NATIONAL QUALITY FORUM

+ + + + +

SCIENTIFIC METHODS PANEL
SPRING 2021 MEASURE EVALUATION MEETING

+ + + + +

WEDNESDAY
MARCH 31, 2021

+ + + + +

The Panel met via Videoconference, at 11:00 a.m. EDT, Christie Teigland and David Nerenz, Co-Chairs, presiding.

PRESENT:
CHRISTIE TEIGLAND, PhD, Co-Chair
DAVID NERENZ, PhD, Co-Chair
J. MATT AUSTIN, PhD, Armstrong Institute for Patient Safety and Quality, Johns Hopkins Medicine
BIJAN BORAH, MSc, PhD, Mayo Clinic
JOHN BOTT, MBA, MSSW, Consumer Reports
DANIEL DEUTSCHER, 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, PhD, Mayo Clinic
JOSEPH KUNISCH, PhD, RN-BC, CPHQ, Memorial Hermann Health System
ZHENQIU LIN, PhD, Yale-New Haven Hospital
JACK NEEDLEMAN, PhD, University of California Los Angeles
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, FACP, FAAP, University of California Davis
SAM SIMON, PhD, Mathematica Policy Research
ALEX SOX-HARRIS, PhD, MS, Department of Surgery, Stanford University
RONALD WALTERS, MD, MBA, MHA, MS, University of Texas MD Anderson Cancer Center
TERRI WARHOLAK, PhD, RPh, University of Arizona, College of Pharmacy
ERIC WEINHANDL, PhD, MS, Fresenius Medical Care North America
SUSAN WHITE, PhD, RHIA, CHDA, The James Cancer Hospital at The Ohio State University Wexner Medical Center

NQF STAFF:
MIKE DiVECCHIA, Senior Project Manager, Quality Measurement
CAITLIN FLOUTON, MS, Senior Analyst
HANNAH INGBER, MPH, Senior Analyst
CHELSEA LYNCH, MPH, Director, Quality Measurement
SAI MA, PhD, Managing Director/Senior Technical Expert
MATTHEW PICKERING, PharmD, Senior Director, Quality Measurement
CHRIS QUERAM, Interim President and CEO
SAM STOLPE, PharmD, MPH, Senior Director, Quality Management
SHERI WINSPER, Senior Vice President, Quality Measurement

ALSO PRESENT:
JUDY BURLESON, MHSA, American College of Radiology
KAREN CAMPOS, CHES, American College of Radiology
DON CASEY, MD, MPH, MBA, FACP, FAHA, FAAPL, DFACMQ, President, American College of Medical Quality
DUSTIN GRESS, American College of Radiology
SRI NAGAVARAPU, Acumen, LLC
DAVID NEWMAN-TOKER, MD, PhD, Johns Hopkins Medicine
JONATHAN SEGAL, MD, University of Michigan Kidney Epidemiology and Cost Center

C-O-N-T-E-N-T-S

Welcome and Review of Meeting Objectives 6

Measure Evaluation
Subgroup 2:
Neurology

#3614 Hospitalization After Release with Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke) (Armstrong Institute for Patient Safety and Quality at Johns Hopkins University) 14

Subgroup 3:
Patient Safety

#3621 Composite Weighted Average for 3 CT Exam Types: Overall Percent of CT Exams for which Dose Length Product is at or Below the Size-Specific Diagnostic Reference Level (for CT Abdomen-Pelvis with Contrast/Single Phase Scan, CT Chest Without Contrast/Single Phase Scan and CT Head/Brain Without Contrast/Single Phase Scan) (American College of Radiology) 61

#0500 Severe Sepsis and Septic Shock: Management Bundle (Henry Ford Hospital) ..... n/a

#0674 Percent of Residents Experiencing One or More Falls with Major Injury (Long Stay) (Acumen/CMS) 92

#0679 Percent of High Risk Residents with Pressure Ulcers (Long Stay) (Acumen/CMS) 94

Subgroup 1:
Renal

#3615 Unsafe Opioid Prescriptions at the Prescriber Group Level (University of Michigan Kidney Epidemiology and Cost Center (UM-KECC)) 117

#3616 Unsafe Opioid Prescriptions at the Dialysis Practitioner Group Level (UM-KECC) 117

Opportunity for Public Comment 154

Next Steps 157

Adjourn 162

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

11:02 a.m.

DR. MA: Good morning everyone. Welcome back to the NQF SMP evaluation meeting for the spring cycle 2021. We had a very robust discussion yesterday. It was a really long day, so I want to thank, again, to our SMP members for your great effort, your time, and great discussion yesterday, and look forward to another robust discussion today. Can we move on to the next slide?

Let's give one second. I see one of the SMP members is still in the waiting room.

A quick reminder, when Hannah calls you for attendance, please again provide a quick introduction of yourself and announce your disclosure for interest for the measures being discussed today. Hannah, do you want to take over now?

MS. INGBER: Yes, thank you. Dave Nerenz.

CHAIR NERENZ: Good morning everybody, and thanks for all the diligent work yesterday and upcoming today. Very rich discussion. I appreciate all the contributions.

I only have a conflict with measure 0500, that actually has been pulled for discussion. So, there'll be no conflicts if that's not discussed.

MS. INGBER: Thank you. Christie Teigland.

CHAIR TEIGLAND: Hi. Good morning everyone. Day two, great day yesterday. So much rich discussion, and we lengthened our list of topics to discuss at our next SMP meeting. But look forward to today and I don't have any conflicts today.

MS. INGBER: Thank you. Matt Austin.

MEMBER AUSTIN: Yeah, good morning to everyone. My only conflict is with measure 3614, which will be the first measure discussed today. I will putting on my measure developer hat for

that.

MS. INGBER: Thanks. Bijan Borah. John Bott.

MEMBER BOTT: Yeah, hi. As noted yesterday, I was on a CMS TEP for 3501e, but I'm not on the subgroup that reviewed that measure, so otherwise, that's all I got. Thanks.

MS. INGBER: Thanks. Daniel Deutscher.

MEMBER DEUTSCHER: Hello, this is Daniel. Sorry for not being able to join you yesterday. But I'll be here today and I have no conflicts or disclosures.

MS. INGBER: Thank you. Lacy Fabian.

MEMBER FABIAN: Good morning. I'm here. No additional disclosures for my subgroup. Thank you.

MS. INGBER: Thank you. Marybeth Farquhar.

MEMBER FARQUHAR: Good morning. I have no disclosures, and I'm here.

MS. INGBER: Thank you. Jeff Geppert.

MEMBER GEPPERT: Good morning. Nothing further to disclose today.

MS. INGBER: Larry Glance.

MEMBER GLANCE: Good morning. I don't have any disclosures. Thank you.

MS. INGBER: Joe Hyder.

MEMBER HYDER: Good morning. I don't have any disclosures for today.

MS. INGBER: Thank you. Sherrie Kaplan. Joe Kunisch.

MEMBER KUNISCH: Good morning. I have no disclosures.

MS. INGBER: Thank you. Paul Kurlansky. Zhenqiu Lin.

MEMBER LIN: No disclosure for today.

MS. INGBER: Thank you. Jack Needleman.

MEMBER NEEDLEMAN: No disclosures for today, but I do have to be off the call at 1:00

Eastern, 10:00 Pacific.

MS. INGBER: Okay. Gene Nuccio.

MEMBER NUCCIO: Good morning. No disclosures.

MS. INGBER: Sean O'Brien.

MEMBER O'BRIEN: Good morning. No disclosures.

MS. INGBER: Jen Perloff.

MEMBER PERLOFF: Hi. Good morning. No disclosures.

MS. INGBER: Patrick Romano.

MEMBER ROMANO: Here. No new disclosures.

MS. INGBER: Sam Simon.

MEMBER SIMON: Good morning. My only disclosure this morning is measure 0500.

MS. INGBER: Thank you. Alex Sox-Harris.

MEMBER SOX-HARRIS: Good morning. No disclosures.

MS. INGBER: Ron Walters.

MEMBER WALTERS: Present. No disclosures.

MS. INGBER: Thank you. Terri Warholak.

MEMBER WARHOLAK: Good morning. No disclosures.

MS. INGBER: Eric Weinhandl.

MEMBER WEINHANDL: Good morning. No disclosures.

MS. INGBER: And Susan White.

MEMBER WHITE: Morning everybody. No disclosures today. Thank you.

MS. INGBER: Thank you. Was there anyone who's joined the call who hasn't announced themselves yet?

DR. NEWMAN-TOKER: This is David Newman-Toker, one of the measure developers on at 11:15.

MS. INGBER: Thank you. All right, good morning everyone. And I'll hand it back to Sai.

DR. MA: Thank you, Hannah. Good morning everybody. We can move on to the next slide, please.

Quick overview of today's agenda. It's a little bit less taxing than yesterday. We have slated three measures for discussion, but one of the measures, after reading the developers' comprehensive response, the SMP member who decided to pull that measure for discussion decided that discussion is no longer needed.

The measure passed both validity and reliability through preliminary analysis. So, we're going to skip 0500. That means we're going to hopefully have a longer lunch break and we will resume at 1:30 p.m. Eastern time.

We have another four measures slated for discussion in the afternoon. All those four measures passed both validity and reliability through preliminary analysis.

However, according to our policy, any

measure can be pulled for discussion for an overarching topic. So they will be discussed in the afternoon, but a re-vote is not necessarily needed unless after the discussion some members feel strongly a re-vote is warranted. Then we can go ahead with re-vote.

And speaking of re-voting, I just want to mention that Hannah just sent everyone another email this morning with the voting link. So, if you don't have the link handy, please let Hannah know.

You can use the chat function, chat privately to her, or you can email us. But you should have the link handy to you in the meeting, if it's needed. We're going to use the link to do the voting.

At 2:30, we're going to open up for public comments. And then, Hannah is going to provide updates on the next steps for upcoming meetings. Then we will wrap up for the day. Before I move on to the particular measure for

discussion, is there any question?

I think one SMP member just joined, Bijan. Do you want to disclose any conflicts for today? I'm not sure if you're trying to talk. We can't hear you. Bijan?

Okay, we'll message you privately. All right, if there are no questions, we're going to move on to the next slide. Okay, go ahead.

All right, we're going to start today's discussion with 3614. At this time, I want to invite our director, Chelsea Lynch, to provide an introduction of this measure.

MS. LYNCH: Thank you, Sai. As Sai said, the measure is NQF-3614, Hospitalization After Release with Missed Dizzy Stroke. The measure developer is the Armstrong Institute for Patient Safety and Quality at Johns Hopkins University.

This is a new outcome measure that tracks the rate of patients admitted to the hospital for a stroke within 30 days of being

treated and released from the ED, with either a non-specific presumed benign symptom-only dizziness diagnosis, or specific inner-ear vestibular diagnosis, collectively referred to as benign dizziness.

The measure accounts for the epidemiologic base rate of stroke in the population under study, using a risk-difference approach. The data source is claims and the analysis occurs at the facility level.

This measure was previously reviewed by the SMP two years ago under a different NQF number, and that history was shared with SMP members during this review cycle.

For this cycle, this measure passed reliability with a moderate reading, but was consensus not reached for validity.

I'm going to hand it over to Sam and Susan to lead the discussion on the issues raised regarding validity for this measure.

MEMBER SIMON: Hi everyone. So,

yeah, I can start it off. So, a few things worth noting here. First of all, the developer relied on data-element validity.

And to test the numerator codes for primary diagnosis of stroke on admission, interestingly, the developer relied on several studies that indicated the validity of ICD-9 strokes -- ICD-9 coding for stroke as a primary diagnosis for admission.

And the developers identified several international studies that supported the validity of I-10 stroke diagnosis in one U.S.-based paper.

So, there wasn't empirical analysis, per se, around the validity of the numerator data element. But this is okay. Or at least, NQF current guidance does permit use of prior validity studies for data-element validity.

This might be something we want to look into later, but it does meet the current guidance that NQF has around data-element validity.

For the denominator diagnoses, the developer looked at the PPV and NPV for discharge diagnoses. And they found very high rates for true positives and true negatives. So, I found that to be interesting.

Another validity issue that came up here and was noted by, I think, a couple of folks in our group, was the significant skew in the performance scores where -- and particularly looking at differences among hospitals.

So, among, I believe, around 900 hospitals, 65 percent of them were ranked as better than the national average in the sample that was used, which does raise some concerns around the validity of the measure score, and in particular, the ability of this measure to discriminate among hospitals, or among EDs.

So, the developer reasoned that more hospitals would be identified as poor performing with more complete data. The current data only has 20 percent of ED discharges in the testing

data.

So while that's true, I do think there's still this overarching concern of limited discrimination within performance.

And then the final issue, which might be the most interesting one that the group kind of picked up on, was the risk-adjustment approach that was just described, which compares the observed rate, which uses the -- I'm sorry. Yeah, the observed rate is the sort of zero from 30 days from ED discharge, and compares that observed rate of hospitalization to the expected stroke rates, which uses the same population's 90- to 360-day post discharge rates, rather than using a more patient-specific clinical or social variables to predict the risk of stroke.

And the developer presents the argument that this approach using the observed and expected rates, because it's based on the same set of patients, there's no need to adjust for sort of more commonly known risk factors,

since they don't change over time in this given population.

So those are some of the, I think, points that are worth considering in this re-vote of validity. Susan, I don't know if you had anything else you wanted to add.

MEMBER WHITE: Thanks, Sam. I just wanted to add one point, and that is around the meaningful difference. And this measure pools three years of data.

And I think it makes it very challenging for a provider to show any difference and to really move the needle. So, even if we were given the skew, and if the skew were to go away, which I'm not sure I agree with, with complete data, I think still having the three-year span is an issue.

And I know why we do that. It sort of has to do with saying the lower bound for the number of cases. But if the measure isn't going to really show and be able to allow providers to

measure improvement, I question the meaningful difference sort of criteria. That's just one other point to add. Thanks.

DR. MA: Thank you Sam and Susan. At this time, I think we can invite the developer to provide a response before we open up the discussion to everyone on the panel.

