDC to align so when you're reporting these things, they're following their same standard, I guess.

CO-CHAIR CELLA: Lacy?

MR. POLLOCK: Oh, I couldn't -- this is Dan Pollock, just very quickly. QRDA is a format. And it's a format that is a representation of a standard called clinical document architecture -- CDA. And our intent or our design is that the data that would be submitted to CDC from participating facilities would be in the form of a CDA message. So it's essentially using the parent standard, and it's

given rise to QRDA. But we are constructed and built in a way that we can accept CDA messages. And so we've focused on CDA.

MEMBER FABIAN: Just a quick note on the standards. So CMS as of 2019 is moving forward to CQL, so different from PDM and the QRDA.

MR. POLLOCK: And those are -- those are relevant standards when facilities are reporting directly to CMS. What we do in our reporting on behalf of facilities to CMS is we receive the data from the facilities, and then on the CDC side, we do the calculations and provide facility-specific data to CMS. We are glad to provide that facility-specific data to CMS in whatever file format CMS wants, be it QRDA, be it the new format. That's -- that, if I may say so, is not that big of a deal. What's a big deal is the front-end stuff -- how the data are extracted, transformed and moved to CDC to begin with. That's the focus.

CO-CHAIR CELLA: Okay, are we back to

Yetunde?

MS. OGUNBEMI: Yes, let's do it.

CO-CHAIR CELLA: Okay.

(Simultaneous speaking.)

(Laughter.)

MS. JOHNSON: We have had -- from our audience -- a question about some of the work that Patrick's done on a -- another measure. So Patrick, do you want to describe what you've done and maybe when you've done it? We're trying to figure out if maybe we should ask Patrick not to vote on this one?

MEMBER ROMANO: Yes, in the interest of full disclosure -- so I serve on the Perinatal Core Measures Expert Panel for the Joint Commission. And one of those core Measures that is in the Joint Commission's portfolio is PC-04, which is called Healthcare-Associated Bloodstream Infections in Neonates. That's a measure that is harmonized -- partially harmonized with another measure that AHRQ developed that I was involved with over a decade ago, which is called NQI-03.

So it's not clear to me whether these are considered to be competing measures, because these measures don't include meningitis. And of course they're captured in completely different ways. But anyway, I just want to be transparent about my involvement --

(Simultaneous speaking.)

CO-CHAIR CELLA: Can I make a suggestion? I mean, I -- do you have a suggestion?

MS. MUNTHALI: So what we're going to do is find out if they are competing or related. Because if it is related as well, you'd have to recuse yourself. So what we're saying now is for Patrick to recuse himself from voting, and we'll find that out and then get his vote afterward.

(Simultaneous speaking.)

CO-CHAIR CELLA: So my suggestion -- my suggestion is that you shadow vote -- can you shadow vote, Patrick? Can you find the SharePoint? And -- pardon?

MEMBER KUNISCH: I am also on that.

CO-CHAIR CELLA: Okay, so can both of you please shadow vote with your -- put your --

MS. MUNTHALI: You were not on -- you were not developing -- I think, Patrick, you had direct measure development -- you were directly developing measures for one of the measures in the core set, right? You developed one of those measures?

MEMBER ROMANO: Well, we both serve on the -- on the Technical Advisory Group for the Joint Commission --

(Simultaneous speaking.)

MEMBER ROMANO: -- the Joint Commission as your specification.

MS. MUNTHALI: And then your -- you had further involvement in one of those measures?

MEMBER ROMANO: With the -- with the architects, that's very remote.

MS. MUNTHALI: Okay.

CO-CHAIR CELLA: Okay, so are you asking only Patrick? Or both Patrick and Joe?

MS. MUNTHALI: I think both Patrick and

Joe --

CO-CHAIR CELLA: Okay.

MS. MUNTHALI: -- because it's the same scenario. And we will -- you can shadow vote.

CO-CHAIR CELLA: Can you shadow vote, please? And then put -- be sure to put your name in the name thing so that they can pull it out if it's needed. We'll see where the vote goes without you included. There are lots of people on this panel, so we'll be able to get a percentage. So -- I think now we're ready to go.

MS. OGUNBEMI: We're now voting on the reliability of Measure 35 --

CO-CHAIR CELLA: Jack?

(Laughter.)

MEMBER NEEDLEMAN: Sorry. I would like to -- Karen asked whether there were some of these four that we wouldn't want to see pulled out. I don't believe the Vermont-based risk adjuster coefficients are appropriate for a national measure. And I would like to see the -- the vote on the SIR measure, anything that relies

upon the risk adjuster done separately.

MEMBER AUSTIN: Can I clarify real quick? It's the Vermont Oxford Network, which is a NICU neonatal register. So people who talk about Vermont -- it's actually hospitals across the nation. So it's not just Vermont. So just to clarify.

CO-CHAIR CELLA: Okay. Are you withdrawing your --

(Simultaneous speaking.)

MEMBER NEEDLEMAN: I withdraw the request.

CO-CHAIR CELLA: Okay. Back to you, Yetunde.

MS. OGUNBEMI: Okay. We are now voting on the reliability of Measure 3528. Your options are high, moderate, low and insufficient. And --

CO-CHAIR NERENZ: Please remind us for this measure, not a discussion, do they have to do either data element or measure score, or both?

MS. JOHNSON: This is an outcome

measure, so they can do score level of testing, or data element testing.

CO-CHAIR NERENZ: Or?

MS. JOHNSON: Yes, it doesn't have to be both.

CO-CHAIR NERENZ: Okay, okay. That's all I want to know.

MS. JOHNSON: I'm correct, right Michael? Yes.

MR. ABRAMS: And they only did data elements?

MS. JOHNSON: And they did data element.

MS. OGUNBEMI: Okay, so our in is five. Yes, perfect.

So we have zero votes high; moderate, one vote at 20 percent; low, four votes at 80 percent; insufficient, zero votes, zero percent. So the measure fails.

CO-CHAIR CELLA: Let me just clarify the math. If -- if both Patrick and Joe were to vote high or moderate, it would still be 4-3,

which would fail -- is that right?

MS. OGUNBEMI: It would be consensus not reached --

(Simultaneous speaking.)

CO-CHAIR CELLA: So -- so you'll have to determine whether it's failed, or consensus not reached after -- after this. Okay. Okay, thank you. Larry, did you want to make a comment? We're going to take a break -- okay.

MEMBER GLANCE: A really quick comment. Just to the measure developers. I think we're -- our discussion here, if necessary, focused on what the materials that they presented us, which was on data element reliability and data element validity. I also think, having looked at the literature from this group, that this is a very strong group of -- this is a very strong measure developer. And I think that if they decide to come back and resubmit this measure at a later time, I think it would be extraordinarily helpful -- a couple things. One would be to provide more detailed specifications.

Two would be to break out the different measures separately. I think it was very difficult -- it sounds like, for the work you -- to evaluate each -- all four measures in one package. Three, give a -- some detailed technical specifications on the risk adjustment models that were used. And four, look -- focus on score level reliability and validity measures.

CO-CHAIR CELLA: I think what you're hearing, Dr. Pollock and company -- and there are other head nods in the room -- is don't take this as a rejection. Take it as an encouragement to -- to bring it back in -- in March. Is that fair?

MEMBER GLANCE: Yes.

CO-CHAIR CELLA: Okay. All right, we're going to take our break now, and we'll have to cut it to more like eight minutes -- seven or eight minutes. So just a quick break and come right back so we can get back on track.

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

11:19 a.m.)

CO-CHAIR CELLA: These are two measures that have passed. But they were pulled up for discussion around a very specific question for clarification and that I wonder if -- and Dave is going to lead the discussion and I'm going to sign off.

But I wonder, Ashlie, if you could do an abbreviated summary. And then Matt will introduce the question why we put forward these measures. And they may not go back for a vote. So we'll decide that. We'll see that based upon the discussion if there's specific questions. But how about a brief summary, Ashlie?

MS. WILBON: Sure. I'd also wanted to welcome Mary, and I think she has some colleagues on the phone as well from NCQA. Thanks for joining us.

So we've merged these two next measures, 3483 and 3484, into the same kind of topic. They're very similar measures that I think share some of the same issues. So just

give me a second here for my notes.

So 3483 is an Adult Immunization Status measure. It's a composite. Both are composites. The Adult Immunization Status measure looks at the immunization status for five vaccines. So influenza, Td or Tdap, Zoster, pneumococcal. One, two, three, four. We're missing one. Pertussis, okay. Or Tdap, right? Okay.

And the other measure, 3484, is a prenatal immunization status. It looks at three different immunizations which includes flu, tetanus, and diphtheria, pertussis as well. And again, also a composite.

Both measures passed reliability and validity. The one issue -- the first issue that was raised and which pulled these measures up for discussion was around the level of analysis. The level of analysis was checked for health plan and integrated delivery system.

NQF states that when a measure developer submitted a measure, the level of

analysis must align with the testing that is submitted. So if you check a box for health plan, you need to also submit testing that aligns with health plan level of analysis.

And so for this measure, because there was two level analyses checked, we wanted to clarify with the developer whether or not that integrated delivery system checked box was intended to be checked.

And if so, how the testing demonstrates reliability for that level of analysis because it appeared that the testing only really focused on the health plan level of analysis. So that's the first issue.

The second issue that came up in the evaluations as well was around the reliability scores. And this wasn't necessarily an issue that Matt identified. But it came up in another subgroup member who suggested that it be brought forward for discussion in that the reliability scores essentially were close to perfect in some instances, kind of 0.99, 0.9 -- in the 0.9 range

all the way up to 0.99.

And I think there was some concern over how -- whether it's kind of mathematically possible to achieve a perfect reliability score and what some of the factors might have been that contribute to that score. Some of the subgroup members were wanting the developer to provide a little bit more color onto how they think that this score could be achieved and kind of what some of the factors that might be that contributed to that.

And I think those were the two main issues. And so I will hand it over to Matt to add any additional commentary. And then perhaps we'll open it up to the developer to provide any responses to that.

MEMBER AUSTIN: Yes. Just for sake of time, I'll actually just say what you captured, Ashlie. And so my concern, this is absolutely right on which is it was clicked off as being specified for both health plans and integrated delivery system. But at least from my review,

the only testing I saw was for health plans.

And so I didn't know I had missed something or if there was an unintentional inclusion of integrated delivery systems. So that's the issue I wanted to approach.

MS. BARTON: I'd actually like to introduce my colleague, Lindsey Roth, who's on the phone who has been the champion for these measures, both the immunization measures. Lindsey, can you hear us?

CO-CHAIR CELLA: We can't hear you.

MS. BARTON: No, we can't. Okay. So what I would say --

CO-CHAIR CELLA: Maybe she's mute. Let's give her a little bit of a chance to unmute. Or she's not mute on our end, right?

MS. OGHUNBEMI: No. But if you are, it's star 7 to unmute.

CO-CHAIR CELLA: Or maybe she'll keep trying.

MS. BARTON: Right.

CO-CHAIR CELLA: Go ahead.

MS. BARTON: So what --

CO-CHAIR CELLA: She's on the web? But we can't hear her. She's probably trying to figure out how to --

MS. BARTON: It's possible. Well, it's a quick answer. So this is --

MS. ROTH: Okay. So can you hear me now?

CO-CHAIR CELLA: Yes.

MS. BARTON: Yes, we hear you, Lindsey.

MS. ROTH: Okay, great. I'm sorry about that. That did the trick. Okay. So, yes, I can address the two issues. I'll just start first with the integrated delivery system issue.

So I'll explain that we selected this category because in our field tests and in our first year pilot results of the measures, it included health plans that are considered integrated delivery systems. So health plan provider groups all are under one entity.

So that's how we were defining

integrated delivery systems in this case and we didn't see a need for a separate analysis for these plans because they're still looking at the health plan level. So that's the intent behind selecting that.

And then for reliability, so we institute a standard approach for calculating reliability in the health plan level. I can describe. It's different between our previous submission and this submission.

So in the previous submission, we heard the concerns about there only being a few health plans in our field tests. So in this submission now we recalculated reliability using the beta-binomial method on a much larger number of plans, over 135. And we also at this time provided the overall beta-binomial scores and then the distribution of the beta-binomial instead of only providing the median which we had done the first time.

I think I did also see a note somewhere about unequal sample sizes across the

plans and whether this could be contributing towards the reliability results. I did want to clarify that each plan includes all the members in their plan who meet the denominator criteria. So we're not looking at a sample of members within plans.

And then we do include the distribution of the measured denominator across the plan. I believe that was the exclusion section of the testing form.

So, and I think we've heard your concerns about our overall approach to calculating reliability for our measures. So it's requested we're looking to providing additional information on our approach as well.

MEMBER AUSTIN: So, Lindsey, this is Matt Austin. So it sounds like in the testing you provided, that included some integrated delivery systems in the results? Is that --

CO-CHAIR CELLA: In our health plans. So the health plans that are integrated delivery systems.

MEMBER AUSTIN: Okay. So an example might be a Kaiser or something like that?

CO-CHAIR CELLA: Yes.

MS. ROTH: That's correct.

CO-CHAIR NERENZ: Dave here. I know we're terribly pressed on time. My suggestion here is I think there's essentially semantic confusion and misunderstanding. Because what I think I hear going on is that this is fundamentally a health plan measure. It was tested in health plan.

By the way, some of them may also be characterized as integrated delivery system. But there's a whole big universe of integrated delivery systems that are not health plans that this measure was not designed to work in nor was it tested in. And is that fair?

MS. BARTON: I think that's fair.

CO-CHAIR NERENZ: You have to have a defined population. If the integrated delivery system has no defined population, you have no denominator.

MS. BARTON: That's correct.

CO-CHAIR NERENZ: Okay. So I'm just -- I'm trying to jump to a potential conclusion. It would seem like this is really specified for and tested in health plans, period.

MS. BARTON: Yes.

CO-CHAIR NERENZ: If by accident some of them are integrated delivery systems, that's just, oh, well. That is -- okay? Is that --

CO-CHAIR CELLA: You can say health plans, including health plans that are integrated delivery systems.

CO-CHAIR NERENZ: But it's not important.

MS. BARTON: Right. I think that's fair.

CO-CHAIR NERENZ: Is that?

MS. BARTON: Yes.

MS. JOHNSON: So we would just ask you to uncheck integrated delivery systems in your specifications.

MS. BARTON: Right.

MS. JOHNSON: And --

MS. BARTON: Yes.

MS. JOHNSON: -- we would be done. You'd be going forward and endorsed as a health plan measure.

CO-CHAIR CELLA: The reason I said including is because I think there may be some reason you would want to be sure to include health plans that are integrated delivery systems that might not necessarily want to be required to report as a health plan, right? I would think you might want to specify health plans including those health plans that are integrated delivery systems or including integrated delivery systems that are health plans.

MS. BARTON: We'll work with NQF on how --

CO-CHAIR CELLA: To get the language so that --

MS. BARTON: -- those boxes have been defined --

CO-CHAIR CELLA: -- you can find your

universe of who's --

MS. BARTON: -- and what's included in which box to make sure that we're lined up.

CO-CHAIR CELLA: Okay. And the reliability issue?

MS. WILBON: So I don't know, Patrick. Did you have a question other than the integrated delivery system issue? Okay. Please go ahead.

MEMBER ROMANO: So it's an interesting question for those of us who are measurement geeks here, and I've had a little back and forth with John Adams over the last week or so to understand better. But the Adult Immunization Status composite measure here is not a traditional patient level all or none composite with the zero, one response.

In other words, the numerator -- the denominator is the number of immunization opportunities. The numerator is the number of immunization opportunities that were actually administered.

So there is a problem with assumption

violation in the traditional beta-binomial model because you have each enrollee could have anywhere from, as I recall, three to five opportunities. And those are obviously not independent of each other.

So I think the reliability estimate of 1.000 may be a little bit inflated because essentially the denominator is inflated because you're counting all of these immunization opportunities as if they're independent of each other where it's actually some kind of complex multi-level data structure where the immunization opportunities are three to five times greater than the number of enrollees.

So it's probably a minor issue. Maybe it's not 1.000. Maybe it's 0.958. But anyway, I think we can accept that it's pretty close. But it may require a little bit of work.

