NATIONAL QUALITY FORUM

Moderator: Kim Patterson
December 12, 2019
2:00 pm ET

Dave Cella: Hi, this is Dave Cella.

Woman: Hi, this is the NQF Methods Panel Team. We'll be getting started in just a minute.

Dave Cella: Okay.

Karen Johnson: Hello, this is Karen from NQF. We will be getting started in just a couple of minutes.

Man: Thanks, Karen.

Man: Hello?

Woman: Hi. This is NQF. We'll be getting started in just a minute.

Man: Okay.

Karen Johnson: So, good afternoon everyone. This is Karen Johnson with NQF. Thank you for joining our December Scientific Methods Panel Call. We appreciate your time. And I think given the emails that have gone back and forth between several of our panel members, I think we're going to have some interesting conversation today.

Just a couple of reminders, we are doing today as a two-hour call instead of a one-hour call just because we felt that, back in the day when we were trying to do our one-hour calls, we just didn't have enough to really get deep into the weeds and the questions that we were trying to answer. So I think today's call having more time, we will - should be able to have plenty of time to talk about at least one of the things that we want to talk about.

Usually we have a few housekeeping items that we go through, so I'll just tell you the ones that I can think of right now, which is, first of all, if you are not speaking, if you would make sure your phone is on mute, just to keep the background noise in check.

Also we know each other pretty well now but it still helps if you could just say your name before you start talking, and I'll try to do that too on my end as well.

So I know several people have contacted us to let us know that, unfortunately (unintelligible) didn't make the call today, but we will go ahead and do our roll call. I think we will have a good number of people on. So, (unintelligible) if you wouldn't mind forwarding us to our list of members. I wanted to just go through the list, and if you're here just say you're here.

Matt Austin? Bijan Borah?

Bijan Borah: I'm here.

Karen Johnson: Thanks, Bijan. John Bott?

John Bott: Yes, I'm here.

Karen Johnson: Hey, John. Dave Cella.

Dave Cella: I'm here. And remember, when you say you're here, to unmute, be sure your phone's unmuted.

Karen Johnson: Perfect. Thanks.

Daniel I know said that he's (unintelligible) and actually so did Matt. I know Matt and Daniel aren't going to be able to join us. Lacy Fabian?

Lacy Fabian: Here.

Karen Johnson: Okay. Marybeth Farquhar? Jeff Geppert?

Jeff Geppert: Here.

Karen Johnson: Hey, Jeff. Larry Glance? Okay, not hearing Larry yet. Joe Hyder? Sherrie Kaplan?

Sherrie Kaplan: I'm here but I have a hard stop at 12 o'clock.

Karen Johnson: Okay. We'll talk fast here. Thank you. Joe Kunisch?

Joe Kunisch: I'm on.

Karen Johnson: Hey, Joe.

Joe Kunisch: Hey.

Karen Johnson: Paul Kurlansky? Zhenqui Lin? He's here.

Zhenqui Lin: Yes.

Karen Johnson: Great. Jack Needleman? (Dave Merens)?

(Dave Merens): Here.

Karen Johnson: Gene Nuccio?

Gene Nuccio: Here.

Karen Johnson: Sean O'Brien?

Sean O'Brien: Here.

Karen Johnson: Thank you. Jen Perloff?

Jen Perloff: Here.

Karen Johnson: Patrick Romano? Actually, Patrick is on vacation. So he's been sending emails but he's not able to join us today. Sam Simon also is unable to join. Alex Sox-Harris?

Alex Sox-Harris: I'm here.

Karen Johnson: Hey, Alex.

Alex Sox-Harris: Hey.

Karen Johnson: Mike Stoto. Christie Teigland? Actually I think Christie also told me she couldn't come. Ron Walters?

Ron Walters: I'm here.

Karen Johnson: Hey, Ron.

Ron Walters: Hi.

Karen Johnson: Terri Warholak?

Terri Warholak: Hi. I'm here.

Karen Johnson: Hey, Terri. Eric Weinhandl?

Eric Weinhandl: I'm here.

Karen Johnson: Hi, Eric. And Susan White?

Susan White: I'm here also.

Karen Johnson: Great. Thanks, Susan. Did I miss anybody or did anybody not get an acknowledgement from me?

Paul Kurlansky: This is Paul Kurlansky. I just joined.

Karen Johnson: Oh, hey, Paul. Thank you. Anybody else?

Jack Needleman: Jack Needleman. I just joined.

Karen Johnson: Oh, great. Hi, Jack. And anyone else? Okay, great. So let's go on. And so what we have on the agenda for today's discussion is really to continue the conversation that started in October at our in-person meeting, to kind of walk through some of the potential options that we might have for updating our evaluation criteria. So, and I'll walk through and remind you kind of where we landed the last time and tee up the conversation for going forward.

Larry Glance: Karen, can you hear me?

Karen Johnson: Yes. Is that Larry?

Larry Glance: Yes, this is Larry.

Karen Johnson: Oh, great.

Larry Glance: For some reason I was having a problem with my phone. Thanks.

Karen Johnson: Okay. Yes, I can hear you now.

Larry Glance: Good.

Karen Johnson: Perfect, thank you. All right, let me check you off. Okay, here we go, Larry. Okay.

Now, the slide deck that I sent you, there's a lot of interesting stuff on that slide deck. We may not get to all of it. And if we don't, that's okay, we will get through it eventually. So we want to just take the time that we need to walk our way through the criteria discussion and then go on to some other of the questions that we have.

If we do have time, we do want to have kind of a conceptual discussion about just endorsement of healthcare performance measures in general, so, thinking about quality measures, measures of access, measures of cost and resources, and even more specifically population health-based measures. And it's something that we had thought we'd get to in our October discussion and we didn't get to it. So we do want to have that discussion, and (Dave Merens) will us through that portion of the call. Again, if we don't get to that today, you will see that in our next call.

And then, when we can, our next item will be thinking about score level validity testing and thinking specifically about what are the appropriate comparators. So we will talk about those just in general and we'll talk about then maybe a little bit more specifically around cost measures. Again, I kind of got - we'll get to that today, but if we do, then there's plenty there to keep us occupied, and I think will be an interesting discussion.

So with that, what we'll do is - well, a couple of things, a couple of other housekeeping things. Today's discussion, when we get to the methods issues on the recommendations, kind of like we did in October, in our in-person, it was quite helpful to have the show of hands in terms of (unintelligible). So we are going to most likely vote today on some questions, maybe, and probably not all of the ones that I sent out in the email a day or so ago, but we will have some voting. And that's really just to help me get my head around kind of the level of consensus that we have on these ideas. So we will do that.

And right before the call, literally like one minute before, I sent out a link that has the instructions as well as the link for doing that voting. So hopefully you guys all have that. It should be at the top of your email box at this point. If you don't have that, if you wouldn't mind just letting us know via the chat function and (Hanna) will help us get that to you. Also we have the chat function and the raised hand function, and (Hanna) is helping us monitor that as well. So, feel free to use that to get in touch with us as we walk through, or as usual, just type up and we'll just have a conversation about these issues as we walk through.

So with that, let's go ahead and get into our slide deck. And this slide here, and this is Slide 6, if you're looking at this not through the Webinar, this is just a current testing requirement. This is just a reminder of how things stand right now. And as you can see here, we have some different requirements, depending on whether or not - or depending really on the type of measure.

So, for kind of what I, you know, kind of shorthand call, the regular structure process outcome measures for reliability, we allow either data element level testing or score level testing. It doesn't have to be both. But for instrument-based measures, we do require both. For composites, we require score level reliability testing. And for ACQMs, it kind of depends on whether the data are stored in structured data fields or not. And so it could change.

The shortcut that I have referenced in this table is one that I think in general everybody hates. Well, I shouldn't say everybody, but there's actually I think a few people are somewhat happy with the shortcut. And that is this idea that, if data (unintelligible) that's really, really testing is required, we don't require additional reliability testing.

Now, I don't think people like that kind of across the board, but we have had some conversations in October and via email about, you know, the utility of asking for data element reliability if you already know something about data element validity. But we'll get to that in a few minutes.

Also, in terms of validity, again for regular structured process outcome measures, right now our requirements say that folks could provide either element level testing or score level testing, or, at least for new measures, face validity. The idea with face validity is that, by the time of maintenance, there would empirical testing of some sort required. Again, with instrument-based, we require both. For composite measures, it's basically the same as the structured process outcome when a measure is new. But by the time it comes around for maintenance, we are expecting score level validity testing. And then for ACQMs, we require data element validity testing. That's not to say we wouldn't like to see score level testing, but it's not a requirement at this point. So that's our current requirements.

In our testing or our discussions in October, we talked about this and we did try to talk about everything together, and that became a little iffy. So we tried to split things apart and talk about score level reliability testing.