DR. NEWMAN-TOKER: Thanks everyone, for your willing to review our measure. We really appreciate the opportunity to be here.

Is there a particular list of things that you -- did you want us to go through sort of one by one the things that you listed? Or would you prefer sort of a general addressing of the concerns?

DR. MA: I think you don't have to repeat the written response. The SMP members should have already read your comprehensive responses. So, I think just to respond to Sam and Susan's comments. And we can also help you by showing the figures you provided in the

response, if you think that's going to be helpful.

DR. NEWMAN-TOKER: Sure. I think maybe I'll just sort of quickly deal with the numerator/denominator data-element validity issues.

It's very clear that ICD-10, which was implemented in 2015, is only now coming online in terms of studies that are starting to proliferate about the accuracy of ICD-10 codes.

This has been discussed in the NQF neurology measures full committee on a couple of occasions. And in general, people are not concerned about the ability to identify stroke.

There are all sorts of still residual questions about the issue of whether stroke subtypes of various and sundry sorts can be accurately coded, or non-accurately coded.

But the issue of identifying acute strokes remains the same. And we've actually done, in the process of sort of doing our own

homework since the submission, have done a comparison of different stroke subsets of codes, and found all of the expected associations in the Medicare data, which includes not only that the association's tighter for ischemic stroke than for any stroke -- which is what one would expect, since what we're mostly expecting to be missing in dizzy patients is ischemic strokes -- but even one run beyond that in ICD-10 data, for the small subset of patients who are coded as having either strokes in the back part of the brain or the front part of the brain -- which is supposed to be done for everybody but isn't -- but for that subset, we find the strokes in the readmitted group after the treatment release ED discharges are twice as likely to be poster circulation stroke patients, while the base rate in the population is that anterior circulation or the front of the brain strokes are five times as common.

So, there's a ten-fold reversal, which is exactly what we would expect because dizziness

is caused by strokes in the back part of the brain.

And so, some of those patients have strokes coming from their heart and their subsequent stroke might be in the front part of the brain, but the vast majority of them have disease in the back part of the brain.

So, there's a lot of internal coherence and consistency within the data that indicates that all those ICD-10 codes are valid and reasonable proxies for what it is that we're measuring.

MEMBER WHITE: David, could I interrupt you just for a second? This is Susan. While we're on coding -- I don't want to get too far away from coding before I ask this question.

DR. NEWMAN-TOKER: Sure.

MEMBER WHITE: So, I think we also need a piece around dizziness coding. Right? So, we need to have the patient --

DR. NEWMAN-TOKER: Yeah. Can I get a

clarification on what the concern is there?

MEMBER WHITE: Yeah. It's really an ICD-9, ICD-10 concern. So, there's a ton of -- I wouldn't disagree with you that --

DR. NEWMAN-TOKER: But we did both ICD-9 and 10.

MEMBER WHITE: You did. But you did not do a validation of the coding of dizziness for 10, I don't think.

DR. NEWMAN-TOKER: We did.

MEMBER WHITE: Oh. Okay.

MEMBER AUSTIN: This is Matt. Let me clarify. So, for the numerator for the stroke, as Sam mentioned, we did rely on studies for the validation of ICD-9 and ICD-10.

For the denominator, which looks at patients who were discharged with the diagnosis of dizziness, we did do our own validation of those codes, to David's point, in both ICD-9 and ICD-10.

MEMBER WHITE: And the four providers

that was just --

MEMBER AUSTIN: And the four providers. Correct.

MEMBER WHITE: Okay.

MEMBER AUSTIN: Yeah.

MEMBER WHITE: Yeah. So, I think there's -- I'm okay with four providers if it's industry, literature-supported. I think four providers is a little light for the denominator specification. So, thank you. That's my question.

DR. NEWMAN-TOKER: Could I just briefly comment on that? I know we responded to this in our written replies, but there are not many options for coding about dizziness in either ICD-9 or ICD-10.

And basically, this coding is unbelievably consistent. And the results that we're showing have been seen everywhere. They've been in other countries, they've been seen in multiple dataset analyses in the U.S., they've

been seen in OSHPD data, they've been seen by analyses done by many other groups, not just ours.

And our accuracy, in terms of data-element validity of the coding of dizziness, is well over 99 percent. It's 99.9 percent and plus in most cases.

So, I just don't see that as being structurally likely to be different, just because we only did the chart-level analysis at four institutions. There's so many reproduced data using these kinds of codes that show the exact same results, that it's hard for me to believe that there's a lot of variation, particularly given the numbers that we found.

MEMBER WHITE: Yeah, if I could just ask a follow-up, David. So, when you say there's so many studies over and over, you mean in general, using administrative data?

DR. NEWMAN-TOKER: No. I mean specifically using dizziness discharges and

stroke returns.

MEMBER WHITE: Okay. So, I would say that the rules around the coding and whether dizziness is captured or not, is going to be highly variable among providers. I won't go into the minutiae of coding rules, but I think there's going to be misses in the denominator because of that, because there may be a lot going on with the patient and dizziness may not be the most important --

DR. NEWMAN-TOKER: Oh, we've studied that very closely. So, you're absolutely right. If you do a structured, as we have, over 300 consecutive dizzy patients coming to the emergency department, and you look at whether they're dizzy or not, through systematic inquiry.

The percentage of the emergency department patients that are dizzy is about four percent, if you -- where the patient says it's an important part of the reason why they're in the emergency department.

If you look at administrative data, it's about two percent of emergency department visits. But that's really not the question. So, that's dizzy in. That's on the way in.

Then, when you talk about the issue of people going out, when people are labeled with benign inner-ear diseases, or dizziness not otherwise specified, there aren't many patients.

We looked at the negative predictive value of whether a patient has dizziness not on the way out, and should have been dizziness on the way out, and it's 99.99 percent.

MEMBER AUSTIN: The other thing I'll add real quickly is, just to clarify, is we are looking at a primary discharge of benign dizziness, right?

So, we're not looking at secondary diagnoses. We're only looking at primary diagnoses.

DR. NEWMAN-TOKER: And I just want to be clear also here, that there is no intent that

this measure necessarily captures every single patient that has a missed stroke, or missed dizziness and a stroke.

It's a barometer. It's a needle. Right? It's an operational viable way of ascertaining whether we're missing strokes in this patient population that is 14 times more likely to be mis-diagnosed with stroke than other populations.

A systematic review we did in 2017 showed that the odds ratio for dizziness and vertigo is 14-fold above something like motor symptoms, where we miss about four percent of motor symptoms and we miss about 40 percent of strokes when they present with dizziness.

And that's simply because -- with the adjusted odds ratio being 14. And that is simply because it's hard. I mean, this is a tough thing to do. There's a lot of sub-specialty expertise and knowledge that hasn't fully disseminated into front-line clinical practice.

And this is one opportunity for us to start moving the needle on diagnostic error, which is something that no one has done. Thus far, there are no shining star examples of quality improvement in this domain.

We need that sort of CLABSI, big-win, made-a-difference, showed the difference. But we can't do that if we don't have a measure that we're looking it.

DR. MA: Thank you, Dave. Before you answer questions around the risk adjustment and the meaningful variation in performance, I see Jack, you had your hand up. Are you having a question about data element?

MEMBER NEEDLEMAN: Yes. I was following David. He had sold me right up until the end there. And I'm not unsold, but I now need more information.

Four percent of the patients coming into the ER, ED, have dizziness as one of the symptoms that have brought them there. Two

percent going out come with a primary diagnosis of dizziness. Is that what I --

DR. NEWMAN-TOKER: No, sorry. Maybe I misspoke, or I spoke too quickly. Let me clarify.

There are about five million emergency department dizziness visits a year in the United States.

MEMBER NEEDLEMAN: Okay.

DR. NEWMAN-TOKER: That's roughly, give or take, two to three percent. Two-and-a-half percent of the emergency department population. Those numbers are based upon CDC's NAMCS data analysis from the most recent years.

Among that -- we'll call it the chief complaint group, or what we refer to, just for ease of shorthand, the dizzy-in group, the people coming in --

MEMBER NEEDLEMAN: I got that.

DR. NEWMAN-TOKER: -- with dizziness. Those patients, only about three to five percent

of them have strokes. So, it's an uncommon subcomponent of that large five million patient population.

Among the strokes, we miss about 40 percent of those overall. Some of those misses are in patients that are told they have benign dizziness and sent home. Most of them we believe are in that subgroup, although the exact details of what percentage of them are in that subgroup, as opposed to called something else, is not 100 percent known.

So, we have focused on the people who were sent out as dizzy, not otherwise specified --

MEMBER NEEDLEMAN: Right.

DR. NEWMAN-TOKER: -- or benign inner-ear disease, to identify how often are they having strokes, and to look at that early rate of return.

Did we send in curves with the response, or one of the documents? Yeah, I think

so. And where it says figure 2A and 2B --

MEMBER NEEDLEMAN: Okay, so --

DR. NEWMAN-TOKER: -- if you could show those by any chance?

MEMBER NEEDLEMAN: Right. So, the risk of misdiagnosis -- of missing the stroke among those folks -- is substantial. But the other question is, those folks who -- I'm trying to figure out how to phrase this. Is it clear that people don't get misdiagnosed into other categories? Not benign dizziness --

DR. NEWMAN-TOKER: Well, we know that there are --

MEMBER NEEDLEMAN: -- not dizziness not otherwise specified or benign inner-ear disorder, but wind up in some other misdiagnosed category? And is that going to vary across EDs, so that the denominator here is a matter of local coding processes in a way that we don't get the same population misdiagnosed across different EDs?

There are other places people could get misdiagnosed into, given the symptoms that they presented, and the fact that they have a stroke.

DR. NEWMAN-TOKER: Well, let me just make sure first that I've understood your question. Because it is certainly the case that people with other symptoms can be misdiagnosed. So, the second one is --

MEMBER NEEDLEMAN: Yeah, I'm talking about the folks who are coming in. Yeah.

DR. NEWMAN-TOKER: Who are dizzy.

MEMBER NEEDLEMAN: Yeah.

DR. NEWMAN-TOKER: The dizzy ones. What are the chances that they're coded as something else?

So, there are, for instance, a couple of case reports of people who were told that they had gastroenteritis, because they had nausea and vomiting.

And they were dizzy and they had

nausea and vomiting, and the dizziness was attributed to low blood pressure from loss of fluids, whatever else. And so, they were diagnosed as something other than benign dizziness, and sent home.

That does happen. But there isn't any kind of systematic bias and coding that anyone is aware of, or there's any literature to support the notion that there's a systematic bias towards one hospital coding all of the people who have dizziness as gastroenteritis, and another where it's coded as inner-ear disease.

We haven't seen anything remotely like that.

MEMBER NEEDLEMAN: Okay. So, what you're saying is, by and large, this is a coherent population, it's going to be relatively consistent across different EDs, and among those who get discharged with benign dizziness, non-specified, or benign inner-ear, a fair number are going to be misdiagnosed. There are going to be

missed strokes.

You're saying the denominator is reasonably clear and the misses are going to vary across different EDs. Because that's the assertion about this measure.

DR. NEWMAN-TOKER: Yes, is the short answer to your question.

MEMBER NEEDLEMAN: Okay.

DR. MA: Great. Thank you, David, for your comprehensive responses. I think Sam had two comments about the risk adjustment model and the performance.

DR. NEWMAN-TOKER: Yes. So, just on the issue of the risk-adjustment methodology. We've struggled a little bit. We've gone back and forth with the staff at NQF about whether to call this risk-adjustment or not call it risk-adjustment.

I think some of the complaints and problems we've gotten into has just been a terminological issue of whether this counts or

doesn't count as risk-adjustment because it's not the standard kind of risk-adjustment.

And so, some people say, well, it's not risk-adjustment at all. And other people have said, it's a clever kind of risk-adjustment for this particular measure.

And so, if we say we didn't risk-adjust, then the people who think it is risk-adjustment are confused. And if we say it's not -- so there's a little bit of a terminological problem there.

But setting that issue aside, we have felt that the sort of observed minus expected methodology -- and actually, I apologize, Sai. If you could just put back up the cumulative incidence curve? I think it's actually almost more instructive than the written response to this answer.

What you can see quite plainly from the cumulative incidence curve is that there's an exponential phase to this problem that's followed

by a linear phase to this problem.

When you plot it out as an incidence rate curve, it's a peak that sort of levels off to a flat baseline. But what's happening here is, there's one rate of events that's happening very early in the first -- mostly, actually, in the first seven days, but sort of tapering off to a stable baseline by somewhere between 30 and 90 days. And then, it's a linear phase after that.

The linear phase reflects the fundamental underlying biological risk for patients. And it's not that that can't change. It's just that it doesn't change that much within a year. It changes over decades. It changes when people get new diseases, or they have the chronic effects of long-term hypertension that progressed ten more years' worth of time, their atherosclerosis has gotten a lot worse.

None of that happens quickly, as a general matter. All of that long-term risk is pretty stable over the course of a 12-month

period, which is the period of analysis.

And the only real risk difference from a stroke standpoint, is this funny short-term risk. And this funny short-term risk looks a lot like the short-term risk for major stroke after minor stroke and TIA.

This is a known biological association, the risk profile pattern matches almost exactly, and it coheres with everything else we know from multiple convergent sources that this is a known problem with dizziness, that we miss strokes in dizzy patients.

It's not a high percentage. It's a small percentage. Obviously, these are small numbers we're talking about, in terms of per 10,000 patients.

But it translates ultimately to somewhere between 50,000 and 75,000 patients a year in the United States who are misdiagnosed, and probably somewhere between 15,000 and 25,000 a year who are harmed as a consequence.

And so, what we're essentially doing is saying, look, this observed high rate is too high, at the beginning. And it should reflect much more the long-term rate. And whatever rate is in between is this short-term stroke risk, except these weren't patients called stroke. They were patients called not-stroke.

So, whatever that short-term stroke risk is, that's the patients who were mislabeled, for sure, right?

There are still some patients actually buried in the long-term risk who were misdiagnosed, and didn't have a stroke until nine months later.

