NATIONAL QUALITY FORUM

Moderator: Yetunde Ogungbemi
June 11, 2019
12:01 pm CT

Poonam Bal: I'm actually going to give a little bit of a statement on behalf of all the project managers that were on this panel. So I'm not sure if you all received a message but I'm actually no longer going to be part of the team. I was promoted to Director in our Quality Innovation Department. So I will be shifting over there. I did also want to say a great thank you to all of you for all the time that we spent together on behalf of (May) and (Miranda). I'm not sure if you actually received the messages but (Miranda) left a couple months ago. But I'm sure she would love to have been part of this. And (May) left literally last week to go to (unintelligible).

So we've really enjoyed working with all of you and I'm glad that I can be part of this in person one last time. And we don't have a replacement for her yet but we will keep you informed.

Jack Needleman: I'm Jack Needleman. I'm Professor and Chair of the Department of Health Policy and Management at the UCLA School of Public Health, an economist by training.

(Larry Glance) Good morning everybody. My name is (Larry Glance). I am a (unintelligible) Analyst. I'll try not to put anybody to sleep today. And I am a Professor and Vice Chair of Research in the Department of Anesthesia at the University of (unintelligible).

Man: Good morning. My name is (unintelligible). I'm a Senior Research Scientist at ES (unintelligible) Medicine (unintelligible).

(Christy Teigland): Good morning I'm (Christy Teigland). I'm Vice President of Advanced Analytics at (unintelligible) Health. We do a lot of the data research and a lot of quality measure development and I really enjoy being part of this group.

Sam Simon: Good morning everyone. Sam Simon. I'm a Senior Researcher at Mathematica Policy Research. I'm also an Associate Director there.

(Jeffrey Geppert): (Jeffrey Geppert). I'm a Senior Research Leader at Battelle Memorial Institute in Columbus, Ohio. And it's good to know that a promotion means you don't have to work for (unintelligible). Congratulations.

(Bijan Borah): Hi good morning everyone. I am (Bijan Borah). I am a Professor of Health Services Research at (unintelligible) Medicine. And I am an Economics and (unintelligible) by training.

Susan White: Hi I'm Susan White. I'm from the Administrative Analytics at the James Cancer Center at Ohio State, Columbus, although (Jeff) and I didn't ride together.

Woman: It looks, like, (Mike) is coming in. So (Mike) we're going to have you introduce yourself as you sit down there. And then we'll go to the phone. Ron is on and hopefully (Jen) is also.

(Jen) Perloff: Yes.

(Mike Stoto) Good morning. I'm (Mike Stoto) from Georgetown University, (Seramley).

Ron Walters: (Ron Walters). I'm from MD Anderson. I've been on a few things at the NQF for about eight years and I'm sorry I'm not there in person. I had a little episode yesterday.

(Jen) Perloff: Oh and (Jen Perloff). I'm a Senior Scientist at the Heller School of Brandeis and Director of Research at the Institute for Accountable Care. And it's graduation season so I'm here in rainy Boston. Sorry I can't be with you.

Woman: Thank you so much (Jen) and Ron. Again welcome everyone. Just a couple of housekeeping things. Have breakfast and drinks and such back there so help yourself. Do whatever you need to do. Restrooms are out the double doors to the right. I'm geographically challenged so I think I'm telling you in the right direction. But you can meander out there and fine them I'm sure. Anything else in the housekeeping realm? If you haven't been able to logon to our Internet site, the login is guest and the password is NQF guest with NQF being all caps. So hopefully everybody's on. If you're having trouble getting on let us know and we'll have (John) come in and help you get on.

I think that's it for housekeeping. And what we're going to do is we're going to start with an update. And (Ashlie) is going to walk us through some updates and then after that we're going to start getting into the real meat and fun of the day.

(Ashlie): So we sent out quite a bit of materials on Friday. Hopefully you've had at least a chance to skim through the slides a bit to get a flavor of what we're going to be discussing today. Some of the meeting objectives obviously I think (Dave Marens) gave a nice summary that we've selected topics both in reliability and validity today that we hope to get some consensus on how the panel would like to kind of hopefully I don't know establish some guidelines for consistency and valuation in addressing some of those issues. Hopefully come up with some potential recommendations to the evaluation criteria as well as guidance from measure developers who are submitting measures and trying to get past the measure panel valuation.

We'll also discuss some of the updates on the method panel evaluation process. We've been working internally to do a lot of process improvements. We're always looking inward at our process to figure out how we can be more efficient, more transparent. And so we'll share some updates on our work to date with those efforts. And then if we have time again this is a very packed day and I think (Dave Marens) and our Co-Chair have about three days of meeting material.

And so we'll do our best to make it through. If we do happen to get to the end we do have one last item that is more of an informational update on some of the other projects that are going on at NQF that are methodologically related that we have some overlap or you may hear about some of the other work going on that may impact our work on the methods panel at some point. So we'll see how we do getting down to that last agenda item.

So a few updates. This first section is really just to give you an update on the work you've done to date. So we've tried to look through and track the measures that have come through the panel and how much work you've done, what measures have made it through, which haven't, and give you a sense of feedback on all the work you've done to date which has been quite a bit. For the spring cycle there were 47 complex measures that were reviewed across all of the subgroups and for the panel as a whole.

So as, you know, the methods panel team when we get your primary installations we kind of sort through those, figure out which there was consensus on, which there wasn't consensus on. And we tee up those measures that there wasn't consensus on or that we may have some more questions for the panel on the calls. And so about 53% of those 47 measures

were actually discussed on the call where again we couldn't figure out whether there was consensus in your preliminary evaluation.

So we had – some very small number three that were pulled for discussion, five that were pulled by staff and again 17 of those where there wasn't initially consensus in your preliminary evaluations. And so after all the calls about 30 passed and went onto the committee so about 54%. There were another about six measures where consensus was not reached even after the subgroup calls. And we do send those to the committee as well to have them kind of review your deliberations and the results of your evaluations and (Dave) will make a final vote on reliability or validity depending on where the issue was. And then another 11 that did not make it past the methods panel altogether.

And again feel free to stop me at any point if you have questions and we'll – sure…

Sam Simon: Yes I was surprised to see it did not pass (unintelligible) or the standing committees. I had thought we were told early on in some charter I saw somewhere that our input was advisory to the standard committee. So I was surprised that the ones we didn't past did not (unintelligible) standing committee. So it seems, like, we're more than advisory. Can you help me understand that?

(Ashlie): Yes absolutely. And I may have (Karen) answered this question because I was a little bit late to the committee. But one of the I guess kind of criteria we built in around the methods panel evaluation is that there is a bit of what we have termed internally as, like, a gatekeeper role where the methods panel there's a stop gap if measures do not pass liability and validity because they

are must pass criteria in our hierarchy of the core criteria that we evaluate measures by.

And in essence given that most standing committees aren't equipped with the methodological expertise it is certainly not as deep as it is here on the methods panel that it would be really hard for us to say that a measure should pass for example after methodologists have looked at it and believe that it did not meet the criteria.

So based on that kind of reasoning we have designed the method panel review at this point to be a stop gap if they do not pass the liability or validity that they don't move onto the committee. And we give the committee some information about why it didn't pass. But they're not allowed to then kind of re-adjudicate it at that level. So hopefully that is helpful and if others have questions we can explain more after that. (Jen) I'm not sure exactly when if that ever changed.

(Jen) Perloff: Yes I don't think I could have said it any better.

(Ashlie): Okay.

(Jen) Perloff: That's really the idea again reliability and validity are what we call must pass criteria. So if a methodologist, like, doesn't make it, then, you know, especially some of their patients with tumors on the panel it really could be putting them in a bit of a bind to have to re-adjudicate your decision.

(Jeffrey Geppert): And one other thing I'll add (John) is that the advisory capacity role really comes into play when it actually lands inside of the committee. In a sense that they're not required to keep the recommendation of the methods panel. So of course inside of the committee there are a lot of expertise because it's very specific to the therapeutic areas that are there. And that when they evaluate

measures if it's something that isn't necessarily a methodological component but a reason that the (unintelligible) especially for reliability then that would be a determination of the committee. Does that make sense?

(Larry Glance: Thanks, sure. Thanks for leading the way (John) with the vertical card. We forgot to mention. Jack has asked to go even though (Laci's) next. Jack wants to follow up on this point.

Jack Needleman: Yes so Sam just to clarify. When we were looking at validity and I know in a number of cases there were a whole series of clinical judgments that were made about exclusions and so forth. And clearly we were not competent to do that. So a lot of validity passes were conditional on substance experts working at the exclusion. So when you say they get a second bite at this apple is that the sort of thing you're talking about?

Sam Simon: Yes that's exactly what I had in mind is that this group really serves as an auxiliary and, like, in an adjunct capacity that just connects directly to the committees. So just see you as an extension of the work. So you're looking at a methodological standpoint really kicking the tires and the clinical aspects of it for the most part.