So, where I think we had pretty strong agreement is that NQF should be requiring score level reliability testing. So that seemed like where the consensus was in our meeting.

In terms of requiring again for these regular measure, structured process outcome measures, in terms of requiring data element reliability testing, I don't think we've reached consensus in that discussion. And then around that shortcut, we had this idea of potentially a new shortcut, not saying that no reliability testing would be required if we had data element validity, but

maybe would - maybe that shortcut would change a little bit. So again we have to talk about that.

In terms of validity, requiring data element validity testing or requiring score level validity testing, we talked about it a little bit in October but we didn't get any consensus. So, and we really didn't talk about instrument-based composite or ACQMs in our last meeting. And today I really don't want to talk about instrument-based composite or ACQMs. I really want to focus on these other kinds of measures, and work on those today and maybe we'll work on - and rethinking some of the elements later if we need to.

So with that, we will actually ask about each of the four - the regular - for the regular structured process outcome measures. We're going to start by talking about validity, and we'll ask you if score level validity should be either required, expected but waived, if there's (unintelligible) justification, or leave as optional, basically that's saying the (year or business) would still apply. So we'll ask that about validity - score level validity data elements, and then we'll work our way backwards to reliability.

Now, another thing, I think we did talk about the, kind of this idea of we'd like to expect it but we'd be willing waive it if there was adequate justification. I think there was a quite good consensus maybe on some of that, but again we didn't have time to do the raise hands and get some counts. So we want to make sure that is (added to that) as well.

The other thing I want to tell you before we really start getting into the discussion is what happens if you guys make some recommendations today? And hopefully we will. Hopefully we'll come to some kind of consensus.

What will happen is that I and my team will actually take those and think about them and put that in kind of context with other things that we've done in the past, and basically make a package to present to our CSAC, which is our governing body. It's the (unintelligible) that oversees the CDP process and oversees our endorsement criteria.

So, what the CSAC could do, depending on, you know, what you guys recommend, they could accept it, you know, as stated. They could reject it. Or they could modify it in some ways. So that, they can basically do what they want to do. But your recommendations are really advisory recommendations to the CSAC. You would actually make the final call on these.

Even if they accept recommendations that reflect pretty major changes, which they could very well do, they may want to suggest the implementation timeframe. So what I mean by that is we may get consensus from you guys today to, for example, let's stick with what we landed on in October, to start requiring score level reliability testing. That wouldn't get potentially approved by the CSAC until April. And then, you know, we won't necessarily expect that measure developers would be able to do that testing in time to turn it around if they are kind of up for maintenance or bringing in a new measure very close to that timeframe.

So what would generally happen is we might decide that we would want to give maybe a one-year grace period (unintelligible) again what the recommendations are. So we would know from CSAC in April if we are going to make changes. And then late spring, early summer, we will start publicizing those, and then potentially adding in a grace period. So it could be as late as August of 2021 if major changes are done. So we might not actually

implement those changes until next summer. I just wanted to let you know what the timeframe is on that.

Now with that, just a couple of other little notes. I sent around an email a couple of days ago, and that's actually (Dave Merens)'s request that, is there any, and I think we need to talk about that first before we delve into the actual questions that we're going to vote on, but we basically wanted to know mathematically, do we have any flavor that, if we have one kind of testing, does that kind of imply or guarantee the adequacy of this other thing?

So in other words, I think as we stated it, if we know that we have score level reliability, does that mean we also know mathematically that we would have data element reliability? And obviously, if something like that were the case, then we wouldn't have to require that as part of their criteria; we could just, you know, be kind of (unintelligible) if you will. So we ask that set of questions, and I tried to do both permutations, and I also had some questions in there that gets to that kind of shortcut question, okay, if we know about data element validity, does that, you know, tell us something or do we know something kind of in turn about that reliability, etcetera?

So we had several emails back and forth that were starting to get to some of these things. And hopefully you were able to read them, and I think a few other people were going to respond via email.

But with that, I think I'm going to stop and I am going to hand it over, if you're comfortable, (Dave Merens), to go ahead and facilitate this discussion by the panel. And we'll - I'll roll us back to the slide - Slide 8, which is I think where we should start. But we probably want to start the discussion about the math and go from there. Is that okay with you, (Dave)?

(Dave Merens): Yes, I'm happy to do it. And I'm just trying to think how to weave the math discussion in. I guess all I would do is ask people to have this on the top of your heads. So the questions here are phrased in the sense of should we require. But just keep in mind, if there's one or more of these, well, I call them mathematical guarantee relationships, that would be a very strong argument for saying that we would not need to require something in the presence of something else. Now I'm not sure that validity is going to be the best example of that, and this is where we start, but that's okay.

And I'll just emphasize what Karen said, that if there are a set of absolutely basic mathematical relationships between one of our criteria and another such that the presence of one absolutely mathematically guarantees the presence of another, that's the basis for saying, okay, if you have the former, then you don't need to tell us the latter, because we know it's guaranteed. But let's see if any of those relationships actually exist.

And then I was going to say, when the slides came up first time, and I'm checking with Karen here, the word testing could be a little loaded here. I just want to clarify that when we use the word testing, we're not strictly meaning testing at the hands of the developer. A developer, for example, could bring forward data from existing published articles by other folks, establishing (validity of a liability), my sense, as we've behaved over the last couple of years, is that that is acceptable.

So, testing here means testing by somebody, not necessarily by the developer. Karen, that's your sense of the language as well?

Karen Johnson: Yes, absolutely the sense of it. And actually also reminds me too that we're talking only about, especially in this first slide, we're talking about what we at NQF call validity testing. So that doesn't mean, and as a matter of fact we

have the other parts of validity that get into risk adjustment, how missing data are handled, exclusions, all that kind of stuff. So, all that stuff remains. It's still a vital component whether or not something should be considered valid for the purposes of endorsement. So I'm talking in our conversation today strictly about the validity testing, which often is some kind of a construct validation, (unintelligible) validation, you know, typically a correlation of some sort, doesn't have to be, but that's usually what we see.

So I just want to make sure that that's what we're talking about. We're not trying to negate any, you know, demonstration of adequacy of risk adjustment, etcetera.

(Dave Merens): Got it. And also one last thing before we open it up. Just remind us, and I know this goes (unintelligible) from this one, the current state here is an or rule, is that correct? That for these measures right now, the requirement is one or the other, but not of?

Karen Johnson: That is the requirement for the regular structured process outcome measures.

(Dave Merens): Okay. So now back to Slide 8. Basically we're asking ourselves when we ask these two questions, do we wish to move from the current state? The current state is or, one or the other.

Karen Johnson: Yes.

(Dave Merens): All right. Happy to open it up for then discussion, whoever would like to lead us.

Gene Nuccio: This is Gene. Quick question about this. If it's a new measure, we're still accepting face validity?

(Dave Merens): That is the current state, yes.

Gene Nuccio: Well, even in Slide 7, which is the revised data, it has an - or face validity. For new elements.

(Dave Merens): Right.

Gene Nuccio: New measures.

Karen Johnson: Right. So this is Karen. So we still do allow face validity for new measures. So if somebody brings in face validity, we don't, you know, that just automatically, assuming that they did it in the way that we prescribed, which is it has to be systematic, you have to have experts, it has to ask the right questions, so there are some rules around that. But if it's a new measure, then that would be sufficient to - again to pass, if you will, the validity testing portion of it.

Again, all of those other things about validity, you know, risk adjustments, the missing data, exclusions, all those other things had to be (right) in there as well. But once a measure comes back for maintenance, what we say is that, at that point, we are expecting empirical testing, not only face validity. Now if we have a little loophole there still and we say, if you absolutely cannot do it, then justify to us why you can't do it, and quite frankly, we'll see if we buy it. If we did, then we'll let it go through.

The history there is about, I don't know, now four years ago, somewhere in that range, we did try to basically drop face validity for new or maintenance measures, just completely say we're not going to take it anymore. And we got incredible amount of pushback from the developers in the field. They thought

that that is, you know, asking for empirical validation is - was just too much. So we ended up backing up to the (accepted maintenance) version of it.

Now perhaps, you know, the time has come and we, you know, do we try saying no again to face validity? I mean, you know, I think we're willing to think about that. But again, that just had such pushback from the field a few years ago that we couldn't push it through.

(Dave Merens): And (David) again. I'd also note, just to try to informally summarize, but I was picking up some (plans to) support for face validity in the email exchange we just had in the last 48 hours or so, on the assumption that it's actually a rigorous formal face validity process, not just asking a couple of your friends.

And also would - just to clarify, and also to check, face validity is set in the context of score level validity. It's basically saying this measure, according to the opinion of 15 (international) experts, is a valid measure of quality of care at hospitals, or doctors or health plans or whatever it is. It falls in the measure score category, not the data element category.