In fact, only one out of every five patients who has a missed stroke gets unlucky, right? The other 80 gets lucky. They don't have a short-term missed stroke. Now, they may get unlucky in the future, but we can't measure that as easily.

So, what we're measuring is the people

who suffer short-term loss. And this is just, for us, a barometer of the measure that matters. So, we could measure process failures and dizziness, and we would only find that 100 percent of patients have process failures at the bedside in evaluating dizzy patients. That's what we found. It's essentially 100 percent.

The misdiagnosis rate on dizzy patients is approximately 80 percent. That's hard to believe, but it's approximately 80 percent for patients with inner-ear disease, which is the stuff that looks like strokes and it's ten times more common than strokes. And for strokes, it's 40 percent.

So, these are huge numbers, in terms of missed rates. And the process failure rates are 90 to 100 percent. But this is a way of having some backstop as a measure.

Obviously, what hospitals are going to have to do -- and this gets a little bit into, is this a meaningful measure for

hospitals -- obviously, what hospitals are going to have to do is, in the short run, they're going to have to fix their processes and they're going to have to measure their processes as intermediaries.

But this is the needle that we actually want to move. We don't actually care if they fix their charts. We care if they save these patients' lives.

And so, from my perspective, the reason for having this kind of a measure is to ensure that our process measures remain tied, anchored, to these critically important outcome measures. And so, that's why I see this as such an important issue.

Anyway, that's why we've taken this observed minus expected approach. Are there other specific questions on the risk-adjustment approach as to sort of why we did it, or what's wrong with it?

CHAIR TEIGLAND: This is Christie. I

have one question. Did I read that right that you found, I think over 65 percent or some number like that, that had better than the average results under this model, under this approach, but none -- zero -- hospitals were identified as worse than the national average?

That was concerning to me. And I wondered why you thought that was the case, and if that has to do with the approach that you took.

DR. NEWMAN-TOKER: Yeah. So, I may call on Matt to help a little bit with this. But let me offer one initial -- actually, Matt, why don't I let you go first. And then, I'll --

MEMBER AUSTIN: Yeah. So, Christie, I think your question is around sort of the skew that folks have brought up in terms of seeing meaningful differences. Right?

So, based on our submission two years ago, the feedback specifically around reliability, was we needed to increase the number of facilities for which we were testing this

measure with, because we had tested it with the original four.

And so, we have been able to access the Medicare fee-for-service data for this analysis, and it's now up to -- actually, 5,000 or so hospitals that we have access to.

The challenge with the Medicare fee-for-service data is that it only represents about 20 percent of all ED visits to a hospital.

And so, we have had to, for purposes of testing, restrict this down to larger hospitals, and we wind up with I think 967 hospitals for purposes of testing.

In a perfect world with a better dataset, i.e. a more complete dataset -- so that could be an all-claims payer, all-payer claims dataset, something along those lines -- we actually would be able to create, or be able to measure more hospitals, and we would have greater precision with the hospitals for which we are able to calculate results. So, it's somewhat a

limitation of the dataset that we've been able to use for purposes of testing.

DR. NEWMAN-TOKER: Yeah. I'll just add that one of the things that we've seen repeatedly is that the biggest problems are in the smallest hospitals.

So, increasingly -- we've shown this, by the way, using H-cup data in prior analyses, that rural hospitals are at higher risk, for example, and EDs that have lower volumes are at higher risk. And we did that in a paper that we published back in 2014 in the journal Diagnosis.

And so, large academic centers that do lots of imaging on lots of patients are not missing as many of these. Right? The rates are significantly lower.

The small rural hospitals are probably missing a lot more. Now, each individual hospital isn't missing that many, because they have many fewer patients, but collectively, they're missing a lot more.

We actually are in the process of graphing that out, just for the full submission, just to kind of prove the point.

But you can see the separation between the curves of the small hospitals if you aggregate them over a longer period of time, just to make the point that -- remember, Medicare data are only one-fifth of the data.

So, when you look at it in the 10-year sample that we have, you can clearly see the differences.

At the end of the day, to me this is a constraint related to the data source more than it's a constraint related to the measure. And I think what we're trying to do is show NQF that this is a strong measure as it's constructed, but we have to use multiple different data sources to triangulate to prove the point that it can be leveraged for the kind of quality improvement purposes that are needed.

And the main thing we focused on in

this particular case was Medicare data, because it was very clear from the last go-around that the presence of four hospitals that were part of the Hopkins network just wasn't enough.

CHAIR TEIGLAND: So are you suggesting this model could be used for, say, commercial payments, or Medicare Advantage patients, even though it was only tested with the fee-for-service population?

Since we only found eight out of 967 that had any harm diagnosis, and none that performed worse. So, how is the measure useful then? I'm still missing that, Matt and David.

MEMBER AUSTIN: Yeah. And I think what we're saying is, the measure's ability to discriminate in terms of high and low performance would be improved upon if we were able to test the measure and to use the measure with more complete datasets.

The challenge there is, our -- the datasets that currently we have access to, are

challenging in order to scale and to prove that point.

DR. NEWMAN-TOKER: In a all-payer database, you would see a much bigger discrimination between the low performers and the high performers.

This is a precision issue on that front. In my view, it's not really validity issue. But more to the point, setting aside the question of our ability to resolve, if we actually had the 100 percent of data that were available from those individual hospitals, rather than one-fifth of the data that were available from those hospitals.

If we had included the smaller hospitals, which we couldn't do because their results were too imprecise when we were using one-fifth of the data, had we included those hospitals you would have seen a much bigger spread in the differences across the hospitals.

We only showed you the largest

hospitals, which are disproportionately places that have lower rates of stroke misdiagnosis on the whole.

And by the way, we did include Medicare Advantage, right, Matt? For this submission?

DR. MA: Matt, you're on mute.

MEMBER AUSTIN: Yes, thank you. Yes, we did include Medicare Advantage as well, at least the Medicare Advantage patients in the Optum dataset.

MEMBER NUCCIO: This is Gene Nuccio. I want to refocus on what quality it is that you are measuring. Is it the quality of properly diagnosing the patient as having benign dizziness? And if that's correct, how would the curve that's reflected for me in figure 3, how would that curve change with better diagnoses? I am certainly not an MD, and so I don't know why it is that you have that large rate in the first 30 days, as compared with what happens after

90 days.

But it seems to me that simply capturing the coding correctly would have nothing to do with changing that curve.

MEMBER AUSTIN: So, Gene, this is Matt. So, it sounds like -- I didn't quite follow the concern. So, we can talk a little bit about -- can you maybe say it one more time?

MEMBER NUCCIO: Sure. Is the quality that you are measuring with this metric the ability of a hospital to properly diagnose the patient as having a benign dizziness?

DR. NEWMAN-TOKER: Let me try to answer that. And then you can tell me whether I'm on track or not, in terms of your question.

What we're after here is better diagnosis, conceptually in the broadest sense, in front-line care settings.

Specifically, we're after trying to resolve a known problem in the diagnosis of dizzy patients, which is done very poorly. Now, the

issue of whether we are focused on correctly diagnosing benign forms of dizziness or avoiding incorrect diagnoses of stroke and missed stroke, is really two sides of the same coin.

The problem we're trying to solve is, how do we diagnose patients with the complaint of dizziness, and how well can we measure our overall performance on a metric that matters, such as a bad outcome after being told you had something benign.

Does that get at what you're asking? We don't really care, at some level, about whether or not they're accurately coding benign dizziness. What we're after is whether they're correctly diagnosing patients with dizziness.

But this is a proxy for that, because it shows us that they thought the patient had something benign, and it turned that they didn't.

MEMBER NUCCIO: Okay. So, given what you just described, how would the curve -- I mean, the idea is that by giving the hospitals

this information, you're going to allow the hospital to make some improvement on their practice.

DR. NEWMAN-TOKER: Yeah.

MEMBER NUCCIO: Based on that, how would the curve -- if the hospitals in your group completely did the right thing, how would the curve in figure number 2 change?

DR. NEWMAN-TOKER: Look different.

MEMBER NUCCIO: Yeah, how would it look different.

DR. NEWMAN-TOKER: It doesn't matter, they all kind of show the same thing. So, what it would look like is, it would cut off that exponential rise.

If you drew a line between the zero intercept of the X and Y axis and the end of the black line -- I can't do it on the screen for you -- but if you could draw that line, essentially cut off that initial exponential rise, and just turn it into a linear rise, that

would be what you'd be after.

MEMBER NUCCIO: Okay. So, that exponential increase in the first 30 days is attributable to them not properly acting on a correct diagnosis?

DR. NEWMAN-TOKER: Attributed to them not correctly diagnosing the patient and failing, therefore, to act.

MEMBER NEEDLEMAN: Okay, this is Jack. Can I -- I actually think figure 1 illustrates this point a lot cleaner.

DR. NEWMAN-TOKER: And by the way, mathematically, I mixed up my points. It had been a point somewhere in between, a little bit below there.

MEMBER NEEDLEMAN: Not from the response. Figure 1 from the original application. Page 15 of the measure testing thing has these two figures, but it has the weekly incidence rate. And this also relates to the risk-adjustment modeling that you're proposing.

MEMBER AUSTIN: And so, this is the figure that has sort of the big --

MEMBER NEEDLEMAN: No, not this one. The one that starts with the line high, and then drops and sort of wanders around at a rate of about 15.

DR. NEWMAN-TOKER: The wandering is just curve-smoothing effects.

MEMBER NEEDLEMAN: Yeah, but that's the baseline you're using for your risk-adjustment.

DR. NEWMAN-TOKER: These are basically the same data that's just represented as an incidence rate curve, instead of --

(Simultaneous speaking.)

MEMBER NEEDLEMAN: So, what you're saying is, if we had fewer people going out the door incorrectly, with their strokes not missed, then that high point at the beginning of this curve would be lower.

DR. NEWMAN-TOKER: Yeah. In fact,

that's exactly right. And that's what I illustrate in my lectures. I put a little arrow that starts at the top and it sort of pushes down to the bottom.

I say, look, if we could push this curve to be flatter, or to be completely flat, we would have accomplished something meaningful. And that's ultimately what we're after, is a meaningful measure of misdiagnosis.

DR. MA: Thank you, Jack, for mentioning this figure. We have a little bit of time to entertain one last question. So, Bijan, your hand is up and please do your disclosure first.

MEMBER BORAH: Yeah, sorry. I was muted then. So, I don't have any disclosure on this or any other measures for the day.

So, Matt and David, in fact, you answered another question that I'm going to ask partially during the course of your discussion in the last few minutes.

So, I'm going to be pushing back a little bit more on the samples that you ultimately will use for this study, for the testing. I think you ended up using 967 ED centers or hospitals.

And I know that you are using -- again, given the beta input, use only one-fifth of the Medicare data. But even then, I think what happened, there seemed to be some sort of selection issues, right? That 967 hospitals that you ended up using in the testing, they happened to be all the large hospitals, right? They typically would have about 40,000 ED visits per year.

So now, I think what you are speculating is that whatever you are finding in that selected sample, the results would be generalizable to all the other hospitals that are smaller proportionally, not smaller times of their ED visit rates.

So, can you shed light on that? I

mean, what is your basis for assuming that the results that you are finding here would be generalizable to all the rest of the hospitals as well in the U.S.?

DR. NEWMAN-TOKER: So, it's not a question of assuming that the results are generalizable. I wouldn't necessarily frame it that way, as to what we're claiming.

But the claim is that this problem is a problem that is ubiquitous. It's probably worse at the smallest hospitals that we couldn't analyze because we didn't have enough data points.

I think in the long run, one of the measures that we were recently evaluating in the neurology measures group was a CMS measure where they take the following approach to measurement, which is -- and this is a measure that CMS already uses, but NQF hasn't quite approved yet.

They take the larger hospitals for whom they can provide precision ratings and

rankings, if you will, based on performance. It's a stroke outcomes mortality measure after inpatient hospitalization.

And all of them that are big enough to fall into the precise-enough category, those hospitals are treated in the sort of pay-for-performance kind of level, from CMS's perspective.

And then the smaller hospitals are analyzed, but they just get private feedback for quality improvement. So, we ultimately envision that that's how this is going to be used.

That is, there will be some hospitals who can ultimately do this for benchmarking and pay for performance, because they're big enough. And I do believe that with better datasets, that will be a much larger swath of hospitals than the ones you saw here using the more limited data from Medicare.

But there will always be some hospitals that are too small to do this kind of

a measure. And they will just get personalized feedback and they will not be ranked because it won't be precise enough for them.

But that's the outcome that we are seeking in the four- to six-year time frame that's required by NQF.

MEMBER BORAH: Okay, thank you.

DR. MA: Thank you, David. I think that brings back to our topics that we have discussed several times, that whether or not reliability and validity can be revealed agnostic of the intended use.

So, that's going to be an ongoing discussion for the SMP. But at this point, I think we have heard a very comprehensive response from David and Matt, and we can move on to the voting for validity. Hannah, are you ready?

MS. INGBER: Yes. I'll just conduct a test vote first to make sure that everyone who's present is able to vote.

DR. MA: Thank you. And it's for

Subgroup 2 members only.

MS. INGBER: That's right. Thank you. So, for those in Subgroup 2 who evaluated measure 3614, you should see a test vote on your screen now. Please just select either A or B.

And I'm not seeing any responses coming in. So, if you're not seeing a question, please let me know.

MEMBER NEEDLEMAN: I am seeing the question and I responded.

MEMBER SIMON: Yeah, I am not seeing a question.

MEMBER BORAH: Yeah, I did it too.

MS. INGBER: Okay, let me try again. Apologies everyone.

CHAIR TEIGLAND: Me too. I responded. We don't have to write -- click on clear vote, right? Or do you?

MS. INGBER: You don't have to clear your vote. No. Okay, if you could try again, please.

It should say a test, and yes or no. Okay, they're coming in now. Thank you everyone. And we're expecting a denominator of eight.

DR. MA: And just for transparency, we need at least six members to meet a quorum.

MR. FORTUNE: And we're just waiting for one more.

CHAIR TEIGLAND: I think it's me and I don't see the voting map. I see a little graph. It says waiting for NQF votes presentation to begin.