MS. BARTON: But I appreciate that. That's a useful point for us to go back and talk to our analytics colleagues in terms of how to make sure that our reliability statistics are

matching the criteria of the measure. I was not aware that all composite measures had to be all or nothing.

MEMBER ROMANO: No, no, of course not. It's a perfect design. I think it's a lovely design for the measure. It's just that you might need to incorporate different assumptions into the reliability assessment.

MS. BARTON: Thank you.

MEMBER ROMANO: And one related -- well, maybe not related. But one question for you to think about. I think that clearly the indirect criteria for validity were satisfied here. I think you reported a score level of validity.

But there does appear to be very severe ascertainment of influenza vaccinations in particular if you compare these numbers like with Medicare population with very solid numbers from the CDC. They run about four times lower. The CDC reports about 60 to 70 percent population penetration for influenza vaccination, and you're

reporting about 15 to 20 percent at the median.

So clearly there are a lot of immunizations for influenza in particular that are being given that aren't beyond the knowledge of the health plan. So that's just another.

MS. BARTON: Which is why we think that this is an excellent opportunity for health plans to learn how to track better the care that their members get so that they can close gaps where they're needed.

CO-CHAIR NERENZ: Okay. I'm detecting not a need for a revote on this with Patrick's comment, if it drops from 1 to 0.95. I'm just looking around the room for nods of confirmation. So I think perhaps we've taken care of this one. Matt, are you okay? You're the one that --

MEMBER AUSTIN: Yes. I mean, if my other subgroup committee members are comfortable with just specifying this as a health plan, I'm comfortable with that.

CO-CHAIR CELLA: Good. Okay, good. Thank you.

MEMBER NEEDLEMAN: Just one quick question and a comment on the reliability estimate. I appreciate Patrick's concern that you get the calculation right. We think that's in there. But what's driving these numbers up to one is basically the large number of members of each of these health plans.

Where we see lower reliability, we've got a big standard error around the estimate because the denominator is small. We've got -- these are accurate estimates of the rate that you're seeing in these plans with the denominators of tens of thousands, hundreds of thousands.

So, yes, the reliability numbers are fine. You're getting accurate measures of what's happening with the plan. Their reliability is not an issue.

CO-CHAIR CELLA: Let's look to the -- thank you, Mary. And then -- thank you. Go ahead and take on the slide. There's a slide that has some questions on it. We're moving

right into our discussion. Dave is going to take it.

CO-CHAIR NERENZ: No, let's just take a quick deep breath. First of all, again, thanks again. We've done the required substantive work reviewing these measures. We've had all kinds of interesting discussions. A tough issue sometimes of how do we think about this, how do we think about that. What are the rules? How do they apply to this strange situation?

And I thank you all for engaging in this, working it through until we have produced the product that we have been asked to produce. I want to just pause and acknowledge that.

Now it's a different mindset. You can sit back in your chairs now. Now we're reflecting on how we do our work and how we have done it well, how we can do it better. And I'm told there are some slides with questions for our consideration. I'm not seeing it.

CO-CHAIR CELLA: What is your opinion of our change to allow developers to provide

additional information after your preliminary analysis?

(Laughter.)

CO-CHAIR NERENZ: Clearly, there's no subgroup division relevant here. So everybody is free to do thumbs up or a card up to comment. But this is now a fully open discussion for everybody. I see a lot of thumbs up. Alex and Bijan have cards up.

MEMBER SOX-HARRIS: Having been through this before, I thought that was great. I think it would be good to have some kind of deadline for them so they're not sending them the morning of this meeting or Sunday afternoon. If we're traveling, there's really no chance to look at it.

MS. WILBON: There is a deadline.

(Laughter.)

MEMBER BORAH: I agree with that point. We need to have some time to explain or to go over those responses.

CO-CHAIR CELLA: John, you were --

MEMBER BOTT: Yes. Well, similarly, while there is a deadline, correct me if I'm wrong, I think we first received it Thursday night. So there's minimal opportunity to review the additional material.

CO-CHAIR CELLA: Gene?

MEMBER NUCCIO: Just a quick comment about the measure compilation document which I thought was great. I really appreciate it. It needed Appendix B, okay, which says, developer late comments. Because there was no link to where the developer late comments were. And so trying to find that as you're reading what our compilation was and then their late responses would've been helpful.

MS. WILBON: We think that's super helpful. We actually went in the actual discussion guide section for each vendor. But your point is valid.

MEMBER TEIGLAND: And some of them were linked incorrectly.

MEMBER NUCCIO: As you went through,

if you click on that. For example, the Group 3 was in the 30s for our compilation. But the developer response was a page like 80-something.

MS. WILBON: Right.

MEMBER NUCCIO: And so trying to find that amongst everything, it's just one more link or one more set of links.

(Simultaneous speaking.)

MEMBER NUCCIO: And we appreciate that. I know it's really great and very helpful.

MEMBER GLANCE: I completely agree. But one corollary to that is I think that the need for developer response would be mitigated if they all send some very detailed technical specifications, especially if it's a risk adjustment measure.

I think oftentimes the problem that we're having as a group is when they sort of give us very sparse details on what they're doing. And so then we have to sort of second guess and ask a lot of questions. So I think at the very beginning of the process if we had detailed

specs, it would be really helpful.

MR. ABRAMS: So I have a question for you all. So what was done was this appendix was made at the bottom. We pasted the developer's response. Would you or would you not like it if staff tried to interweave the responses after the questions that you posed in the actual section that's relevant? Just asking how you would prefer to digest the information from.

MEMBER NEEDLEMAN: In an ideal world, yes. But frankly what would be more useful for me in terms of where the staff spent their time is a lot of the issues were over the reliability, testing some of the other data. And we get refined measures.

Frankly, I would rather have the staff get some of the stuff from the testing appendix into the discussion so I'm not running around looking for the testing appendix on the SharePoint site and then finding the right page in it to go look. That, to me, would be more valuable in terms of understanding what the

developers are responding to than having the staff spend time trying to interweave the data in to the comments.

The thing that a number of the developers did that was very useful which I would tell them if you're giving us response, do this, was to put the comment down and then the response. And that makes it very easy to go back and forth between the two documents.

MEMBER FABIAN: I'm just going to add on to Mary's comment. But I think that the -- I see that as the middle ground of being able to submit information after the fact and continuing to improve what's required in the form. So I think it's already ideal and best practice is what if the form is laid out in a way that's clear enough for us to get it the first time as well as clear enough instruction from the reviewers, the developers being able to supply the right information.

So I think we're always doing it in that way hopefully. And then we have this

secondary submission while we're still working through that catch up between what processes we're trying to add in and what expectations are reasonable for people hatching up the best process.

CO-CHAIR NERENZ: Joe?

MEMBER HYDER: I'll echo Jack's point. I think the detailed and highly formatted specifications in the initial documents were very helpful. I think that there's some really specific technical things about reliability and validity that we'll want to see programmed in.

Having the developer response come in a little bit of a jumble frankly is fine by me because I'm going to read the whole thing anyway. And I may learn something along the way. But having it up front the way it was, was great and probably could be improved a little bit.

CO-CHAIR NERENZ: Sean?

MEMBER O'BRIEN: Yes, I thought the whole process was great and the material provided by the staff was very helpful. Just one thing

about, like, timelines from the developer's perspective, I think they had a week or less to prepare responses including new data analyses and a holiday in there. So I think if it's possible to extend the timeline, then you'd possibly get better data from the developers.

CO-CHAIR NERENZ: Terri?

MEMBER WARHOLAK: I thought that the process went really well. And as somebody who is new, I thought it was explained very well and that was very helpful. So I really appreciated the staff.

I got to tell you, though. I'm not really a fan of having multiple measures on one form just because, like, if I can't figure it out, then how can possibly calculated reliably? So I would like to see those separated. Not -- I mean, both for, like, so that it's not, like, an all or none, they all pass or none do. We'd have to pull them out. But also just for clarity sake. I think it really helps.

CO-CHAIR NERENZ: Just a quick

response to that. We got lucky actually the previous cycle with all the CAPs measures. Those of you involved in some of those subgroups you can do. Say, the hospital CAP survey, just kind of off the top of my head yesterday, said 5 or 6. It's, like, 10, 11. There are 11 measures and they're all under the same number.

And what it means, if you're in a subgroup or if this comes up in this setting, you're basically having a discussion -- well, either you have to say we're having a discussion about number three, period, and that's all that this is about. Or you're basically discussing and voting a package of 11.

And we got an image of how tough that was just a few minutes ago this morning with trying to figure out, is this one okay, but this one is not okay. I said we got lucky with CAP because they all went through and they all had some of the same properties and they were tested in the same way. Okay. That's all fine.

But myself would want to echo that

concern. If we could somehow have things come to us as 1223A and 1223B or something and then distinctly separate them for review and discussion. So if we're going to have something in front of us to sort out, at least we know it's that one we're worrying about and not that one that live under the same number.

Okay. Patrick, you're up, and then Jen next.

MEMBER ROMANO: Yes. I think that the challenge we had here today was to pinpoint what is the measure. I mean, because I think that the performance measure, the accountability measure that we were discussing this morning is the SIR.

And so everything else are just components needed to construct the SIR or else they're ancillary things that are reported, the registry programs like Sean's program. I mean, they report all kinds of things back to the participating organizations.

And it's wonderful. You get a whole dashboard of measures from STS. But you need to

be clear about bringing to this committee the measure that is intended for accountability, not all little pieces of it.

CO-CHAIR NERENZ: Okay. Jen and let's try to stay focused on the question. And then some of these things about how the measures are done may pop up again after lunch.

MEMBER PERLOFF: My quick corollary is that measures that are distinct also can be very related and have a pattern. So it's not just the half measures when they're 11 and 1. Even today with the measures we were just talking about, they were two closely related immunization measures. And looking at them together and thinking about those patterns also is very revealing. So it's both within a complex set, but they can be distinct and talking and looking at them together.

CO-CHAIR NERENZ: Yes. And I guess my point just to be clear is I don't mind as a subgroup member getting these ten together or three together. Having them together because of

their similar structure is a good thing. It's just clarity and discussion so that if we have a problem with one of them and it comes to this forum, at least we know we're talking about this one specifically. Okay? Good.

How about the second question, the pulling of measures? We did do that for, what, four or five measures total. I guess I'm sort of looking -- I have to look at Matt and Sherrie a little bit. But personally I thought it was good. Other comments?

MEMBER KAPLAN: I didn't even pull as many as I wanted to. The CAPs measures may be the same, but that's -- I put the brakes on that one.

CO-CHAIR NERENZ: Jeff?

MEMBER GEPPERT: I think if we pull a measure, we need to be more explicit about that's what we done and why we did it because sometimes it was -- it seemed like we were evaluating the measure in total, but we weren't, in fact. There was some specific issue that we were supposed to

be focused on.

CO-CHAIR NERENZ: Patrick?

MEMBER ROMANO: Right. And similarly, I think if the measure is pulled, it should be fair that the measure might require a revote because there might be some information that not everybody appreciated. So if a measure is pulled, I mean, a few times people said, well, the measure already passed, the measure already passed. Well, that's fine. But at least the idea of pulling, it should open up the possibility that there might need to be a revote.

CO-CHAIR NERENZ: Yes, just a quick response. I know, Sherrie, you got a card up.

MEMBER KAPLAN: I just wanted to -- so I get dinged and it's accurate for evaluating measures on what I think should be the standard versus what actually NQF gave out as the standards and where we are with adjusting the standard for the current evaluations.

So I think the pulling opportunity gives you the chance to say, well, here's the

things I was concerned about. And maybe the developer has a response to that that actually would reflect my concerns about it, enough to get through this issue until the standards move along and change. But I think the point of why are you objecting to this and what do you want to raise as an issue would help.

CO-CHAIR NERENZ: And I think as we come into this room, assuming we keep doing things this way, we can probably do a better job of targeting and streamlining the discussion. As I think to some extent we did.

But actually sort of in writing formally in advance and say, here's the issue upon which this was pulled. Say, here are the two or three outcomes plausible for discussion of that issue. Which of those may require a revote?

Just so everybody sort of knows in advance, are we debating something that if settled in a certain way will lead to a revote? If so, let's know that in advance. Or we just say something that we just need to clarify but we

probably don't need to revote.

It's like the thing that we just did about to specify for health plans. So I hope I didn't jam that through too far. But it just seemed like it had an easy solution. It's just kind of essentially an error in the check box. Fix the error, go on, move on.

So those that are pulled, I think we can frame them a little better in future cycles. But still I thought we did a pretty good job as it was with focusing, what's the key issue, what do we need to talk about, and then do we or do we not need to take action?

All right. In person versus webinar?

CO-CHAIR CELLA: I mean, travel is not fun for busy people. But I thought it was much better to do it in person. That's my opinion.

CO-CHAIR NERENZ: Terri and then John.

MEMBER WARHOLAK: I've come pretty far, but I think it was worth it, absolutely. It's so much easier to follow the conversation with a group this big in person than if we're on

the phone. I really appreciated meeting everybody face to face.

CO-CHAIR NERENZ: John, then Sam, then Matt.

MEMBER BOTT: Just maybe on the other side of the fence. I think with a constructive facilitation guidance, I think it could easily be just as well conducted by phone. I mean, traveling pretty far for essentially weighing in on two measure assigned to my team seems like an exorbitant use of my time.

CO-CHAIR NERENZ: Same order, Sam, Matt, then Gene.

MEMBER SIMON: Yes. I mean, having done it both different ways, I think the quality of the discourse is way higher having it in person. And really I think it was important to have it this way.

I think it also might -- I think it goes a ways towards promoting some consistency across the different subgroups. So I think that's something else to consider.

CO-CHAIR NERENZ: Matt?

MEMBER AUSTIN: Sam just made my point about I think hearing the conversations of the other subgroups is helpful in terms of promoting consistency and learning.

CO-CHAIR NERENZ: Gene?

MEMBER NUCCIO: I learned a lot on how the other groups reflected on what the problems were and seeing the information. And that's absolutely missing with the webinars in terms of you never hear what the other groups have done.

CO-CHAIR NERENZ: Jeff?

MEMBER GEPPERT: Is there a technical capability in this facility to more kind of video conferencing type participation? That would at least give some people the option to do this remotely.

MS. JOHNSON: I think yes. The answer is yes. Although I haven't learned how to do that. Let me just say I haven't learned how to use it yet.

MEMBER KAPLAN: Yes, we used it once

for another quality measurement meeting, but this was just last week and it was for the entire day and it worked. So we could use that in the future as well.

CO-CHAIR NERENZ: Elisa?

MS. MUNTHALI: I just wanted to clarify the process then. So we do the in person now for the measure reviews as well. We would do that twice a year plus the in person to discuss the -- or we would just do away with that and then just have these?

That's helpful because I see both sides. I agree. Like, in John's camp where traveled and was here two days to have two measures and we briefly discussed which is -- it's a hard justification of time relative to all the practicalities that are happening.

But of course, like, the discussion was face to face. It's better. You're kind of forced to pay a bit more attention than if you were on the phone. You're able to get distracted with the other things that are happening.

However, we're still going to have do all those things tomorrow now that we weren't doing today. So we should ask a question once we go back to all of it tomorrow.

CO-CHAIR NERENZ: Just a quick response, and this is looking forward so we don't know for sure. It may turn out that if we get into a pattern of two meetings, one spring, one fall, largely like this.

If we get more consistent in our own thought processes, if the measures -- requirements of the measure developers are more clearly stated, we may find that in a day and a half or a two-day block of time we can spend less time with controversial measures and more time with deeper, broader issues.

But we don't know. We'll have to see. And it also depends on how many come in a given cycle. If we get double set of measures in, we may have twice as many to talk about here. So we'll just have to see how it goes. Larry?

MEMBER GLANCE: Having done both, I

think that the in person hearing was vastly preferable. I also think this is going to segue way into the third question that you had or the third -- actually, the fourth, the one about the shadow voting.

And I think the reason it's really important is right now you can have a small group of five people who decide to pass the measure and then it doesn't go in front of the whole group. And it may mean that small group of people achieves consensus. But maybe the rest of the group would not have agreed with that decision.