Gene Nuccio: (David), this is Gene again. That was my question. I mean, if we say that we want score level validity, if it we say it's required, then are we bumping out face validity, a developer who offers face validity the initial - during the initial review given that they may not have had time to collect sufficient data on the score in the real world, kind of thing for us.

(Dave Merens): And that's why I wanted to be conscious about the word testing here, because, if for example our general view is that we should require some information about score level validity, a second branch of that recommendation would be whether face validity in a rigorous fashion would do the job. But establishing face validity doesn't quite meet my semantic (test) of the word testing. So the

words are going to be a little slippery here. But at least for purposes of framing the discussion, one might say as a branch of the proposal that face validity, if you can establish it in a rigorous way, is a way of establishing score level validity. And if we require score level validity as a concept, face validity could be a way to get there.

But we could say, yes, we want score level validity information, and face validity is not an acceptable way to get there. So there are a couple of branches on this one.

Sherrie Kaplan: This is Sherrie. I'm having trouble because I still think a lot depends on the stage of development of the measure. And you don't want to, you know, impair the development of new measures, especially in places that we don't have much information on quality of care like pediatrics and some of their - some areas in other specialties.

I think if a measure is in the early stages of development, face validity is actually (assessment) of content validity. Did you get the content right? But then if you're going to push it for the next purpose of measurement, now we're going to start comparing different units of the healthcare system one to another's using just "Oh yes, we got the content right," isn't going to do the job. Well, does it accurately reflect the performance of, for example, individual physicians or, you know, different purposes of measurement?

So I know Karen's reluctant to have this get too complex, but I think (unintelligible) into a different era of quality assessment now that CMS has upped the stakes. And I think that, you know, for example, and we just had this discussion, establishing testing for one purpose, or comparison of one unit, doesn't guarantee the results for another purpose or unit of comparison.

Jack Needleman: Right.

Sherrie Kaplan: So we need to kind of bound this a little bit and maybe (tier) it up with respect to stage of development and purpose.

Jack Needleman: Yes. This is Jack Needleman. Sherrie, thank you. By the way, before I enter into the current discussion, which is terrific, I think we all need to acknowledge a real debt of gratitude for all the folks who've contributed to that email conversation that preceded the meeting. It's an extraordinarily detailed, thoughtful, reflective and interactive set of emails that have certainly helped clarify my thinking about the content of today's call.

Getting back to where we are in the conversation, Sherrie, you're - I heard what you said and I don't know if you intended sort of (slipping) into the issue of reliability. Does it allow you to make distinctions between units?

When I think about the validity testing, the first question is simply, is it measuring what it claims to be measuring? And there's an element, at the data element level, do we - is what's being measured what's claimed to be measured? The reliability issue is, can it - is it being coded and recorded and collected with reliability? But is the data element, if it was collected, reported, and (math-ed) accurately, would it be measuring what we think it's measuring?

And I do think data element validity is required. In some cases, it can be established by face validity, if you've got a - if the measure is of readmissions and you've got a measure that says we're counting a readmission within 30 days, and you believe 30 days is a reasonable window, then the measure is inherently valid, because it's measuring what it says it's measuring.

When we move to score level validity, there are three kinds of ways in which something goes data element to a score. I think it's three, it may only be two. One of them is we're applying risk adjustments. And again, the same issue of, are the elements in the risk adjustment accurate enough to use, is a reliability question. But is the risk adjustment model correct for adjusting for the other reasons why we might see this outcome? That's validity question. And it ties to the risk adjustment.

The other place where we see score level, and you can only do that if you've got enough data to do it, but all of our risk adjustment models, including the first wave stuff, has included some empirical analysis of the risk adjustment model. So, somebody who's been able to get that data for those measures that are risk-adjusted.

The other kind of score level validity we've seen is where there's an instrument that consists of many individual components. Sometimes it's a survey instrument, sometimes it's a composite measure of three or four different clinical measures. And there the question is whether things aggregate up to something that you believe is measuring what they claim it's measuring. And I think a lot of the issues about whether we should allow face validity without a lot of empirics had been around those measures, where we've got individual data elements there being aggregated up, or an individual survey that's being aggregated up.

And on initial reading, face validity might be sufficient to let us move forward with the measure on that. I would argue that cognitive interviewing, a low data, low volume in kind of measure, is important for any kind of instrument-based data, and it's - you can make the argument that it went through a cognitive interviewing process, and yes, people ask the respondents what did you think you were answering here and how did - what influence - what your

answer meant, what did your answer mean, then I think we've got validity at the individual element item and potentially at the score level as well.

So I think we need to be clear about what we mean by score level validity and the different contexts. Risk-adjusted measures have one kind of score level validity issue. Instrument-based measures have a different kind of score level validity issue.

Karen Johnson: And this is Karen. Ron has his hand raised. So, Ron, do you want to either respond to Jack or have another issue or point you want to make?

Ron Walters: No. I've been sitting here, dreaming, we might have just heard the answer. But I've been one who has been, and we aren't into reliability yet, but on validity alone, I'm still not understanding, and I'm trying to think of a specific example that we might just have heard an element of that, is how data element validity and requiring that should not be a regular part because score level validity guarantees data reliability. I'm still having a hard time expecting that. I'm having a hard time not coming to the conclusion that is expected, but can be waived if adequate justification is provided.

So then my mind went to the face validity discussion that occurred for the data elements. And it's interesting, I think -- anyone can say if I'm way out there or wrong -- but we kind of take face validity as an argument for the data element validity, in the sense of a whole bunch of experts said that that data element does relate to what we're - the concept we're trying to measure. And we say, yes, okay, these are experts, that's fine.

And so that alone may be what justifies the expected but waived, if adequate justification is provided, it can be that simple. But I do not - I'm not one in favor of leaving it optional. I just haven't thought of the perfect

counterargument where we have an absolutely high level of score level validity testing (perform) and yet the data elements (are crap), because nothing has been done.

There's got to be one out there, I just haven't thought of it, and I think it might have been alluded to in the previous discussion. But nonetheless.

Woman: Can I offer an example? Let me attest - let me see…

Ron Walters: Sure.

Woman: …if I understand. So my understanding in claims data, oftentimes the diagnostic coding is inaccurate. But if you pull together enough sort of diagnostic information, the score itself may behave or move in such a way as to kind of get at the underlying construct. So I do wonder if claims data and diagnostic information is an example. It's a test question, not a - I'm not asserting that.

Because the errors are distributed randomly among providers and, you know, when you pull enough of it together, there's sort of a latent concept that's correct. Is that a possible…

Ron Walters: I have previously made many references to not only the documentation that's in the chart but also the claims - coding of that information that ends up in claims data. And it's those kinds of things that I am concerned about when we talk about both reliability but also definitely validity.

Woman: Validity, yes. I guess - yes. I'm thinking - mine is not really a validity statement. It really is not valid if it's not valid. Yes. Okay.

Eric Weinhandl: This is Eric. I mean, I think that the validity, if we think about claims, like that's the data source that I'm most familiar because I spent so many years dealing with it, I mean this last cycle, I think the standardized transfusion ratio (unintelligible) involves myself in the conversation about it. But, you know, you look at the analysis on the ecological level, and transfusion (unintelligible) poor outcomes. And so you say, oh, yes, this measure has great face validity.

The reason that that association exists is because transfusions occur in hospitals, not in outpatient setting, at least from the (unintelligible) patient population. And so there's obviously face validity. But then you actually look at the details of what the claims data reveal, and you realize that there's, I won't go into all the details of this, but there's lots of problems with the quality of the documentation of the data. And so, like, I've struggled with this in part because, if you don't have an in-depth knowledge of the datasets itself, you can be easily misled into thinking that there is face validity on the basis of some of the associations that were recorded in the measurement information forms.

But in fact, underneath it, at the element level or at the documentation level, there's significant problems with the data. And I think this is - maybe it's unique to claims, maybe it's not, but it crosses all sorts of domain. Readmissions are a great example of that, right? I mean, on some level, yes, readmissions are obviously - 30-day readmissions are very clearly documented in a Medicare client. You can count how many days went from discharge to readmission.

But if you know all the details of Medicare policy, you recognize that, well, providers may or may not have changed the nature of the claim. And so if there's observation (unintelligible) that are being documented as outpatient

claims, then 30-day readmissions rate goes down but the actual intervention or delivery of acute care hasn't changed.

And so all these details, they bother me a lot. I know I'm new to this committee so I'm not sure if this is (unintelligible). But I struggle with the fact that face validity seems so unpersuasive to me sometimes.

Man: I agree. And we'll get to reliability in a little bit, but everything you said was true.