MS. INGBER: Oh.

CHAIR TEIGLAND: Oh, wait. Okay, it just popped up. Let me try it.

MS. INGBER: Great. And I got you. Thank you. Thank you everyone. All right, we can conduct the vote on validity.

All right, voting is now open for validity on measure 3614. Your options are moderate, low and insufficient, as data element testing was conducted.

Okay, thank you everyone. All right, you can show the results, Caitlin, if you're able. Okay.

So, as you can see, for measure 3614 on validity, we have five votes for moderate, two votes for low, and one vote for insufficient. Therefore, the measure passes on validity.

DR. MA: All right, thank you, Hannah. Thanks for everyone's participation. We can move on to the next measure.

DR. NEWMAN-TOKER: Thank you all for your time. We really appreciate your hard efforts.

I know how tough it is to review these measures. And I've done it myself. So, appreciate all your hard work.

DR. MA: Thank you, David, for your comprehensive response --

DR. NEWMAN-TOKER: Okay.

DR. MA: -- and calm demeanor, very much appreciated.

DR. NEWMAN-TOKER: It's my pleasure. Thank you. Bye, bye.

DR. MA: All right, we are moving on to the next measure. And just for the record, we have only -- this is our last measure to be discussed in the morning session.

Matt Pickering, our senior director at NQF is going to lead the description of this measure.

MR. PICKERING: Great. Thank you, Sai. Can you hear me okay?

DR. MA: Yep.

MR. PICKERING: Excellent. Well, hello again everyone. It's good to see you again on day two.

I'm going to be talking about 3621, NQF 3621, which is a new composite measure of three different process measures, as you can see listed in the title here, the Overall Percent of CT Exams for which Dose Length Product is at or Below the Size-Specific Diagnostic Reference

Level (for CT Abdomen-Pelvis with Contrast/Single Phase Scan, CT Chest Without Contrast/Single Phase Scan and CT Head/Brain Without Contrast/Single Phase Scan)

This is a measure that uses registry data. It is at the level of analysis of the clinician group and practice level, as well as the facility level.

The measure is not risk-adjusted, it is stratified. As indicated, the three process measures, so it is not risk-adjusted, but the developer indicates it is stratified.

And as you can see listed here on this slide, the SMP subgroup did pass the measure on reliability.

It also passed the measure on the composite construction, so we are focusing our conversations today on the consensus not reached, which is for validity.

And for validity, the measure developer had indicated that they conducted a

systematic assessment of face validity of the component measure scores.

Before that face validity, there were some concerns from the SMP that it wasn't systematically assessed in what was provided in the testing attachment.

And the developers really relied on current use in alignment of the national guidelines, as proof of face validity.

The developer also uses approvals by CMS and their contractors of evidence for validity of the measure. However, it was not clear to some of the SMP members whether -- that the composite score, individual component scores of the measures within the composite, were tested for that face validity.

That's probably most of the concerns related to validity, with respect to that face validity component.

The developer did provide a response to that issue on page 146 of the discussion

guide, indicating that they did actually, recently complete a face validity assessment with a panel, and they listed out within the discussion guide as a response to the SMP concerns related to the questions that were asked the panel members that were seated, as well as the stakeholder groups that were represented. And then, percentages of respondents rating against those questions.

So, I will definitely stop there and I'll turn it over to our lead discussants, Marybeth and Matt, to see if they have any additional supplemental concerns, or supplement what I've been talking about with any additional issues.

MEMBER FARQUHAR: Yeah, I'm going to take the lead while Matt changes his hat there.

So, I just had a question -- I'm going to bring up the reliability a little bit because I just had a question about, you know, the method that they use. And I just wanted to, for future,

to let them know that they really should be, you know, looking at some other options.

The reliability for the signal-to-noise ratio is .9995. And, again, you know, a little bit more testing, one of the group mentioned a split sample reliability analysis might have been really nice to confirm and might help us understand a little bit better about the reliability there.

But also, you know, they did respond back and they basically said that the reliability's .7 or higher for, like, samples of 20, 20 people or more, but we would have -- it would have been nice just to see the data if they would have given it to us.

And I do have a question about the eligible patients and the reported patients. And I'm just kind of curious as to why there's a difference between the two if they -- is it just the people that are in the registry that are using, or is this -- did they estimate across the

USA of those that are not participating?

So, that was kind of strange because there was, like, 300,000 people missing from what they stated as eligible. So whenever they come on, that would be nice to know.

With regard to, yes, they did do a -- they convened a panel recently. They had 21: ten physicians, nine physicists, one patient, and one value-based purchasing person. They did ask three questions specifically that I thought were of interest.

The first one was do you think monitoring radiation dose indicates a good, worthwhile activity for advancing and maintaining safety and quality?

Ninety-five percent -- 20 members -- agreed. One member brought up -- which I am going to bring up again as an important issue, and maybe it doesn't belong in the SMP Council but it belongs with the importance to measure and, you know, meaningful differences here.

I have always associated dose along with clinical quality imaging. And I know that there is no quantitative way to evaluate the clinical quality imaging, but it seems like it's not complete if you don't have some aspect of that within this measure or an ability to look at that.

You could have a very low dose and you could have a terrible image quality, have to repeat it, and get a higher dose but still have a terrible image quality. And, you know, you still -- and then you're exposing the patient two, three, maybe even four times getting exposed to radiation that they really don't need to.

So, that's one area that I think that would be appropriate.

I also take note that they do use the registry data and they do use elements that are basically transmitted right to their registry. So, again, I'm questioning, you know, is it just the registry respondents or, you know, or is it

a larger group that they are missing out on?

And is there any difference there if they're not capturing the whole eligible population versus the registry population?

The second question that I thought was interesting was describe -- is this measure described as a reasonable or appropriate way to assess performance? And, again, 71 percent of the panel -- 15 members -- agreed that it was reasonable.

But, again, image quality is needed here to maintain a significant diagnosis. So, you know, again, it's image quality coupled with the dosage that I think they're kind of missing the point here a little bit

Yeah, six of the members -- let's see. Excuse me for a minute.

Twenty-nine percent -- six members -- while not specifically stating that the measure was not reasonable or appropriate did not agree that the measure is the best way to assess

performance quality. So, that was interesting.

And then the third question that they asked the panel was will the scores obtained for the measure as specified reasonably differentiate clinical performance of cost providers and separate out high performers from low performers?

And, again, this is a little low for this. This 62 percent -- or 13 members -- agreed that the scores obtained from the measure would differentiate clinical performance.

Three panelists indicated that the age of the scanner had important information that related to the image quality as well as the dose.

Another panelist also noted that the direct reported levels were not meant to differentiate performance.

And the measure -- let's see -- yeah, this measure collects the CAT scan or radiation outputs specific to patient and exam and compares the actual dose indices to benchmarks. So, they did do the piece with regard to stratification,

which was good.

What else? We also had an issue with regard to the literature review. We had noted that there wasn't kind of any systematic way or methodology that they reported on what they decided to include and what they did not include but, rather, they put -- they gave us a response with regard to the qualifications of the person that did the review. And then also cited the national guidelines and whatnot.

It would have been useful to see what kind of methodology and what kind of search terms they were using, what they included and what was not included just for future reference.

The other thing that they were saying about the older scanners, the older scanners don't have a direct access to the registry from what I gathered. So, they were using what they call OCR -- optimal character recognition -- software to read the data that comes out of the scanner, and it captures -- it's a secondary

capture.

I'd like to know what their -- what the accuracy of that is. Having worked with registries before and currently, the accuracy of the OCR stuff is not very high. So, again, I just want to point that out and see if they did any testing with regard to that, and any verification that that was accurate.

And I think -- and then the last question I had for them was the risk stratification analysis is --- they say that it's good for the group level and the facility level. And I need a little clarification as to why they think it belongs with the group.

Matt Austin, do you have anything else to add?

MEMBER AUSTIN: Yeah, the only thing I would add to that is they do specify the measure at both the group/practice level and the hospital/facility level. And NQF's guidance is clear that testing needs to be provided for each

level of analysis that the measure is being specified for.

At least for the face validity question it was not clear to me whether that applied to the group or to the facility or to both. But it did not appear as if they provided testing for both levels. But maybe I missed that or misunderstood that.

DR. MA: Thank you, Marybeth and Matt.

Do we have the developer on the call? And I would like to invite you to provide your response at this time.

MS. CAMPOS: Thank you all so much for allowing us to discuss our measure. We are grateful to have it be going to the SMP. There were a few issues addressed and I tried to remember all of the ones that were listed.

So, in terms of the reported -- number of patients eligible versus the number of patients reported, we do use our registry --- this measure as a quality improvement measure in

our dose registry, but also it is a qualified clinical data registry measure used in MIPS, which is a CMS payment program.

So, I believe that column is just differentiating how many patients were reported to CMS for that. So I just wanted to point that out.

In terms of the face validity -- or the --- of the literature search, I'm going to go ahead and let my colleague Dustin answer the search engine question for that.

Dustin, if you're on.

MR. GRESS: Sure, yeah.

No, those -- those documents are large consensus bodies that, you know, essentially synthesize the information themselves of the National Council on Radiation Protection and Measurement, the International Commission on Radiation Protection, the American Association of Physicists in Medicine, ACR -- the American College of Radiology -- and others. So they do

their own work to synthesize the evidence. And that's available.

Now, to be honest, this measure is, you know, evidence-based. There's a publication in 2017 that, you know, this is -- this is novel in several ways, largely the size information that is worked into the measure.

So, we did compose -- and I guess Karen would need to let me -- would need to chime in -- we did do a review of the systematic review, but that paperwork was submitted after the discussion guide. So, there was further information there.

But, you know, in terms of the literature search, you can do a literature search, computed tomography and diagnostic reference levels, you'll pull up about 1,900 exams on -- excuse me, studies on PubMed. Of those, about 80 of them will be useful. And you will see some widely disparate methods for trying to develop and establish diagnostic reference

levels. And so, you know, that's not particularly helpful. You know, I just went through the studies in the last month.

So, those tiny shreds of evidence that are around are gathered by experts, and experts look at their own data and they come to consensus guidance and publish them in documents, like what I tabulated for the discussion guide.

MS. CAMPOS: Thanks, Dustin.

In terms of the face validity survey that we conducted, I will say they were a panel of experts. I don't think a lot of them have a quality measurement background, so I do think that maybe they got a little too into the weeds of what the questions were in terms of what the measure is trying to capture.

There aren't any standards for quantifying image quality at this time. So, you know, this measure is going to be best ways to measure scientific exam level DLPs and, you know, be able to obtain diagnostic quality images

across patients of different sizes for just indexed differentials.

So, you know, there is a little bit of wiggle room there. I'm not sure that they really understood exactly what we were asking. We do think -- you know, we agree that the DRLs alone are not an appropriate way to calculate performance, but it's closely related to the dosage received by patients, so.

MR. GRESS: Karen, would you like me to speak to the image quality question as well?

MS. CAMPOS: Yes, please. Go ahead, Dustin.

MR. GRESS: All right. So I -- you know, the question being asked about incorporating image quality with radiation dose is really kind of the holy grail of medical imaging. But, you know, as Karen mentioned, there are no standards for quantifying image quality. They just don't exist.

So, there are a number of research

groups who have published various methodologies for, you know, very small subsets of clinical indications or types of exams. And that's something that our organization, of course, is pursuing in different -- different avenues, and other organizations are also.

But it's just, you know, being able to monitor one's performance with how you apply radiation dose to your patient population is one very important element of quality safety in a clinical practice that's using computer tomography.

It's not all-encompassing. If we could have clinical indication information and quantify image quality for all of those with size information and correlate that to radiation dose, that would be great. But those things just don't exist at this time.

And so, the one thing that we do have is very good data and size-specific data on radiation dosage use for some CT exams. And

that's what this measure aims to utilize.

MS. CAMPOS: Thanks, Dustin.

I think I forgot one of Marybeth's first questions about where the data is coming from. So, it is coming exclusively from our registry. It is data, you know, across different regions in the U.S. But it is exclusive to the ACR.

In terms of the direct submission of data to the registry, so, yes, OCR software can be a little finicky. It is a very low percentage of facilities in our registry using secondary capture to submit data.

We get data from the scanners, from PACS, and from a radiation dose screen, which gets transmitted into TRIAD, which is a cloud-based web server that goes through our registry. That's why we're saying that it is a direct transmission. And really there would not really be any missing data to the registry.

And then I believe the last question

was in terms of the risk stratification, why we did it at a facility versus group. And so I'm going to let Judy Burleson, our other colleague, jump in on that question.

MS. BURLESON: Hi. Thanks, Karen.

Can you hear me?

MS. CAMPOS: Yes, thanks Judy.

MS. BURLESON: The structure of the registry in terms of facility and group participation is very similar. But in terms of groups associated with facilities is why they're pretty close to what facility numbers would be compared to group numbers. The Group 10 is associated with facilities in the registries. So, it's pretty much a mirrored rate, facility versus group.

We have provided performance data, facility versus group, because of the Group 10 level submission to CMS through the QCDR. But really it's just a way to slice and dice the information that's submitted from the facility.

DR. MA: Thank you, Judy. I think that that answered Matt's question about the two levels.

Zhenqiu, do you have your hand up?

MEMBER LIN: Yes. I just wanted to follow up on this one.

Are you saying that you treat -- you would equate hospital with group level testing? Because in the testing form you checked both group and tested in a hospital facility.

So, my question is do you treat them as equal?

MS. CAMPOS: No.

So, in the testing form the facility/hospital is, like, kind of lumped together, so we just checked that box. But it's per facility. But I don't know if the NQF team could clarify on that, but that's how it's written in the form.

DR. MA: So --

(Simultaneous speaking.)

DR. MA: --- two different, two levels. I think what you are saying and Judy was saying is it's really at the practice, the facility level.

MS. BURLESON: That's right.

MEMBER LIN: So, facility, do you mean hospital or you mean just a group practice? And I was confused. I thought that we saw mostly is pertaining to group and practice level, not the hospital level.