And if that's the case, then you really do need to have an in-person meeting. And I think that's something that we have to discuss because, I mean, if you think about other types of review, like, one of the standing committees, for example. I've been on a standing committee for readmission measures for many years. And we all talk about and we all vote on every single measure, not just a small subgroup.

And I think that that gives a lot more

credibility when you have a large group of people as opposed to a small working group who weighs in. And there's no way you can do that through the small working groups talking over the phone obviously.

On the other hand, I think it does demand a bit more of an investment on everybody's part to do some kind of review of every measure before they come to these meetings. And that's a big deal because we look at a lot of measures, and I think it's something that we have to talk about.

But I don't think you can separate out those two issues. The one is do we have an in-person meeting? And two, do only the working group members vote on individual measures or should it be the entire group?

I would favor, A, the in-person meetings. I think those are incredibly valuable. And two, I think we really should seriously consider having everyone vote, not just a small group because I think then it becomes a lot

easier for us to go to the standing committees and say, look, you had 30 people -- 30 methodologists who voted on the scientific acceptability. This group did not think this measure was scientifically acceptable.

So honestly, you should really, really take that into account in your decision making versus if we go to the standing committee and say there were five people who voted and we don't think this is a good measure. Then they kind of look at us and say it's five people. Honestly, we have 30 people on our standing committee. We're not necessarily going to take your thinking into consideration as much.

CO-CHAIR NERENZ: Good. Thank you. Lacy then Jack.

MEMBER FABIAN: I just had a quick question. I may have missed your point. So if we all voted, though, on this, we would only be doing that for specified measured that got pulled then, not the ones that automatically -- or had consensus reached. So that would only be the

challenge. You would still only have the committee members voting on those measures that got all or nothing. But then they wouldn't have the opportunity to have the whole panel at 30.

But I agree that if -- I would find that more valuable in this meeting if we were able to all go. But I'm not sure how it syncs up with the measures that don't get pulled.

CO-CHAIR NERENZ: Yes. That's a good question. Just a quick response where we're headed. You could almost imagine some kind of reverse triage process where if the subgroup all thought it was wonderful, then we don't spend time on it. Kind of flip, look at the other side of the spectrum. Or maybe it's only the ones in the middle.

If the subgroup all thought it was awful, we don't talk about it. If the subgroup all thought it was wonderful, we don't talk about it with the exception anybody can pull anything for discussion. That overrides. And then we talk about what's in the middle.

So there would be some ways to do it. But the result of what I just said would still be that only a subgroup would do the really goods or the really bads. There would not be a full group vote. Larry, your point was, if possible, it'd be nice to have a full group vote on everything theoretically.

MEMBER GLANCE: No, I agree with this and with the approach that you're suggesting. I think it's very similar to what an NIH study section does during triage. The ones that are absolutely awful, there's no point in really bringing them up for discussion.

It's a little bit different from an NIH study, the ones that are absolutely terrific where everybody agrees. We don't discuss those as well. But it's that gray zone. And I think we have to sort of fine tune that gray zone so that we can make it practical and so we can make it work, right? I mean, obviously, we got through I don't know how many measures here. But we can't do 40 measures in two days.

CO-CHAIR NERENZ: Absolutely. Okay. Jack, Patrick, and Dave.

MEMBER NEEDLEMAN: Yes, I don't think it's feasible to do 55 measures with all of us reviewing them nor have I seen any measures where everybody was in agreement. The really bad ones the staff is doing a terrific job never let it get to us.

So I do think -- I think the subgroups work. I think the subgroups have worked very well. I suspect when we actually see the shadow voting stuff, the shadow votes are for the most part going to closely match the final subcommittee votes, not perfectly but pretty close.

So I like the consensus not reached. Perhaps some consensus not reached coming up to the full committee. I like the ability of any member of the committee to pull a measure up. And I would extend that beyond the individual subgroups. So that requires having everybody seeing the results of all the subgroups to be

able to say, gee, I'm not sure the subgroup voted the way I would have here and I'd like some more discussion with this.

Those are two feasible things. But everybody voting on everything and we don't have the weeks that would be required to be able to do that.

CO-CHAIR NERENZ: Okay. Let's do Patrick and Dave. And then just a time check. I noticed that straight up noon there's a schedule for public comment. Just because members of the public may have arranged their own schedules, we may need to do that essentially on time. So let's take a couple minutes and then we'll see where we are, how many public comments there are.

MEMBER ROMANO: Yes. So I think overall the process will move fairly quickly. I share John's frustration a little bit at essentially having two days out of my schedule for effectively one vote.

And so the question is, is there a way where members here can feel like their time is

being utilized more effectively when there's shadow voting and so forth? How would the shadow voting potentially be fed back to inform the real voting?

So I worry about the reliability of the subgroup folks, especially five people in a subgroup. If you get three, moderate or high, you're at 60 percent which is not greater than 60 percent. And so you're in consensus not reached.

So we end up with a fair number of CNRs that I think could be avoided if there are a way of marshaling more votes. So somehow there has to be a mechanism to marshal more votes, whether it's through making the subgroups larger or whether it's through in selective cases allowing everyone to vote on particular points based on the discussion that's happened here.

I love the idea of the sort of NIH reverse triage. If there were specific criteria for bringing those measures back to vote on the entire version.

CO-CHAIR NERENZ: All right. Dave and

then Zhenqiu. And after Zhenqiu, we're going to do public comments. So if anybody is teed up on the phone for public comment, we're going to get to you in just a couple minutes.

CO-CHAIR CELLA: So I got a sneak peek at yesterday's results in the shadow voting, and eight out of ten times there was agreement and twice there was no agreement. My hypothesis was not confirmed by the two that didn't agree. I thought that the reviewers, the subgroup members, would be tougher than the listeners and the shadow voters.

Turned out it was the other way around on the two where there wasn't agreement. And I have a different new hypothesis about why. My first hypothesis was the reviewers -- remember these discussions, we were kind of tough, right? I mean, there was a lot of -- most of it was asking questions about things that involved a critical view or concern.

And I'm thinking that the reviewers who were more immersed in the material were

actually bringing out their concerns but on balance wrote in those two examples a little more positive. But those of us listening and not as immersed in the review might have been more influenced by that discussion.

So one way to fix that, I agree with Larry. My personal view is I hope we'll move away from shadow voting and have whatever comes out of this committee come from the committee as opposed from a subgroup. I agree with that argument.

But one way to get around that phenomenon, if it's true, that by listening and not being as immersed in the detail, listening to the discussion that actually in a couple cases tipped the people on the shadow vote more negatively is to review with the subgroup at the end of the discussion. So where are you now in your vote?

Now that does mean it's a public statement of your vote, so it's maybe something to consider. But that's one way to actually get

kind of a recheck on where the reviewers are with the vote.

CO-CHAIR NERENZ: Okay. Zhenqiu, your card is down.

MEMBER LIN: No. So Patrick already touch on that. I thought it's a good idea to enlarge the subgroup a little bit more. They also solve various concerns, right? I mean, you have more members vote on it and carry more weight. And it may be hard for whole committee to review everything. But at least you have more people involved.

And also I saw another problem. I say for a subgroup if you have six people and then two or three couldn't make it, now you're down to three, right? Based on three votes, right? I think it's a good idea to acknowledge the subgroup.

CO-CHAIR NERENZ: I'm seeing nods. I think that's a process issue we have to deal with. That gets pretty tough from a developer, particularly if it goes south. But three or four

votes.

Tell you what. We may have some time. We may not get much public time. But let's -- since it's on the schedule and we have people waiting either in the room or on the phone, I'll let staff here manage public comment.

MS. OGHUNBEMI: If there's anyone on the phone that would like to provide a public comment, you can do so now. If you're muted, you can press star 7 or you can submit a comment via the chatbox.

CO-CHAIR NERENZ: Just give it a couple more seconds. I'm not hearing anything so far.

MS. OGHUNBEMI: And we're getting ready to bring up the shadow voting for all measures.

CO-CHAIR NERENZ: All right. I am not hearing any public comment.

MS. OGHUNBEMI: Yes.

CO-CHAIR NERENZ: All right. So let us -- we're about to see some shadow voting

results. But let's go Matt and Bijan, and then we'll see if we have shadow voting to look at.

MEMBER AUSTIN: Yes. Real quickly. So sort of tagging on to the previous comments about how to expand the number of votes. I mean, one possible idea is to -- if Subgroup 1 had originally reviewed the measure, asked Subgroup 3 to review the measure if that measure did not pass or did not have consensus, right?

So it's not necessarily opening it up maybe to the entire group, but maybe it's assigning it to another subgroup who combines with the original subgroup. So there might be a way to sort of even out workload too because we could have actually a lot of measures in this meeting. So if you had to review every single measure, that could be a lot of work. So that might be a way to sort of balance.

CO-CHAIR NERENZ: Yes, nice idea. Thanks. Bijan?

MEMBER BORAH: So I think it's going back to that workgroup issue. So if we were to

enlarge the subgroups, that does not mean that with the number of measures that we would have to evaluate would increase. So how do you kind of reconcile or how do you actually deal with that data?

MEMBER PERLOFF: I'm going to lead into the next question on the list that was there before. I will confess publically to everyone here. I was bad on the documentation side of this round. In some sense, I've evolved past the form. I'm going to put this in a positive spin.

I've done the form so many times now, I write down sort of critiques and then I have to go fit them into the form. And I find myself spending so much time just navigating. But I'm learning the criteria. I'm getting really good at critiquing measures.

So if we have more measures, larger committees, more measures, we have to be more streamlined in how we're able to provide some feedback, I think. So there is this sort of interconnectedness between what we're looking at

and what we need to get back to do.

And to Larry's point, I think there's an -- we need an intermediate set of information. If we were to skim all of these measures more quickly, the documentation takes a long time to go through. So there needs to be sort of an upper level abstract that gives you a lot of the texture that we need but in a snapshot.

So that's a hard ask. But I think that these two sort of issues will help kind of address some of the other things that we're talking about.

CO-CHAIR NERENZ: All right. Shadow votes, do we have something to look at? We're going to need some motivation.

MS. WILBON: This is a little bit small. We tried to fit everything on one slide. So largely, there was agreement. We didn't have time to calculate the number of people who were not comfortable submitting ratings for the measures we read this morning. We'll work on that.

But you can see -- and actually, I don't know if Karen will want to review some of this. But essentially, on all but two measures, the shadow vote corresponded with the subgroup vote. The exception was the med rec and the 0018.

So I think for those us that were a part of this 2456 Measure discussion, that might not have been such a surprise that there was some discordance there. And 0018, I'm not as sure. I don't know if others might have --

CO-CHAIR CELLA: It was very close. If it was a one column to one column comparison and it was very close, then it probably would've passed overall.

MS. JOHNSON: That one, just FYI, the numbers from the shadow vote was right at 60 percent. And remember we need 60.0-something or other.

CO-CHAIR CELLA: Yes. But if you combine that with the subgroup vote, it would've passed.

MS. JOHNSON: Yes.

CO-CHAIR CELLA: So there's really only one where the vote would've been different had you used the whole group.

MS. JOHNSON: Yes, Yes. So it was pretty good. One of the things that I thought was really interesting -- and again an idea hasn't been that we want all of you to look at all the measures, certainly not in glorious detail, right? So that's too much.

So we were really interested in whether or not you felt comfortable submitting a rating after looking at whatever you felt like looking at and after hearing whatever.

So to me, a little bit what's interesting but I haven't had time to think about it very much is did people feel -- lots of people felt pretty comfortable on several measures and quite a few felt pretty uncomfortable on several. I'm not quite sure what the difference between the measures were. So there is a pretty wide variation there.

MEMBER NEEDLEMAN: Karen, can you ask the group how many felt comfortable when it was a shadow and sort of an advisory vote versus how many would've felt less comfortable if the vote had been for purposes of endorsement or not -- or moving the measure forward or not as opposed to shadowing this?

I voted on some things because I'm an aggressive person. I voted on some things that I'm not sure I would've. Had all the background and had not fully absorbed the documentation, I'm not sure I would've been prepared to vote if it was a vote that counted.

MS. JOHNSON: Okay. So you were comfortable shadow voting but not necessarily comfortable real voting. Okay, okay. Yes, and I didn't think to ask it in that way, yes, yes. I don't know if anybody has any thoughts on what you're seeing here.

MEMBER AUSTIN: So for me, I was sort of a mixed bag. Sometimes I didn't feel comfortable and sometimes I did. Some of it was,

at least for me, sort of the complexity of the measure and how much I sort of could follow in terms of the conversation and quickly sort of absorb from the documentation. And for other ones, sometimes I could do that and sometimes I couldn't. And where I couldn't, I felt less comfortable.

MS. WILBON: Do you think that we could work on maybe the way that we're structuring the introduction? So at this time, we had, like, staff do kind of a higher level overview and then, like, the lead discussant do more in depth maybe overview of what the issues were.

I'm wondering if that maybe that intro was structured in a way that even if you didn't review the measure, if you had all -- I mean, we thought we were trying to sort of do that. But maybe in some cases it wasn't as clear or consistent about how the measure was structured or about the specifications or something like that.

I'm just wondering if maybe that would help or if it was maybe more the nuance of the things that at a high level you really can't get to that level.

MEMBER AUSTIN: For me personally, it's I'm better with visual than audio, right? So to have someone sort of verbally describe the measure, it doesn't quite sink in the same way as literally as you see in the numerator and denominator.

CO-CHAIR NERENZ: Okay. Something struck a chord. We've got Christie, John, Sherrie, bang, bang, bang. Just like that.

MEMBER TEIGLAND: Okay. I'll start. I think the discussion guide is the key here and I relied on it heavily. I did have to sometimes go back to the measure specs. But I think we need that measure spec. I do, because that's how I look at them. I want to see the detailed definition. I don't want to see the description that you provided usually wasn't very useful for me, right? It was not that accurate. It was

ambiguous and overbearing.

So I had to really -- but we had the documents. But they're hard to find and sometimes the internet would stop on me. So I think that if we had a good discussion guide really and we can quickly find the -- we have the specs, I feel confident in the subgroup's committees to, Yes, analyze these measures and tell me what the issues are and then look at the information, hear the feedback and most made a decision.

There were a couple where they were really complicated. And I said, no, I'm not comfortable. But I think if we really think hard about how we do those discussion guides and really actually spent a lot of time on them. So I won't say spend more time, but maybe there was some stuff there that wasn't all that relevant. And I don't know how to do that.

MR. STOLPE: But having easy access to the specs would be useful for you.

MEMBER TEIGLAND: Absolutely, yes.

MR. STOLPE: Let me ask you this then. So this isn't something that's hard for staff. So we could easily put in, say, an appendix, for example, with a link --

MEMBER TEIGLAND: Yes.

MEMBER O'BRIEN: -- at the front of the discussion of the individual measure. So if you want to click on that, it will lead you directly into a link to the appendix. Is that something --

MEMBER TEIGLAND: Yes, that would've helped because I had to go search for it in my own files, right? I had it somewhere.

MS. JOHNSON: Really funny thing is that we didn't even think of doing that. It would've been easy to do.

MEMBER TEIGLAND: Yes, yes.

MR. STOLPE: Well, I tried to summarize some of the specs and some of the discussions which was an effort for me. But if it would just be easier to --

(Simultaneous speaking.)

MEMBER TEIGLAND: It'd be better for me to look at the specs.

MR. STOLPE: For example, just put them in a link.

MEMBER TEIGLAND: I can read that spec quicker than I can try to figure out what you're trying to tell me.

MS. WILBON: That's helpful.

MEMBER BOTT: Yeah. My comment is a hybrid of Christie and Matt. Like Matt, I often -- if I get lost in the verbal conversation, I like to fall back and look at a document. And I resisted going to SharePoint because I figured I fart around so much on SharePoint and miss more of the conversation.

So if there was a humongous document like what I saw on the map coordinating committee, perhaps in the discussion guide if I can click on the testing form and then there's the full testing form for that measure or that MIF, that would've been golden. I would've had the information I possibly would have needed.

CO-CHAIR NERENZ: Sherrie?