Man: So when we flag somebody needs to check the clinical logic here. Those recommendations were highlighted as it went to the…

Sam Simon: Correct. Those are passed directly onto the committees.
Man: Thank you.

Jack Needleman: (Laci) then (Jeff).

(Laci): I'm trying to recall from last year I think we talked a lot about what metrics were kept before the scientific methods panel versus what are being tracked now. And I was just curious if you had any insights on overall changes to endorsement based on now having the scientific methods panel. Are there similar numbers being endorsed overall or fewer or changes to measures?

(Karen): That's a great question. I don't know if I have the hard data that I should have. Maybe (Alisa) does. My impression is you guys are a little harder so maybe we're endorsing fewer.

(Laci): I think we are over allowing endorsement to (Karen's) point. What we're seeing coming in more are the complex measures that you guys see. So we're seeing less of the process measures coming to NQF overall.

Sam Simon: Yes so it's not on this slide but we had 72 measures total. So the fact that 47 came into methods now says a lot about the co-complexity measures.

Jack Needleman: (Jeff).

(Jeffrey Geppert): Maybe I'll get to it but I'm just wondering for the 11 that did not pass did the developers have any feedback on the process. Did they feel, like, they knew the expectations (unintelligible)?

(Laci): Another great question. Yes and no. I think we've had some kind of back and forth with the various developers. Most of them really do appreciate your deep dive and the advice that you have given. We are also in the process now of doing a formal evaluation what we call our redesign of our endorsement process. And of course you guys in the formation of the methods panel is a big piece of that redesign. So we just send out and most of you responded to your surveys. We did a lot of surveys. We have some developer responses

from there. Yesterday we had developers in the room for an all-day workshop. We got some feedback there.

So I think yes in general, you know, everybody it hurts when your measure doesn't go through. But I think people do appreciate that they have put a lot of work into the effort. So having an expert look at what they've done I think has been appreciated.

(Ashlie): So kind of dovetailing on (Jeff's) question a bit about some of the reasons why measures did not go through. That the testing that was needed to demonstrate reliability and validity was not done or wasn't, you know, in the measure packet at least not in the way it was submitted. The tested methodology was unclear or inappropriate. There was inadequate data in the testing samples. So too few states or the population state measures for example inadequate or low testing results for reliability and then a lack of risk adjustment.

And I would say just one addition to (Karen's) comment earlier I think we'll talk a little bit more about this as one of the focuses of process and proven efforts. And one of the main pieces of feedback we've heard is that developers want more of an opportunity to interact with you guys during the evaluation process. I think over as we kind of have gone through each cycle we've kind of opened up the process a little bit more to developers.

So initially they weren't allowed to talk and then now they're able to talk. And then so we will be as before kind of increasing that opportunity. I think they'd like to have an opportunity to dialogue more, to offer answers or clarifications as they come up in the discussion. So we're certainly trying to balance that with the time that you guys have been reviewing packets and

working with developers to make sure that their submissions are their best foot forward right?

And so that there isn't really a need hopefully for a lot of back and forth and that you have everything you need in front of you to make your evaluation. So we'll continue to discuss that especially the improvement section about how we're trying to think about that and certainly think of your feedback on how we might, you know, (unintelligible) that…

Jack Needleman: Sam.

Sam Simon: Just a follow up what (Jeff) mentioned. And then also following up on this point. It does seem, like, the feedback, you know, this analysis in particular could be critically useful to developers. You have this information were provided to them. But these were the sorts of things that trips up the panel because I feel, like, a lot of the panels I was on things were just not clear. And that they have a sense that they're going to be very clear about what methods they're using. That really helped a lot of our deliberations and all.

(Karen): And I think I'll add to that a couple things. We did a session yesterday. (Matt) helped us out in a session yesterday. And we were trying to start saying hey these are the views that we're seeing. So please pay attention and, you know, a lot of what we were saying is add more detail, you know, tell us your story. Put that stuff in there. So that's one of the things that we're trying to think about especially after those discussions we're going to have additional concrete guidance and start really putting out there and making more available and a little bit verbal point. So I think we're getting there.
And the other thing I will mention for the measures that do not go forward to the standing committee so you guys have said, you know, one of these things or something similar, you know, kind of stopped the aggression. We do a

fairly detailed summary and provide that to the developer. So we, you know, point by point would say, you know, this was the concern, this was the concern. They get written summaries from us as well as to of course be able to participate on the calls and listen to your conversation.

(Man 1): Okay (Karen). There was I should remember your answer to this because I asked you this on a call we had between these two meetings. There were a few cases where the reviewers thought they probably have the information. It's not in a submission. Can't we just ask them for that information and let it go forward between the time of the meeting and the parent committee with (unintelligible)? And I think the answer was no and there was a reason. Could you comment on that?

(Karen): Sure. One of the reasons is this loss of time and trying to be really respectful of your time because if we allow that kind of thing, we would want you to have time to actually look at the stuff and reflect on it. So you've already put in a lot of time and now everybody's saying hey we did put this in but now look at it.

So having taken the time to do that and bringing you back on the call all within the timeframe that we needed to do seemed a little too short for us. But then this afternoon we'll talk a little bit about how we're trying to figure out is there a way to make this happen. I think the other thing too is just consistency. To be frank with you we get pushed back sometimes because some people think well you allow us one group to do this and not everybody. So, you know, once we said hey we can't do it we have to be consistent across.

So, you know, in that particular situation I think you were probably right. I think they did have that table laying around somewhere on their desk and, you

know, but didn't put it in. Yes so I think that is – those are the two main reasons that we have for that.

I will say that I think you guys are much more sympathetic. And I don't want to be mean. Going to the six-month – every six-month opportunity for evaluations to us that doesn't seem, like, that long. I know it does to you guys but say three or four years. So we're thinking hey if you don't make it, it's not that much…

Man: (Unintelligible).

(Karen): …yes. So we think there was a big, you know, I do realize that that's kind of (unintelligible) thing that (unintelligible) doing and you guys aren't. So I applaud you for being sympathetic with who we are sometimes.

Man: I think more, like, grant submitters, you know, (unintelligible) of the year, okay. No, no I'm jus saying it's all relative to your vantage point I understand. Thank you.

(Ashlie): All right so we'll keep rolling here. So again a few metrics here. We've been trying to track over the cycles that you've started. I won't go through column by column. But you can see that the number of measures that have been submitted for review that you've reviewed every cycle has gone up every cycle. And I would say with the number of measures that are not passing has been relatively – the number has been really relatively consistent I would say. The percentage varies again by the total number of measures that were submitted have been pretty consistent the number of measures that don't make it through.

And I'll just maybe point out a couple of other things we've started to pull in some metrics about where there was alignment with the standing committee

where measures made it through the standing committee valuation after having reviewed the methods panel recommendation. And certainly we've had a little bit of fluctuation there. And we'll talk a little bit more about some of the examples of measures that didn't make it through the committee after the methods panel made a recommendation. Yes.

Sam Simon: When you say there was lack of agreement is this specifically around the validity, liability judgments or did they reject something we had let through on the usability of one of the other standards?

(Ashlie): So we'll talk a little bit more about that. But I think it was mostly validity which makes a lot of sense right because we expect, you know, their clinical experts and as you mentioned there's a lot of other sub criteria within validity that addressed clinical logic that the methods panel doesn't evaluate. So we would expect in many cases for them to re-adjudicate those elements outside of the (unintelligible).

Sam Simon: And could you help me with the various rows? They don't seem to be mutually exclusive. So trying to get to 39 say for fall of 2018 I am not tracking if 17 passed and 2 did not pass, does that mean that there were 20 of some other variety?

(Jen) Perloff: Yes I'll help you with this table. This was my creation. I realize it's a bit of a mess. And the unanimous task did not pass – does track the cost, the four cycles that we have. So remember the first two cycles when we first started this effort we had this idea that we could kind of put you in your little cubicle and make you do things kind of, like, (unintelligible). You kind of do your own thing, buy your stuff, you know, when there was a split decision we would have our co-chairs (unintelligible).

Beginning of 2018 we sent to a subgroup model right? So the 17 and the 2 whatever's left are things that it wasn't co-chair arbitrated. But it's not necessarily that we did or did not discuss it on the call because on the call things that were not unanimous – let me rephrase that. Things where there wasn't a complete consensus were definitely discussed on the call. But also other things to be pulled.

So I just didn't try to put all that detail in those lines. So basically you're 39 in the fall. The fourth line down number that we see lower in sufficient ratings, 10 of them did not make it through. So you guys served as our "gatekeeper" for those 10. That means 29 did make it through and were evaluated by the various standing committees.

So of the 29 that they started out with 39, 10 did not pass your review. That leaves us with 29. Of the 29, 23 were passed by the standing committee. The remaining six were not passed. And that's what we mean by not aligned because by definition we had sent to the standing committees only those ones that you guys did pass through all right? Does that make sense or are we still a little bit confused?