Bijan Borah: So (unintelligible) I remember going back to that issue of -- this is Bijan -- going back to the issue of readmission. And I'm really confused. So, for readmission measure, Karen and others, can you sort of help me understand what - readmission is a very hard measure, you know, whether you have readmission or not, so, within 30 days. Now in that context, what could be an element - what would be the element - data element level validity and score level validity, and what (unintelligible) be looking at? For readmission.

Karen Johnson: This is Karen. I think what I'm hearing you ask, and tell me if I misunderstood, is, if people were doing data element testing for a readmission measure, is that what you're asking, they would…

Bijan Borah: Yes.

Karen Johnson: …they would show that, you know, they would show that, you know, the readmission that's on the claim actually matches, you know, the hospital records, you know, the medical records. They would hopefully be also looking at the data elements that are used in the risk adjustment model. So, basically the idea is the critical data element should be - that are used in the

measure should be (matching to) the gold standard, which is typically the medical record. So that would be what they would be showing.

Ron Walters: And I'm still…

Bijan Borah: Okay.

Ron Walters: …making a case, this is Ron, I'm making a case that it should be expected but it can be waived if adequate justification is provided. I'm not as - I'm not quite going to require, I could be require but we'll get a lot of pushback. But for exactly the question that was just asked, it would require not more than the sentence that Karen made as far as adequate justification.

(Dave Merens): This is (Dave Merens). One other thing, just give another example, essentially in support of where you end up, there may be data elements that are just so obviously valid that no serious human would ever question them. Now, whether death is one of those, I don't know. But it's - we could probably find examples where the data (unintelligible) level. There's just no serious doubt that it means what it says it means.

And I guess as I put my reviewer hat on, if somebody brought in and said, okay, here's the three key data elements and we claim that we don't need to actually do formal testing, because there's no serious doubt that they mean what they say they mean, and nobody's ever questioned it and they've been used for this kind of purposes for 50 years.

Would that be a sort of example of the direction you're trying to go?

(Dave Merens): Yes, I think when we get to reliability, we may talk about that (unintelligible).

Ron Walters: Right.

(Dave Merens): But for validity, I'm reasonably okay that the social security (dead) index, it measures exactly what it's supposed to do. Now if you start to say, does it measure it within 10 days? Does it measure it within - how real time is it, and so on and so on, and are there occasional mistakes? Of course, all of that's true. But it is the accepted standard for the definition of being dead, I guess.

Larry Glance: Hi, this is Larry Glance. I'm going to weigh in a little bit. I just want to start off with the question, which I think is, should score level validity testing be required, expected or optional? And I'm going to go out on a limb here, I'm going to say that it should be required. And I think that there are several components to this.

The first one being face validity. I think face validity is necessary but not sufficient. I think that there ought to be some measure, some consensus that the clinical experts, the clinical content experts agree that the measure is valid. So for example, when you consider readmissions, is that a - does that capture quality or does it capture access to care? So you need to first get that consensus that the outcome that you've chosen to measure is a valid reflection of quality.

And then moving on, I think that Jack gave us a nice structured to think about validity. I think there are several components to validity other than face validity. I think that there's construct validity, which is where you compare an existing measure to other credible measures. I think that's a little bit weak oftentimes because it's not always clear that the credible measures really measure what you think they're measuring.

I think that the key to looking at score level validity testing, once you've established face validity, is, at least in the case of risk-adjusted outcome measures, to look at predictive validity. To look at the risk adjustment model itself. Because if the risk adjustment model does what it's supposed to do, then you, in a really perfect world, you could, with a high level of certainty, predict what the outcomes for individual provider, be in a hospital or physician, would be conditional on their patient cohort that they're treating and conditional on the risk factors for those patients.

So then if you could predict the individual outcomes for those patients and you can compare them to the expected outcomes of that physician or that hospital in aggregate, then you've got a pretty good way of looking at the performance or the quality of that provider.

So I think that predictive validity, at least for risk-adjusted outcome measures, outcome measures, is really at the very heart of score level validity testing. And I don't think it's reasonable for NQF to endorse measures that do not have score level validity. I don't think it's enough to have face validity because, honestly, that's just too low of a bar. CMS is using these measures as a basis for redesigning the entire healthcare system? And it likes to use NQF-endorsed measures. And I don't think CMS should be using measures that are endorsed purely based on face validity alone.

So I would suggest that, one, yes, score level validity testing should absolutely be required. Two, that the basic components of that validity testing should be one face validity, two predictive validity, and also we should as well look at missing data and exclusion and other things that affect validity. But at the end of the day, yes, it should be required.

Ron Walters: I'm sorry, Larry, this is Ron. My comments were about data element (unintelligible) I agree with you about score level validity.

Sherrie Kaplan: This is Sherrie. I'm going to agree with Larry, most of what he said, but come back to the issue of face validity at an early stage of development of these measures. When they probably shouldn't be used sometimes of things you were raising wherein I totally agree about that. But face validity assesses content validity at some basic level and probably necessarily, as you said, insufficient.

But, and this is coming back to Jack, I was meaning discriminant validity, Jack, not reliability at the unit comparison level. If measures don't discriminate between the units that are being compared accurately, that's a validity issue, then they really shouldn't be used for that purpose.

Ron Walters: Fair enough, Sherrie. Thanks for the clarification.

Sherrie Kaplan: So, discriminant validity, Larry, I think - projected validity is a little dodgy because you have to be out there long enough to have longitudinal assessment empirically of that kind of - or depend on the literature for it. But I would think discriminant validity is what we're looking for in these comparisons at some basic level, in addition to other aspects of validity. But in terms of concept validity, I think if the measures don't discriminate, then they shouldn't be used for that purpose.

So I would say that, while I agree that the - I still am fuzzy about data element validity, but score level reliability, I'm going to come down on, it should be expected but waived if, for example, we're in the early phases of development of this measure, and there should be a qualification until you get to the

empirical assessment of score level validity, shouldn't be used for kinds of purpose - specific kinds of purposes. And I'm probably (unintelligible).

(Dave Merens): Yes. (Dave Merens) here. One thing, if I could suggest a possible compromise between Larry and Sherrie here (unintelligible) required but (unintelligible) early phase, which maybe the initial (submission) process, face validity, subject to certain stipulations, may be an acceptable way to establish score level validity. It is required (unintelligible). But that we move as quickly as possible (to test) it.

I find myself (unintelligible) Larry in saying all these measures are going to be endorsed and used out there to switch millions and millions of dollars (unintelligible) the bar should be pretty high. I'm also sympathetic, to Sherrie's point, we've heard from developers that (unintelligible) early on in the process where it'll be chicken and egg, that if you don't (unintelligible) to be used, sometimes the (unintelligible) score level.

So (unintelligible).

Larry Glance: Let me just come back a little bit. I think sometimes Sherrie and I differ because we have different - we use terms somewhat differently.

So when I talk about predictive validity, I'm saying that you have a risk adjustment model which allows you to predict the individual outcomes, conditional - for a patient conditional on their risk factors. So that does not require longitudinal data per se. It requires a database that was used to develop and validate the model. And I don't think measure developers should be coming to NQF with risk-adjusted outcome measures that - where they haven't done anything to validate the model. So I think that that should be an absolute requirement for risk-adjusted outcome measures. I don't think these

things should be released out in the wild and be used for reimbursements and pay for performance and accountable care and all that stuff, without having been validated, without having been shown to have, A, face validity, and B, predictive validity.

Sherrie Kaplan: That's a great point, Larry, because I would call that discriminant validity because it discriminates between people who are sicker and less sick. So, you know, we are probably talking same and similar language but in different kind of situation for what constitutes, you know, the groups we're studying.

Larry Glance: Thanks.

Sherrie Kaplan: Karen, this is Sherrie. I have to sign off, but thanks for a very lively discussion as usual. And bye, you all.

Karen Johnson: Thank you, Sherrie.

Sherrie Kaplan: And happy holidays.

Karen Johnson: Happy holidays to you.

(Dave Merens): Karen, again I just noticed (unintelligible) we may not (unintelligible) have on this…

Man: Hey, (Dave).

(Dave Merens): Yes.

Man: You're kind of - your audio is a little weak. I don't know if it's because people aren't on mute or maybe you're running low on battery or something.

(Dave Merens): (Unintelligible).

Man: I don't know to others, but it's coming across to me as a little kind of fuzzy.

Man: Same for me.

Man: Same here too.

Karen Johnson: Yes, (Dave), you were kind of coming in and out, so I don't know if you maybe were walking around or something. You want to try again?

Man: I think he's switching phones, he said.

Karen Johnson: Ah. Okay. Okay. It looks like maybe Paul has - Paul, do you have your hand raised? If you want to jump in to this while we're waiting on (Dave) to hop back in.