MEMBER KAPLAN: So having sat on a bunch of NIH committees, the prior reviewer really gives you that snapshot. And the staff actually does that for you. So the primary reviewer's role in this case is being served by staff, of course. And the synthetic kind of comments of everybody who actually did review it and summarized those in a single place is great with the links would've helped a lot back to the original proposal.

Because honestly when those proposals -- they come in, in two parts. And the testing part is the one you really want. And that's often -- when you read the first part and then it's referring you back to the second part and then that's got an appendix to some third part. If you could streamline that, I think the whole process would go down a whole lot easier. And for me, that staff summary is critical because I don't want to have to do that.

(Simultaneous speaking.)

CO-CHAIR NERENZ: Just quickly. We just have our scheduled lunch for you. People are probably ready to pass out from hunger.

(Laughter.)

CO-CHAIR NERENZ: We'll do Patrick, Daniel, Lacy, and maybe we can set this to rest until we come back after lunch.

MEMBER ROMANO: I just wanted to comment on that one clear instruction set because maybe that could be viewed as a process failure or a process improvement opportunity.

So this was a case where perhaps there needs to be a higher level committee discussion about what constitutes appropriate evidence for validity testing. Because basically the developers were saying the measures stay valid because nobody has ever shown a better way to do what we're doing. That is using pharmacists to measure med rec.

And second, that the score is valid because we can improve it. Well, yes, they demonstrate that they can improve the score. But

is that a strong enough construct? It's not really a quality construct because it's not linked to outcomes in any way.

So I have a feeling that this is an example where this could come back to the large group and say, is this is really construct validity if the construct is simply that we can improve? I can improve lots of things. That doesn't mean that it's meaningful. It doesn't measure -- it doesn't mean that it's a quality concept just because I can improve it.

So that's -- I mean, I'm not trying to revisit us in history here. But just to sort of explore how when we find these discrepancies, there might be a process to bring them back and say, well, what issues are raised?

CO-CHAIR NERENZ: Okay. Daniel, Lacy.

MEMBER DEUTSCHER: Just a very brief comment. I think I would have not felt comfortable actually voting on the measures I haven't really reviewed in depth. And thinking of what you said that on this cycle we didn't

have that many measures. So we could possibly have many more which means less discussion time on each one of them.

And during that small amount of time getting into the documents I haven't really spent time looking at and digesting that and giving a valid vote, I would have not felt comfortable.

I think maybe a compromise, I would probably prefer to have, like, a minimum number of members on the subgroups, maybe a little bit higher than what we had today by one or two, maybe seven members or as a minimum which would mean maybe probably to our review a little bit more measures than we have but not too much of that. Maybe that's a compromise I would have preferred.

MEMBER FABIAN: I also wouldn't have felt comfortable having the vote count. I felt okay being harsher on the measure because I knew it wasn't counting. But it was a way to say --

PARTICIPANT: Speak louder, please.

MEMBER FABIAN: I really think that

there's opportunity to make the improvement. But to the point of the visual, so more details and access to the details is great. But the reality is being able to review like all of the measures against ahead of getting those details or even during is not going to happen.

So to the point that I think Matt made about being visual, it is really nice to have the slide that overviews, like, where the consensus wasn't reach and then what numbers were there. I wonder if there's opportunity to take that, like, with an extra couple of bullet points.

And these are the particular issues that are being discussed around consensus not reached because that was the other challenge. Like, some of the measures, we departed away from the part we were voting on. Not to say the discussion wasn't valuable for other things. But as a person who didn't review the measures, it made it that much more difficult to be confident on what I was actually voting on.

So if I had that cue as a part of the

slide to remind me, oh, we're just voting on this piece, I would've felt more comfortable paying attention to that and then having my vote potentially count.

CO-CHAIR NERENZ: Yeah. And just a quick reply and I'll get out of the way and watch. One thing conceivably we could think of is trying to even bring out in sharper detail what's the reason for, say, the consensus not reached. We had the distribution in the books. But if three people said low and two people said moderate, why?

In the discussion, it kind of comes out. But if we could set that up in advance somehow and say, okay, here's the key issue. These people felt this number was okay. These other two people looked at the same number. They interpreted it the same way. They thought it was not okay. It's simply a matter of where the low-moderate line gets drawn. And now let's see what we can do with that.

And that's different from saying these

people looked at this number and then led to the rating. And these people looked at this whole different number and went to the rating. And now let's see if we can sort that out.

So I was sort of looking at that during our two days. Can we figure out as tightly as possible why something is in front of us? And then can we help the group and all of us resolve that? But we'll see if we can do that.

MEMBER TEIGLAND: Wouldn't that be in the comments? I mean, whenever I rate something low, I carefully say why the heck I rated it low.

CO-CHAIR NERENZ: Yeah. But some of that didn't flow through --

MEMBER TEIGLAND: So maybe that needs to be --

CO-CHAIR NERENZ: -- the computer, yeah.

MEMBER TEIGLAND: -- in the discussion, yeah.

CO-CHAIR NERENZ: Or even some selection of that sort of to highlight the

disagreement. Not all comments, but just to say --

MEMBER NEEDLEMAN: Why doesn't it in the discussion guide?

MEMBER TEIGLAND: Sometimes.

MEMBER NEEDLEMAN: Sometimes.

MEMBER TEIGLAND: Sometimes.

(Simultaneous speaking.)

MEMBER TEIGLAND: And I don't know if everybody ranks them, if they just say low.

(Simultaneous speaking.)

MS. JOHNSON: I think sometimes it was pretty clear that there were two or three things that were problematic across the board. And other times, there were lots of different things. It was Wednesday, Tuesday. So I think in general, we've gone a little bit more for comprehensiveness rather than synthesis.

So maybe we'll get you guys to help us on that a little bit. We don't want to hide any concern. Of course, you can still bring that up even if it's not in the discussion guide. So

maybe we do need to synthesize more.

MEMBER KAPLAN: Karen, could you add to this afternoon revisiting your story? Because high-moderate don't mean anything.

MS. STONE: We put it on the list.

MEMBER KAPLAN: Is that on the list?

CO-CHAIR NERENZ: Yeah. Okay. We have -- what do you want to do about return time? We're five minutes over. Do you want to take the full half hour? Do you want to come back in 42 and eat fast? What's the pleasure of the group?

PARTICIPANT: Forty-two.

CO-CHAIR NERENZ: Okay, 42. So let's go.

(Whereupon, the above-entitled matter went off the record at 12:23 p.m. and resumed at 12:49 p.m.)

CO-CHAIR NERENZ: These are a little more finite methodological issues, and they have a couple of flavors. And then eventually I'm going to turn to Karen to kind of walk us through.

One of the things we want to do -- now reflecting back on how we've spent the last day-and-a-half -- is identify methodological issues that we should try to settle over a period of time, like in a December webinar call or something, where we've had some time to think about it.

Not likely to settle it in five minutes, but we've seen now -- particularly in this set of measures -- that this is an important issue in order for us to do our work. To communicate with developers, to be consistent, we need some thought and some resolution. So there's that kind of thing.

There may be a couple of things that we can settle quickly, meaning right now. And we'll see if that can happen. So for at least this block of time, it's not so much about the mechanics and workflow of our process. It's about: how do we think about this approach to validity, or how do we think about this particular problem with risk adjustment?

So it's moved a little more into substance. And with that, I will let Karen take us. Because there's all kinds of things we could talk about.

MS. JOHNSON: We have an ongoing list. Let me tell you a couple of things that we know we'd like to tackle and that we'll tackle eventually, and I'm not sure where we'll get. But we've often talked about we should identify thresholds for certain things. We agree, but that's not an easy conversation. So let's just know that we're not going to do that today. Okay? We'll come back to some of those things in the future. And if those of you on the phone wouldn't mind muting your phone, we'd appreciate it.

So one of the things that might be really fast to chat about today is kind of some faces that were made when we said, oh yeah, remember, NQF allows data element validation to kind of stand in place additional reliability. So this was something that has been an NQF thing

since before I came. I've been here nine years. It's been here before I came, and probably before you came, Elisa.

And so we don't really know what the thinking was behind it. I suspect that some the thinking was a way to make things a little bit easier on developers. And also many years ago, a lot of our measures were more basic process measures that were paper-based, so people often were working with fairly small testing samples.

I don't know, Marybeth, maybe you remember. Do you? Marybeth used to work at NQF. Maybe you have some history that would give us just a flavor of why that is.

MEMBER FARQUHAR: I think you're right. I think it had to do with measure developers and how much information that we were asking at the time. They did have a spreadsheet or a worksheet, but it was kind of not like what you have now. It was very abbreviated, and a lot of questions were asked and a lot of information was passed to the steering committee. They

didn't have this panel at that point.

MS. JOHNSON: Right.

MEMBER FARQUHAR: So some of -- you had maybe one or two methodologists on the steering panel to do some of that work, and it just felt like they needed to do some more.

And remember too, it was evolving as we were going along. So this was an experiment -- -- NQF's experiment. So they came along and they started listening and started with a consensus, and then they started to get folks more involved in that. But I believe you're correct with regard to the measure developer, and the burden that they felt that they had to submit all this information.

MS. JOHNSON: Okay. So Michael, I'll come to you in a second. So I think my feeling is that you guys would like to formally recommend that we recommend to our CSAC that we change that.

So maybe the way to have this discussion is does anybody think that we should

not move in that direction? So should the status quo -- you know what I'm saying? And Michael, do you --

MR. ABRAMS: Well, let me offer this connection to our meeting today to try and seek clarification from you about why data element validity might stand as a substitute for reliability.

And so if you recall the measure that came up today where we talked about this was the sepsis meningitis measure. Okay? And they, in proffering to us the reliability experiment that they did, they took their measure as it was run from their normal kind of chart review, and then they compared it to this online -- what they call this online calculator. Right?

And they did this in 320 cases and showed the positive predictive value and since then discussed with you those kind of statistics. And so they were high, so this is reliable. And they didn't do a very good job explaining this, but I'm intuiting that what they were trying to

convey and why they thought it was reliability, is because they viewed it as a kind of test/retest. Right? They used the same 320 observations.

And the test/retest -- or maybe I'll modify it. They did a test and a retest with a different test. Right? And they used that as a way to say that they could reproduce the measure. And the fact that they did that they felt suggested the reproducibility overall could be accepted. And that was kind of the case they made.

I'm wondering if that might characterize -- and you all tell us -- if that might characterize why somebody would take data element validity like that and say that it tells you something about the reproducibility of the measure.

CO-CHAIR NERENZ: Patrick, and then Jack.

MEMBER ROMANO: I wonder --- and again, this may be part of the history that you

folks may be more familiar with. But would it make sense for this to be permissible one time for a new measure because of the fact that if the measure is not yet in use, then the developer will not have access to data with which to measure score-level reliability. And then they would be defaulting anyway to the reliability testing of patient-level data elements.

And validity testing of those data elements could be at least as good as reliability testing of those data elements under certain circumstances, depending on how the measure is collected in practice. And I think that the argument was made today, for example, that what they are calling reliability testing is really a validity testing of their mechanism for capturing the data.

But anyway, my point is that if we don't have score-level reliability, which we don't require, and it wouldn't be available initially, so in this particular case would it still make sense to allow validity testing with

the idea that if there's a very high level of validity, and if the method for validity testing is similar to what is being done in the field anyway, then it's effectively a reliability test at the same time.

CO-CHAIR NERENZ: And let me just comment. We'll get to Larry. What you said right at the end was getting to where I thought this needed to go. And I've been looking at the high-level math folks in the room who are more adept at this than I am.

Is it simply mathematically true that valid data elements must be reliable? That seems to be the logic upon which this is based, and I'm seeing at least two negatives shakes of heads. So Larry?

MEMBER GLANCE: So reliability and validity are two different things. You can show that administrative data is reliable without it being valid.

So in order to show that administrative data is reliable, you would take

the medical record and you'd have two different abstractors abstract the same medical record and show that their reproducibly getting the same ICD codes.

But those ICD codes may not validly capture the data concepts. So for example, we know that ICD coding for myocardial infarction is not very sensitive. It may be very reliable. You get the same ICD codes for MIs from the same bunch of medical records, but it may still not be valid.

CO-CHAIR NERENZ: But it's the other one.

MEMBER GLANCE: So having said that --

CO-CHAIR NERENZ: It's the other one.

MEMBER GLANCE: You know, having said that, is the data -- if it's valid, is it necessarily going to be reliable? Well, the way you get at validity is you basically compare what you get using the data elements that you have, and then you compare them to the authoritative medical record and see if they agree. And that's

validity.

So, you know, I guess that would make an argument that if you have validity, you have reliability. But --

CO-CHAIR NERENZ: That seems to be the core question.

MEMBER GLANCE: Yeah.

(Simultaneous speaking.)

CO-CHAIR NERENZ: Could I -- no, keep going.

MEMBER GLANCE: But to me -- and maybe I'm going to be out of turn here, and maybe Karen is going to hit me for this. But I think really the bigger question is not about whether it's okay to have data validity without looking at data reliability.

I think really the bigger question is, under current NQF guidelines right now, if you pass the data reliability and pass the data validities, then you can get a pass on scientific acceptabilities.

And to me, I don't think that should

be acceptable. I don't think it's enough for the data quality to be high, for us to be able to endorse a measure as being scientifically acceptable. I don't think under any circumstances a measure should be endorsed in terms of scientific acceptability, unless you can determine that score -- at the score level, it's both reliable and valid.

I think we need to throw out base validity, even for a first-time measure. And I think we should throw out data quality as being the only requisite for endorsement for scientific acceptability. I think that should be -- to my way of thinking, that is like the core issue for our group.

CO-CHAIR NERENZ: Just before we get into that -- if we do get into that -- let's just -- I'm curious about the negative head shake simply on the more narrow issue of: does data validity imply or guarantee data reliability? I have a lot of negative head shakes down here.

Can we just sort of --- but then I

promise we'll come back to that one. So I had Sherrie -- well, Sherrie was a very visible, negative shaker. And let's just move down the --

(Simultaneous speaking.)

CO-CHAIR NERENZ: Yes, this is the -- the California group doesn't like this one.

MEMBER KAPLAN: So some of this is absolutely disciplinary training. And that's who I am, so that's kind of how I was trained. Measures are only valid and reliable for the purpose they were put to and the context they were tested in.

And so are you going to re-validate a thermometer? No, because it's been tested so frequently, and we know what the intervals mean. We know which are probably reproducible, in whose hands they're reproducible, blah blah blah.

And one of the measures of glycemic control, for example, has gone through those kinds of transformations really. You can ask the question, is it valid? Is it reliable? For what purposes and what circumstances?

But that's why it gets me -- it grinds my back teeth when people say I'm using a valid tool. Because we don't know the circumstances -- you use the measure of functional status that requires literacy of adults and now you're using a four-year-old. No. So can you have a valid measure that's not also reliable? It depends on the circumstances. And can you be accurate without being reliable? Can you -- yes. You can be accurate, and then not be able to reproduce that accuracy because you change the circumstances of testing.

So for me, it's goofball to ask the question can you have a valid measure that you can't reproduce? Why would you want to do that?

CO-CHAIR NERENZ: And again, let's just do the foreign language translation. In your context, in your life, a measure is what we are calling around here a data element, like a survey score for a patient.

MEMBER KAPLAN: Yeah, close to the IPO.

CO-CHAIR NERENZ: Okay. No, I understand. It's just --

MEMBER KAPLAN: Can you have an accurate data source without it being reproducible? Yeah.

CO-CHAIR NERENZ: Yeah, got it. Okay, sure. Jack.

MEMBER NEEDLEMAN: Okay. So we're dealing with lots of language, and sometimes we're very precise and technical, and sometimes we're very colloquial. So for me, the validity question colloquially is: are they measuring what they think they're measuring? That's the validity issue.

How well it's measured, how accurately it's measured, how representative this measure is of the unit on which we want to analyze it, is a reliability question.

MEMBER KAPLAN: How accurate it is.

(Simultaneous speaking.)

MEMBER NEEDLEMAN: Well, let me -- I'm going to be much more specific about that.