Sam Simon: So the bottom lines four, five and six do sum to 39?

(Jen) Perloff: Yes hopefully they do. If not I should return back my math degree right?

Ron Walters: (Jen) the only thing in the two cells where it says N/A there really are numbers in there. It's just because the process changed. They're not co-chairs on every trade. These are the ones that were not unanimous.

Sam Simon: And we appreciate the fact that we have given you an awful lot of (unintelligible) you know. Okay. So I was just trying to figure out if there was a relationship between those three numbers let's say for fall and the first two numbers understanding there's something in there.

Another quick question. Was there any kind of pattern that you saw between the new measures that were submitted? Were they more likely to fail? Or were the maintenance measures more likely to fail?

(Jen) Perloff: That's another really good question. I don't think we live it out quite that way. In general, you know, I would almost say half and half. My gut feeling would say newer measures had a little bit harder time but that's not necessarily the case. Sometimes the maintenance measures coming in they're kind of used to what they did and now people are looking a little loser to things that did pass through. So it's been some of both. It's certainly not been all, you know, one side versus the other. (Ashlie's) writing it down so we'll try to go back and get those numbers for you.

Susan White: Sorry this is Susan. I just have one quick question. I don't think there's a slide in here unless I missed it. But did you do any analysis on one subgroup versus the other? I know you sent us an email about that which it had a summary. It looks, like, one of us might have been – we were loaded with measures that weren't great or I don't know which right so…

(Jen) Perloff: Yes we haven't actually done a formal analysis by subgroup. And it really is non-systematic. So, you know, some of, like, the cost subgroup I think we did mainly cost measures. And I think that may have had the highest rate of return.

Susan White: I was on that one and I think the reason for it was there was one submitter…

(Jen) Perloff: Yes.

Susan White: That did the same thing over and over.

(Jen) Perloff: Yes.

Susan White: So I know it's hard to tell. I mean it's not, like, you signed randomly right?

(Jen) Perloff: Yes. So we haven't done – we've kind of thought about it. We thought a little bit about (unintelligible) are really hard graders. But we don't have, like, names. We just kind of know well, you know, so yes and is that fair right? But I think so far I think we haven't – I don't think we've landed in the, you know, subgroup, just happened to get all the, you know, the really hard graders that would make it unfair.

(Ashlie): And to your point Susan to the nature of the measures because often what we do is try to group, you know, like measures together. And sometimes there's problems who have done similar methodology to try to make it easier for, you know, one viewer to kind of apply the same logic going forward. So one group may be survey measures, one group may be cost measures, another may be I don't know another kind of topic area. So it makes it a little hard to compare. But I think it's a good question. I've written it down and see if we can take a look back, thanks.

Woman: And there's a question on the phones.

Ron Walters: Hi it's Ron. Thank you. Sounds like I'm in a group of doctors. Indeed I mean because they want to split the data over and over. Thank you very much for collecting all the data. I think it's very good and the other analytics that can be done on it have been suggested. So I'm not opposed to this. We probably will get to numbers that are uninterpretable anyway because there's so many dices that are possible. But I do very much thank staff for collecting

this data because it gives us a feel for how our process is going. And so thanks again.

Man: (Christy).

(Christy Teigland): Yes we're going to talk about this. But I want to say that many times I wanted to be a much grader and fail people regarding the social deterrents of health adjustment and the fact that they didn't do it. And I ripped them but I couldn't do anything about it. I had to pass them and that killed me so – and, you know, even from the economics, I'm an econ nutrition, you know, I just disagreed with their, you know, they had the evidence. And they said but we're not doing it. So I really want to talk about that. It should have some weight. I mean this is a huge move that the NQF made and now we're just giving them a pass on it. And I'm not happy about that.

Ron Walters: We will get to that in a couple of points later.

(Karen): Yes. And just so you know the same kind of summaries of your discussions including concerns etcetera that we give to the developers that do not pass, we also make that kind of thing available to the standing committees. So while our decision has been that the standing committees get to make that final call, we definitely if you guys talk about concerns on that it's there in black and white. They see that.

(Ashlie): So we'll keep rolling here. This next slide is around kind of the consensus not reached step. Some of this was in one of the previous slides we already presented. But just to give you some sense over the last few cycles how many measures in your initial valuations came back to us when we kind of put all the ratings together where measures kind of fell in the consensus outreach zone.

So relatively consistent about half – a little over, so a little over half went down maybe a little bit over the fall and spring – fall of 2018 and spring 2019. And then it looks, like, I mean after we're having the calls this ability for you guys to kind of discuss these issues about those measures that are consensus not reached and then again voting again and hopefully and actually for most of them coming to some consensus I think is a really good outcome. And for us is a good sign the group calls are really working as a way to adjudicate those issues. So I think that's a win there for sure.

So this is – the next few slides are kind of around those measures that passed the methods panel but the standing committee then voted again and it did not pass for some reason. So let's see I don't know the names of these measures. I don't know (Karen) if you remember or which but again we can just see that to Jack's point earlier all of these are on – most of these are on validity. There's one regarding reliability at the end that we'll talk about. But what's the best way to go through these?

(Karen): Yes.

(Ashlie): I don't have the measures.

(Karen): And, you know, to some extent even though it would be lovely to be able to delve into these and really kind of dissect everything, this is really more for information to give you a flavor of what's going on. And apologies I had notes at one point in time about what these measures were and I actually don't remember what they are and I didn't write them down for today. The first one – the S&P did pass the measure. So it wasn't a CMR, a DNR or a consensus not reached. It was an actual pass. But you see that the methods panel – I mean sorry the standing committee really was concerned with the risk

adjustment. And really feeling, like, the risk adjustment was not appropriate for that measure – the 3456. And these things were raised by the methods panel.

So (Christy) to your point if we – you referenced that one. So what that means is you guys are actually adhering to our desire for you not to fail all things directly. And what our advice here is or our guidance is don't fail it because you disagree with the particular factors that are or are not included in the risk management model. This does not mean that if you feel, like, the model itself was inappropriately done or, you know, statistics or calibration whatever were not appropriate or good enough you can't sell them on those kinds of things. So that did happen and the standing committee actually kind of did agree with you.

And the next 13366 you passed. The standing committee and we have little asterisks there. These were to date and actually we're kind of still in the middle I think of our comments. So these are not final decisions just so you know how it works if you're not familiar with our process. Standing committees making their initial judgment. We write the report out for public comment. We bring the standing committees back together for what is called those comments all. And at that point they could revote on things.

So we're kind of – when these slides were created this was kind of in the middle. So it came back for any revotes. And these may have changed but this is what it was going into I think the April, early May timeframe. We presented these to our (CSEC) Committee.

So the dual status adjustment was at least some of the discussion of the standing committee. They also and this was kind of interesting to me. They were not happy with that point 61C statistic which generally you guys have –

and we don't have thresholds. So that's another – that's a difficulty for people. And the other thing that's a little trick was it really the C statistic that they were uncomfortable about or was it the risk adjustment and they're not saying necessarily that's what it is because sometimes it's kind of hard to seek out the actual reason that a committee may go up or down.

(Ashlie): So just a quick (Shendae) very nicely sending us. And so 3456 was admission to an institution from a community is the measure – 3356 was hospital visits after urology ambulatory surgical center procedures.

(Karen): Okay and then these I'll just go through these if you're okay with these. And these measures were ones that even after separate calls there was still not complete consensus amongst the subgroups. So when we say CNR I just mean that it wasn't a strong majority of the subgroup saying either pass or fail. So we do go ahead and send those on.

The first two cases again risk adjustment strategy. And the underlying populations I believe these will probably – I think these might have been the Medicaid. Yes these were the Medicaid measures of the really multiple comorbidity for Medicaid I think is what these were.

So it's really some concerns about the differences in the populations and then 0964 this was one that was interesting because the methods panel looked at what was a fairly low correlation result for validity. And again we don't have thresholds. So some of you thought that it wasn't good enough. Others thought that it was okay. So we sent it through based on CNR and that particular measure 0964 – do you know what that one was?

(Ashlie): (Unintelligible).

(Karen): Frantically looking to see what those are. The committee was okay with those. They like the measure. And then 0753 again kind of a split decision on liability that the committee did pass it. Almost the same kind of thing, you know, it's one of those things where it's probably a fairly equivocal estimate. And, you know, we don't have thresholds. It's a judgment call as to whether you feel, like, you know, it's good enough. So that's what happened on those last two. (Shendae) have you found those yet?

(Shendae): No.

(Karen): Okay.

(Larry Glance): Just maybe we don't have to have that information. But I just – I know it's a small sample. But it's probably a good thing that of the four where we didn't give consensus, they went yes twice and no twice which kind of indicates that our, you know, we're both teetering on the same fulcrum. Maybe we don't need to this. Does anyone feel, like, we need to know? We're probably behind already so…

(Ashlie): So we attempted to try to give you a little bit of peek into what we're expecting for the upcoming cycle. Just in looking at the measures that are queued for maintenance for this coming cycle and just a really kind of cursory check of how many of those are complex and haven't been reviewed in a while they would likely come through the methods panel.