MS. BURLESON: So, the structure is basically a facility being a hospital or an imaging center where the equipment -- CT equipment is typically placed and used, utilized at the center, at the facility. And radiology groups are associated with facilities, whether they're a hospital or outpatient imaging centers.

So, in the registry structure the registry facility identification is where the equipment is. And facilities will have radiology groups associated with their facility in the

registry. So, when that is available, the performance data for a group is based on the imaging information received from the equipment at the facility.

MEMBER LIN: So, you are not differentiating different type of facility, even if potentially there could be some kind of difference; right?

MS. BURLESON: Yes. We can do that by different types of facilities. But I'm not sure that is what NQF had looked for.

I mean, we can break out data by location, census region, type of facility, academic community and that sort of thing.

MEMBER LIN: Thank you.

DR. MA: Matt?

MEMBER AUSTIN: Yeah. And I -- I guess I don't want to keep repeating the same point over and over again.

I guess I'm just still really confused on that they checked the group/practice box and

the hospital/facility box. And then what I heard was that the two are sort of interchangeable, but then yet different data were presented for each.

And so I guess I'm just sort of confused on how to understand this and how to evaluate it, I guess is the bigger question.

MS. BURLESON: Okay.

MEMBER LIN: I share Matt's because I had the same reaction, so I was confused about that as well.

MS. BURLESON: I'm sorry, I'm just looking back at the last --- so that we provided performance data for the facility and for group using the structure that I described. And breaking it out because we see the measure as potentially appropriate in both the hospital program, accountability program, and a physician-level program, which it has been for several years in the MIPS program.

So, we do have that TIN facility -- the tax ID group information associated with

facilities, and performance data for the group, which for the most part would be overlapping. But it potentially could be slightly different if a facility has more than one group practicing, more than one radiology group associated with the facility, imaging and radiology section.

And that's very unusual, but it could occur that the facility would have more than one group associated with it and may not -- we may not have all the groups registered in the registry.

So there could be slight differences in the number of physicians/groups per facility. Or, potentially, the group may register and have -- pull in data from some of their facilities but not all of their facilities.

So, there is not necessarily a one-to-one match for all facilities in the U.S. that are in the Dose Index Registry. Not all of them have associated group Tax ID Numbers that have been registered and associate themselves with

facilities when they want to use the measure for MIPS reporting. But not all facilities have groups that are interested in that. So that's why we have a slight difference in the performance for facility and group.

But -- and maybe this is Matt's question. So, we did the risk stratification by facility and not group, and stated that it's similar. That's because of the -- the data that we used to do that risk stratification comes from the facilities.

If a group -- if groups were using the measure, their -- the stratification would be the same at their facility. It's very similar information.

MEMBER AUSTIN: And just to clarify, the face validity survey you did, was that with the intention of measuring groups or measuring facilities? That wasn't clear to me.

MS. CAMPOS: It was with the intention of just measuring performance. It could be

applied at both the facility and group level.

DR. MA: So, this is a tricky situation where TIN is the deciding factor level in this mix of the practices and the facilities.

Any other comments?

MEMBER WALTERS: This is Ron. I was in Group 3. I'm in Group 3.

I think everybody has a clear understanding of the issues. They've been all brought up. And so this is a valuable measure. The people who use it, use it.

Whether or not it matches to the process is what we have been talking about. Not quite the same as yesterday, but there's no doubt that between the group versus individual practice versus practitioner and the face validity and the way that was conducted, this is a hard one to score. And I had the same difficulty going through this too.

So, I mean, it either -- I mean, it's a valuable measure. Either it's endorsed or not,

either it's passed through or not, either it's insufficient or not, or possibly low. But that's kind of the difficulty I think I hear going on.

DR. MA: Thank you.

MEMBER LIN: I just have a follow-up question and I'm going to develop it.

Do you worry about an unintended consequence, like if it's, like, too low dose, right, that is not good either.

MS. BURLESON: Let me start by answering -- start answering that, and then maybe Dustin would want to chime in.

I think that because the way that we scored is DLP that's at or below the diagnostic reference level, that takes into consideration some variances where the dose may be higher or lower, and so that there's some standardization there in terms of unintended consequences.

I think you would think -- I'm not sure what you are specifically referring to, but maybe the too low of a dose. Is that what -- where

you're concerned?

Dustin, do you think that you can answer that more?

MR. GRESS: Sure. I can take a shot at it.

So, I hear this concern from other folks, including medical physics colleagues. You know, the concern about, you know, ever assessing for diagnostic reference levels to being some sort of race to the bottom. I think it's a legitimate question.

So, you know --- but I think this goes back a little bit to what I said before where we don't know, we don't have any evidence-based criteria for assessing image quality.

So, you know, part of the assumption is that if a physician, a radiologist sees a set of images that are not of diagnostic quality, then they will reject them.

And so -- and we have, of course, an accreditation process that -- it's separate. But

the image quality issue is addressed through different -- different processes, and that really is the cost. The cost of having too low of a radiation dose is that a study may be non-diagnostic and need to be repeated which is, you know, I would assume suboptimal.

But, you know, the backstop is not, is not this measure setting a lower threshold, because that's presuming something that we cannot presume.

Does that answer your question?

MEMBER LIN: Thank you.

MR. GRESS: Sure.

DR. MA: Marybeth, Matt, do you have any other questions? If not, we can move on to the voting. All right, hearing none, we can pull up the vote.

And as a reminder, since unlike -- face validity is used, the highest rating can be moderate.

MS. INGBER: Thank you. Yes, voting

is now open on Measure 3621 for validity. Your options are moderate, low, or insufficient. I'm again not seeing any results come in. So let me know if you're having any trouble voting.

(Simultaneous speaking.)

DR. MA: Only for ---

MS. INGBER: Yeah, they're coming in.

DR. MA: Only for Subgroup 3. Yeah, I think there's a delay on that website today.

MEMBER WALTERS: This is Ron, the voting's working.

MS. INGBER: Yes, I see them coming in now. Thank you.

All right. Thank you, everyone, I've got all the votes. Caitlin, you can feel free to share the results.

MS. FLOUTON: So for scientific --- for validity testing on Measure 3621 we have four votes for moderate; two votes for low; and three votes for insufficient. Therefore, consensus is not reached on validity.

MS. INGBER: Thank you, everyone.

DR. MA: Thank you, everyone.

And I also want to, since we have a few minutes left for the morning session I do want to call on the members who reviewed this measure, if you can provide some helpful guidance for the developers as they are working on the internal analysis, what would it be -- if you can offer some tips that can help them.

For example, I hear a couple of members who are saying hospitals and independent facilities should be separated out in the analysis.

MEMBER WALTERS: That was one.

And also the face validity was -- I think it's just -- and I don't think -- I think it's clearly described, the face validity, but I think there are a lot of questions of whether that face validity was portrayed accurately or not, or the best that it could have.

So this was hard to evaluate on

validity. And I knew it was either going to be a 5-4 or 4-5 vote. And I think if you go over the transcript of the discussion you'll get plenty of recommendations for how this could be made better in presentation.

I do think at the end, like I said, this is an important measure. It's utilized. It makes sense. But there's a lot about the mechanics of how it was presented that I think could be improved.

MEMBER AUSTIN: Yeah, I would echo Ron. I think the concepts are there, I think it was sort of maybe better organization of the information. And if they are going to specify for two levels, you know, including both of those consistently throughout. And I think there's -- obviously it sounds like the face validity questions may not have been asked of the right people or maybe asked in the right way. That may be worth revisiting in terms of that exercise.

MEMBER FARQUHAR: Yeah, I would have to

echo that as well. This is Marybeth. I was -- read it through several times and just had a very hard time following the logic. It may have been me. You know, I'll admit that.

But it just was very, very hard and difficult to piece it together. So, if it could be a little bit more clear and a little bit more detailed would be really, really helpful.

DR. MA: Thank you, everyone. As the --- validity again is voted as consensus not reached. This measure will be moved forward to the respective standing committee.

And we are done for this morning. Thank you everyone for your participation. We are going to take a lunch break and resume at 1:30.

Thank you.

(Whereupon, the above-entitled matter went off the record at 12:42 p.m. and resumed at 1:32 p.m.)

DR. MA: So, welcome back to the

afternoon session of the second day.

We have four measures that are slated for discussion for this afternoon. All four measures passed reliability and validity during the preliminary analysis. However, they are pulled for a discussion for some overarching concerns.

So, the first two measures, six -- sorry, for these two measures 0674 and 0679, their results are a little bit on the lower end for reliability test, and --- as some team members have discussed about this very topic over a year now.

We decided prior to this meeting -- actually prior to we even revealed any measures, a decision was made that all the measures that their -- if their reliability results are at the lower end, they will be pulled for discussion to make sure our criteria are applied consistently across all three subgroups. So, these two measures, even though they passed, they fall into

that category because their results are at the lower end.

So, this is the context. I am now going to pass to my colleague Matt to very quickly describe what those two measures are and the tests and the results.

MR. PICKERING: Great. Thank you, Sai.

So, like Sai mentioned, I'll just touch on both of the measures quickly and then I will turn it to Alex for discussion.

So, the first measure that you see on the slide, 0674, as 27 and 28 of the discussion guide is the Percent of Residents Experiencing One or More Falls with Major Injury (Long Stay).

So, this is a maintenance measure, it's an outcome measure. The data source is assessment data, specifically from the Minimum Data Set. And it's specified at the facility level of analysis.

I'm just going to touch on the

reliability portion here and then we'll go to the next measure.

So, the testing was done at the measure's score and data element level. For the data element level they use inter-rater reliability testing with gold-standard nurses, or those nurses that are trained to do with the MDS instruments. So, they did some inter-rater reliability testing with gold-standard nurses and gold-standard nurses to facility. And they used kappa statistics.

And the results were fairly high: .967 for gold-standard to gold-standard nurses; and .945 for facility nurse to gold-standard.

The developer also did score-level reliability testing where signal-to-noise is split half statistics. And the signal-to-noise reliability result was an average of .45.

And then the split half correlation for the measure, it was positive. And it -- with the --- provided limited evidence of internal

reliability.

I also will then go to 0679.

If we can just go to the next slide. So, now we're going on to 28 and 29 of the discussion guide.

It's a similar measure. It's Percent of High Risk Residents with Pressure Ulcers. It's also a maintenance measure, an outcome measure using the minimum data set as the data source and specified at the facility level of analysis.

For the reliability results or testing that was done, similarly it was data element and measure score, with the data element using gold-standard nurse as the abstractor and facility nurse as the abstractor, using kappa statistics. And the values were also high here, so .92 for gold-standard versus gold-standard, and .97 for facility nurse versus gold-standard.

For the signal-to-noise and split half reliability testing the split half correlation,

the relationship was deemed to be moderate, suggesting modest evidence for internal reliability. But then for the average signal-to-noise reliability score it was .5. Whereas, again, the 0674 was .45.

So, this is the reliability results for these -- both of these measures up for discussion. So, I will turn it over to Alex to lead that discussion.

Alex?

MEMBER SOX-HARRIS: Thank you, Matt. You had a great summary of these two measures.

Before getting into this, though, I want to thank the developers for really a nice set of analyses to address both reliability and validity using several different lenses on each of those constructs. I really appreciated the quality of the submission.

There --- a couple of things I want to focus our attention on. One has to do with the main reason we're discussing these measures,

which is the relatively modest score level, entity level reliability, but also the way we think about our reliability algorithm, as we touched on briefly yesterday morning.

So, the item level reliability of these measures depends on the reliability of the data extraction at the nursing facility level. And as Matt mentioned, the way this was tested was with a fairly large, four -- roughly 4,000 patient sample representing 71 community nursing homes in eight states and 19 VA nursing homes, so what I consider a very large, very representative sample from the data set, which then got evaluated by gold-standard nurses.

So, I see this as a well-designed validity analysis, item-level validity. So community, real world abstractors to a gold standard, which is fine. I don't see it as a reliability analysis.

A reliability analysis would look something like community abstractors compared to

themselves on the same cases, or two different community abstractors doing the same cases and seeing the correspondence of that.

So, just as a technical design point, I consider the item-level analysis to be item-level validity analysis. But in this case, this might be a distinction without a difference because NQF by the guidelines say if you do -- if a developer does item-level validity analysis, that absolves the need for item-level reliability analysis.

So, I characterized for both of these measures, I think, based on the item-level validity, the excellent. Both the numeric value of the reliability analysis and the design -- which I think we need to always consider both of those things. Sometimes we see big numbers but the underlying sampling is quite, you know, limited.

But in this case I think the item-level validity analysis is excellent.

Therefore, that propagates to the item-level reliability judgment. So, I think the lowest we can rate this measure is -- on reliability is moderate. You know, we can't do high because it's only at the item level at this point of the discussion. But it's excellent.

So -- and based on the algorithm, as my updated understanding of it from yesterday, even if this -- if the entity-level reliability is missing or terrible, we still can't rate the measure lower than moderate on reliability, which I think -- I think is a problem that we had a good chat agreement yesterday that we need to revisit that, that structure of the algorithm.

So, but that's the implication of the algorithm that this measure, these measures, because they have excellent item-level validity evidence, can only get a no lower than a moderate on reliability judgment, no matter what else is done. That's the way I'm reading it.

So, that's one point I want us to

discuss.

The other -- the other main point is about the -- so, let's just say the criteria and guidance for measure evaluation are changed and we were allowed to use score or entity level reliability as our primary criteria for judgment. In my view, having a median signal-to-noise ratio of .45, which is -- you know, that means half of the entities have lower than that, I consider that very low.

The split sample reliability on the first measure -- on 74 -- was .18. It's very low, in my opinion.

And then something that we've been discussing over the last, you know, year-and-a-half at least is what kinds of analyses would we like to see to bring to life the meaning of these reliability statistics in terms of classification stability, which is the thing I believe that we're most concerned with.

And another thing I loved about these

submissions is there was a stability analysis included by the developers. It was included in the validity section of the application, but there was a analysis of performance, a quarter-to-quarter performance stability.