Paul Kurlansky: Just want to support what Sherrie was saying in that, when you first introduce a metric, it may be very difficult to establish validity. There may be tremendous face validity to it. I'm thinking something as simple as, you know, mortality. You know, I don't think you need to get to go far to establish the, you know, the validity of mortality as a measure for quality in cardiac surgery. However, you know, if you were presenting NQF with a new mortality measure, then you would not have validated it other than through face validity.

So, and even with the recent mission, I think the (SDS) got into a sort of a bind this year because they had a composite measure which was - it was predictive, so it did fulfill that capacity. But the validity of it was, I mean the

initial submission was based on face validity. And then once you resubmit it, there's nothing else out there that corroborates it or correlates with it because it was a unique and innovative metric. And so you're kind of - you can't use face validity for resubmission. So it becomes like a catch 22.

So I think that there is still a role for face validity in terms of particularly when you are presenting something that is new and that has serious face validity.

Larry Glance: Paul, can I just respond to that a second? I think that there may be a little bit of confusion. So I think that when STS develops measures, risk-adjusted outcome measures, and they (unintelligible) out there in a peer-reviewed literature, they don't just talk about face validity, they also talk about predictive validity. They validate the risk adjustment model.

And what I'm saying is that, at the heart of score level validity is how good is the risk adjustment model. And so what I'm saying is that, when measure developers like STS submit new measures to NQF, they should include information on, A, face validity; B, predictive validity; C, how they handled missing data. I actually think that the construct validity is a relatively weak type of impure validity testing. And I would think that that should be optional.

But the other pieces that I just mentioned I think are very important. And I don't think measure developers again should submit measures to NQF for endorsement without some type of information on predictability, how well the risk adjustment model works, in other words.

Karen Johnson: (Unintelligible) this is Karen. Just a couple of things that I want to make sure that we're all on the same page on. We do ask for, you know, discrimination

and calibration of risk adjustment models for those outcome measures that are risk-adjusted. So I completely agree with you to be good, but I think we're talking about more along the lines of the concept validation potentially, or some things like that. And I do want to remind everybody that, you know, even if we're looking and being very careful, which I think we try to be on the risk adjustment for the outcome measures, we do get a lot of measures in that are structure or process measures.

So maybe the discussion needs to, you know, think about a process measure and think about construct validation or something along those lines for process measures. You see what I'm saying, Larry? We have to think about the process measures and structure measures as well.

Larry Glance: In other words…

Karen Johnson: Because they won't have the…

Larry Glance: But Karen, I thought that the primary goal of our methods panel is to look at complex measures, primarily risk-adjusted measures, not really process measures. And I also thought that NQF is trying to get away from process measures. But again, I thought that our committee is primarily lenient on complex measures, not process measures. Am I wrong?

Karen Johnson: Yes - well, sort of. The methods panel is helping us to evaluate those complex measures. So, absolutely the ones that you guys get are those outcome measures or (unintelligible) measures, composite measures. But you're also acting as an advisory group to NQF for the criteria overall. So I'm actually asking you not to think just about the measures that will be coming under your purview as you do the evaluations, but to think about all of, you

know, all the measures that come in to us, and help us think there's a criteria for everything.

Larry Glance: So what I would say then is I think that construct validity is a valid way of evaluating the validity of process measures. But I think risk-adjusted outcome measures, I think predictive validity is much, much more powerful than construct validity.

Karen Johnson: So, what would you say if you're thinking about, let's just - what would be your opinion on required versus expected, versus optional for a process measure if they're doing construct validity? Would you still like to see that required? Would you be okay if it wasn't, if they did just data elements? Or how would you feel there?

Larry Glance: So I think for process measure, that you should require construct validity. It's not enough to have face validity. I think that for a risk-adjusted outcome measure, you need - I would not require construct validity, because it's never clear to me that the new measure needs to necessarily correlate with existing measures because those existing measures may not be as good as the new measure. So I don't think construct validity is a very powerful way to establish the validity of the measure, unless you have to do it that way, which you would for a process measure.

(Dave Merens): (Dave Merens) here. I'm wondering if we could slide a little bit back to what I think is a more basic question. It seems to me that we may be overshooting the mark just a little bit by distinguishing the different types of validity. The question in front of us isn't exactly which type of validity, it's whether validity testing at all should be required.

I'm inclined to think that we could spend a long, long time weaving around through the various types of validity when in fact, when measure developers come to us, they occasionally bring different sorts of validity information to us that may be tailored to the circumstances and the type of measure they're bringing. I just don't know if we're going to get to solid ground here with focus on specific types of validity where I think we may be actually close to some agreement on the simpler, more basic question of yes/no, should score level validity testing be required. I'm hearing a lot of yes on that.

Jack Needleman: Yes. Yes. This is Jack Needleman. I'm also hearing a lot of yes on that. The one other thing I've been struck by in the conversation is what we mean by validity. And that one level, and this is where content validity is particularly relevant, is the question of, is the measure measuring what it claims to be measuring? And then the second element, often this comes up in reapplication, is, is it measuring quality? Is it a valid measure of quality? And I think in dealing with that second question, we've seen a lot of efforts to correlate measures with other measures that are accepted as measuring quality. And if it's correlated, then it must be measuring some element of quality.

And I think the element of the conversation that we haven't had is whether that broader question of not "Is the measure measuring what it says it's measuring?" but, is the measure measuring quality is within our purview, or whether that's a (unintelligible) in committee's judgment related to importance and usability.

(Dave Merens): Yes. Jack, (Dave) here again. I strongly support what you just said. I've been feeling all along and almost put it in one of email chains earlier that it seems like in our hands the key validity question is the second one, does this measure quality?

Karen Johnson: So, a couple of things. This is Karen. A couple of things that come to that. I think we need to talk about the quality thing a little separately, and that was what we had teed up for the next conversation. So I think we definitely want to get there.

I guess a question for you guys though is the question about does it measure quality. Statistical testing, I mean, isn't that more of a conceptual question? You know, whether - no matter what kind of testing you do, that's not really going to answer that question, is it? And that's a question I'm posing to you guys.

So, is that more of (unintelligible) question and not necessarily going to be answered by the result of some correlation or what-have-you (unintelligible)?

Man: Well, I think…

Jack Needleman: Karen, one of the things I've been struck with -- this is Jack again -- one of the things I've been struck with on the resubmissions has been the effort to find some correlation with some other measure. And that becomes a validation of the measure. That clearly is a validation that's measuring quality, not that it's measuring what it says it's going to measure.

And we've seen some weird stuff there on the last go-round. There were correlations with (unintelligible) and the 0.05 level, statistically significant because the big samples, and we have the developers telling us, well, it's statistically significant, it doesn't have to be a large correlation, so long as it's statistically significant. And that just struck me as fundamentally wrong. But my intent here was find the measure that you think is measuring similar things that are considered measuring quality, and find a correlation that's high

enough to convince you that they're measuring comparable things, if not the same things, with enough (unintelligible) difference that this measure is useful, separate from those other measures.

Karen Johnson: And going to that question, sometimes it's really hard for developers to find something to correlate with, and I think sometimes we get these really weak demonstrations, if you will. And again that's going to be one of the things that we want to talk about a little bit later in the call or our next call (one). But I think that is what makes me hesitate with the making score - making score level testing, validity testing required.

I think what we'll end up getting, and this is just my opinion, you guys can take, you know, take it as you will, I think what we'll end up getting if we require it, is these really weak tests that maybe, you know, check our box but don't really tell us much. And maybe that's okay. But I do want to (show) that up here.

((Crosstalk))

Alex Sox-Harris: This is Alex. Go ahead.

Man: Go ahead.

Alex Sox-Harris: So this is Alex. For an outcome measure, first, I just want to emphasize, I think it came up in the email chain that, the way I think about these issues for outcome measures differs from process measures. So I think it might be a mistake in the rubric to lump those two because to me it curbs my ability to think well about the different situations. So, talking about outcome measures, what I want to know is that the outcome is being measured properly, call that element level validity. Sometimes that's obvious, sometimes that's not

obvious. Sometimes it looks obvious that, I think it was Eric said, when you poke on the data, you find out you'd been misled. So I think we have foundational, as Larry and others have said, that the quality of the - or the accuracy of the risk model is essential.

And an idea that (Dave) raised, a method that he suggested, on those email chains yesterday was another thing that would be persuasive to me is, if you had risk-adjusted outcomes and then you check if process quality explain some of the outcome variance, so, in other words, things that are happening during the course of care are actually related to the outcomes that are observed, the adjusted outcomes. That would be interesting and persuasive.

What I don't find persuasive is, as have been recently mentioned, these, you know, correlations of outcome measures with other related outcome measures. That's just not helpful. But the things I mentioned previously I think are all important ingredients to me having confidence in an outcome measure.

Karen Johnson: Okay. Other thoughts?

Sean O'Brien: This is Sean, and since no one else is talking, I guess I'll try to make a few points.