Again this is really just an estimate. That number may change. Just to give you a sense. New measures again we're always kind of working in our teams to reach out developers in our topic areas to see what measures are coming in. And we're still working on that. But we'll certainly kind of keep you abreast

if you have a sense of the total number of measures that may be coming to the methods panel.

A couple of kind of more logistical items. If you remember when you were first seated to the committee that there were -- to the panel -- half of you that were given a term of three years, half were given a term of two years. So those that were given the two-year terms your terms are expiring. Technically I believe it would expire at the end of this cycle.

So what we'll do shortly actually for those folks who terms is I think it's about 10 of you who are expiring we're going to send you an email with the link to SurveyMonkey with two questions, your name and whether or not you'd like to continue to reup for another term. If not that's fine too. We'd just like you to let us know.

And so if you could send that back to us by the end of today that would be helpful so we have a sense of how many seats are going to open up and we'll be as you see here with our third bullet point we'll be expanding the group quite a bit going forward. We have as the number of conflict measures has increased and we've been learning as we go to see kind of the workload for the methods panel, that we'd like to expand the group quite a bit and play with how we might maybe divide folks across cycles, maybe, you know, 20 due spring, 20 due fall to lighten the load a bit given the number of measures that are coming through.

So we're playing with several ideas but regardless of how we kind of end up distributing folks we will be expanding the group quite a bit as we're trying to figure out seats and all that stuff. And certainly have appreciated everyone here. But wanted to also kind of let you know that we are changing things up a bit with the hope of lightening the load a bit for those of you who are

volunteering your time. So it's certainly a concern of ours that we're looking to make some improvements on.

The other piece that we are looking to do for this next cycle essentially going forward is to have an in-person meeting for measure review. It's one of the strategies that we'll talk about a little bit later this afternoon that we're hoping to implement as a way to enable developers more time for back and forth because we spend almost a two-week period scheduling multiple conference calls.

By having a one-day in-person meeting we think we'll be able to get through the review of most of those measures as well as give you guys an opportunity to hear all those sessions across all the subgroups. And so we'll potentially be moving some in-person meeting setup as opposed to conference calls and we'll be working on dates and so forth to come. But just another heads up and kind of peek into the future of one of the things – a couple of things we're looking to do for the fall.

(Karen): (Sherry) did you have a question?

(Sherry): Yes for those of us who are slowly expiring anyway do you have a maximum number of expiration cycles that you can – you reach a toxic level of…

((Crosstalk))

(Ashlie): Yes so I will say initially we had thought – so we aligned our term renewal with our regular standing committee policy which is three years. So you can renew for up to one time I think for three years. That policy was in place before we did two cycles a year. So that was when we were doing maybe one project a year right? So the term you would do, like, maybe – in a term you

would do three evaluations. And now what you're getting with the term is double that.

And so what we've done is to kind of change that term renewal to a shorter period given that you're actually doing more work in that term. So it's one term renewal for two years which is four cycles which is quite a bit. So, you know, we've talked about, you know, if you kind of lay off for a couple of years and then you want to come back certainly we would consider that. But we want to give, you know, want to be mindful of time. Obviously you're volunteering your time and also make sure we have an opportunity to invite other folks into the fold as well. So (Karen) do you have questions or…

Jack Needleman: Can I just make a comment? Maybe I'm alone in this but as one of the people that drew the three-year straw, to those who drew the two-year straw understanding that apparently we're going to be adding 10 or 15. Even if all 10 people that are rotating off stayed, we would still be adding another 10 or 15 people to the committee.

So just as a Chair and probably NQF staff I get a little worried about how that will lose. I think we're actually converging, coalescing and creating a culture. So this is a plea to the 10 of you if you're willing to do two more years I'd be very grateful personally not to put pressure on you. But I think that would be great just to get this committee, you know, I think in another two years the committee will really have its leg.

(Karen): And if you have guys throughout the day about, you know, this idea of expansion, again we're doing it mainly because we really feel, like, the workload is bigger, longer, harder than we initially anticipated it to be. That's the main driver - the expansion. By expanding and bringing new people in we'll have to bring them up to speed. And you guys have been doing this now

for two years and you're getting up to speed. So there will be some difficulties there.

So if you want to chat with us throughout the day this isn't written in stone yet. This is just our thinking of a way to improve things for you guys. So we're happy to chat offline about that.

Ron Walters: Note that (Paul) has joined us. Do you want to just give a very quick introduction for him?

(Paul): Okay.

Ron Walters: That should do it.

(Ashlie): This is kind of our next agenda item here or sort of our next agenda item and then we'll go onto the update. We wanted to just give a brief update on how the whitepapers are going. And I think we were going to tap a couple people who've been leading these efforts to give us an update on where they are in terms of the kind of writing status and plans or efforts to submit to date. Yes (Larry) do you want to be the first one?

(Larry Glance): So we had a writing committee that was tasked with creating a whitepaper on looking at the scientific validity of risk adjusted clinical outcome measures. And this effort actually started about a year or so ago I think. So it's taking a little bit longer than anticipated. The purpose of this whitepaper was essentially to sort of lay out a sort of best practices for looking at scientific acceptability. And the idea being that we could use this as a platform for coming together as a committee to coalesce around these best practices. And also and this is also as important to provide measured developers for the set of best practices and to encourage consistency in measure of submission.

So the way we did this is we had a core writing and then of about I think it was, like, five or six people. We went back and forth a whole bunch of times. I think I lost count. And then after that it went to the overall writing committee. And it went back and forth a couple times. And since then it's been submitted. And it is now currently under review at the Annals of Internal Medicine. And it's back with the editors. So we're hoping to find out within the next couple weeks where we stand with that journal and keeping our fingers crossed. This panel is a fairly high impact journal. And I'll let everybody know what happens with that.

And I do want to thank everybody who participated in this. It really was an amazing amount of going back and forth. And I especially want to thank the people in the core writing group on all their help, thanks.

Ron Walters: Okay the second one I guess…

Jack Needleman: I thought you were going to – I just want to recognize (Larry) in all that you did. I mean you're throwing a lot of credit to the writing group and back and forth. But you really drove this forward so thank you on behalf of the committee for driving this whitepaper through and getting it submitted. I thought the timeline was actually fast. In terms of (unintelligible) it was faster than I expected it would be so well done.

Ron Walters: (Unintelligible) appreciate it. The second one essentially had their same kind of structure and process – the core writing group and eventually the panel will have a chance to review and comment. This one I think we could describe this as the intro paper – describes us and what we do. It's not about particular issues. It just says here's this panel (unintelligible) what kinds of issues are we working on.

So even though it was slightly behind (unintelligible) whitepaper in sequence I think if they were somehow, that every series with end up with a number, this might be the person who says (unintelligible). And again I think (unintelligible) comments and eventually (unintelligible) some comment.

This was submitted to (unintelligible) last month. I just wanted to tell you it was rejected there not because it (unintelligible), they just didn't think it was appropriate for their journal. And I think we suspected that going in, like, a lot of things that you try to shoot high and go for the high impact journal.

It's not quite clear and I don't know if we should have it open for discussion now where we go next but I think we'd talked about that a little bit among the group. I think we're probably going to go have to (unintelligible). But I think probably it's going to end up in some journal more tightly focused on issues of quality measurement. And there are two possibilities with being (unintelligible) quality could be a second choice. There is some support for that. But we'll turn it around. I'll see since we're not really making changes (unintelligible) and let you know.

And then let me just mention the reliability paper at the moment is an outline. It's not really a fully drafted paper. (Shay) and I have been partially taking the lead on this. But what I wanted to point out and this could be a transition (unintelligible). We really want to (unintelligible) this morning's discussion and forum, you know, what the chunks of that might be. There's some things that perhaps are important issues that may not after this morning especially be important. But we may come up with things this morning that actually need attention.

I think (unintelligible) as it is in a series of whitepapers we agreed to essentially use the writing of it as a way to share a higher degree of consensus

and consistent use of terminology among ourselves. But the also be able to convey that with developers and anybody else that's interested to say this is what we want to see. And then we use this journey. This is what we mean even if the whole universe doesn't use it that way. And then when things come in this is how we'd like you to use the term. But I think our discussion over the next couple of hours will, you know, have a great deal to do with sort of reshaping and refining where this goes.

(Karen): So I've written the notes to myself that methodologists want more data. So I'll remember that for our next in-person meeting. Yes for some reason I thought we'd blow through this and we didn't so we're running a little bit behind that's okay. The way we set up the day is really to try to focus the most of the morning and the early afternoon to the methodological substantives, things that we would like to get through.