And what that found was quarter to quarter where you would not expect the true quality to change in an entity based on, you know, fall avoidance, for example, facility performance -- 25 percent of facilities jumped at least three deciles in performance quarter to quarter.

So that's a -- that looks to me like a lot of instability in measurement when we have -- you know, changing three or four deciles just based on a random or measurement area is quite large.

So, I see that as an opportunity for us to see what are the implications of a median signal-to-noise ratio of .45, a split sample reliability of .81. What that means is measurement is really, really jumpy. And that's

low reliability.

So, personally, if I were just judging it -- if I were able to judge these measures -- the reliability of these measures solely on the entity level data presented, including the stability analysis, I would -- I could not do any -- I would not assign anything higher than low.

And the same for the other measure, the discussion of the other measure is completely parallel. So --

DR. MA: Alex?

MEMBER SOX-HARRIS: Yeah?

DR. MA: Sorry to interrupt. I just want to add one clarification, which is both 674 and 679 conducted a bolstered element and performance score-level reliability test as shown here, the select sample and the signal-to-noise, they are at the performance score-level, so the highest rating -- so the start point for you is high, not moderate.

Then you'll bring in your concerns and

your judgment. Just wanted to add that clarification.

MEMBER SOX-HARRIS: Yes, yes. Right.

So, I may have misspoke at the beginning. I think the point I was trying to make is even if we only had the item level, that's as high as we can go. But I don't think we can go any lower than moderate based on the fact of the item level is excellent.

So those are my two -- those are my two -- sorry for being so long-winded, but the two things are the algorithm and the judgment of the entity-level reliability based on these results.

So I'll stop.

DR. MA: And before other SMP members chime in, I just want to pivot a little bit about this table. So, the top two measures are the ones we are discussing. And we listed another two measures that also are at the lower end for reliability test results.

Both 2881 and 3612 were discussed at yesterday's meeting. So, the voting results was one star is based on the preliminary voting result. And then yesterday's voting results are added at the bottom. So you can see what happened to those measures as a reference.

At this time I think I am going to invite Dave to chime in.

CHAIR NERENZ: Thanks.

Alex, that was a great summary. And I was actually sort of responding so I could get clarification. If I was following Alex correctly -- and it's a really tight hair-split distinction -- I thought what he said is that since they compared the data element abstraction to a gold standard, it was really validity, not reliability.

And so we end up -- so they really didn't do data element reliability. They were excused from doing it because they had data element validity.

So I think -- and so I think what's really worth discussing here is this interesting logic where in this example -- which I think is fascinating -- that data element validity trumps all. Meaning, given the list of requirements in the algorithm, if you've got data element validity, you don't need to test data reliability. And if you -- well, you're -- no, you're given a pass on it. That's actually more accurate. You're given a pass on it.

And then once you're given a pass on it, since the requirement is only either data element or measure score, effectively you pass reliability.

So, I think it's a wonderful example to bring up for discussion. You know, clearly if the NQF moved in the direction that we've discussed many times in the past of requiring both levels of reliability and both levels of validity for everything, this funny anomaly would go away. But, boy, it's an interesting example

and I'm glad you brought it up.

DR. MA: Jeff, you're next.

Thank you, Dave.

MEMBER GEPPERT: Thank you. So, I think of reliability often as sort of a proxy for preventability, the idea being, you know, if the poor-performing hospitals, facilities, or clinicians performed as well as the high-performing, then these events would be prevented.

And so I guess I'm just wondering as we start talking about reliability thresholds whether the fact that these are -- the top two are patient safety measures and where there might be sort of a no-fault presumption of preventability, whether that matters at all.

Maybe it doesn't. But historically it has sort of mattered. And there's been sort of less emphasis on sort of between entity variability because basically the presumed threshold was zero or near zero.

But the other ones, you know, the EDAC

and -- you know, where there's more of a quality sort of interpretation, that -- you know, that wouldn't apply.

DR. MA: Thank you, Jeff.

Joe, you're next.

MEMBER HYDER: I wanted to make a comment to follow up on Jeff's and then offer a question.

You know, Jeff's comment about reliability having the component of preventability is really interesting for these measures because the measure developers specify that these are -- at least in the case of falls with major injury -- they're essentially never events. And so they're entirely preventable. So, preventability is 100 percent in their minds. It's curious.

Alex and David articulated, I think, the struggle between the data that are presented and then using those data to navigate the algorithm and come to a satisfying conclusion.

I know this has been discussed previously, but I wanted to make sure that since they articulated it right up to the edge that they had an opportunity to comment on what that next iteration may look like because the current scenario appears to be unsatisfied.

DR. MA: Thank you, Joe.

Should we at this point invite the developer to provide a response?

MR. NAGAVARAPU: Sure. Sai, this is Sri Nagavarapu from Acumen. Can you hear me okay?

DR. MA: Yes.

MR. NAGAVARAPU: Okay. Great. So I think the important clarification that we wanted to make -- and we really appreciate the chance to jump into this discussion here. The important clarification that we wanted to make is that the way that this measure is presented on the Care Compare website for public reporting, it actually is a four-quarter average.

And the reason that CMS did that was precisely to ensure greater stability of the measure for the public reporting channel that most people would access. And so you can think of this measure as being shared publicly, primarily through Care Compare where beneficiaries can visit, and they would see a four-quarter average there.

And then for the downloadable files that are available on data.cms.gov separately, those files present both a four-quarter average as well as the measure results by quarter for four separate quarters together.

And so in the public reporting sphere, mainly what people are seeing is the four-quarter average on Care Compare, and we have reliabilities that you pick for that measure, if people are interested, because the reliability, you know, like CMS intended, is dramatically higher for the four-quarter average.

And then on data at cms.gov, people

see both the four-quarter average and the full context of measured movement over four quarter, so that they can kind of assess and take into account movement that happens over the four quarters.

To give you just a very quick feel for the reliability of the four-quarter average, for the falls with major injury measure, the four-quarter average has, just applying the standard signal-to-noise metric, the four-quarter average has a median of .8 for reliability and a mean of .77.

And the split-half Pearson correlation is .626. Split-half ICC is .625. So that's for the falls with major injury.

And then for the other measure, for the pressure ulcer measure, the analogous four-quarter average has a mean reliability of .78, a median of .81. The split-half Pearson correlation is .69, and the split-half ICC that's forecast forward is also .69 there.

And so, yeah, this is definitely like part of -- part of the intent of CMS in presenting these four-quarter averages publicly for these measures has helped put the stability question in measuring these important patient safety areas.

I'll stop there in case folks have questions.

DR. MA: Thank you, Sri. Are there any other comments, questions, for the developer?

MEMBER SOX-HARRIS: I would just say thank you for those clarifications, and I think the measure should be specified with a four-quarter average because what you just said is the -- the reliability statistics you just quoted would be, you know, really good and would lead me anyway to pass the measure, if that's the way it's -- if that's the way it's specified.

The way it was specified and tested in the application gives very different numbers, which would lead me to fail on reliability, if I was allowed to do it just on the basis of the

NWL.

I don't want to -- I want to respond to Joe's question about what another iteration -- but I don't want to cut off other questions or comments that people have.

So I think my preference -- and Dave alluded to it I think -- is to have, especially for maintenance measures, at least for maintenance measures, to have the requirement that both item, or as we're calling it, patient or encounter level reliability, and score or entity level reliability, so having both. And maybe score them separately, but at least have the score level reliability be referenced.

So if that's poor, then the measure is unreliable. Something like that. That would be my -- that's where I'd like to see it go.

MEMBER PERLOFF: What was the logic -- I was just going to ask the logic of the item validity trumping all. Was there a thought process to why the old rules weren't there? I

was just curious. Or maybe we don't have a historian who knows. Okay.

DR. MA: Patrick, you are raising your hand. Can you provide a response to this question?

MEMBER ROMANO: Well, I'm not sure I can provide a response except that there has been, you know, concern that measure developers are often in a situation where they are building measures based on registries or based on a particular set of data. For example, ECQMs usually have test data from a limited number of sites.

And so, you know, at the first submission, it has been very difficult often for measure developers to have that entity level reliability. But, of course, that doesn't apply at maintenance, and I think we have gradually come to understand that we need to raise the bar at maintenance.

And this is also a natural progression

because I think effective 2019, measure developers are now officially encouraged to submit both. I think that's the wording that's used in the guidance documents.

So the next step from Encourage is required. So that's a logical next step I think for us to take.

DR. MA: Thank you, Patrick. I see a lot of head nodding on my screen.

So just before I summarize the next step, is there any other comment from anyone?

MR. NAGAVARAPU: This is Sri from the measure developer. Just a quick note. I would think for your comments there -- what we could do is, you know, originally this measure was endorsed as a one-quarter measure. And so we've just kind of, you know, been focused on the four-quarter nature of the public reporting.

But we can definitely make that clear in the submission materials. But, yeah, thanks for that suggestion.

MEMBER SOX-HARRIS: Yeah. And it sounds like the way things may go is that next time this comes up, the standards might be different. And so your four -- and it sounds like your four-quarter data, you know, might do well under a new standard, so that's a good way to move.

DR. MA: Thank you, Sri. Thank you, Alex. So just to summarize and play back, since our current guidance allows either data element or patient level, or accountability entity level reliability testing, this one has data element, as Alex mentioned, as excellent on the data element reliability test and results.

Actually, looking at the preliminary analysis rating, both measures passed with a moderate rating. So it's very consistent with what Alex just provided. I think that the guidance that SMP members have provided through last few meetings is we are moving towards to the entity level reliability because that's how the

measure is going to be used.

And as we are moving towards that reaction, we are going to update our guidance, hopefully in June-July. And of course we can't change the policy here in the middle of the review cycle, and that would be not be fair. But going forward, we are going to update our guidance and seek public comment, and those updated guidance will be applied to future cycles.

Okay. And also, we have talked about reliability for quite some time, and hopefully at our May advisory meeting we can wrap up and provide clear guidance for the developers.

Okay. I think we can move on to the next measure. Okay. So we have -- next, we have a very similar situation. We have two measures, 3615 and 3616. They both passed the reliability and the validity, but the SMP members have identified some overarching topics to discuss.

So before we move on, I want to ask if the developer, University of Michigan, UM-KECC

folks, are on the call already because we are a little bit ahead of the time.

DR. SEGAL: Hi. This is John Segal from UM-KECC, and we're here.

DR. MA: Okay. Great. Thank you. Thanks for being flexible. Very rarely we run ahead of time. So --

MEMBER ROMANO: We're under good management today, Sai. That's why.

DR. MA: Thank you. I also want to check if my colleague Sam is on the call now.

DR. STOLPE: Hi, Sai.

DR. MA: Hi, Sam. Do you mind briefly describing these two measures? Since the concerns are very similar, we're just going to discuss them together.

DR. STOLPE: Absolutely. Thanks very much. And hello again to our SMP colleagues.

What we're looking at here is a pair of measures related to unsafe opioid prescriptions inside a dialysis facility. These

are for dialysis practitioners.

What you'll note, and as Sai pointed out, is that both of these measures have passed. But there were a couple of concerns that were raised that are summarized at the bottom of the slide here.

For this measure, it may be helpful for us to actually just discuss the two of them together. Perhaps it will resolve all of the concerns associated with the two under 3615.

But I will just briefly read the measure description for this measure, and it's the percentage of all dialysis patients attributable to an opioid prescriber's group practice, who had an opioid prescription written during the year that met one or more of the following criteria, either duration of greater than 90 days, morphine milligram equivalents, or MMEs, greater than 50, or an overlapping prescription with a benzodiazepine.

So, UM-KECC, appreciate you being

present today for this discussion. The main thing that we wanted to touch base on -- and there may be some other concerns that are raised, and I'll hand it over to Dr. Romano in a moment -- is around this question. And it's this: to what extent is the validity analysis confounded by unmeasured case miss -- case mix, excuse me, considering that dialysis physicians with sicker patients such as comorbid cancer, have higher mortality rates, higher hospitalization rates, and higher opioid use.

Okay. With that being said, I'll hand it over to Dr. Romano to provide any other additional context and start off the discussion.

MEMBER ROMANO: All right. Thank you very much, and good morning. Good afternoon to folks on the east coast.

So this is a very interesting pair of measures. They are labeled as measures of unsafe opioid prescribing, and unsafe here is determined based on three criteria. That is, the duration

being over 90 days, the dose being greater than 50 morphine milligram equivalents at any point, or any degree of overlap with a benzodiazepine. So it's three criteria for unsafe prescribing.

So this is clearly a process measure. And as we discussed in the discussion guide, it's quite unusual for process measures to be risk adjusted. Usually, the concept of process measures is that if there is something called unsafe prescribing, that it's unsafe for everyone.

So the use of a risk adjustment model implies that there must be some appropriate indications for, quote/unquote, "unsafe prescribing." There must be some conditions under which this type of prescribing is at least not contraindicated.

So this causes us to put a lot of attention on the risk model and how the risk model is used. Again, process measures usually use denominator exclusions or stratification to

handle heterogeneity in the population. Very unusual to do a risk adjustment model.

So in this case, the developers developed a very sophisticated risk adjustment model that has 178 risk factors in it. It includes a wide set of demographic characteristics as well as comorbidities. It has an overall c-statistic of .74.

So it's comparable in discrimination to mortality models that this committee has reviewed, much better than readmission models that we reviewed yesterday. And, further, the calibration plots show that across deciles there appears to be an eight-fold difference in risk.

The overall -- as far as I can tell, from the data shown here, the overall failure rate in Table 2 of the submission is about 40 percent. So that's the average across all of the accountable entities is a 40 percent failure rate.

So overall we have -- we have a

measure with a high failure rate, a process measure, with a risk adjustment model, and the content of the risk adjustment model shows, for example, that cancer is a very important risk factor, as you'd expect. Cancer patients are more likely to need opioids for management of their cancer-related pain.

Malignant metastatic cancer is also in the model. Again, patients with metastatic cancer are particularly likely to need opioids, rheumatoid arthritis, and so forth.