So we measure -- the reliability issue comes up at two very different levels. It comes up at the individual measurement point for one-by-one-by-one, where the measures wait. So if I'm measuring blood pressure, I know there's going to be variability there. So the question is: is the measure I got reliable? Is it a reflection of what I think the blood pressure is going to be?

In the case of the issue today it was: if I go through the medical records, if I go through the electronic health record and I use this algorithm, am I always going to come up with the answer, this patient had sepsis, or this patient didn't have sepsis, or meningitis?

And one of the issues of reliability is at the measurement level. Right at the measure -- where the measure is first applied. The second question that we have about the reliability: is the aggregation up? So when we're asking: does this clinician, does this hospital have a good score or a bad score? How

do they compare to others? Is that a reliable measure given the inherent variability in what we're measuring across patients?

And that's a separate issue from the question of: is the measure accurate enough to be called reliable at the measure level? What we were discussing today and what we were talking about was that the algorithm was being tested for accuracy at the individual patient measurement level. And the question was: is that reliable? And they were saying: do we get the same answer out of the HR that we get when we have a formal abstractor doing it?

But that doesn't answer the question -- the second reliability question -- which is, when we begin applying these measures of what proportion kids got sepsis, whether or not -- and we say, this place is doing better than this place, whether that's a reliable comparison given the inherent variability of the phenomenon at both places.

And so we've got two different levels

at which we're analyzing reliability. And this look at the score level says, we're ignoring, at least in the first round of the testing, the second issue.

CO-CHAIR NERENZ: Yes. Sean?

MEMBER O'BRIEN: To me, this is an issue of, can score-level -- can the item-level properties be replacing the score-level properties assessment, to meet some recommendation, or there's something.

But just to come back to this issue of reliability versus validity, I would say those are fundamental constructs and useful and helpful for organizing thinking in a lot of contexts -- but we're sitting around talking about the same issue, debating, is this reliability or validity, then those terms are not as helpful as they could be.

So in the case of item-level reliability, if you randomly sample, say 10 abstractors, from some population and get different results for the assessment of whether

an infection occurred based on chart review.

You could say that's --- and if there's discrepancies, say, describe that as a reliability issue. But if those 10 abstractors are associated with 10 different providers, and now those 10 abstractors are going to pool charts off to infinity for, now, one of the measures we used to compare across those providers, then you're now going to say, well, that's a permanent, fixed kind of systematic difference. And now in systematics, so let's call that validity. So I think it's context-dependent, and I'm not sure those distinctions are always that helpful.

(Off mic comments.)

MEMBER GLANCE: I guess the terminology does get really confusing. And Sherrie and I always have this back-and-forth. But if I could make it really concrete in terms of risk-adjusted outcome measures. So how to separate out the score from the data elements.

So if you look at readmissions, okay,

the data elements are whether or not you were readmitted -- yes or no -- and what the risk factors are for readmission. Those are the data elements. So when you're looking at reliability and validity at the data element level, okay, that's very different from looking at reliability and validity at the score level.

So when you're looking at reliability and validity at the score level, you're using something like the signal-and-noise ratio for reliability. For validity, we're looking at the risk-adjustment model. If it's a logistic-regression model, you're looking at discrimination and calibration.

And so the point that I was trying to make earlier is that currently under NQF rules, okay, you can pass a measure as being scientifically acceptable --

(Music plays in the background.)

(Off mic comments.)

MEMBER GLANCE: So under current NQF rules, a measure that the entire package can be

endorsed as being scientifically acceptable by only passing it in terms of reliability and validity of the data elements as I've defined them. And that, to me, should be one of the key discussion points for our group -- whether data quality, as measured using the reliability and validity of the data elements as I've defined it, should be enough for a measure to be called scientifically acceptable. And I think that should be a core discussion for our group.

CO-CHAIR NERENZ: Got it. Daniel? And then we'll see where we go from there.

MEMBER DEUTSCHER: I agree with Larry about the need to always look at the score level, since this is kind of the purpose of this group is to get to that level.

But going back to the reliability or validity question, I think that trying to determine or to state that a measure is reliable just because it's valid is kind of a non-valid statement, if I could say. Because --

CO-CHAIR NERENZ: Okay, if I could just

fine point. It's not that the measure is reliable, because it's the data element. Does that make any difference?

MEMBER DEUTSCHER: Right. Well, I was thinking of data elements. Yeah.

CO-CHAIR NERENZ: Okay.

MEMBER DEUTSCHER: And why is that maybe not a valid statement, because reliability and validity are not dichotomous. So you could have a measure that's valid under certain circumstances, but it's valid with a certain level of validity, and it might be also reliable with a certain level of reliability. So there's a certain level here and a certain level here. And one does not determine the level of the other, which is why they always need to be looked at. Both of them.

CO-CHAIR NERENZ: If I could try to weave a couple of things together and maybe just put out a straw person sort of proposal, and I'm intentionally sort of parking Sherrie's comment about reliability and validity being not

fundamentally -- just simply properties of measures, that's a big issue here that I agree with absolutely.

But that's -- our whole structure is we assign those labels or those pass-fail characteristics to measures. So we're going to have to get back to that.

What I think I'm hearing and saw in the head shakes, is that we are not happy as a group, because I heard nobody with this current rule that says, if you pass validity for data elements, you therefore get a pass on -- or you can use that to pass reliability.

I saw head shakes again, so I haven't heard anybody speak in favor. So I think --

MEMBER KAPLAN: Can you repeat that?

CO-CHAIR NERENZ: Well, I'm trying to figure out -- you can impute -- I'm not sure quite what the verb is. You can bring forward to us right now evidence of data element validity. And if that is acceptable, you do not have to separately bring forward evidence of data element

reliability.

And our scoring instruction is to take the validity pass and bring it over to the reliability category and give that a pass. That is the current rule.

And I have not heard anybody argue that that is a good rule. I have heard people shake their heads or say that is a bad rule, because validity does not guarantee reliability. And if that's the sense of the room -- and now I'm seeing nods -- we should convey that back to CSAC or whoever.

PARTICIPANT: You want to vote?

CO-CHAIR NERENZ: Okay, then -- well, okay. I mean, do we have a show of hands on that?

PARTICIPANT: We're agreeing that it's a bad rule.

CO-CHAIR NERENZ: It's agreeing it's a bad rule. Okay. Then there's the second thing that Larry has been speaking about is that -- and again, I didn't see any -- I'm watching the

non-verbals here. We currently have a state where particularly for new measures, if you pass the data element test, you do not have to pass measure score test, meaning if your data elements are reliable and valid, you're good to go. That's what the current rule is.

Larry objects to that. I personally share that objection. I'm watching the non-verbals. I see other people. And the proposal then presumably would be essentially there are four boxes, and you have to check them all.

You have data element reliability and validity; you have measure score reliability and validity. The sense of the room that I think I'm seeing is that you must do all of them.

MEMBER NEEDLEMAN: So I want to revisit that sense of the room.

CO-CHAIR NERENZ: Okay. No, that's good. I just wanted to lay it out there to talk about it.

MEMBER NEEDLEMAN: Because -- and I

was not present at the creation. But the sense I had was when that bifurcation -- do we have data good enough to use as a measure if the data reliability issue was sufficient?

And I think the model was, if we're comfortable that the data is reliable, now let's go out to the field with this measure, get some more data from the field and decide whether at the score level, at the comparison level it's being used at, there's enough reliability to continue the measure. And that required -- but people were not going out into the field until they had an endorsement of the basic measurement concept.

And times have changed. And the question is whether we can expect the measure developers to go out in the field at this point and collect data from more and different groups based upon -- before they've been told at the measure level the data is reliable.

And that seems to me -- that was the split that seemed to be implicit in the data.

We're going to make sure the data is reliable, go out into the field, go collect data from a large number of providers, units, whatever your unit of analysis is.

And then come back with data on whether or not the scores are stable enough, consistent enough that we are comfortable that it's a reliable measure applied in those cross-comparisons among whatever the units are that are being compared.

To say let's not go there -- let's say we want to see both the data reliability and the score reliability, is to say to developers, go start collecting data from multiple units before you come to us with the specific measure in hand to ask us whether we think it's an inherently reliable enough measure at the data collection point. That's the question that you're asking.

CO-CHAIR NERENZ: Yes, thank you. And that's a good framing. I'm not an authority on the creation of this either. That sounds like a fair summary. And that would be the implication.

We would be telling developers, if you have to touch all four boxes first time in, you have to do that multi-provider, multi-unit data collection before you come in the first time.

Okay, so let's go Joe, Sean, Patrick, and then back to Larry.

MEMBER NEEDLEMAN: So my question would be what would be the consequence? Because I think we'd all love these bulletproof measures where they touch all four boxes, to use that phrase. What would be the consequence for the measure development pipeline if we tell developers that?

CO-CHAIR NERENZ: I'm implying in Jack's comment, it might be negative. But developers would have to speak to that. Sean?

MEMBER O'BRIEN: Yeah, so your proposal, or a specific direction which is requiring the developers to check four boxes, I think my proposal is scaled back a little bit from that, because I think the question of, what should developers be required to demonstrate

empirically, is a different topic. And I have some different thoughts about when validity testing makes sense and what types of analyses I want to see as a reviewer.

So a possible alternative or version of recommendation would be to revise the NQF evaluation criteria so that you don't have a box that says reliability, and those two fundamentally different things are stuck inside that box -- the item-level reliability and the score-level reliability.

They have the same word in common, and so because of some historical accident they were together. But they don't belong together. So split them apart, and at least I can get behind that very well, in terms of what -- and then I'd say, what do we require for first time submissions, things like that, for a different topic.

CO-CHAIR NERENZ: Okay. Andrew.

MEMBER ROMANO: I'm sorry, but I just want to go back --

CO-CHAIR NERENZ: Oh, and I'm sorry. I missed -- I'll get back to Jeff then. Sorry I didn't see your sign up. But Patrick first, then Jeff.

MEMBER ROMANO: I just want to go back to the previous question just for a second. So you asked basically if there was any justification for the rule whereby data element validity testing could substitute for data element reliability testing.

CO-CHAIR NERENZ: Right.

MEMBER ROMANO: And I do want to put forth one scenario. Okay? So a lot of the measures are based on administrative data. Essentially we could call it claims data or coded data. Right?

So when we're looking at the data element then, reliability testing means we have two different coders, given the exact same medical record in the same circumstances, independently reviewing and coding the medical record.

Data element validity means that we have a gold standard -- a criterion standard, if you will -- with maybe physicians or nurses, or a group of people coming to some kind of consensus decision about what the true value of that data element is.

Now, it could be argued that if the validity is very high -- close to 100 percent -- and if the sample is representative of the population, then do we need to do another study -- a separate study, an extra study -- in which we test reliability by having two different coders reviewing the same record?

In other words, if we know the data element is true and valid and it's measuring what we want, and that's based on a sample that's representative of the population, then what is the additional value of doing a separate study with two different coders reviewing the same record?

CO-CHAIR NERENZ: Well -- yes, and that's a wonderful way of framing it. And that's

-- again, it sort of draws the question, is the scenario that you described possible?

Meaning does that validity that you have tested guarantee the reliability that you have not tested? And I keep seeing negative shakes every time we frame it that way.

MEMBER ROMANO: My point is to you --

PARTICIPANT: Because it doesn't add any value.

MEMBER ROMANO: It doesn't guarantee it. But the question is: does it add value to know that the reliability is high? In other words, if you know that the data element is valid, do you also need to know that two coders would get exactly the same results? Because presumably if the reliability was below the bar, then that would show up on the validity test.

CO-CHAIR NERENZ: No, I think --

PARTICIPANT: No, no, no, no. No, no.

MR. ABRAMS: It would be helpful to clarify --

MEMBER ROMANO: Yes.

MR. ABRAMS: -- what we're talking about in terms of the data element.

MEMBER ROMANO: We're talking about a very specific situation, where reliability is based on two coders independently reviewing the same record. Validity is based on the coder being criterion-checked with a gold standard. So how could it be that you get 100 validity?

MEMBER NEEDLEMAN: You're shifting things that are in the reliability category into the validity category measure.

MS. JOHNSON: Now, I would say that Patrick's scenario is probably the thinking that happened back in NQF back in the day. Right? That somebody has done a test and they said, hey, we've looked at the gold standard, it's accurate, so why do we need to have somebody go back and show that two people agree on the thing that you know is accurate, right?

MEMBER NEEDLEMAN: Let me take Patrick's example and make it a little bit concrete and why I think there's a difference

between reliability and validity. There's a lab test somewhere in the -- from a culture somewhere in the medical record that says you either had an infection or you didn't. Right?

That's a gold standard. We know exactly what to look for there. That's the measure we're looking for to say: did this patient have an infection? That's the validity check.

The reliability check is somebody going through the records, do they find that test and get it accurately recorded? Do they record things positive where the test was negative? Do they miss tests entirely and report no test, no infection? That's a reliability check, not a validity check.

MEMBER ROMANO: Right. I understand. I'm not saying that they're not different. What I'm saying is that the reliability check doesn't add useful information to the validity check.

MEMBER NEEDLEMAN: If you can't --

MEMBER ROMANO: If the validity is

very high and the sample is represented --

MEMBER NEEDLEMAN: If I know exactly what I'm looking for in the medical record, but whoever is abstracting it routinely fails to find it, that measure is not reliable and shouldn't be used.

MEMBER ROMANO: But it wouldn't be valid either.

CO-CHAIR NERENZ: The data element isn't. Okay. Jeff, you looked like you were going to follow right on Patrick. So let's get --

(Laughter.)

MEMBER GEPPERT: I think part of the challenge with sort of -- it's really hard to think of an actual example of where a measure is valid but reliability adds value.

So that's probably like the key thing, right? We need to think like an actual concrete case where a data element is valid but a data element is not reliable. That'll help clarify our thinking. And there's some confusion about

whether what Jack just said is actually a true example.

CO-CHAIR NERENZ: So to my ear, for what it's worth, Jack just gave us one. But I --

MEMBER GEPPERT: I agree with Patrick that what he described is actually not a valid data collection process. That actually wasn't the point I was going to raise.

CO-CHAIR NERENZ: Okay, go ahead.

MEMBER GEPPERT: It was more to the point of the pipeline question. So isn't it not true that eCQMs have sort of a different process where they don't have to demonstrate score level reliability and validity? That's not --

MS. JOHNSON: No, that's not quite right.

MEMBER GEPPERT: Okay. What's the criteria for eCQMs?

MS. JOHNSON: For eMeasures we actually require data element validation --

MEMBER GEPPERT: Right.

MS. JOHNSON: -- for eMeasures. And

we say that if you have enough data, we would love to see score level. So it's pretty much the same as any other measure. There's really no difference other than we are actually requiring data element validity.

We also say that for an eMeasure, you don't have to show us reliability right now, right, because by definition if the eMeasures work the way that we thought they would work, you're doing it the exact same way every time. And therefore by definition it'll be reproducible.

MEMBER GEPPERT: Which is kind of the example that Patrick gave about the CDC with their automated process. So if we were to change the rules where now we require both data element reliability and validity, and measure score reliability and validity, that would be a change for eCQMs.

MS. JOHNSON: It would be a change for eCQMs. And what I would say is let's not worry about eCQMs right now because we might have to

say we need to do something a little different maybe for eCQMs. Or maybe at some point -- what I'd love to get to is the ideal. And then we can play around at the edges with: are there circumstances where we'd be willing to accept something a little bit less? Okay?

So I'd really like to go back to Patrick's. Because if you get all four, I think -- does anybody disagree that you would love to see NQF require score-level reliability? Does anybody disagree with that?

(Off mic comment.)

MS. JOHNSON: Okay. So that's one of the four boxes -- right -- that we would love to see.

MEMBER GEPPERT: The only question I think for me on the table is whether it's required at the first endorsement opportunity.

MS. JOHNSON: Right. So we'll come back to that.

MEMBER GEPPERT: Okay.

MS. JOHNSON: Yeah, we'll come back to

that. So we would love to see score-level reliability. Nobody's fussing at that. Okay. How about score-level validations? So some kind of either criterion or construct, or something like that?

MEMBER AUSTIN: Face validity?