There's a few things on the agenda that we may not make it through. They're more nice to have. So the bullets are kind of a little bit further down. We may not get to that. That's okay, you know, we're get to them eventually. But we wanted to spend, you know, a lot of our time today really talking about method. And we also know that some of you at least are going to have to leave us a little early this afternoon but that's all right.

So while the things that you would miss would be I'm sure interesting to you. It won't hurt us for you not to be here necessarily. We really want you here for these major bullet points. So we're going to spend more time relatively we say on reliability as compared to validity. But again a good amount of time on both. And we want to talk a little bit about data element reliability and then what we call four-level reliability. But with that we are going to spend relatively less than I think on data element reliability and really focus

on the (unintelligible) discussion on four-level reliability and what we expect. But again I might be surprised so we'll see how it goes.

So just to remind you guys of our criterions reliability (unintelligible) the measure of specified produced assistant. Results of the quality of self-care delivery and under reliability that takes into account the patients which we want to be precise and complete as well as assessing that it's being done. Again you guys really know this well. Many of our – it depends on the measure type in terms of what kind of testing we're expecting. Some types of measures we expect more testing or specific types of testing and others maybe not. And we may talk about that a little bit later today.

Last year we talked about our definitions of reliability. And I don't think necessarily that we have changed our minds about the general definitions. And we don't want to work on this today. But if we need to work on this we will. But really we're trying to get from the (unintelligible) data element perspective it's that repeatability or consistency. Let's not use the word stability because that kicked us off on a rabbit trail last year. So we're not going to use that word right?

But it's the idea of repeatability and then if it's four level it's really the precision that we're interested in. So we want reproducibility of the data elements with the same population in the same time period. From a score level we're thinking about the variation in scores due to systematic differences of cost to entities in relation to random (unintelligible). So this is mainly background. I think if anybody has any concerns about these definitions this is probably the time to raise them.

(David): Okay I guess this may work. But I'm not sure I'm fully happy with sort of a number of features here. For example just the placement of the Xs and 5s that

somehow repeatability is not an important feature of the metrics score. I find that a strange idea. You'd have to think that at least in principle you'd say, you know, if I measure this property in the hospital or doctor. And then if hypothetically I could go back, like, a week later and something I get the same thing.

Now I know in practice this literal concept of test, retest is impossible – might be. But just since we're talking about concepts here it seems repeatability it's just as much an important property of measures for a level as it is at the data element. Now and then, you know, the precision, you know, I'm a little happier with the text at the bottom than I am with the word precision itself.

You know precision to me implies something, like, you know, how many digits after the decimal point of some grade are we talking about. I mean is the mortality rate 6% or 6.5 or 6.51. When I see the word precision and that's not really what we're talking about here.

So I have at least those two concerns and have others. And the question is, like, the (unintelligible) whatever, you know, we spent a lot of time worrying about the Wordsmith. But at least what I want to make sure is we have the concepts right. So when we develop this we say this is the concept we're shooting at. And then hopefully we can use the same words from there.

(Mike Stoto): I think this is really important. One of the things that struck me in reading this paper over time is that people use these words. They come to us with different conceptions of what these words mean. And I think that we need to recognize that and then try to say here's what we want. This is really our opportunity. And then not to try this, make everybody happy with their current understanding. But someone saying that what's NQF mean.

The other thing too is that I think we sometimes confuse the concepts with how we measure them. And I think one of the nice things about the (Don Adams) paper and so on is that he lays out those concepts and then talks about how to measure. And I think that's to the extent that we can do that going forward, you know, the analog of that going forward I think we'd be better for it.

(Christy Teigland): Not surprisingly I agree wholeheartedly with (David) which has been kind of working back and forth on this issue. But for me reproducibility, you know, repeatability, precision of the (unintelligible), all of those things are synonymous. And what we mean is we can't reproduce whatever observation we make. And I love my bathroom scale. It is consistent. It is wrong but it is consistent. And so I absolutely (unintelligible) and that's kind of one way to think about this. Reliability in any score is only (unintelligible) a valid tool driving thing because it's only scores of that. Only the populations and only under the circumstances which we study.

You can't ever completely assume that. And now nobody's ever going to validate, you know, certain measures of physics and other things, like, that. But, you know, the next slide actually even is worse in the sense that we're all error – all measures have errors and figuring out what the magnitude the area you can tolerate for which purpose it's being tested is kind of the goal of this exercise.

And I don't think that's quite selected. And people in, you know, the non-measurement world interchange the liability and validity which is probably the next two slides. But giving that kind of a better solid asking what is meant by reliability leads to the purpose that we kind of put it to. So trying to estimate physician performance is a different issue from trying to estimate the performance of a hospital. It's a different issue from kind of the performance

of the healthcare. So any one of those can have different kinds of thresholds for example that you would put this to. And I think that I agree with (David).

(Jeffrey Geppert): We've got some real substantive issues about the standards we're going to use for each of these. But want to make sure we get there. I agree about the not Wordsmithing. I would simply note that the use of precision here as (David) noted is not the standard colloquial use. And I'd like to see us get a better word but I don't think we need to get a better word at this meeting. It can, you know, the concept can go with whatever we mean it to be.

The other thing is just the language there about signal to noise that has been sort of the more standard. But we look at a variety of other ways to measure and reliability that are not strictly signal to noise. And again don't have to Wordsmith the language here. But we ought to as we go into the next discussion substantively of standards we need to recognize we're looking at a variety of different measures not all of which are signal to noise measures.

Jack Needleman: (Jeff) and then (Christy).

(Jeffrey Geppert): The only issue with the concept of reliability as described there is what if there's only one measure. And does that measure not have a score of reliability is a measured entity in reference to some sort of threshold and whether we can get distinguished performance relative to that threshold or benchmark. And for that you only need one measure of entity.

(Christy Teigland): I'm still kind of (Alice) down in the (unintelligible) ability. And the reason is that for a lot of measures (unintelligible) around measures. And they didn't have enough to do the same calculation, same time period correlation to show that it was reliable. And so they didn't rely on any (unintelligible), you know, coefficient. But they relied on, they reverted to the tech. And then

sometimes the technological panel would just be, like, six people. And, you know, three of them said it was good or better. And so they passed it. And it just didn't feel that strong to me.

So those kinds of measures particularly we're doing more and more measurement at provider level those can get small really, really fast. And it's really hard to show, you know, in the same time period.

(Ryan): I think this is a great discussion. I think it's really important to attach a method to the concept. So to me, you know, you can define the liability in my mind in two different ways. One of them is precision and the other one is reproducibility. And so when we're talking at the score level if you're talking about precision, you get that with the signal to noise. If you're talking about reproducibility with split sample reliability testing where you're looking at the interclass correlation proficient. And as it turns out the ICC is a conservative estimate of the signal to noise ratio. So it kind of converged.

But I think what we need to do as a panel is we need to give very clear guidance to the measure developers what our expectations are for the methods that they should be using to evaluate score level reliability and data element reliability. And I think the concepts, the words are important but probably what's more important is saying, you know, these are the methods that we would like you to use.

And that way when we're evaluating these measures we're saying okay so we asked them to look at either this, this or this and they did that. And so let's go ahead and move on and look at what the results were. That's the first piece.

The second piece and this is where I think we're not quite on as firm ground is coming up with thresholds for what we think is acceptable for reliability

whether again you're using the signal to noise ratio or you're using the ICC. And I think that's a lot more difficult because there are some skills that people have put into place. They're out there. They're very arbitrary. They're really not evidence based in any kind of way.

And short of doing some really fairly rigorous stimulation exercises I don't think we're going to be able to really have a very evidence-based way of establishing what the correct thresholds are. But I think we will need to have some idea. That way you don't have one subgroup coming up with one set of thresholds that they think are must pass. And the second one having another. I mean I think as a group we need to have some consensus.

Sam Simon: All right (Christy) I think you're up from before but so it's Jack and then (Gene).

Jack Needleman: Okay so…

Sam Simon: Oh…

Ron Walters: I'm sorry. I just – no offense to Jack but just for clarification (Ryan) is your use of the precision – term precision the way you just did it I think ICC to precision.

(Ryan): (Unintelligible)…

Ron Walters: Single light.

(Ryan): …with precision and the split capped liability test could reproduce (unintelligible).

Ron Walters: Okay now in the first one does your use of the term precision there in that particular (unintelligible) approach equate to this issue of how many digits after the decimal. Is that a difference sense of term? I would think it did.

(Ryan): Like the (unintelligible) you get it the precision.

(Karen): You have to use your mic sorry so we can catch it yes.

(Ryan): So for precision again to me it's embodied by the signal to noise ratio. You know how – so if the signal is – if you knew what the point was – whatever the point estimate was for the measure for a particular provider, you know, how often are you going to get around that target that's the precision piece.