But then there are other factors in the model that are clearly endogenous, things like drug dependence, substance use disorder, anxiety disorders, previous opioid poisoning. These are factors that are in the model, and yet clearly they are tied in. Patients get the diagnosis of substance abuse disorder because they are on chronic opioids.

So the risk adjustment model looks very good in terms of performance statistics, but

there is no conceptual model or theory underlying the selection of factors for the model.

But then the real problem, as it shows on the slide here, 53, is that the validation of the measure is based on dividing provider groups into tertiles showing that the top tertile has a failure rate over 46 percent, the middle tertile 30 to 46 percent, the best tertile under 30 percent. And they show that the patients in the highest tertile, or the worst tertile, have a slightly higher hospitalization rate, 1.49 versus 1.41.

They have a bit more hospital days per year, 6.1 versus 4.1. And they have a higher death rate. But this, of course, is not surprising because this is, as far as is reported here, an unadjusted analysis.

So, clearly, patients with cancer, with chronic diseases, are more likely to get chronic opioids. They are more likely to be hospitalized. They are more likely to die.

So the other members of the subgroup pointed out that the tertile analysis is also simplistic in terms of missing the overall distribution of the pattern. But the fundamental problem is that we are -- that the sole basis for validating the measure here is an unadjusted analysis when the developers have in fact shown that risk adjustment is essential for the application of this measure.

So there is a muddling of concepts that made it very difficult for a couple of us I think to evaluate this measure, because we have a process measure with a risk adjustment model that includes endogenous variables, but then the risk adjustment is not factored into the validation of the measure and the construct validity tests that are applied.

So I'll stop there and see if my colleagues want to add anything.

DR. MA: Eric, you can unmute yourself.

MEMBER WEINHANDLE: All right. Hear me now? Okay. Yeah. So this is an interesting one. Both of the measures are interesting. I mean, I've mentioned it a few times to the SMP or to the committee that dialysis is my domain, too, and I do know the measure developers, though I'm not involved in this one at all.

I noticed some of these aspects, and in particular the endogeneity threat I think is an interesting one. I would say that some of these conditions -- anxiety disorder, substance use disorders -- given that they are being sourced from claims, have pretty low prevalence in my experience.

So it would be surprising to me if they exert a substantial influence on the model, but conceptually I must admit I was troubled by them as well.

You know, to the bigger issue, I struggled with this with all of the literature that is developed on opioid, benzodiazepine,

gabapentinoid use in the Dallas population. There is a lot of it. Every observational study that is done in this population shows that all three of those classes, and combinations of those classes, are associated with increased risk.

So what's confounding and what's, you know, actual directly attributable safety issues with the medications or the mixes of the medications is I think inherently unclear. And, I mean, we could debate it over and over, and I'm not sure that we'll ever get to any resolution, given the limitations of administrative data.

I will say that my impression has long been -- maybe the measure developer feels similarly or not -- that the amount of variation across physician practices and across dialysis facilities, whatever unit you want to look at, it's just really extreme, given I think anybody's experience with this dialysis patient population and the sort of prevalence of moderate pain/severe pain that you would encounter in a

typical dialysis facility with 70 patients.

So there is unexplained variation that I think very strongly in my experience hints at, you know, physician preferences and physician practices. And I wouldn't expect that any, you know, adjustment to the model would actually change things in terms of, you know, individual observational units that are either higher or below average.

So I guess that's my first thought about it all.

MEMBER ROMANO: I'll point out Sam did put something in the chat indicating that NQF has other endorsed measures related to opioid prescribing that are not risk adjusted. And they have exclusions for cancer. They also use a 90 MME threshold rather than 50 MME.

But, of course, the particular choice of the threshold is more of an issue for the Standing Committee. So I think here we are just -- we are really focusing on the validity of the

use of a risk adjustment model, and then the validation analysis that is not based on risk adjustment.

DR. MA: Larry?

MEMBER GLANCE: So I thought that was a great presentation by Patrick. And I more or less agree with him about the issue of endogeneity. I think it may not be quite as straightforward, though, in the sense that if you consider other outcome models, other risk-adjusted outcomes, so, for example, an AMI model, or acute myocardial infarction.

In that model, you are going to have risk factors, such as a history of previous myocardial infarction, history of angina, history of previous cardiac surgery. And in some ways it's similar to the current model that we are discussing in which a history of substance abuse is in the model.

Is it endogenous? Yeah, to some extent. Absolutely. But then if you have an

AMI model as a history of coronary artery disease is a history of a previous MI, could you see it in somewhat the same way? You probably could. So although I agree about 80 percent with Patrick, I don't think it's completely as black and white as he -- as he presented it.

MEMBER ROMANO: I mean, I'll just say that this is the difference between an outcome measure and a process measure. Again, conceptually, we are looking for the risk factors in a process measure to be more about the indications or lack of contraindications for the preferred therapy here.

And so when I see a lot of factors that seem to be immaterial, like cataracts and glaucoma, in addition to the factors that are endogenous, it makes me wonder what the conceptual framework underpinning the model is.

DR. MA: I think it sounds like now is a good time to invite Jonathan to provide a response.

DR. SEGAL: Sure. Thank you for the opportunity to provide some clarification in the testing form that you have all reviewed. And I'm not sure if it would be helpful to start with a little bit of background in terms of what -- the differences between these two measures and the rationale for them, just because I know some questions came up early on about that.

When we initially had conceived this measure it was designed to be done for the dialysis provider, what's called the monthly capitated payment physician, or MCP physician, because dialysis patients have told us that they trust their dialysis docs and the staff to look after their medications and be the guardians for safe prescribing.

When we had our technical expert panel, the patients and patient advocates agreed with that concept, but other members of our technical expert panel felt that because 90 percent of opioid prescriptions that are

written for dialysis patients are written by other providers, non-nephrologists, that a measure that looked at the opioid prescriber would also be appropriate.

And so this was our compromise to try and make everybody happy is to develop one measure that looks at the dialysis physician, who is maybe not writing the opioid prescription but has a general sense of responsibility to make sure that patients' medications are safe. And then a second measure that looks at the opioid prescriber, since they are the one who is actually putting pen to paper and writing for the opioid.

So that's the rationale and the difference between these two measures, which are -- otherwise essentially look at the same concepts.

You know, in terms of the title of the measure, which uses the word "unsafe," I think we were -- when we came up with that concept, we had

been obviously looking through the information that the CDC has, and which really speaks towards trying to make for safer opioid prescription and pain control.

And the reality, obviously, is that there is probably no such thing as safe versus unsafe. It's really a matter of risk, right? So these are high-risk opioid prescriptions as opposed to opioid prescriptions that might be considered somewhat lower risk, particularly given our patient population or dialysis patients.

So in terms of our rationale for risk adjustment, rather than an exclusion criteria, which, as Sam mentioned, are present in other NQF-endorsed measures, we felt that it was really critical to try a risk adjustment strategy to mitigate against the unintended consequences of under-treatment of pain in patients that have multiple comorbidities that we know have a significant pain component.

So we mentioned, obviously, cancer, sickle cell disease, there is a whole host of these.

And so we wanted to be able to empower physicians to be able to adequately address analgesic needs and prescribe opioids and do so in a way where they didn't feel like every time a patient needed a long-term opioid prescription or at a dose that's higher than, say, 50 morphine equivalence, that they were going to somehow get dinged on a quality measure and then not treat patients' pain or rapidly or inappropriately taper pain medications just because they are trying to meet the metrics.

So to be clear, this quality measure, to end up in the numerator, it's not based on individual prescribing events, right? So we're looking at the totality of a prescriber's experience over the course of the one-year measurement period, and then we're comparing that to the prescriber group's peers.

So you mentioned that, for example, 40 percent of these opioid prescriptions, that was our mean, right? So 40 percent of the opioid prescriptions fall into this category of high risk by meeting one of the three criteria that we have outlined to be in the numerator statement.

So being above 40 percent doesn't necessarily mean that's a failure for the measure. It just means that that's up slightly above whatever the average is at the time in the performance year that we're measuring.

So to really get called out -- this measure -- means that you have to be in the very far extremes relative to your peers. So as prescribing practices change over time -- and we know that opioid prescriptions are already a little bit lower than it has been -- people are judged against their peers. And so that's where we think the comorbidity adjustment helps, in that providers who are taking care of sicker patients with more comorbidities are essentially

giving credit relative to providers who are taking care of healthier patients.

So, and then to be clear, because this gets to the specific question about our validity analyses, our validity analyses do use the adjusted measure. So if it was not clear that we were doing risk adjustment but that that didn't factor into our validity analysis, I apologize.

We in fact do use the adjusted rates in the validity measure and then look at crude hospitalization and crude mortality rates. So we did tie the two together for our validity analyses that you have reviewed.

So let me stop here and see if there is additional things I can help clarify.

MEMBER ROMANO: That's very helpful -- thank you -- to clarify that. So, but if I understand correctly, the death rate and the hospitalization rate are not risk adjusted. Is that correct? Those are crude rates.

DR. SEGAL: Correct. We don't have a way of adjusting hospitalization and mortality at a provider group practice. I mean, we haven't -- that hasn't been developed and fleshed out and stuff. So we chose to use the adjusted opioid prescription rates and then use simple hospitalization mortality.

MEMBER ROMANO: And then could you explain what the performance score is exactly? Because you describe cut points for the performance score, T1, T2, and T3, and those are presented as absolute percentages. So could you clarify what exactly the performance score is? I don't think it's stated anywhere in the documents.

DR. SEGAL: Right. So we used the -- we used those tertiles just as an example to try and show the relationship between provider groups that have, you know, low, medium, and high levels of high-risk opioid prescriptions and how that factors into hospitalization or mortality.

Those cut points were just simply for the validity analysis itself and are not actually part of the measure performance score in any way, shape, or form. So we are -- for our measure performance score, we are basically looking for extreme outliers relative to their peers.

And so that rate is actually more like three and a half percent essentially of provider groups that we feel comfortable identifying as having a significant portion of their opioid prescriptions different than the rest of provider groups.

So that's essentially our performance scores using outliers, and we use a null -- we use an empirical null technique to try and limit the number of extreme providers that we identify with -- in our performance score.

MEMBER ROMANO: Okay. That's fine. But, again, what you're showing in TV13, with these tertiles of -- where you put -- you do -- your entire validity analysis here is based on

putting providers into these tertiles.

And, again, I'm not seeing evidence that these tertiles are based on risk-adjusted scores. Correct me if I'm wrong.

DR. SEGAL: Well, the proportion of the opioid prescription part is what's risk adjusted, right? So it's a direct standardization technique, so we'll adjust their percentage up or down based on the risk adjustment model, and then use that for hospitalization and mortality.

MEMBER ROMANO: Okay. So you did direct standardization before you put the provider groups into these tertiles.

DR. SEGAL: That's correct.

MEMBER ROMANO: Okay. Very helpful to explain this more, but I wonder if you could also address the concerns that we have about the model and the fact that the model seems to include factors that would ordinarily be considered appropriate exclusions, such as metastatic

cancer.

And it appears to include things such as chronic kidney disease that every patient in the analysis should have, and it appears to include things that are themselves indicators of opioid use, such as substance use disorder.

DR. SEGAL: Sure. So in terms of a little bit about our kind of -- our comorbidity selection process, so we started by looking through the AHRQ CCS categories, and we looked for frequencies of categories that were above one-tenth of one percent in our dialysis population.

And then we used a forward stepwise technique to identify ones that were associated with our opioid use. And so we used -- when we looked for claims -- when we looked for comorbidities in Medicare claims, we used these in the prior year from the performance period. So these are not kind of current comorbidities; these are comorbidities from the past year.

And we did this comorbidity selection for the two different measures separately, so we looked at the prescriber and did a comorbidity selection and we looked at the dialysis doc MCP group and did the comorbidity selection separately.

But we wanted to have one common list of comorbid variables to use for both measures, and so we took the union of the two. There was a large degree of overlap in terms of the comorbidities that were identified, but there were a few small differences, and so we just ended up taking both of them.

With regards to comorbidities like kidney disease, while obviously that doesn't make a lot of sense for our MCP measure -- and those would normally be filtered out -- for the prescriber measure, it's not at all unusual to have -- to see a claim for kidney disease for a non-nephrologist opioid prescriber in the past year.

So we decided that because we were looking for a model with the best predictive value, you would leave all of those comorbidities in, even though we recognize that some of them at face value don't necessarily make a lot of sense.

DR. MA: Eric?

MEMBER WEINHANDL: Yeah. Just a quick question for the measure developer, and so I guess two parts to it. The first part is just to confirm that the comorbidities on the prevalent -- the prevalent comorbidity is a long list. Those are coming from inpatient claims. I believe I see the note, so tell me if I'm wrong.

But then the second part of it is, if they are coming only from inpatient claims, I mean, I am curious about the measure developer's opinion if -- if, for instance, just previous year hospital admissions and/or days have been included in the model, in place of all of this comorbidity, would you end up with a relatively similar result?

DR. SEGAL: Yeah. That's a great -- that's a great question, and I don't know, but I can see -- I can see your point in terms of a more simple and slightly different approach at tackling this.

MEMBER WEINHANDL: You know, it just erases the question. We probably wouldn't be having this endogeneity discussion if the two covariates in the model were admissions and days. And I just have a feeling that we're seeing a lot of odds ratios that look interesting, but actually they are probably all -- on average, they're positive, and you're just looking at different versions of positive from weak to strong.

DR. MA: Are there any other comments, suggestions?

MEMBER ROMANO: Again, I would just encourage the developers to take a more careful and thoughtful approach to selecting the features of interest. Again, the conceptual framework

guides us. As Larry mentioned before, an AMI outcome model for mortality, we are looking at, what are the factors that would be associated with death following an MI?

Here I think we are looking at the factors that would be associated with appropriate use of moderate to high dose opioids. And so from that framework I think you could be a lot more thoughtful in how you select the variables.

DR. MA: Thank you, Patrick, for your very insightful comment.

If there is no more comment, I think we are going to wrap up this session. And I want to thank Jonathan for your presentation and the response you offered.

DR. SEGAL: Thank you for the feedback. We appreciate it.

DR. MA: And feel free to reach Sam for additional information or technical assistance as you move to the Standing Committee Review.