MS. JOHNSON: Let's just talk about if -- that would in sections, so let's put that on the table --

MEMBER SIMON: Yeah, that's one of my questions.

MS. JOHNSON: -- for a few minutes. Yeah, let's not fight about face validity right now.

MEMBER SIMON: Yeah. The concern I have there is that if you're developing a measure that doesn't necessarily correlate with something, something that's fairly novel, and you're trying to do a cutting-edge measure -- you're looking at care coordination, and you cannot validate the score -- you might be able to do a very nice analysis of face validity, but you

may not be able to get score-level validation. And that's something we've bumped up against with some measures. Looking at preventative care, for example, can be very tricky.

MS. JOHNSON: So what you're saying is it sounds great, but we know that sometimes there's a problem finding a reasonable, good enough comparator. Because we all fuss about the comparators that people come in and bring in. Right? And a lot of them do it because we say, hey, this is an X-Y-Z kind of measure. You have to do it. So here you are. I kind of know it's not great. But I'm willing -- this is what I can give you.

MS. WILBON: So my question would be: is construct validity the only kind of validity? I mean, I think known groups or some other types of validity came up as -- I mean, is --

MS. JOHNSON: We're using the term construct validity as an umbrella term.

MS. WILBON: Oh, as an umbrella. Oh, okay.

MS. JOHNSON: So don't worry about it's actually construct versus concurrent versus discriminate, versus known groups. It's all kind of validity. You're coming up --

PARTICIPANT: Non-face validity.

MS. JOHNSON: -- with some kind of relationship on --- okay. Yeah.

(Simultaneous speaking.)

MS. JOHNSON: Non-face. Other body part.

MS. JOHNSON: Non-face. Yeah.

MEMBER GLANCE: So I would challenge the idea of construct validity as being what we should be looking at. Because by definition, when you're doing construct validity, you're assuming that whatever score you come up at the score level, that it's valid if it demonstrates a reasonable level of agreement with some other measure that's capturing the same concept. Right?

So for example, if you were introducing a new mortality measure, one way to

demonstrate construct validity would be potentially to show that it agrees with some other mortality measure, or some other readmission measure, some other outcome measure.

And one could argue that you may not get very good agreement, but it doesn't mean that the measure is not valid. It just means that you're capturing a slightly different domain. So I would argue -- and I've argued -- you and I have had this discussion many, many times -- but for risk-adjusted outcome measures, really the way we should be looking at validity should be predictive validity.

And by predictive validity, I mean does the risk adjustment model work well. Because if you had -- in a perfect world with perfect risk-adjusted models, you could predict exactly what should be happening to a patient, what their predictive probability of death is, what the predictive probability of readmission is.

And then you'd know for a provider's

cohort, what is the level of agreement between the observed and the expected? So I would argue that predictive validity is what we should be talking about, not construct validity.

MS. JOHNSON: So can I stop you there just for a second?

MEMBER GLANCE: Sure.

MS. JOHNSON: And ask: can we shelf that one for a few minutes?

MEMBER GLANCE: Okay.

MS. JOHNSON: The reason I'd like to shelve it is half of our measures coming through right now are not risk-adjusted measures. All right?

MEMBER GLANCE: The complex measures though? The ones that we're looking --

MS. JOHNSON: The ones that you guys are seeing, but I'm talking about all of our measures. So we need something that we can say kind of across the board, not necessarily for just the complex measures that are coming here. So that's --

MEMBER GLANCE: Okay.

MS. JOHNSON: Yeah.

MEMBER GLANCE: Okay, got it.

CO-CHAIR NERENZ: Okay, we have a few other cards up and we probably have two or three different things on the table. It's so hard to keep it tight, and I know you're trying to work through a little grid. But let's go Sherrie and then Sean.

MEMBER KAPLAN: Well, for me, when you do psychometric testing, you're only validating the score. You're not -- the data sources you can check out, et cetera. But measures don't exist in the abstract. They exist in the application. So when you're testing them, it's how they're applied. So they're applied to someone and you get a score back. And then you do the psychometric testing on the score you get back.

And so those circumstances can change. The validation exercise there is really pretty simple one and it depends on the measure.

Because construct validity may be all you have. What I was going to ask you, Karen, to address on the NQF side is data collection costs money. And so if you can't get the money to do the empirical assessment of the score that you're trying to get the approval for, what processes does NQF have for different phases of development, which is what Sean was addressing?

If there is a tiering opportunity, and you absolutely can't get money to do the data collection without NQF endorsement, does that ever happen? And then can you consider kind of an opportunity in phased development, like the FDA has -- your Phase 1 measures have this kind of process?

CO-CHAIR NERENZ: Yeah, I was going to just think of that when Jack made the comment about the history of -- you know, first time through versus later time through -- that part of the trouble I think a lot of us have -- and I'm looking at Larry a little bit -- there's only one endorsement. It's either endorsed or not

endorsed.

And the first time through it gets endorsed and it goes out in the -- released in the wild has been our phrase. And all kind of crazy things happen. Who knows? This might be a more comfortable, low-key discussion if there was a provisional endorsement the first time, labeled explicitly that -- and then a second full endorsement, second or third or fourth time through, based on, I would say, touching all the four boxes. But there may a historical reason why that doesn't happen.

MS. JOHNSON: Yeah. Well, there's two things. I want to hand over to Lisa for something that we're hoping to institute. I will tell you that back in the day when we first had this first thing that we were talking about and how we would rate, the rating used to work like this.

If you did one level of testing -- it didn't matter which one -- then you were eligible for moderate. And then the idea was, as Jack

suggested, go out, get some more data, get some more money, do whatever it is you got to do. When you bring it back for maintenance, we would expect more testing at that point and hopefully at a different level, in which case you were eligible for a high. Okay?

So that was kind of the thinking 10 years ago. Right? And at some point maybe five years ago, we started saying, wait a minute, we do a lot of accountability things. Score becomes very important. So score-level testing, we changed the way the algorithm works.

(Telephonic interference.)

MS. JOHNSON: So the algorithm now works that you're eligible for a high if you didn't score, and only moderate if you did data element. Right? So it's trying to get the field to go in one direction or the other, possibly in the wrong direction because both of them tell you something different, and we'd actually like to know them all I think.

But it still -- is one more important

than the other? Do we need to expect some kind of different testing as we go along? I think that's one way we could think about it.

Historically, we haven't been able to do that. Developers will say, I had a contract to do this, this is what I was able to do, and that's -- I don't have another contract so I can't add to that. Right? We would have loved to have new testing every time they come back. Right? Maybe a different population and different something or other, to your point, but we very seldom get that. And it's a resource requirement usually. So then that takes us to another thought that we've had.

MS. MUNTHALI: Yeah, it -- this conversation is -- this is what we struggle with almost on a daily basis. It's really challenging because we want to make sure the field knows that our measures have stood the test of time, they've gone through a rigorous scientific review.

But we also are concerned about the measure development pipeline. And if we are

causing undue angst and obstacles to measure developers who may have resource constraints, that may not have the money to do the testing, and so our first thinking on this, we introduced for eQCMs, in particular, what we call a trial approval process, in which short of the testing, if they could meet our evidence standards, our feasibility standards, our use and usability standards, we would give them NQF approval to go out for three years to go out there and test.

The thinking was that that would promote uptake of use of the measure, and perhaps they would also then get the opportunity to have access to the data that they needed to test. This was about three or four years ago. I can tell you we probably have had one or two measures get fully endorsed. It's challenging still with the NQF approval.

So we realize that we probably need to provide developers access to multi-stakeholder committees, to some sort of NQF staff technical assistance, as we do for fully-specified measures

throughout the development pipeline. And so we're in the process now of trying to see if we can secure funding for this. It's a process in which we would give approval for developers at the ideation stage, so even at the evidence stage.

So we would bring together multi-stakeholder groups to see if this idea is strong in terms of its evidence. We would also do it for testing. And anything more than that would have to go through our full approval process. So I can't say more about it, but we've actually spec'd out this idea and developed it because we see the tension as these measures come forward to us, and the tension and challenges that developers have coming to us as well.

CO-CHAIR NERENZ: Sean, your tag has been up for quite a while. And then I've got Joe.

MEMBER O'BRIEN: Yeah. So definitely a lot of different perspectives and backgrounds. And I'm looking to change my mind on this, but Karen raised a good question of: do we all agree

that it's important to always require score-level validity testing?

In my case, I would say no. As a reviewer, the things I want to see are analyses that address the concerns that come to mind when I review the measure. And I'm really thinking about specific things. So in an outcome measure, it's almost always of interest to be sure of the differences you might observe not explained by case mix.

There's a measure yesterday looking at prostate cancer where they're measuring the intensity of symptoms, or presumably they're trying get at some type of underlying assessment of the intensity of symptoms by using three number claim days that had a diagnosis code, a particular diagnosis code, and you could say that's another example where your reporting the measure could vary from what you're actually measuring. And there probably is 10 other examples.

So when I want to see empirical

analyses, or when specific issues have come to mind -- and there's some data from the data at hand that could address that issue -- sometimes there's issues that come to mind, but the relevant data are not the data that the measure developer is collecting and analyzing; they come from other literature and other sources.

So I would just try to prioritize and make a plug for careful, critical thinking about where could this measure go astray from the difference between what it's measuring and wants to measure, and focusing on those issues.

As a developer, I would say when it comes to filling out the analyses to address that requirement for score-level validity testing, I'm often in the position of saying, what will get the job done and get the box checked here? I've done all the analyses that I think are important and are relevant. But now there's still some box to check off. And I think the box gets checked off by developers in different ways that don't always really get at anything too important.

So when you're measuring something, you cannot say, well, here's what we're literally measuring. But there's one step away from that, we're trying to get at something else. But even that is still trying to get at something else. So for example, if you're measuring something with claims, claims number, admissions for such-and-such diagnosis, or we're going to use claims data, then what about how did that measure if we had actual clinical data, and say, well, that's the thing we wanted to measure.

And even that, there could be related some other underlying measure of quality. So as a developer, all you can do is pick one step along the way and provide some analysis showing that step one is associated with step two. Check. You're done. But we don't know necessarily about how you get from step three to step four, in terms of more and more underlying extrapolations of more fundamental constructs of quality.

And then another example where you can

get the box checked is if you have a measure that's not directly measuring anything in particular, you can demonstrate that it's associated with some gold standard, then you can provide that analysis, demonstrate high correlation, and everyone -- all their review panels -- is happy.

But if your measure that you're putting forth is that particular gold standard, and you don't have anything else to compare it to, now you're stuck. So it seems to me a contradiction that you could use one measure and say it's the gold standard and serves to validate some other measure, but that alone is not going to stand by itself.

So I think there are a lot of paradoxes and just this requirement to always require -- it's tough for developers, and I'm not sure it always provides useful information. And so I would think about requiring less rather than more.

CO-CHAIR NERENZ: All right. Thank

you. Joe and Jack.

MEMBER KUNISCH: So speaking to the eCQMs specifically -- but this may apply to all quality measures -- when you're developing a new one and you put it out there, or put it up for endorsement, you did your testing and everything, so you have limited data. And so you want to make it kind of provisional and let them go out and collect more.

The challenge with that is healthcare organizations are inundated in the quality metrics that they collect and report under registry, joint commissions, CMS, and so forth. So there is a reluctance to say, hey, look at this new measure. Let's adopt it and put it in our environment and start actually reporting out. But we're only going to use it internally, because right now it's not required.

The other challenges were dependent on EHR vendors, especially on the eCQMs, to build these out for us. Now, if they don't do it and we decide to do it on our own, do we go by the

eCQM specifications, or kind of follow it because we now have to write our own custom queries to extract this data?

So if it does get adopted in CMS, now we have to go back and use this eCQM, official NQF version. So there's going to be a lot of challenges to say, okay, you've got three years to come back, go out and actually test this, because you've got to get people to actually adopt it.

And I think even manually abstracting measures or whatever, now you might get lucky and it's opioid crisis and we want to measure some of these things, so we adopt it because it has some value to us. But in general, I would just say it's going to be a significant barrier.

CO-CHAIR NERENZ: Jack?

MEMBER NEEDLEMAN: Yeah. I think as I process the conversation, I think all of the requests for the way we score puts a very sharp line between validation, is it valid, and is it reliable. And that climbs there. But I think a

lot of us have internal definitions of either the validity or reliability, to begin shading one into the other. And sometimes it matters, and sometimes it doesn't. I thinking of some of the measures we talked about today. So we had, for example, the CMS and Yale folks coming in with their opioid --

(Off mic comments.)

MEMBER NEEDLEMAN: Thank you. ED -- thank you -- overdose. Thank you. So the first question about validity is do they have the right codes? Have they captured all the codes? It's a code-based measure. Have they captured all the codes? They would capture an overdose. Have they got any codes in there that aren't overdose?

Well, that's a validity issue. It has nothing to do with the quality of whether the coding's right. That's a group -- but are the codes right? That's a validity measure.

And then you've got the issue of, okay, so now we've aggregated this up and we're calculating it by state or by county or whatever.

And if you told me that the question here is, what's the overdose experience of Medicare patients, on its face it's a valid measure because that's exactly what the codes -- the codes are collected from that population, and that's who it's being applied to.

If you then ask me is it valid as a population measure and as a whole, then you get the issue of, can this stand if you had CMS data only -- the Medicare Part A and B data only -- would you use it as a substitute for a measure of population rates of overdose, or is it surrogate? Or is it a proxy for that?

And that's a different validity question. Still about validity questions. And it gets answered by construct validity. In the case of the prostate stuff, we're being asked to assess as a matter of validity, does the number of claims days you had, is that really a measure of how dysfunctional you are? As either erectile dysfunction or incontinence, is that a measure of dysfunction? And I didn't see a lot of data to

discuss that. But that's the validity question.

When you get to the reliability question, now you get to the question, does the code -- the way the codes are implemented, the way people code into those codes -- do they overestimate? Do they underestimate? Is there systematic difference in that? And that's the reliability question. Somewhere in there the risk adjustment question comes. If you're making comparisons across different groups, is the comparison accurate? And that is sort of -- sometimes I think of that as validity, and sometimes I think of that as reliability.

But we see things shading one into the other between is it inherently -- are they measuring what they think they're measuring if the measurement is good? And is the data good, which is the reliability question.

CO-CHAIR NERENZ: Yeah. And we're going to go to Larry and Christie. And I'm just trying to think how we can respond to that or resolve that because it sort of starts back with

Sean's comment.

I mean, somehow as long as these words are going to be in the common vocabulary between us and the developers, I think we need to have definitions and boundaries in saying, when we say this word, that we mean this and that.

I'm not sure -- although somebody can suggest it -- how to proceed without that kind of labeling and categorization. I mean I appreciate the difficulty, particularly when we come from different disciplinary backgrounds, and we tend to use words in somewhat different ways.

I don't know how to end up in a workable fashion without just declaring, for purposes of the work we do and the developers we do, here's what it is. But I'm open to -- I just don't know where to take this line.

MEMBER NEEDLEMAN: One of the issues, for me, is whether -- and I find myself looking at the validity section before I look at the reliability section, because I understand what -- because the validity is about what the intent of

the measure is and how it's going to be -- how that intent is going to be realized in theory.

And the reliability is how good is it realized in practice, if I agree that they're measuring what they think they're measuring.

CO-CHAIR NERENZ: Okay, good. Larry, then Christie.

MEMBER GLANCE: Yeah. So just for a minute, I just wanted to bring back the discussion to the point where it started. And I think initially we were talking about whether or not it was enough for a data element to be valid. And I think we all kind of more or less agreed that we'd like to see both data element reliability and data element validity.

And then we kind of leapfrogged into saying, well -- and I sort of pushed this discussion -- I said I really think that we ought not to pass or endorse a measure based only on data reliability and data validity. And I think we all sort of agreed that we need to have score-level reliability and score-level validity.

And then I think -- then we leapfrogged to saying, okay, so then we should basically check all four boxes. So data reliability, data validity, score-level reliability, and score-level validity. And I'd like to respectfully push back on that. Okay? I think that -- and this goes back to the issue of pipeline in terms of how much we can expect from measure developers. I think that we should expect from measure developers to see information on score-level reliability and score-level validity. And the reality is, for the majority of the measures that we look at, that's what we see.