Reproducibility with split half the liability testing where you split your dataset into two halves you're basically, you know, on one half you're estimating say risk standardized mortality measures. And you're doing it for the same group of hospitals in the second half. You're seeing how close they are to one another. And that's the reproducibility piece.

Ron Walters: And then I'll get a…

(Crosstalk))

(Larry Glance): Go ahead (Jay).

(Jay): (Larry) I'm not going to agree with you on that because I think we're both trying to get the same issue with those. And I think it's useful to reflect on the use of these measures. In some cases we're looking at absolute measures. What's your score? Is it good enough or bad enough against some other external standard? And in other cases we see these measures being used to

compare places either based upon, you know, scores that are considered above average or below average or in other cases ranking everybody and doing very real things to those rankings, like, paying people from where they are in the ranking.

So for me this issue of measure score reliability is one of we got an estimate. That estimate we know has implicit uncertainty based upon the size of the sample and the underlying variance in the phenomenon being measured. But we've got a point estimate and we know that there's variance around that in terms of what the value will be from time to time. Or how well we've measured it with this particular sample.

And we're trying to figure out whether the variance is tight enough that we're confident that the next time we did this measurement or if we split the sample and did it once in one half and once in the other half that we get basically the same results. And I was one of the folks, like, (Christy) who throws us down the stability rabbit hole. I'm not going to do that again.

But this underlying concept of a tight enough measure that we're confident that if we did it again we'd get the same result is I think implicit in what we're talking about in terms of measures for reliability whether the same results are the ranking stays basically the same or whether the absolute score stays basically the same. And so for me the repeatability is or the split sample is there's not something else it's just another way of testing how likely are we, what's the evidence that we're going to get the same result.

So the signal to noise is about how there's one way of measuring how tight the estimate is for a given provider. And split sample's another way of measuring how tight the estimates are because when you split the sample you get the same results. But that to me is the essence of what we mean by

precision that we basically have a tight enough estimate we're going to get the same results if we do it over and over again in the same data with the same kinds of patients.

Jack Needleman: We're going to start a wording rabbit point for people who as they're speaking a card goes up. You get one rabbit point each time that happens. (Gene) and (Sherry) and then (DQ).

(Gene): This is perhaps a question to (Larry) and his team. In your paper you do a wonderful job of describing or introducing the concept of risk adjustment when you start talking about scores. There are no such magic words under measure score. Would you suggest or would your team suggest that we highlight the fact that we're talking about your risk adjusted score and its reliability as opposed to just, you know, a non-risk adjusted measure?

And I agree with Jack that what we're talking about here is whether or not there is some level of consistency or repeatability okay in our measures that we were not getting wildly different things when we split up.

(Sherry): I think with 40 people on this panel the likelihood of consensus drops to near zero. And so this is probably just as a base that we should probably take offline because we're not going to achieve. But I think that we don't want to be reductionists because the precision estimate's or the reliability estimate's going to change with the nature of what you're observing.

For example cap of statistics or reproducibility across raters. And it depends on the nature of the measure, depends on what you're doing with it. The purpose of the measurements involved. I think that setting thresholds is a mistake because measures for example are not going to have that kind of

evidence base to build on. And it depends on where we are in developing the measure. I think getting past this is a real goal for this group.

But I think the debate and the discussion is proof. So I just think that this is the kind of thing that is also prevalent in the community developers. And so just kind of continued conversation as it relates to change and NQFs criteria. And so I think if this definition is going to then change what you're doing, it's worth additional discussion. On the other hand if we're going to have debates about when to use what, what threshold is the best we'll be all morning here on that issue.

(DQ): So when I look at the IP I come from two perspectives. One is (evalerative), one is discriminative. Now for (evalerative) to give you an example my admission in medicine, the hospital admission, those you should want to measure them as (unintelligible) as little measure of error as possible. But it's not whether your rate is different from others. Also extra rating matters. That's what discriminative measures. Sometimes it's more interesting the different the (unintelligible) entity.

But those kinds of measures it can tolerate a little bit of measurement error and which entity. It's the variation and amount of entity are allowed you can still get a good signal. That's how I tend to look at that.

Sam Simon: So I just want to…

((Crosstalk))

Sam Simon: …the measures in your use of those terms (unintelligible) in one group or another or are they measures that can be used in both ways?

(DQ): I think sometimes not clear separation can be used both ways but the intent now you can see which one is more on, you know, kind of continuous spectrum right? And you can see whether this is really more on validity. Even in looking at the data element right for lab testing or vitals you want to be precise. You want a measurement of error.

Now for rate or rating right based on interview, based on initial assessment, then it's more consistent right? Not message goal standard. That's for lab test, for vitals you have those standards. You want to look at that from a validity perspective. You want to get a precise measurement. And then for certain other things it's more really about, you know, from a discriminative perspective.

(Dave): Okay. So I'm going to just make a couple observations and see if I got some things right, right? In terms of the table in front of us does everyone agree that it's fair in terms of this 2 by 2 grid they put an X in every box? Does anyone think that's a problem? So we that could solve some of the maybe superfluous differences about this or that. And you can look at repeatability and precision, how you can rename those but need it later in the paper for both data elements. Okay so that's one thing.

And that's just I heard kind of three sequential needs that we have achieved at least in the paper. And then I'll make a suggestion about what we could do this morning. One is terminology. We have to, you know, to (Mike's) point, you know, we have to put our flag in the ground as to what we're saying these terms mean. And it may not be that anyone outside of the room agrees with but at least there will be a paper in the draft.

I'm hearing that we don't necessarily have that conversation today because that could take up the morning. And it's maybe it's better to do when (Dave)

takes the outline and produces a first draft that we can all react to. Is that reasonable? So that's terminology.

The next is methods. And that's probably where we should spend our time right? And then the third is thresholds which maybe we need to dodge for a bit but the paper's going to have to speak to thresholds even if the paper says we're not going to give you a specific number but here's a range that we tend to like and here's a range we tend to not like and there's something in the middle that's conceptual and that's probably how that's going to end up. But what's the point of talking about that right?

So let's focus on the methods. Does that make sense? Yes? And we'll actually have terminology methods and thresholds in the paper. But we really need to dig down in the methods because I think, you know, (Larry) said something around it was quite helpful. People have lots to say about just that topic, (Jay).

(Jay): Totally I love your farming of it (Dave). With regards to the thresholds I do hope we'll have some discussion today and I love your framing. They're clearly thresholds. Everybody seems to agree high enough. They're clearly thresholds. Nobody thinks is high enough. And there's…

((Crosstalk))

(Jay): …in the middle. But one of my frustrations in this last cycle was I was looking at the results. I was trying to make my own decisions about whether it was reliable enough. And it would have helped to have some sense of thresholds or a common consensus among the committee about what the statistics mean and which ones are clearly, you know, we all agree. And it's based upon the degree of consensus and the results. I think we have some

commonality in the threshold. And not perfect but if I look at the way we acted in the decision making, people were drawing very similar judgments about when the number was acceptable and when it wasn't. So I do hope we'll get to some element, some discussion threshold before the end of the day.

(Larry Glance): I like the framing in the plan. The one thing I would suggest is that as we talk about the method, that people try to articulate as clearly as possible what are their assumptions about what they're actually trying to measure. So to help us to communicate and it will also help us to I think ultimately to frame the (unintelligible).

((Crosstalk))

Jack Needleman: And then (Christy) and (Dejon) next.

(Larry Glance): Hi so I think this was a great discussion.

((Crosstalk))

(Larry Glance): So one of the points that we ought to think about and we put together in that first whitepaper is the idea that before you look at measure reliability you need to look at risk reduction because if you don't risk investment to the extreme case, there may be extreme variability between the providers because of differences in case mix okay. And so if you do a lousy job with the risk investment then you're going to have a more reliable measure. So you can never think about these separately. That's the first piece.

And then the second piece and this is corelated and I don't know that we always think about these as much is that, you know, in order to get more – the simplest way to get

a more reliable measure is to have providers who have more observations. And more observations they tend to have – I mean one of the factors is reliability. The more observations you have the more reliable the measure will be in general okay.

And so in my little corner of the world in surgery okay the way measure developers do that is they aggregate lots and lots of different surgical procedures together. So for example the American Cardiac of Surgeons in these (unintelligible) models. They take everything. Every non-cardiac procedure and they put it into one model. So now they have a lot of observations for a hospital. So you can create more reliable measures.

But there's a tradeoff with that because the risk adjustment becomes a little bit more problematic. So when you're putting in breast surgery and thoracic surgery and in a whole really big spread of surgeries into one model. And then the risk factors in that model have the same co-efficient regardless of what surgical procedure you're undergoing. And, you know, on the face of it it's not really valid. And if you test it it's going to turn out that it's not going to be valid. And yet so you create a more reliable measure at the cost of having worse risk adjustment.

So it's, you know, my point being these are really complex issues that we're talking about. But they're really is this interaction between risk adjustment, between validity and reliability. They can't really be considered separately in the sense that, you know, one does affect the other.