MEMBER ROMANO: Sai, may I ask, is there -- is there any interest or support for revoting on this measure on validity? Or is that -- I know there would have to be a specific sort of change of perspective, if you will, but just opening that up for discussion.

DR. MA: Yeah. I think it's a question for your subgroup to kick off here. Like after hearing the discussion and the response, do you want to change your vote?

CHAIR NERENZ: And I guess, Patrick, I was thinking the same thing, although I would make a little more fine point on the question. Presumably, in order to revote this, we'd have to have -- what would it be?

Like two or three people, maybe a number of people, who would change their vote in a more negative direction, because if people have heard the last discussion and actually feel a little more comfortable with the measure, a revote does nothing. Although maybe for the

record it would pass a slightly different message to the Standing Committee.

So I guess I would only step to a revote if two or three of our subgroup members would be willing to indicate, either by private chat to Sai or by however means, that they actually feel now much more negative about this and actually would indicate that by voting. Otherwise, it doesn't change anything.

DR. MA: That's a good suggestion, Dave. I will give two minutes to the Subgroup 1 members. If you feel you are going to change your vote unfavorably, you can chat me privately. All right. I will give you another half-minute and see -- so far I've got one response indicating no change is needed.

MEMBER ROMANO: I will say I think it just raises some issues that maybe we need to discuss about measure score validity and how we interpret measure score validity. Again, my overwhelming sense is that some groups just have

sicker patients. And so they have -- even after adjustment, they have more opioid prescribing, more deaths, more hospitalizations.

And so we should discuss as a group how to improve our confidence in these validity -- entity level validity testing when confounding is such an obvious consideration.

CHAIR NERENZ: Yeah. Patrick, two points on that, and then Sai can tell us if anybody really wants to push a revote here. Coming out of yesterday, it seemed -- and monitoring the chat as well -- we have some very clear issues to discuss about sort of the standards and criteria for measure -- entity level validity. What correlations are allowed? What level of correlation is expected?

You know, if there is a situation of confounding, how do we deal with that? A number of related issues. But I think we have found through these two days that we've had a number of really troublesome issues about validity.

And if our subgroup is to be faulted, perhaps we may have been a little -- not critical enough about this particular one, in the same way we were, for example, in the first three we talked about yesterday where we said, you know, you've got two measures correlated that have the same numerator events in them. Therefore, we don't accept what you are sending us. Do it over.

We didn't provide that sharp a review or comment here. And so, therefore, now we don't have any revised redone measures to look at.

So I guess we just have to keep getting better and better at this in terms of our own processes, but also to keep getting clearer and clearer to the requirements to measure developers that in this domain, here is what you can do, here is what you should do, here is what you must do, and then here is what is not going to sell us at all, and so don't even bother.

MEMBER WALTERS: I would like to second that. And we seem to uncover a lot of

issues in the measure review sessions, and we have moved forward over a few years now. We finally might be getting close to where a reliability of 40 percent is not reliable.

But in the off -- between the measure reviews is when the real work has to get done for formal recommendations to CSAC and the Board about what, as you said earlier, is no longer suggested but what should be required. And I agree there has been about five topics yesterday and today where it's either very strongly suggested or required.

And you're right, we don't uncover these until we talk about specific examples involving a given measure, but then we fall. I wouldn't say that. That's too negative. But we have room for improvement in making progress on those discussions to getting them actually encoded enough ahead of the time so when the measure developers come to us for the next round, they aren't totally hoodwinked by a bunch of

things they haven't -- didn't even know about at the time they were redoing their measure.

So this does require a certain amount of planning, probably three or four months in advance of the next measure review meeting, to know what we really want to be ratified by the chain up from us, and communicated, more importantly, to the measure developers that when you come back with your measures, whether they be new or maintenance, here is what we expect to see.

And, you know, we're getting there, but we still keep on covering issues. And that will probably always happen, but then we need a plan to take care of those issues and make them become actuality.

Sorry about that.

CHAIR NERENZ: No, no. Ron, that's really good, and I saw Patrick nodding. I know if, Patrick, you want to come back in on this one, but yeah.

DR. MA: Thank you, Ron, for -- go ahead.

CHAIR NERENZ: Just to follow up on Ron. I think our last couple of what we call the between cycle meetings, like say January-February, I think we have made some nice progress on reliability, and then we had to just set it aside while we turned to the actual review of these measures.

And, you know, about an hour ago, sitting here thinking, you know, where we did we leave that on reliability, because -- you know, because I know where we left this, we had to turn around and do this work of specific measure reviews.

But, you know, for what it's worth, I started outlining myself last night a set of issues about validity that at least in my mind were prompted by some of the discussions we had yesterday. You know, does it make sense for -- is it a priority, as one of our developers told

us yesterday, to correlate outcome measures with other outcome measures? I don't see any sense in that.

But we heard it yesterday, and, you know, so if it's not that, what is it? So I actually have some confidence that if we can pull back up what we were talking about on reliability before we then got into these, and if we then turn our attention a little more sharply to validity, between those two measure domains, we could probably have something to put forward to CSAC in these next couple of interim reviews.

Now, of course they have to go through a process. It may be this time next year before we review measures that came in under some different set of expectations. But that would be a good thing.

DR. MA: Larry?

MEMBER GLANCE: So I agree with all of the points that have been made. I think that, you know, there is a lot of room for improving

what are already very good algorithms, but could certainly be made better. I think that -- I kind of think that we need to also in a more -- in a more general way rethink what validity testing means.

I think that David sort of touched on this. The idea that we view empirical validity testing by looking at the correlation between a new measure and existing measure has never struck me as a very strong way of doing validity testing, because a lot of the existing measures aren't necessarily all that good.

We don't have a gold standard per se. I think that people have talked about, well, gosh, if we know that a certain practice which is captured in a process measure represents a true best practice, then we ought to be able to see a correlation between adherence to the best practice or good performance on that process measure and a new outcome measure that is being proposed.

The problem that we have is that there are very, very few process measures that are going to be linked necessarily to really best practices, because we know -- and people have looked at this in the literature -- is that over 50 percent of so-called best practices are backed up only by level of evidence C, by expert opinion.

So it's really hard to kind of base our validity testing on empirical validity testing. And I would suggest -- and we've spend a lot of time as a group, and I know that not everybody here was part of the methods panel that contributed to the White Paper that we published a year ago. But we spent a lot of time struggling with this and talking about this.

And I think, as a group, we came to the -- I don't want to say conclusion, but I think that we thought that predictive validity was extremely important. And by predictive ability I mean that if you have a risk-adjusted clinical outcome measure, okay, and you look at the

performance of that prediction model, if you can show that that model performs well -- and we can get into the details of what exactly good model performance is. I mean, that's a whole other discussion.

But if you have predictive validity, then you can make a really strong case that your measure, your risk-adjusted outcome measure, is valid. And the reason I say that is let's take the example where you have a perfect prediction model. So you have an outcome, say, live-die after CABG surgery, and you have a bunch of clinical risk factors in that model.

And we can all agree that those are preexisting risk factors, that we're not doing anything stupid like putting stuff that's endogenous.

So if you have a really good model, and if it was near perfect and you could predict the outcomes of patients conditional on their risk factors, okay, and in that perfect world

where you have a near-perfect prediction model, then you could take a cohort of patients in a particular hospital and you could predict each one of their predicted probability of death, and then average them together and get their expected mortality rate.

And then you could compare the observed to the expected, and then you really would have a very, very good way of knowing whether or not a particular entity's performance was either above average or below average.

So I would advance that predictive -- and it gets -- obviously, it gets a little bit more complicated when instead of using non-hierarchical models you are using hierarchical modeling, and you use -- looking at PE ratios instead of OE ratios.

But the point is -- the point that I'm trying to make -- is I think that our panel should move a little bit away from spending time -- I don't know that we'd want to spend too much time

with empirical validity testing. And I think we should start to shift our focus to predictive validity testing.

And I think that some of the issues that we discussed yesterday and that were addressed in a much of emails overnight about the type of testing that we need to do to ensure adequate predictive validity, I think this is something that is an important thing for our committee, our panel, to look at.

So just to summarize, I think that validity testing is extremely important. I think we still have some room to go in terms of figuring out what that should be. I think we should, as a group, decide how much emphasis to put on empirical validity testing versus predictive validity testing.

If we are going to talk about predictive validity testing, there is lots of issues to talk about. We talked a little bit about the issue of out-of-sample validity testing

yesterday. You know, what thresholds should we use for model discrimination, for model calibration?

There is a lot of things that we could be talking about, but I think that we should start that discussion.

DR. MA: Thank you, Larry. We are on the side putting together a running list of topics that we want to discuss at the next advisory meeting for the SMP members.

But for the record, I don't want to keep our developer here. I just want to announce that nobody asked for a revote. So we're not going to revote on these two measures.

I think at this point we can wrap up the discussion of the two measures. Great. Thank you.

Now we are going to provide the public an opportunity to provide comments. Don Casey?

DR. CASEY: Yes. Thank you very much. I actually could not make the entire call

yesterday, but I have been on for the whole rest of yesterday as well as all today.

And I want to just say -- it's the first time having been involved with standing committees, most currently the Patient Experience and Function Committee -- how impressed I was with the great expertise and thought that the staff and the members of the committee have put into these measures.

It was really eye-opening to me to see how much time you spend on this because, you know, as a Standing Committee member, we just received the reports. So, and I appreciate the -- I can't remember who it was, but someone mentioned this notion of how we might consider resolving differences between what the Standard Methods Panel puts together as well as the committees.

What I did do is put in the chat room a thank you to Dr. Nerenz and the authors of an article that appeared in the American Journal of Medical Quality in December on the Scientific

Methods Panel. And I think this was a very elegant overview, and it really helped me actually summarize in very clear terms.

You know, it's hard -- no offense -- to figure things out sometimes from the NQF website. And I just want to bring this to the attention of the audience, especially the panel, I am the new senior associate editor of AJMQ.

And perhaps maybe, Sai, NQF may want to consider making this article more widely available to the membership, because I think it really helps to solidify a much better understanding of SMP. So, Dr. Nerenz, thank you very much.

And I put my email address in there, too, if you want to follow up with me, any of you, about this. But thanks a lot.

DR. MA: Thank you, Don. Thanks for mentioning that article. That was a very well-written article, and I would say if you haven't read it, thanks for putting the link in the chat.

Anyone else from the public? All right. Hearing none, we can move on to the next section.

All right. I now will invite my colleague, Hannah, to provide next steps.

MS. INGBER: Thank you. So our initial next steps include moving forward with the spring 2021 cycle. Measure submission deadlines vary depending on the topic area of the measure, but they will be in the first three weeks of April. So this Friday, April 2nd, then April 9th, and then April 16th.

The NQF staff will summarize all of the information that the SMP gave us and provide that to the various standing committees. That goes into the PAs. It goes into the final -- the reports that get drafted for comments. The standing committees will certainly get that information.

The measure evaluation meetings for spring 2021, for the measures that were discussed

yesterday and today, will take place in the June-July timeframe. And the CSAC meetings to render the endorsement decisions will be held on November 30th and December 1st.

The intent to submit deadline when measure testing and specifications are due for the fall 2021 cycle is August 2nd. Again, we strongly encourage developers to reach out to NQF staff for technical assistance ahead of that deadline, as we are always here to help you and understand -- help you understand the submission process and answer any questions.

Next slide, please. Thanks.

So at the May 4th meeting, the SMP will continue discussions on reliability guidance that we have had at previous meetings. The SMP is working on producing a guidance table on reliability for the developers that will list multiple methods of testing reliability along with their appropriateness for the level of analysis that their group is testing, and an

accessible range of results.

We are also planning to have a discussion on risk adjustment as it relates to other NQF work. So the best practices for developing and testing risk adjustment models project began last year, and a technical expert panel has been convened to review the development of some technical guidance for measure developers on social and functional status, related risk adjustments.

So since this technical expert -- since the technical guidance is meant to help developers with submitting measures to NQF, the SMP's input on that guidance is extremely valuable.

So the risk adjustment project will hold a web meeting on May 13th from 1:00 to 3:00 p.m. Eastern Time, and we'll invite SMP members to join in on that conversation. But We'll get it started at that May 4th meeting.

Of course, all members of the public

are also welcome to all of these meetings.

Regarding the July meeting, we are working on rescheduling this date because July 20th will be during NQF's annual conference. Because of COVID, the annual conference meeting date was changed, so we'll be rescheduling the SMP meeting.

Just some information about the conference. Every year NQF puts together this annual conference where we bring all stakeholders together, and of course we welcome everyone on this call to join. This year our theme for the conference will be The Care We Need: Driving Better Health Outcomes for People and Communities.

Within this theme, we will explore five topics: ensuring appropriate, safe, and accessible care; implementing seamless flow of reliable data; paying for person-centered care in healthy communities; supporting activated consumers; and achieving actionable

transparency.

So we'll look for a different day for the SMP to meet because we want to encourage everyone to go to the annual conference.

The fall 2021 measure evaluation meeting will be on October 26th and 27th, and then the final SMP advisory meeting for the year will be on December 14th.

Thanks.

You can always feel free to reach out to the Methods Panel Team at methodspanel@qualityforum.org, and our project webpage is here in the slides as well. And the SharePoint site, the SMP number is used to review documents.

Thank you, everyone, for your great attention and contributions today. I will hand it back to Sai.

DR. MA: Thank you, Hannah.

I will pause here before we wrap up and see if anyone has any questions about the

next step.

And, of course, for the upcoming SMP advisory meetings we have collected this running list of topics for discussion. But if you have any burning comments, questions, that you wanted to discuss, always feel free to reach out to us.

My only suggestion is always copy the team mailbox, making sure your correspondence is not lost.

With that, I think we can wrap up for this evaluation meeting. I want to thank everyone for the insight and for illuminating comments, suggestions. I know I have learned a lot. I hope our developers feel the same way.

With that, I think we can give some time back to everyone. Take care. Stay safe.

(Whereupon, the above-entitled matter went off the record at 2:55 p.m.)