One is just when Larry was advocating for requirement of predictability, I thought like I didn't know what he meant by that, and it turned out that Sherrie and Larry were, you know, using the same words and meaning something different. And that's just an example of why (unintelligible) the language of reliability and validity is really not very helpful for me. So I wish if when we say, you know, if we (unintelligible) risk adjustment, just a risk adjustment (unintelligible) concrete, critical issue that recurs all the time that we ought to

always be looking at, thinking about. I think the evaluation criteria call attention to that issue. But I just, for me, I don't find the language that helpful.

So that's why, if I was going to vote on should score level voting testing be required, I'm going to say, no, less than optional. Even though I heard the testing can be interpreted broadly to refer to prior literature. I just think it's such a broad, undefined concept to say what is score level validity, that I'm not in a position to say, yes, I really think that needs to be tested because I fundamentally don't know what we're talking about and I don't know what it is.

And I do think it's probably more distinguished in cases where a measure, at least on the surface, is the thing that it's literally measuring, is something that on the surface is inherently interesting and meaningful in itself, compared to cases where there's always an extrapolation from the thing that's literally being measured, to some other proxy. And I think in the cases where, you know, you're expecting to make some leap from the literal interpretation, some other thing, you know, those number of claims for condition X you see is going to be (unintelligible) or something like that. In those cases I think we ought to always be looking for evidence, you know, looking for validity issues and looking for evidence to help address them.

But I just think that the really interesting and critical issues are the ones that take, you know, turning on your brain and thinking about, and the critical thinking and analysis go a long way, and that just leaves (unintelligible) statistical analyses that just, you know, address some correlation, usually just don't get at the heart of the most critical methodological issues for the measures.

I think, you know, instead of saying, "Should score level validity testing to be required?" I would say, should score level validity be required? And I'd say absolutely.

Karen Johnson: Yes, okay.

Sean O'Brien: I think that an informal process that uses, you know, critical arguments and analysis and also gives a little flexibility for situations where no one in the world is questioning an issue, doesn't necessarily benefit from having additional statistical analysis of that issue, I think if you require, all the reviewers would be convinced that the measure is (unintelligible) (important issues), but I just don't think it should be "Yes, we expect score level validity to be done" when we don't even always even know what we mean by that.

Karen Johnson: Thank you, Sean. I think you're right, we still are a little bit off in terms of our terminology. So, (Dave), do you think it's worthwhile to go ahead and do the voting and see where we are? And let me make sure that I make it very clear, these votes are really just to help me understand where people are in their thinking, not really (mining) in any particular way.

(Dave Merens): Yes. I think, I suggest we do that. I'm watching the time and realize (unintelligible) equally rich discussion on reliability, we haven't touched yet (unintelligible) interesting optional things that we probably won't get to, so. And as you say, this is to get a sense of where we are right now. There may be some email follow-up we could do and - when we see where we are.

Karen Johnson: Okay. And (Hanna), you're going to help us with voting. She's bringing up the screen right now. And we're going to start out with the question about score level validity testing. So again this is, you know, in addition to any

checking of risk adjustments, adequacy and all that kind of stuff (unintelligible) that additional testing.

And right now when we talk about face validity, we are talking about score level, thinking about the score and whether or not the score can differentiate different score quality. So right now face validity would be included as part of an option. So we're not trying to vote on whether we want to keep or delete face validity, okay?

So, go ahead and cast your votes. While you're doing that, I'll give a few minutes. Let me see how many I expect.

I'm probably looking for, if I counted right, I'm looking for about 17 votes. And I can't tell - can you turn in, the numbers, just so we can see how many, right, voted.

Bijan Borah: Hey, Karen, this is Bijan.

Karen Johnson: Yes.

Bijan Borah: So this also includes the situation of new measures, right?

Karen Johnson: I'm sorry, can you repeat that?

Bijan Borah: This question does include situations where measure developer is coming up with the new measures. Correct?

Karen Johnson: Yes. Yes. Right now we're not really trying to differentiate between new versus (maintenance).

Bijan Borah: Okay.

Karen Johnson: Yes. Which makes it even more complicated, right?

Bijan Borah: Yes.

Dave Cella: Karen, yes, this is tough, but this is Dave Cella. Is there a vote on the screen? I'm looking at the questions to discuss slide. Is it supposed to be a vote on the screen?

((Crosstalk))

(Dave Merens): There's a link in the email she just sent that takes you through…

Dave Cella: Oh. Oh, yes, okay. Yes, you mentioned that earlier. Okay.

Man: It's in a PDF.

Karen Johnson: And we know Dave is still working on it. And I can't tell from the screen. Is there any way you can share this with the audience, so they can see where (unintelligible)? We are very fortunate to have (Hanna) helping us out today. I don't know how to run any of this stuff. So, (Hanna) is doing this for us. She's trying to pull it up on the screen so that everybody can see where we are.

Man: Your assessment of your ability was extremely reliable and valid.

Karen Johnson: Yes. As (unintelligible).

Man: I'm getting a screen that says (unintelligible) just hang in tight, you're ready to go. Nothing there.

Man: That doesn't sound very reliable.

Man: It's (unintelligible). And I'm logged in under PK and it's…

Man: Try refreshing the screen.

Man: Yes. That was there until she opened up the testing. So you might have been logged in a little early or something.

Man: So when we do vote, it just sits there on the screen until you clear it…

Man: Right.

Man: …say you're done, and then we clear? Okay.

Man: That's what I'm seeing. That's wrong. It might - it's done and I'm just waiting. But yes, I voted.

Karen Johnson: Okay. That's okay, (Hanna), don't worry about it if you can't get it up on the screen. Can you tell how many people were voting?

Man: To "Response recorded." "Response recorded." Tells you you're good.

Man: Yes.

Karen Johnson: Okay. Okay. Good. And can you tell how many (unintelligible) has voted? Do we have something close to…

Man: Fifteen.

Karen Johnson: Fifteen. Okay, that's pretty close. Okay.

So I'll read this out. We're not quite sure how to get that on your screen. So, of the 16 votes that we captured (unintelligible)…

((Crosstalk))

Karen Johnson: …this is for score level validity testing, seven voted for required, six voted for accept it but waive, and then two suggesting (unintelligible) is optional. So I think we still are not at complete (unintelligible) and that we're probably not quite as (unintelligible) so it looks like people are suggesting that we try to maybe strengthen our requirements a little bit. So we'll keep going as Dave suggested in terms of some emails back and forth, etcetera, to try to maybe get a little bit more clarity. But that's where we landed. So, seven required, six expected, two (unintelligible) optional.

Man: Well, if we went with the electoral college, required would lose even though it got the popular vote.

Karen Johnson: Oh. Dave has spoken (unintelligible) sort of he expected that waive, right? Expected that waive.

Let's go ahead and try the data element one, because I - we had that discussion a little bit earlier today, and let's see where we are. So, should data element validity testing be required, expected to waive with adequate justification, or (leave) as optional.

Gene Nuccio: Just a quick question here, we're still talking about - we're not talking about instrument or any of the other ones. We're talking about just that first column, if you will.

Karen Johnson: Yes, absolutely. Yes.

Gene Nuccio: Okay.

Karen Johnson: They're both required for instruments.

Gene Nuccio: Right.

Karen Johnson: Yes.

Man: Yes, that's an important distinction. Thanks for bringing that up. That was Gene, right?

Gene Nuccio: Yes, it was.

Karen Johnson: Yes. Yes, thank you. And are we close to our 15 people yet, (Hanna), can you…

(Hanna): (Unintelligible).

Karen Johnson: Oh, okay. We can just leave it as (unintelligible). Okay, 14.

(Hanna): (Unintelligible).

Karen Johnson: Okay, we've got our 15 people again. So, for data element, we have data element validity testing. We have three think it should be required, 11 think it should be expected but potentially waived, and one would like to leave it as optional. Okay. So, much more consensus on the data element validation. Okay. Thank you guys for that. Let's go now to the discussion about reliability. And let's see, I just advanced the slide, but we're waiting on (Hanna) to walk us into a regular slide deck.

(Hanna): (Unintelligible).

Karen Johnson: What's showing on the screen?

(Hanna): (Unintelligible).

Karen Johnson: Okay. Yes. Okay, good. So, questions to discuss. Revisiting the recommendation regarding score level reliability testing. So this is harkening back to our October discussion. I think we did come to consensus in that meeting, that we would like to require score level reliability testing. So if that is still the case, if people still feel that that's the case, would you be willing to wait until maintenance for that? So in other words, not necessarily require it for the brand-new waivers - brand-new measures, but wait until maintenance. Or would you be willing to grant a waiver for the…

Man: I think you want to advance - can you advance one slide? I think. That was two. Yes, there you go.

Karen Johnson: Yes.

Man: Okay. Thank you.