(DQ): So going back to the point of (unintelligible) I think one of the points that had been brought up earlier with the new (unintelligible) something, like, (unintelligible) and everything. Can we think of any sort of (unintelligible) for at least for maintenance measures. For example when the maintenance measures come for sort of reassessment, can we at this point (unintelligible) as

to how this particular measure has been used and then sort of use some I don't know from the particular measure to come up with some sort of (unintelligible). I am (unintelligible) on that and (unintelligible). Do we know enough about the measure that has been (unintelligible) for the three years and then, you know, that information (unintelligible), you know, that's all?

(Laci): Let me jump in here just from the (unintelligible) perspective. It is a question and it is something that comes up. What we have tried to do in the past is we tried to ask that you even though we know sometimes what the use is and what the intended use is, we try not to let that influence whether we think it's kind of good enough for you. And that's really hard to do. We we've tried to steer away from that knowing that different groups may use things differently. Somebody might use one measure for internal QI only. Somebody else might be using it for a major payment program and anywhere else.

So we have tried to stay away from that. It's very difficult to do. So I think it's kind of on the table generally. I'd like to cool it off the table for today's discussion but I get your point that in the threshold world when we get to that discussion maybe a little later we might want to revisit it.

We did try this a couple years ago and we really had no good evidence basis to say if it's used over here in this kind of program we, you know, you want to, you know, you're humanness wants to do that. But it's, like, well why would a payment program – why would you need any better numbers for payment than for public reporting right? I mean it feels, like, maybe you should but we couldn't come to that – we couldn't come to agreements on that.

(Larry Glance): Just very quick response on that. There's another dimension closely related. It's not exactly the type of program but it's more of statistical mathematical. You know if you're going to use a measure to identify the worst 10% of some distribution with just 10% you get certain issues that have missed classification that are different than if you're in a public reporting and you want people to make choices to help the distribution.

And you want people to say well is this hospital any better than hospital C and they're both kind of literal distribution. It's just that it's a different decision issue and has different reliability issues (unintelligible). But that's at least one that I found that is math and fiscal distributions than it is about the qualitative nature of the program.

Jack Needleman: Did you still want to say something? You card was up and then down.

(Christy Teigland): It was. I just had a really quick comment that, you know, there are already site thresholds in a lot of the measures I reviewed. And they still say NQF says 0.4 is good enough or the CMS 0.4 says 0.4 is. They do it already. And they present that as though this is the threshold that people have said and that's apparently not true but that's what's I follow off.

Jack Needleman: Precedent, thank you.

Man: So just to follow on what (Larry) just said I've been looking at the listed model – university model right? And it is (unintelligible). Sometimes you combine the patient on multiple specialties. Then you can even get a good model. The model performance and that can be good. It doesn't work well for certain specialties. So now if you look at different entities. Some entities only, you know, have certain type of, you know, specialty procedure right. And thus you have a different kind of impact.

In reality it's driven. If you had a combination, you had some specialty, you had (unintelligible) or a lot of (unintelligible) you're going to get a very good performance (unintelligible). Everyone likes these because you're going to get very high. So, you know, you have rare event right? Sometimes (unintelligible) but you just needs to be mindful if there's nothing.

Sam Simon: Well (Larry) as I was listening to, you know, your comment about whenever you use the term reliability I kept going to my, you know, safe place. And it wasn't really at all what you were talking about. And then I began to think all right so what in my terminology if you will what are you really referring to. It seems to me that you're referring to differentiation. You're saying that you differentiate and be full if you don't deal with risk adjustment. Is that – you kept that term differentiate for the example you were giving?

(Larry Glance): Well I'm just wondering if maybe getting back to I know when I'm talking about terminology. We agreed on that. I even suggested it. But maybe this is kind of like a differentiation reliability. If we could give it a modifier that refers to a very important aspect of reliability that has really nothing to do with the reliability of a data element or even other aspects of reliability of a performance score, you know, is it you too?

Man: I think what he's trying to get to is that there's (unintelligible) suppose you (unintelligible) and model right? When you don't adjust for anything you're probably going to find loss between (unintelligible) and variance right? So there appears to be a lot of variance but that may not be fair because you need to account for patient because mixed difference, you know, across provider. So when you adjust that you can refuse, you know, the variation and (unintelligible) entity.

Sam Simon: We could call it variance management. It's a part of me that doesn't want to call it reliability.

Man: Yes.

Sam Simon: But maybe I'm, you know,…

Man: You don't want to over adjust either right so it's kind of – you don't under adjustment. You don't want the over adjustment.

Sam Simon: Yes I think I understand the goal.

Man: Yes.

Sam Simon: I just don't think of that as reliability. And, you know, if that's me then I'll stop. But I'm just wondering about how we can get the terminology that everyone says yes we know what that is, you know, in a way you're talking about variance management or…

Man: No I think you calculated it based on our obvious model (unintelligible) a variance estimate on the model. (Unintelligible) but it's still you're gaining.

Sam Simon: Okay. Sorry (Sherry) and (Mike) had a photo-finish over here and then (Jack). Oh and then (Mike) okay. Heard her first. Were you up? I didn't see you over there.

Man: I'm ready whenever you want.

Sam Simon: Anyone on this particular point and then we can move on yes.

(Jack Needleman): So let me just make – I heard (Larry) say something that I think is useful here okay. He said and this is a little bit out of step with – no I wasn't trying to be ironic at all. Okay.

((Crosstalk))

(Jack Needleman): I wasn't trying to be funny at all. It was that you need to…

((Crosstalk))

(David): I will. So, you know, I think when our developers fill out the applications they're very form oriented okay. And the reliability piece comes first and then the validity piece and then the missing data etcetera comes at the end. That's when we start to have them talk more about risk adjustment. But (Larry) said something important. He said you can't do, you know, good reliability without doing risk adjustment first.

And that was in the context of talking about this issue of precision. And talking about signal to noise. And what signal to noise is about is differentiating between providers with and I'm like trying to quote as best I can the Adams R tutorial that we circulate so widely that says the key assumption there is that the only variability that's left over is the stuff that relates to quality. If it relates to case mix you're in trouble right? It's going to fail. And you're going to see variability. You're going to see signal that you think is real that isn't real. And it's going to give us a fallacious idea that we're reliable in being able to differentiate between providers when we're not.

So the key thing that I see and why this thing seems to be tripping us up between repeatability and precision, if repeatability is about you take one measure, you take your bathroom scale, you get on it and you get on it a

second time, you get the same weight. It's a single point in time. When we get to precision we're talking about doing that procedure across a whole bunch of different bathrooms right? A lot of different households doing it repeatedly and then saying we're willing to tolerate random noise as long as it doesn't overwhelm the signal we care about.

And we can only believe that if we risk adjust first. I think it sort of reorients things a little bit. I think, you know, the paper tries to talk about that. So that was, you know, sort of the key thing that I heard coming out of this discussion or one of the key things.

(Larry Glance): Thank you. (Sherry), (Mike), (Jack) thank you.

(Sherry): So I was a little worried about discriminate validity and the confusion that, you know, that caused. Discriminate validity means you can tell the difference between some groups that you've identified. So that's validity, that's accuracy. So I think where the variance components get to the issues that (Dave) raised is if the standard error of measurement is the standard deviation times the square root of one minus the reliability. There's that containing term of variance and it contains reliability.

So if you put error bars, if your thought means and say you did some generalized estimating in places you get a noise, thought everybody meets with the standard error bar, the standard error bar reflects the amount of error in the measure. The difference between whether or not you can tell any point estimate from another point estimate you then use that standard error bar to kind of gage how much of the variance is not, how much is error. Let's just call it error.

((Crosstalk))

(Sherry): Precision yes. And (David) the only thing I disagreed with you about was precision. And your use sounded, like, calibration to me. That's, like, the points after the decimal. But then precision gets to the issue of how much of the variance is noise and how much of the variance is actually some table phenomenon that you can call whatever it is you want for validity.

But I think the thing comes together around the standard error of measurement. So if NQF1 is to write down some guidance, then have everybody support some estimate of the standard error of measurement.

(Mike Stoto): So when we talk about differentiation that pulls to mind what I think ultimately what are we trying to do with these measures. And I think that there are two properties, really one. One, is that when quality improves or is better between one and the other that the measures are able to pick. And the other one is that when the measure shows some change – improvement or a difference between two we want that to actually reflect the difference in quality.

Reliability, precision, validity all help us to understand those properties. I think that having in mind that goal about what makes a good measure – those two things maybe a useful way to think about it.

(Jack Needleman): I just want to double back to (Larry's) comment about the risk adjustment. And I do think risk adjustment is an inherent part of this reason that (Michael) said which is that you want to eliminate sources of variation that don't have anything to do, that have to do with the patient mix at the individual provider level. But don't reflect the underlying quality. So if we were to take them – but I do think that the issue that (Larry) raised about using the NQF in this

example of the precision of the risk adjustment as you get more diverse patients into the mix I think can be overcome.