I think that when you're talking at looking at data quality, data reliability, data validity, that can be a pretty big ask. If you're telling a measure developer, a CMS contractor, that you've got to go out there and show that the administrative data used to construct that model, both for the risk factors and the outcome piece, that that -- if you want

to go out and show reproducibility, data reliability, that that may be kind of hard, not so much for the outcome in question -- although it may be hard for the outcome if it's a rare outcome -- but for a lot of the risk factors, they're pretty uncommon. So to be able to show reproducibility is hard.

Validity becomes even more complicated because for most of these measures -- the ones that are claim-based -- what you're doing is you have these mapping algorithms that take lots and lots of ICD codes and map them into diagnostic categories as risk adjusters. So my point being, without going into spending a lot of time on this, is that it is fairly resource-intensive, I think, for measure developers, to look at data quality, to look at data reliability and data validity.

So I don't know that that should be part of our -- and it's certainly not part of our initial ask. But I think that, again, going back to my initial point, is that for a measure

developer just to simply show that the data quality is high, and then that ought to be enough for that measure to be endorsed as being scientifically acceptable, that should not be acceptable.

And I guess if there was one outcome that I would like to see come out of this group today, is that we all agree that as a minimum, if we are going to look at scientific acceptability, we need to have a pass on both score-level reliability and score-level validity. And I was wondering if we could have a vote on that.

CO-CHAIR NERENZ: I think that's clear enough. Now, I just need to do a quick check. Christie, Sean, and Jen all had boards up. Are you seeking to speak to this point, or -- okay, let's go Christie -- I mean, that is a tangible proposal. It is not where I thought you were going to go with this, and it's going to take some discussion probably. But it's a tangible proposal. So Christie, why don't you say what you had, and then --

MEMBER TEIGLAND: Yeah. I'm basically agreeing with that. And I keep coming back to this measure yesterday that Jack just mentioned, which was the measure of claims days that had a visit for incontinence or erectile dysfunction. And then I totally believe that those claims, they probably had a visit. I mean, I don't really need to see data element reliability there. We could probably do that. But it's claims data. It is what it is.

I have a huge issue with the fact that that measure score could mean either thing. Right? It could mean that this person was too embarrassed to go to the doctor before the person got prostate cancer, had no claims for incontinence of erectile dysfunction. Then when they had prostate cancer, they had these conversations with their doctor. The doctor said, no, we can treat this. This is okay. It's okay. And then they have like 20 claims. And you're saying that's a bad thing? That's always a bad thing?

See, if you don't validate that score as measuring what it's supposed to measure, which is this person got this bad stuff happened after this prostate surgery, that is not a valid measure. We should not be judging people based on that measure. So -- and I wouldn't necessarily have to see testing, but I need to see some theory, I need to see some justification. I need to see something to tell me why. They didn't convince me that was a good measure.

MEMBER TEIGLAND: Right.

MEMBER PERLOFF: I would say on the flip side, I live and breathe claims data. And I can tell you every reliability call, it never is reliable in my book, because I know way, way too much. The arguments people make are either a waste of my time because they're wrong and I don't believe them, or they've gone to extreme lengths to prove what's already very well known. So that's not a good investment in my mind.

So I'm concurring from the opposite

side of the coin that the item-level reliability is nothing to me compared to the score level.

CO-CHAIR NERENZ: All right, so that sounds to me a little bit like Larry's, and we're running out of time in this time block anyway. But since Larry put a proposal out, let me just try to restate it and make sure the essence is correct.

MEMBER NUCCIO: I just want to --

CO-CHAIR NERENZ: Okay, Gene, go ahead.

MEMBER NUCCIO: Some counter -- point. I agree that we should look for score level reliability and validity. I disagree that we should ignore a data element level reliability and validity. And I'll give you an example.

CMS is the owner, ultimately, of many of our measures. In its instrument-based measures, like MDS and OASIS for post-acute care, they have a great reluctance to ever do any data element-based testing at the reliability or the validity level, to the point where they go

through multiple generations of these instruments.

And then as we had in our group, we had reliability data from the late '90s, okay, on an MDS instrument that has since gone through a couple generations. And so to make believe that the score from the most recent data are based on reliable and valid items is a little bit -- tense. And if CMS understands that they must pay for data element reliability and validity information, then, hey, add it to the pot of money, so.

CO-CHAIR NERENZ: All right, so that seems like there's a kind of a point, counterpoint here. Dave, you'll sort this out, right?

CO-CHAIR CELLA: Well, I might make it worse, because, you know, Jen, I don't live and breathe claims data; I'm more of an observer of claims data work. But I have begun talking to more -- more providers. And I hear often about how bad the current quality measures are. And,

you know, I, and they have all kinds of reasons.

But I actually think a lot of times it comes down to the fact that the garbage that goes in is -- what I'm hearing here from a few is we have to kind of ignore it. Because we all know it's bad, and if we pay too much attention to it, nothing will pass. And there'll be a revolution, and everything will go away, and NQF will dissolve, and we don't want that.

So I guess, you know, what I'm looking for is like what's the 20-year plan. I know we're not the Board of Directors here, but is there a 20-year plan for getting rid of bad measures and replacing with good measures?

And if so, I would be on the side of advocating for the quality of the data elements. Because I think a lot of times what people object to, and I'm talking more about the, you know, the clinicians that are judged by it, is that they know the data quality is not -- you know, you gave an example, Christie. But there are so many examples like that.

MEMBER TEIGLAND: Yeah, there are so many.

CO-CHAIR CELLA: It's not really measuring what it's trying to measure. And what I'm hearing here is that some of us are saying we really should ignore that because if we don't, if we pay too much attention to it, we're not going to like anything. So what's the 20-year plan?

MEMBER PERLOFF: First a quick, friendly amendment. It's not that I want to ignore it, but there's no one developer that can solve that problem alone. It is a large, systemic problem, and we need to take on these issues of quality in claims data.

But it shouldn't be put on any one development team alone. And that's why I want to take it off of them. Not because I want to give them a pass, that problem has to be resolved somewhere else.

CO-CHAIR CELLA: So can we fix the 20-year plan with small, with small improvements here and there? Or does it need a --

MEMBER PERLOFF: But they can't make small --

CO-CHAIR CELLA: -- wiping off of the table.

MEMBER PERLOFF: -- improvements in the whole CMS data processing system, and that's our challenge.

MEMBER ROMANO: So the one-year plan --

CO-CHAIR CELLA: I want to hear the 20-year plan first.

CO-CHAIR NERENZ: Okay, let's do that, and then we'll go to Patrick.

MS. MUNTHALI: So you're not going to hear the 20-year plan, but you're going to hear why the Scientific Methods Panel is so important.

And it was important for us to establish and bring methodologists together, because we recognize too, our measures, the measures that come through, you know, sometimes they're not where we'd like them to be. They're sometimes not in a place where policy has gone

ahead of the science. That's a challenge we often have sometimes. So we are looking to the Scientific Methods Panel to help inform what that plan is.

We're already starting to think about the value of measures that are out there, to think about measure burden. Trying to learn about measures. There are so many things outside of the methodology that we still don't know about with these measures. And part of it is trying to get the feedback from the field that we don't get in the rest of the evaluation of measures.

And so we do have an initiative that was spearheaded by the Board to start to look at that. So this is part of that, you know, progressive, why we don't have an answer right now. We're hoping that we can do this in collaboration with the Scientific Methods Panel and the CSAC.

CO-CHAIR NERENZ: All right, let's, we can probably spend a few more minutes, but we do have to watch our time, especially with people

having flights scheduled and whatnot. There are some update things we have to do.

Let's go Patrick, Sean, Sherrie, and then may just have to nip this off at that point with a -- because I -- with respect to Larry, I don't think your proposal is just going to sail through unopposed, so that means that we'll still have to keep talking about it.

MEMBER GLANCE: Why don't we have a straw vote?

CO-CHAIR NERENZ: No, we can do that; we can do that. Let's end with that, but I heard a couple of -- let's just see how it shakes out, let's see where it goes. Okay.

MEMBER ROMANO: I just want to be, I just want to agree very strongly with what Christie said, that -- but I also want to observe that we're collectively interpreting the idea of a score in two different ways. And so if we're going to say that we're requiring score level validity, we need to be clear about what's a score.

Because what Christie described is a score at the patient level, whereas others talk about a score at the level of the accountable entity.

CO-CHAIR NERENZ: Yes.

MEMBER ROMANO: Now I agree completely with Christie that the problem with that measure was that it wasn't clear that it was valid at the patient level. In other words, the patients that had this stepped increase in the number of claims actually had worse erectile dysfunction. So that's the fundamental validation question there.

But when we talk about reliability of score level, we're talking about reliability of the accountable entity score. In the case of validity, I don't care so much about that.

Because if this is a valid measure of a relevant patient-reported outcome or an outcome that matters to patients, and we have other evidence that this is an outcome that is partially preventable or can be improved, then why do we need to show that it has some construct

validity with something else? Because that becomes a trivial exercise that people show two variables are correlated with each other, but it doesn't tell us that either one is valid. They could both be invalid, and they're both correlated.

(Laughter.)

MEMBER ROMANO: So in this case, what I care about is what Christie said, that's validity. Now at some point down the line we'd like to see that provider organizations can improve these scores, known groups validity, other kinds of things. But I think I just worry about applying a single definition of score and saying that all the measures have to meet score level validity, when there's really different ideas of a score.

CO-CHAIR NERENZ: Well, let's just not try to set policy on one word that has two different meanings. When we say measure score level in this room, it means, as you just said, Patrick, it's at the level of the accountable

entity. I know Sherrie, all of us who come from psychometrics struggle with that, because if you give one patient a survey, that patient gets a score. But that is not the meaning of the word that is relevant in this discussion.

And you, Patrick, you said it. When we say measure score, we're talking about at the accountable entity. It's 50 patients, it's 100 patients, it's 10,000 plan members, it's whatever it is. It's an average typically. Okay, so let's just be careful about the word. Sean and then Sherrie.

MEMBER O'BRIEN: So ultimately it's the score level in the sense of measuring the entity of interest, but it's the score level reliability and validity you ultimately care about. And so I think that item level properties can be subsumed into the score level aspects.

And I think we just heard examples where item level issues could be an issue for score level properties. And if those issues for a specific measure were of interest and were

concerns, those should and could be addressed empirically.

And I could give examples of where, other examples of item level validity issues or reliability issues impacting score level validity or reliability. But I'll skip that and say I'm definitely ready to vote yes on Larry's proposal. Ultimately, I'd like to maybe make a placeholder for a possible further proposal.

In terms of what is the checklist that we look at when we're assessing measures, and right now it's we look at validity, possibly at two levels, and reliability at two levels. And is that the most useful for helping us to get information from us in terms of what's important to pay attention to and to provide feedback to measure developers?

From my standpoint as it's currently formed, validity and reliability, it's all kind of mushy and we might as well just get one big acceptable. And teasing them out, I don't know what to make of one versus the other that much.

Ultimately, and when I review journal submissions for -- in the peer review, I'm not categorizing my thinking -- between reliability and validity. But there are other types of checklists.

So some things that I think could be a different way of organizing criteria, and this is something I sent to NQF ten years ago. At the end of a meeting, we're really too late to bring it up. But I'll just say here was my list, I might not have time to go through it all, but here is my list of checklist.

Measure topic, does the measure topic have a direct and unambiguous relationship with quality? Two, inclusion/exclusion criteria, the proposed inclusion/exclusion criteria justify an appropriate data quality. Are the required data elements captured consistently, accurately, and reliably?

Did the data -- did clinicians faithfully capture that and rely on clinical variables? Case mix suggests -- applicable -- when applicable -- are they risk -- sufficient

variables measured with adequate detail or were there any important risk factors excluded. Were variables under the provider's influence appropriately excluded? Has the fit of the model been demonstrated empirically?

Sample size considerations, getting at reliability, analytic considerations, getting at missing data and other issues. Reporting issues in terms of whether uncertainty measures are presented in technical documentation in terms of whether it's been specified precisely and unambiguously.

So to me a more concrete, paying attention to specific concrete issues that could vary from measure to measure is more helpful for me. I think that reliability and validity fundamentally useful constructs, but they're just too broad to be useful for what we're trying to do.

CO-CHAIR NERENZ: Okay, thanks. I just at first glance hearing, I sort of like that, although it's kind of a radical overhaul of

what the developers get. But I -- maybe that could be sent around for further discussion.

All right, to close out, Larry, why don't you clearly state a -- oh, I'm sorry, Sherrie.

MEMBER KAPLAN: Just I wanted to react to Larry's proposal --

CO-CHAIR NERENZ: Okay.

MEMBER KAPLAN: Because I think one size fits none. I think there are circumstances in which it makes sense to have the data element and, the way you call it, and score reliability together, and you have to both. And you've been in, you know, around for a long time so you have empirical data, and you should put it out there for us to look at.

And there are circumstances in which it makes no sense at all. And by the way, Patrick, you don't get to be stupid about which construct allocation variable that you use. Yeah, you could correlate, you know, quality of life with the number of Kias on the street, you

know, but.

So the idea that you've got to have some kind of conceptual basis for doing what you're doing. And that rationale should be shared. If you left one of these out, you should have to tell this group why.

And then make a case for, you know, you got a certain -- certain kinds of measures need it. You can't just kind of do this checklist, tick-box thing. There have got to be categories of measures that should have certain requirements of them. And I think NQF is going to have to now be more nuanced about and more -- less kind of all measures have to have this.

So, Larry, I think I would vote against your proposition, only because I think there are circumstances in which you'd want both, and there are circumstances in which the evolution of the measure is so new, and we need measures in this area, but we'd be willing to accept less than the score level.

CO-CHAIR NERENZ: So, Larry, just so

we can do the straw vote, if you can just clearly state --

MEMBER GLANCE: Sure.

CO-CHAIR NERENZ: -- that you proposed we would require X and Y; we would not require P and Q and so --

MEMBER GLANCE: Sure. So currently, a measure developer can have a measure endorsed for scientific acceptability if they pass only the data reliability and data validity piece, okay. But hold on. So what I would -- I would put two parts to this, though.

The first one is that that is not enough, okay. Yes, or no, is it -- the question is is it enough for a measure to be endorsed as scientifically acceptable if it only passes on data reliability and data validity. That'd be the first vote.

The second vote, two separate votes, the second vote would be to replace that, but they're two votes, would be with the requirement that a measure can be deemed to be scientifically

acceptable if at the score level, as I have defined it, not as you've defined it, as I have defined it. If at the score level --

CO-CHAIR NERENZ: The group is asked to work with.

MEMBER GLANCE: Exactly.

CO-CHAIR NERENZ: The measure score at the accountable entity.

MEMBER GLANCE: Yes, exactly. It is both reliable and valid, okay. So those are the two questions on the table.

CO-CHAIR NERENZ: All right, I think we might have done the first one already, but why don't you -- so the first one is would people agree that passing data element reliability and validity only is not acceptable?

There's a funny void down at that end of the table. I take that to be -- okay, people don't disagree. We've got maybe three-quarters of the hands up, three-quarters, 80 percent. But not all.

All right, so the second, it seems to

essentially flip it. Is says now if you have measure score reliability and validity, you can pass. Who would support that? This is more looking like half, a little under. Okay, so as usual we --

(Laughter.)

PARTICIPANT: Consensus not.

(Laughter.)

CO-CHAIR NERENZ: No, but I -- we do need to close this out. I think that this has been useful discussion because we just need to keep to plugging away at this because we have concerns and objections about what goes on. We don't want to make the measure development pipeline so tough and so impossible that nothing comes through, but also I like this term release into the wild. You've got CMS particularly moving millions and millions of dollars around into here and away from there based on measures that may have some pretty weak properties. So it's balancing that that's the issue. We don't have the whole full view of the balancing, but we

play a role. So maybe to be continued. But I --

MEMBER O'BRIEN: Can I say something real quick: You mentioned something about the burden to developers. When I voted yes -- and I was listening to the wording that Larry put forth and I thought about asking this. I think measures should and could qualify for endorsement if they were deemed to be valid at the score level and reliable score level. I didn't say if they had empirical analyses demonstrating those. I think that a review assesses it and then convince a group of people to think about it carefully, that's good enough. And it might require the empirical analyses, but I don't think that's always going to be necessary for me.