Karen Johnson: Yes. So I guess first of all, does anybody, is there kind of - any kind of ground swelling of resistance against requiring score level reliability testing? If not, then let's just talk just briefly about, you know, are we willing to wait until maintenance, and are we willing to grant a waiver?

Larry Glance: Hey, Karen.

Karen Johnson: Uh-huh.

Larry Glance: This is Larry. It seemed like - it was interesting to me that we got consensus so quickly on data validity testing. Would it be okay if I just address that just for a second?

Karen Johnson: Sure.

Larry Glance: So, many, many of the measures that NQF evaluates for CMS are measures that are based on administrative data. And I think that before we require data element validation, which essentially requires measure developers to go back to the medical record and re-obstruct the medical record as the authoritative medical - as the target source, and look at the agreement between the administrative data, the ICD codes, ICD-10 codes that have been coded, and the medical record. That is a very, very resource-intensive process.

Karen Johnson: Yes.

Larry Glance: And it seems that we really haven't discussed that at all. We have never - as far as I can see, the measures that I've reviewed, the measure developers typically do not address data element validation. And primarily I think it's because of a resource issue, because of feasibility issues. So I think we ought to have a little bit of a discussion around this.

\ I'm not saying that we shouldn't want to have valid data. I think valid data is a good thing. I just - I'm questioning the feasibility of it for many of our administrative - for many of our measures that are based on administrative data.

Jack Needleman: Yes. This is Jack Needleman. You know, the CMS measures that have come in have all had pretty standard stock language in it in which they - I can't quite pull up the specific one right now, but they basically say we had not looked at the individual coding of this data element for purposes of validating this data element in this measure. However, CMS has this process for validating the coding of its claims, which is the major source of the administrative data, validating its claims through - by going back and looking at the coding of the - in the original patient record and how it's been claimed and (billed) to us. And we're relying upon that as the check on the data.

And for the most part, that language has been acceptable. It goes back I think to the argument, the discussion earlier about whether, if there is a literature, something else that one can point to, or process, independent of the measure, that should lead to competence in the data used in the measure, it would be acceptable. And that's what people have been doing with the CMS measures, and by and large I think we've been accepting that.

Jen Perloff: But that's exactly wrong because that validation is, that auditing, is done for financial validity. It is not diagnostic coding validity. That's a secondary issue in those audits, having been a participant. And so that justification is like fingers on a chalkboard to me, because that actually does not at all get at the issue Larry raised.

I think to truly validate claims measures for diagnostic information is incredibly labor-intensive and costly. And I've only seen one measure developer do that correctly. It may happen to be at a medical school with great access to EHR data. So that one drives me crazy, because the audits have a very specific financial goal. And the measures are not often financial.

Ron Walters: So this is Ron. And that's why I felt slightly differently about reliability than I did about validity but I ended up at the same place, for all the discussion. And I realize that I'm not sure how reliable the data (unintelligible) are either. But the justification we'll get, I expect it, but waive if adequate justification provided, as opposed to the other three options. And I'm sure the statement that we'll get most of the time is the same statement that was said earlier, that we probably will become almost boilerplate as time goes on. And it'll be up to the committee again as they review each measure as to whether to accept that justification or not.

But even more so than for validity, the amount of effort that would take to assess the reliability would just almost totally show off - shut off measure development for a while.

If we made it required, and yet I can't go to leave as optional. I just can't get to - can't get to optional. So, you know, that's my opinion.

Larry Glance: It's a really difficult issue because it really goes to the heart of the measure development process. And I think the validity issue also, you ought to think about it differently for when you're talking about the outcome versus the diagnostic codes that are used for risk adjustment.

So, for example, if your outcome is a hospital-acquired infection, I would think that one would really want to carefully look at whether or not the ICD

codes validly or accurately capture clinical infections. And so there, I think that measure developers should make some effort to go back to the medical record as the authoritative source and re-obstruct those medical records and see, look at the level of agreement between the ICD codes that are being used to code infections versus the - what you're getting when you actually look at the medical record itself.

I don't feel that strongly about looking at the data validity of diagnostic codes that are being used for risk adjustment. I think those are still important, but I, again, for the reasons that we've discussed, I don't think it's feasible to look at all the diagnostic codes. But I think in the case of the outcome itself, I think we should consider whether or not we should ask developers to validate the outcome data element if it's being taken safe from administrative data.

Dave Cella: Yes. I - this is Dave Cella - I guess I thought of administrative data as - I was one of the middle voters that as - an explanation that would be satisfactory for not providing reliability data.

Karen Johnson: Yes.

Dave Cella: So, yes, I think - and maybe it could even be stated explicitly that this does not include medical chart obstruction to test through the accuracy of the administrative data.

Karen Johnson: So this is Karen. In the past we have - and I agree with you, I think that would probably be, you know, the votes on data element validity did come out as expected, not required, so the waiver might be exactly what Larry is getting to with the claims data and the (burden), and Ron as well.

We have - so, just a couple of things. In the past, every now and again we do get people who have gone to the literature and there have been studies where people have looked at specific, you know, ICD, well, 9 in the past, you know, codes to see if they're really reflecting in the claims what's in the medical records.

So, people on occasion those happen, and maybe for some of those outcomes, data elements that Larry is kind of referring to, you might have more luck finding those actually in the published literature. So, you know, it could be out there and people should probably be encouraged to look.

I think to Jen's point about the auditing, what I will say is, I know that we do have developers who talk about that and they say, you know, it's audited and therefore, you know, they're valid. That statement on itself, even now with our requirements, wouldn't stand on its own, because we would have to, you know, we would actually need to see the results of that, not just a kind of a blanket statement.

So, you know, I think the thing you're talking about chalkboard, I get the frustration with that and, you know, they would have to actually show (unintelligible). So usually what they do is they say that but then they also show score level validation or something along those lines as well. So they still…

Woman: Yes.

Karen Johnson: …you know, can do both. But if that's all they said by itself, that would not meet our requirement.

So I think, Larry, what we should do is, with the data element validity vote and where we are, we could have either some additional discussion about what might be a reasonable justification, you know, if we wanted to lay it out, or we could just let people, you know, say what they will and we could talk about it later. So I think maybe we could do that maybe in an email exchange or something along those lines.

So if it's okay with you, let's go back a little bit and talk about score level reliability testing and just whether or not people are willing to wait until maintenance or willing to grant a waiver with adequate justification.

So I think this is really thinking about, now that I read the question (unintelligible) I think there's probably a hole in my logic (unintelligible). And if it's a new measure coming in the door, you know, would you have to have score level testing for that new measure or are you willing to wait till maintenance? Or maybe the other way is to say, we'd be willing to grant a waiver, and maybe in your mind you might be more willing to grant the waiver for a new measure and less willing for maintenance measure. So that's kind of what I was getting at. And I don't know if people have any thoughts on that.

Man: Karen, just real quickly, what would be the rationale for waiting until maintenance if this is an important concept?

Karen Johnson: I think the only rationale might be that score level reliability maybe could take - you have to have enough providers and enough patients I think to get a reasonable estimate of reliability perhaps, and that's a question for you guys. And sometimes people develop measures and they're starting with a pretty small number of medical records and patients and that sort of thing. So they

would do typically a data element kind of IRR analysis with their 300 records or whatever they've got. And that might not cut it for a score level analysis.

Dave Cella: Karen - sorry.

Karen Johnson: Yes. Okay.

Dave Cella: This is Dave Cella. I'm just doing a sort of a time check here. It looks like we only have about eight minutes until we're supposed to do public comment and then wrap up. So we want to make sure, if there's anything you want to be sure to get voted on it done, we got about eight minutes.

Karen Johnson: Thanks, Dave. And why don't we just plan on not getting to the data element reliability testing, we'll not worry about that today, and let's just think about the score level reliability testing and, yes, the justification that I just threw out, that may hold water, it may not. So if you guys want to talk about that.

Jack Needleman: So, Karen, this is Jack Needleman. Just real quickly, I'm - I'd be willing to wait on some measures but not others. The CMS administrative data measures, they've got the data to do score level testing, even as they're proposing the measure the first time.

While we had an example of - in one of the last rounds, and I forget which one, where people had done a lot of development on the measure, they wanted to get indoors, but they did not have the large-scale data yet to demonstrate, you know, the score level reliability in differentiating (unintelligible) units because (unintelligible).

So those are different circumstances. And I think the decision about whether to wait on score level reliability measures should depend upon whether the data is in fact enhanced in large enough quantities to do the testing.

I'm with Sherrie, I don't want to discourage new measures where - on the basis of data demands that are unrealistic. But that certainly doesn't apply to some of the folks who are proposing measures.

Ron Walters: Yes. So this is Ron. Why wouldn't that be just the grounds for a waiver? And we still require score level reliability testing. The rule is you don't wait for maintenance, you expect it at the start. But you are willing to grant waiver with adequate justification. You just gave a perfectly good one.