MS. JOHNSON: Can I get one more just straw vote? What if we modified Larry's thing and said do you guys believe we should ask our CSAC to bless our changing requirements so that we require score-level reliability? I'm not saying anything about anything else, but right now we don't always require score-level

reliability. I just want to know if that -- if you -- right now I've got a thing and I've got a check mark there that I thought that's what I heard you say, but I don't know that -- I don't know --

(Simultaneous speaking.)

MEMBER ROMANO: Can I ask a question to clarify?

MS. JOHNSON: Yes.

MEMBER ROMANO: So in the first submission under this proposal --

MS. JOHNSON: Right.

MEMBER ROMANO: -- would it be adequate to simulate score-level reliability? In other words, knowing the distribution of how many patients would be eligible across accountable entities and how many accountable entities there are, developers could simulate what they would think that the score-level reliability would be before they've actually collected multi-site data.

MEMBER KAPLAN: What if you added

unless you can find a compelling reason for not --

MS. JOHNSON: We could add that.

CO-CHAIR NERENZ: So it's kind of where we have now with the social risk-adjustment that we essentially ask for it except if you can convince us you don't need to do it. Then you can go through. So it could be structured similar to that.

MEMBER NEEDLEMAN: I'm going to suggest that when this gets to the next committee which actually will set the policy, there will be a lot of pushback, some of which we've already heard in this room. So I think it's perfectly reasonable for the Scientific Methods Committee to be a little bit more aggressive in pushing for the highest level of reliability and validity that we think is useful for assessing measures and let the committee tell us, no, you're going to have to settle for a lower level because that's not practical and not consistent with the broad measure development process.

And in that spirit I'm going to probably lean to voting yes, but even though I've got real mixed feelings about it. Especially if there could be -- revive some concept of provisional endorsement or temporary endorsement, because that's the key step is getting the data from multiple sites to assess score-level reliability. If you can't simulate it, how are you going to do it? So there has to be some preliminary mechanism. All these medical schools are getting provisionally accredited.

(Laughter.)

MEMBER NEEDLEMAN: They're popping up all over the place.

(Laughter.)

CO-CHAIR NERENZ: Matt and then we're truly running out of time. We do have some updates to do.

MEMBER AUSTIN: I want to maybe just go on record as -- I think what Sherrie was offering I'd maybe want to echo that. I like the idea of maybe having the expectation that they

need to provide all four unless they can explain why that could not be done. All right? So in Patrick's example for some reason they didn't have the testing where they could do score-level reliability testing, then they would need to explain something. Maybe that they would expect to have that in the next numbers of years.

So I kind of like the idea of communicating that all four is what would be sort of appropriate, but then getting them sort of ways -- sort of what we do a little bit with the empirical validity testing for maintenance measures, right? There's the expectation that they'll do that unless they explain that they couldn't do that.

MS. JOHNSON: And the only thing that I'll give you on that, Matt, is the way all of our stuff is written we already have that expectation, right, and people don't do it for us. So what we're talking about is are we actually saying that if you don't come in, we don't even look at it, right? Or maybe with

Sherrie's thing.

So it sounds to me like score-level validity is still a little bit in the testing, might still be a little bit off. And we certainly haven't figured out the data element. Validity, we have to have it, whatever. Final score-level reliability. Would you be in favor of us trying to get that as a requirement regardless of whether it's new or maintenance of whatever? Again, CSAC can push back at some point.

MEMBER KAPLAN: If you added the unless the -- could you do the conditional thing? Unless you provide a really compelling case for not doing it?

MS. JOHNSON: Okay. We'll add that. Is everybody --

MEMBER HYDER: The concern is if it's a requirement, then it gives this committee less flexibility to do what they may think is reasonable, to Sean's point about his own checklist. If it's a requirement we lose the

flexibility there.

MS. JOHNSON: Potentially.

MEMBER FABIAN: Right. That was my clarification is if we had it in there now that we'd really like to see it, then you're not going to get dinged if you don't have it. That's different than we'd like to see it and if you don't have it, you need to have a compelling justification. And I would support that. I think that --

(Simultaneous speaking.)

MS. JOHNSON: Okay. Let's let Larry be the last because we really want to tell you about this really interesting project that we've got going on. Poor Nicolette is ready to tell you. So --

(Laughter.)

CO-CHAIR NERENZ: Well, we really have to move on into the updates. Unless, Larry, you know that what you say is going to bring this to closure, let's -- I mean, this is --

(Laughter.)

(Simultaneous speaking.)

PARTICIPANT: Do we get to vote? I thought we --

MEMBER GLANCE: Give me 30 seconds. I think we can vote for what Karen proposed. And I think it -- I mean honestly, if we've all decided as a group that it's not enough to have data reliability and data validity, then the next step has to be we have to add something. So that's what we're talking about doing right now.

CO-CHAIR NERENZ: Yes. Yes, okay.

MS. JOHNSON: All right. Show of hands?

MEMBER O'BRIEN: Did we vote yes? I know it was 50/50, but that initial question just about reliability, I mean before everyone in the room except a couple people raised their hand, so I hope that doesn't get lost, but there at one point was a motion yes, we think --

MS. JOHNSON: That's what I thought, yes.

MEMBER O'BRIEN: -- score-level

reliability is important.

MS. JOHNSON: Yes.

CO-CHAIR NERENZ: Yes. And now the same question, score-level reliability.

MS. JOHNSON: Yes, score-level reliability with -- it's an expectation and if you don't do it, we want to see a good explanation as to why.

MS. WILBON: Of -- no, requirement, not expectation.

MS. JOHNSON: No --

MS. WILBON: Requirement, right?

MS. JOHNSON: Requirement, but potentially an --

MS. WILBON: Yes.

MS. JOHNSON: -- out.

MS. WILBON: Okay.

MS. JOHNSON: Reliability.

MEMBER ROMANO: Or potentially that there could be a provisional endorsement.

MS. JOHNSON: We're not going to condition endorsement.

MEMBER ROMANO: Not going --

MS. JOHNSON: That's not on the table, no.

MEMBER GLANCE: But if you allow measure developers to say but we don't think it's important we solve for SES, 90 percent of the measure developers don't include SES. I think we should require score-level reliability. Let's take a vote on that.

MS. JOHNSON: Let's take a -- require score-level reliability. Let's put that on there as that's worst -- that's hard-earned, Sherrie. Let's see if anybody even goes there. Require score-level reliability. Show of hands.

(Voting.)

MS. JOHNSON: Okay. And then Sherrie would be -- you would have yes if we had put the expectation in -- thing in there?

MEMBER KAPLAN: I would, yes.

MS. JOHNSON: Okay.

MEMBER KAPLAN: But it would have to be a requirement for a really compelling case.

MS. JOHNSON: It would have to be really compelling.

MEMBER KAPLAN: It wouldn't be like, oh, well, we don't drop dead to the --

(Simultaneous speaking.)

MS. JOHNSON: I didn't have time?

MEMBER KAPLAN: Yes.

PARTICIPANT: So it's required unless it's not? Is that the --

(Laughter.)

MS. JOHNSON: And that's our conundrum.

MEMBER GLANCE: Karen?

MS. JOHNSON: Yes?

MEMBER GLANCE: Could we have one more question? Score-level validity?

MS. JOHNSON: Score-level validity. Before we were doing it with an and. Now we're doing it separate.

MEMBER GLANCE: But we've already agreed that we're going to do score-level reliability.

(Laughter.)

MEMBER GLANCE: This is like Brexit.

(Laughter.)

MEMBER GLANCE: One more vote.

MEMBER O'BRIEN: Larry, are you saying score-level reliability or score-level validity testing?

MEMBER GLANCE: Score-level validity testing, yes. Up or down?

(Laughter.)

MS. JOHNSON: All right. Score-level validity testing. Should we require it?

(Voting.)

MS. JOHNSON: Okay. It's about the same as it was before. Okay.

Score-level validity testing with an exception of Sherrie's wording?

(Voting.)

MS. JOHNSON: Got a little higher. Okay. But not unanimous. Okay.

MEMBER O'BRIEN: My score-level validity and my hand is up. It's just that I

think we should vote that -- I mean, as review we said yes, we're convinced it's valid.

MS. JOHNSON: Right.

MEMBER O'BRIEN: And the empirical analyses might be part of what contributed to that.

MS. JOHNSON: Got it. Okay. Thank you, guys.

Nicolette, do you feel like telling us about --

(Simultaneous speaking.)

MS. WILBON: Karen, is there enough time to actually make it worthwhile? I don't know that it's actually -- 10 minutes, by the time she gets into it and we have next steps and stuff at 2:30. It's --

MS. JOHNSON: You don't think there's 10 minutes with -- you can tell them about the project? You don't have to go into detail.

MS. MEHAS: Whatever -- I mean whatever you all prefer.

MS. WILBON: I don't know that it's

really --

MS. JOHNSON: You don't think it's worth it?

MS. WILBON: No.

MS. JOHNSON: Okay.

MS. WILBON: Honestly.

MS. JOHNSON: All right. We're doing some real neat stuff.

(Laughter.)

MS. JOHNSON: I will tell you that it's trying to go beyond thinking about the individual measures and thinking about how you group measures and how you use grouped measures and programs and how you might want to --

(Off mic comment.)

MS. JOHNSON: Yes. So it's really exciting, yes.

PARTICIPANT: Is that some of the score-level type stuff that sort of overlaps with some of this?

MS. JOHNSON: It totally overlaps. And I think some of the genesis of this was

something that happened -- I don't remember who did it, but we were talking about reliability at the score level and we were talking about how does that work if you're -- if it's being used in different ways and different programs, right? So you have this risk of misclassification and -- but what if you're using a threshold versus you're pulling based on say the top quartile, or something like that? And does reliability feel different depending on how penalties are applied, right? That was kind of the germ of the idea and it has gone since this. It's really kind of interesting stuff.

So we'll tell you about it -- maybe that will -- maybe we'll just tell you about it on one of our monthly calls. We might want to do that because I think you'd really enjoy hearing about it.

Okay. All right. What's next? Public comments?

CO-CHAIR NERENZ: No, we're -- there is no public comment.

MS. WILBON: Public comment and then just next steps. We have our monthly call in December.

CO-CHAIR NERENZ: While you're -- there was a request just for a quick update on the white papers.

Larry, the one that you led, just quick 10 seconds where it is?

MEMBER GLANCE: Sure. So our white paper has been accepted by Annals of Surgery. It is in press. It was reviewed by CMS and approved by CMS. Thank you, everybody, for all your help.

(Applause.)

CO-CHAIR NERENZ: The internal white paper was also completely written contribution, evaluator submit. We learned about the CMS review process. It is still to my knowledge hung up in there as it went in there early May. It may never emerge.

(Laughter.)

CO-CHAIR NERENZ: I'm serious. I have no idea. I have no idea what's going on.

MS. JOHNSON: It -- yes, so it is still in the clearance process. So the other day when Larry's came out, I got an email that said congratulations, it's out of clearance. So I expect that kind of email hopefully.

CO-CHAIR NERENZ: Like I say, there's a -- there is --

MEMBER GLANCE: David, I just made a little contribution to CMS and it's --

(Laughter.)

MEMBER GLANCE: Then I -- it was a little tweet accompanied by --

(Laughter.)

CO-CHAIR NERENZ: Be nice. No, there's a third white paper that Sherrie and I have been working on through the outline and first draft stage that I guess I'll just say is flat out dead in the water in my view because of the CMS review process, because we can go through, put a ton of work in this and it could go to CMS and die. And I will not commit to the work until I understand its process. It's just

-- makes no sense.

PARTICIPANT: What's the topic?

CO-CHAIR NERENZ: Reliability standards. And the problem is that in this paper we may take issue even in passing mention with certain CMS current policies, and if we do that, they'll probably kill it. And then we will have done all the work for nothing. So it's an utter impediment to writing these white papers. But that's where we are.

PARTICIPANT: Unless we're academics in which case we let them --

CO-CHAIR NERENZ: No, as individuals we can write whatever we want, but then it doesn't speak for the group.

MEMBER PERLOFF: But there's a backup plan that a group of folks in their academic worlds could write a paper that informs the field if -- as a worst case scenario --

MEMBER KAPLAN: But it doesn't carry the --

(Simultaneous speaking.)

CO-CHAIR NERENZ: No, any of us can do that, but the idea when this started now a year-and-a-half ago or so was that the group would speak in these papers. And in fact the authorship would be set up so the core writing group would be named, but then there would be a list at the end, essentially everybody. And in fact PubMed would pick everybody up as an author because it would speak for the group. And I think everybody liked that idea, but when we started down that path we had no idea that CMS review is part of the process.

MEMBER PERLOFF: But it could still be valuable to the field, just saying. Like thinking of John Adams' paper, we all see it --

(Simultaneous speaking.)

CO-CHAIR NERENZ: Oh, sure. And that's true. It's just a different --

MEMBER PERLOFF: Yes.

CO-CHAIR NERENZ: -- approach.

MEMBER PERLOFF: Right.

MS. JOHNSON: Okay. I don't think we

have to do the public commenting period, so let's just tell you about our next steps and let you guys go. We're -- well, it will take four minutes to do next steps, so we'll be right on time.

MS. WILBON: So next steps. Basically at this point you guys have done a great job over the last two days, so thank you.

The main thing that we'll just let you know is that we have our next kind of bimonthly call on December 12th, which is a Thursday. It's a -- it will be a two-hour call. We'll kind of sit down and look at our notes from this meeting and bring back some of the core methodological issues. Maybe just pick one to see if we can get through it and make some decisions.

I know that we are hoping to put forth potentially a set of recommendations for the criteria for CSAC in the spring, so we'll probably use the next set of calls with you guys to start working through what those recommendations would be so that we can have kind

of a cohesive set of kind of guidance or recommendations that we'd like to put forth and figure out kind of when that would be implemented if it's approved by CSAC.

So that's our next kind of hurdle to cross is to put forth that guidance so that we can kind of get things cemented in our criteria and in our guidance so that developers can start to adjust and start submitting things that is more in alignment with what you guys would like to see.

MS. JOHNSON: And we also have a few other next steps just in case you're curious. Starting next week -- no, the end of this week, somewhere in there, developers will be sending to NQF the remainders of their submission forms. So this is not things that you see, but if you wondered where's the evidence form, where's the feasibility stuff, they're sending that in the next couple weeks. And that's stuff that will be going to the standing committees.

We've already been summarizing the

information and your discussions. And this is stuff that will go to the standing committees. So it won't be exactly a cut and paste of the discussion guide, but it will look very similar. And that's the kind of stuff we will be telling the standing committees.

The measure evaluations by the standing committees, somewhere in January-February, depending on the topic, CSAC in June. So it's still kind of a long process. So things that you've done today and yesterday we won't know the final disposition until a few months down the road, but we will try to come back and let you know so that we close that loop for you.

And then again, the next intent to submit deadline, which means the next bolus of specifications and testing attachments, due January 5th. So that means that we'll be probably sending assignments to you guys for the next round probably early February. I think that's what our time frame is.

So with that I hope you have a great

holiday.

(Laughter.)

MS. JOHNSON: Enjoy not doing any evaluations and we'll catch you after the beginning of the year. Thank you guys so much for coming. I so much enjoy these discussions and I know it's a lot of effort on your part, and we really appreciate it. Thank you.

CO-CHAIR NERENZ: Karen, your boss has left, but please communicate just how appreciative the committee is on the quality of the staff work that has been done.

MS. WILBON: And the --

CO-CHAIR NERENZ: I feel incredibly well-supported in this process.

MS. WILBON: And the two Chairs.

MS. JOHNSON: And our Chairs.

(Applause.)

CO-CHAIR NERENZ: Thanks everyone. Safe travels.

(Whereupon, the above-entitled matter went off the record at 2:30 p.m.)

U

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

Before: NQF

Date: 10-29-19

Place: Washington, DC

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