I'm trying to put the logic together in my mind of, especially if this happened to be a pay-for-performance, it depends a lot on the usage or comparative type measure, that we're going out there and saying to people, "Yes, this measure is important to use, but we don't know it's a reliable measure." I mean, it just doesn't - that just doesn't sit well with me. But perhaps that's more in the adoption and application of the measure than the measure itself.

But I still think we can accomplish the same thing to say that we do want score level reliability testing, but we'd be willing to grant a waiver with adequate justification. When it comes around again, we'll see what work is done.

Larry Glance: So this is Larry Glance. I'm going to be a little bit black and white. I think that we should require score level reliability. I don't think we should be granting waivers. I think that when NQF endorses a measure, it can then be used for public reporting. If a measure is not reliable, I don't think it should be used for public reporting.

Ron Walters: Well, it does have a certain logic to it.

Man: Agree.

Man: Karen, do you want to get a quick vote before we run out of time, just to get a sense of the room?

Karen Johnson: Yes. Let's get a sense of the room. Let's get a sense of how I wrote that voting question. I think it's exactly like that, so it might not - Larry, you might be having a hard time…

Ron Walters: Yes. I don't know how to answer…

((Crosstalk))

Ron Walters: From what's on the screen, I don't know how to answer it.

Man: How do you answer if your answer is no?

Ron Walters: Yes.

Woman: (Hanna) can create one.

((Crosstalk))

Man: Because you either have to be willing to wait until maintenance or you have to be willing to grant a waiver, and we just - there's a third option that's missing there.

Man: Yes.

Woman: Yes. So, (Hanna) added a no.

((Crosstalk))

Man: There you go.

Woman: Yes, that was fast too. She learned how to do that. The no means that we wanted, and we're not willing to wait, and we're not willing to (grant) a waiver. You got to have it.

Man: (Unintelligible) no is if you agree with Larry's position, I would say.

Man: Got you.

Woman: (Unintelligible)

Man: But Larry, you can vote.

Karen Johnson: And we're probably still looking for about 15 people, and (Hanna), if you could - (Hanna) likes these percentages, I tell you.

Larry Glance: By the way, I did not take any money from large corporate sponsors.

Karen Johnson: Okay. We've got…

Larry Glance: But I will accept grassroots.

Man: Yes.

Man: You know, you're getting free healthcare, which is kind of emolument.

Man: Yes. Well, you read all the evidence.

Man: I'm confused (unintelligible) no response. This is John.

Man: No is you don't like either those. You want to require the reliability (unintelligible).

Larry Glance: Basically you're positioned with me if you say no, is that the reliability is required without - pretty much without exception. It has to be reliable…

((Crosstalk))

Man: Thanks.

Karen Johnson: And it looks like we have a person who was able to vote this time that wasn't. So we're at 16 votes, and zero…

((Crosstalk))

Karen Johnson: Yes. Not quite sure what happened there. So, nobody is willing to wait till maintenance. Five people are willing to grant a waiver. And 12 people -- oh, we're at 17, okay. So I hope that - is that working right here? Just to make sure. I had counted 17 but we only had 15 votes on the other ones. So again, this is - we're not going to, like, change the world with this vote. This is just giving us a flavor.

It looks like 12 people with Larry, and they want it and they want it now.

Woman: I think that should be okay.

Karen Johnson: So, no waiver. Okay. So we think that's pretty reasonable. So it looks like a two-to-one margin of requiring it versus expecting it versus - and granting a waiver potentially.

Man: Okay. Now that Larry has a posse, we're at the 10 minutes to the hour. Did you want to try to squeeze something else in or go to public comment?

Karen Johnson: I think we should go to public comment because I know us and we're not going to get anything settled in five minutes, for the next one, I don't think. Do you think we would get anything settled and be able to vote on the next one, the data element reliability?

Man: No.

Karen Johnson: Okay.

Man: I vote no.

Karen Johnson: Oh, well. And these two hours has gone by so fast for me. I don't know about for you guys.

Man: We could follow in email, maybe get a sense on email.

Karen Johnson: Yes. We'll try to do that. We'll definitely try to do that. Okay. Why don't we - before we hand it over for public comment, does anybody else from the panel have anything, kind of burning issues or things you want to either reiterate or have us think about via email or on our next call?

Okay. Hearing none, let's go ahead and open it up for public comments. Do we have anybody from the public who would like to weigh in, either with just comment in general or comment on where the (sailboats landed)?

And you could, if anybody has a public comment, you could do it via phone or via our chat function.

Okay. Hearing none, I think we will go ahead and wrap up our call for today.

So, what are our next steps? Well, our methods Webinars, we have a poll that is forthcoming, and forthcoming, (Hanna) is ready to send that out right after this call, I think. Right, (Hanna)?

(Hanna): Sure.

Karen Johnson: Great. So in this Doodle poll, it's quite scary to look at because there's a lot of (8's) on there. And what we're trying to do is get people's availability for the full year of this year. And we had, and John actually had asked us about this, and it's a valid point, we had said that we wanted to go to bimonthly Webinars with you guys rather than monthly, and they would just be longer, two hours instead of what.

What we have realized is that we don't want to have Webinars with you in the months that you are either coming to us for in-person meetings or you're doing the evaluations of the measures. So what we're going to do is we're going to backtrack just a little bit and, rather than asking you to do bimonthly Webinars, we're going to ask you to do Webinars in the month of January. And then I think in the Doodle poll would tell you for sure, I think maybe May, June and July, and then we'll do one in August, and then another one in

December, or something like that. So we still have six but it's not going to be like every other month. It'll be kind of in the months that you're not doing a lot of other work with us and for us.

So you'll see that in the Doodle. And also in the Doodle is a couple of sections, and I think it's for late March, early April, and then also in the month of October. And those, the Doodle is asking for the full day. So we're trying to come up with days for our in-persons. And it'll be, like we had this past October, it would be a two-day in-person meeting. So we look forward to hearing your feedback on those.

We still have our methods panel mailbox you can get a hold of us. We will be following up on some of these things via email. And then in January, probably towards the end of January, I think is the dates that we are putting forward. We'll revisit for another Webinar, we will try to finish up our data element reliability discussion, maybe clean up anything else we need to that we haven't done through email on these issues. And then we'll get to that question about is it really a quality measure, etcetera, etcetera? And then talk about comparators in score level construct validation specifically, but maybe even outside construct validation. So, lots of interesting things to come up with.

And just so you know what's going on at NQF in terms of our cycles, our evaluation cycles. January 6th is our deadline for measures coming in for our next evaluation cycle. So it is right around the corner. And we as staff will take probably about - I think our timeline is about four weeks to process those measures and get them in shape to send to you, which means that we'll be sending you measures for your evaluations probably early February-ish timeframe. So that's what we're planning on.

So with that, anybody have any questions or any final comments?

Ron Walters: Karen, I don't know how many - it's Ron - I don't know how many people in this group are planning on attending the annual conference.

Karen Johnson: Ah, yes.

Ron Walters: But if you could ask that in this infamous poll you're sending out, it's a captive audience, we'd have to find the right time because those days are pretty full and everything. But it's a March meeting, and depending on who's there, could be done.

Karen Johnson: It could be. Ron, that's a really great idea. And we should have (flashed up), and I didn't think about, we should have (flashed up) our annual conference numbers, the dates. They are set and I've forgotten when they are. But we're - (Hanna) can do that.

Ron Walters: Twenty-third to the 25th, I have it right here at my left-hand side.

Karen Johnson: Oh, okay.

((Crosstalk))

Ron Walters: I mean, I'm just thinking of an opportunity. I don't even know if six of our members go to that or ten of them do, or whatever. But, anyway.

Karen Johnson: Yes.

Ron Walters: We can find out pretty quickly.

Karen Johnson: Yes, we could. If we did that, the 23rd to the 25th is a Monday through Wednesday. So if we did try to hang it on, you know, we have decided we'd probably do the Thursday and Friday. But anyway, we can go from there, let's see. Yes, thanks for that, Ron. Good idea.

Ron Walters: Okay. No problem.

Karen Johnson: Anybody else have any last-minute comments? Dave and Dave, thank you so much for facilitating today's call. I know it was a little haphazard on my part, I appreciate you taking over and making it a good conversation. And thanks to all of you for joining for these full two hours. Go ahead, Dave.

(Dave): We should be thanking you, Karen, you did a fine job facilitating. Thank you.

Man: Thank you. I think we made good progress. Good use of the time.

Karen Johnson: Thank you all. We'll talk to you....

Man: Bye-bye.

Karen Johnson: …soon. Bye-bye.

Man: Bye-bye.

Woman: Bye.

Man: Bye.

Man: Bye.

END