All the risk adjustment models will run on basically the whole sample or some other sample. So you've got the aggregate of all the patients that are there. Now given that you've now aggregated a bunch of different surgeries, you might want to put in a dummy variable for each type of surgery to capture the underlying mortality rate of that surgery. And allow for that kind of variation to be sopped up in the risk adjustment.

You were concerned that some of the other things you control for say age that the gradients of mortality across age is different for surgery A than surgery B. But that can be dealt with by interacting the age variable with the surgery. Assuming there are enough cases across the whole sample we will get reasonably accurate estimates of that interaction. That's not the problem. Then we go to an individual institution which has a quarter of the surgical types that are examined and a different age distribution for each of those. The risk adjustment model which is based upon the total sample will capture the expected mortality for each of those individual patients.

So I think the issue you raise is one of the sophistication of the risk adjustment model not an inherent problem. And one of the ways it gets dealt with is through using much more thoughtful interactions about things, like, age and so forth that may interact with a specific element of this very diverse pool that you pull together. I think that's a solvable problem.

(Mike Stoto): So I don't want to get too much into the weeds here. But it is a really complicated problem with a lot of procedures and for modeling. You know there are literally hundreds and hundreds of different procedures. And that has been dealt with in a number of different ways. Initially they tried to adjust

for surgical complexity by using (unintelligible). And then they said, you know, that doesn't really work that well. What part of the use are completely cured (unintelligible) and basically expert based.

And then they actually introduced surgical procedures because of its random effect. I think this gets very complicated. So then when you try to do interactions between 300 different procedures and each one of the 20 or 25 different patient level risk factors it's not something that is a prevalent problem. I think people have really thought about this. And there is no easy solution to it. It's not, like, you have three or four different patients. You have hundreds of different patients.

So it goes back – I mean and I don't want to get into a back and forth. But my point being is that you can aggregate lots of different diagnoses, lots of different surgeries and come up with bigger populations and therefore more reliable measures. But there may be some cost involved in terms of just how good the risk adjustment becomes or maybe starts to fall apart. That was the only point that I wanted to make.

Ron Walters: Oh thank you. So in my confused mind risk adjustment was really a function of validity more than liability. By which I mean that, you know, if you have a person biased because the SDS does it by procedure rather than lumping all the procedures together but they have that problem in the pediatric side. So it's a big problem.

But I understand that the question to me for reliability is if you took this same group of hospitals, so the same group of surgeons and measured them again and again, would you come up with the same answer. And the question of whether or not the distribution actually reflects quality is not an issue of the reliability of the measure. It's a question of the validity of the measure. How

well does that difference distinguish quality. That's a problem of validity. It's not so much a – I don't see it so much as a problem of reliability. I think provided your, you know, un-risk adjusted or poorly risk adjusted measure comes up with the same answer in the same distribution every time it does it, then it's functioning reliably. It just isn't a good measure.

I was going to try to respond although (Larry) is (unintelligible). I think my understand to this point was that if you're thinking about reliability you're signaling noise from it. If you don't do adequate risk adjustment, you just pick up a stronger signal than you would otherwise. And to your point which I also agree is it's both stronger and wrong. And the issue about reliability is you want to get down to its proper level of strength. And then you switch over to validity and we say now is it correct. Is that all clear right?

(Paul): In general you want to make sure your risk adjustment is right first because the risk adjustment attenuates the amount of – it accounts for case mix. So basically if you don't account for case mix you can have a lot of variation across providers. And so then your reliability will come out pretty good. And then when you adjust the case mix, okay then the spread between providers is going to narrow a lot. And so then your reliability is going to go down. So I agree there's clearly two separate concepts. I'm just suggesting that before you evaluate reliability you ought to evaluate validity.

Ron Walters: So the reason I said what I said was because I see the world the same way your comment reflects (Paul). And I don't think of that as reliability and you articulated that thank you. If you go back to your scale (Sherry) and let's say we have nine planets. We'll keep Pluto in. That scale I think is very reliable but it would give you a really different number on each of these nine planets, right? So it's not valid for comparing the planets to one another because you haven't gravity adjusted. But it's a reliable scale.

(Sherry): I don't want us to get into a big back and forth either because reliability – actually risk adjustment can do the reverse. It can spread the distribution. So you can't guarantee that just because you risk adjust at the front end it improves precision, it might not. I'm not saying when you should do it. I'm saying you should definitely do it. The question is when you do it. And do you need to actually establish the precision risk adjusted and un-risk adjusted.

And could you look at the variation and absolutely because it can spread the distribution. It can cause you to bear more noise in the measure. And it depends on, you know, if you establish the scale on planet Earth where there's X gravity and whatever, and then you transport it somewhere else, it has to be reestablished or it's reliability on that planet under those circumstances. And that was the point I was making.

But if you're looking at it for different purposes, you know, then the risk adjustment gets to be dodgier and dodgier because all healthcare is nested. And if you're adjusting first at the hospital level and then at the physician level? Are you adjusting at the physician level only? Are you adjusting level for composite measures at the patient level and then the physician level and then also that level?

So, you know, we can actually have kind of issues about when to do it but should you do it absolutely. When you do it I'm not so – I'm kind of agnostic.

(Jen) Perloff: This is (Jen) on the phone. Can I jump in?

Ron Walters: Sure go ahead.

(Jen) Perloff: As you're talking about the fact that risk adjustment can cause harm as well as benefit, price standardization came to mind for me. And I spend most of my time in resource use and cost measures. And so standardizing the prices that's another kind of variation. It feels, like, we're decomposing the variance here some of which, you know, we want to get rid of because we know what it is. And some of which is, you know, that residual random noise.

And so just as you were talking that idea that price standardization can function, like, risk adjustment, it can be effective at taking an unwanted type of variation out of a measure. Or if done poorly it can be adding, you know, more noise. So I wonder if that helps clarify or generalize the concepts at all. But that was just a quick thought.

Ron Walters: A clock observation here we're coming up on scheduled break time. And we should probably stay pretty close to that. We have folks on the phone and others who are sort of adjusting their schedules for our topics of discussion. I'm just wondering either as a potential closure but also something we can think about in the break and come back.

You know we're talking at least at the moment about whether risk adjustment perhaps primarily lives in the domain of reliability and validity. I guess I'd offer a suggestion. It has a role in both at least I've heard sort of arguments on both those points. And also in the current NQF instructions of the developers it has its own separate heading. It's not listed as a subpoint under either reliability/validity. Is that correct? It has its own standing as a heading.

Woman: It does but it's under – it's nested under validity.

Ron Walters: Okay. Well then I'll pull that (unintelligible) but – but one possibility could consider is first of all give – give it its own separate level of heading and then

perhaps have it come first to suggest that it's something that does have potential effects on both reliability and validity. At least I think we would address (Larry's) point on paper with which I agree, that risk adjustment does have a financial effect (unintelligible) at least convey the signal in terms of where it falls into the form that it's not just a sub-point of validity with that. It doesn't require us to say it's just this or just that. I'm not sure. I assume consensus (unintelligible).

Man: So just to clarify I don't think I was suggesting that we consider risk adjustment under reliability. I still think it should be under validity. I'm just suggesting that when we evaluate measures we first evaluate validity (unintelligible) because risk adjustment is so much a complicable piece of that. And then once we've evaluated the risk adjustment then we evaluate the reliability.

First we evaluate validity of the measure then we evaluate the reliability. And again the reason being is that in many cases if we don't view proper risk adjustment we don't count the replacements, we're going to have a bigger spread in terms of the signal. And so the measure is going to look more reliable than is actually (unintelligible).

Man: Yes. Clearly different suggestion and I just may (unintelligible). Now is there any basis just because of the current sequence in how to worry about (unintelligible) reliability? Are there situations with people (lack) of judgment or (unintelligible) where (unintelligible) should be established first? It is now, I mean, just in terms of the sequence. Your suggestion that, that be (unintelligible), that (unintelligible).

Man: Yes. So we haven't been talking about (CROs). I think maybe that (unintelligible) that comes from multi-item data and the…in my (unintelligible) and then look at its validity, (unintelligible).

Man: Maybe we can leave that hanging and (unintelligible).

Man: (Sherri's) got her face (unintelligible).

Woman: (Unintelligible).

(Karen): So this is (Karen). I think this is something that we have thought about many times. Should we do validity first versus reliability? I think internally what we'll have to think about because we do think about specifications as part of reliability and we have to think about whether we should think about that separately first, talk about steps and then maybe validity and then reliability. So that would be my…I would hate to wait until after you have all of that discussion and then talk about that.

Man: Sorry. I know we're all ready for the bathroom break but would it work to say for data elements reliability before validity and for performance measure score validity before reliability?

Woman: I don't think that works for the following reason. If you're valid but you're un reliable, so you show up 15 minutes before the – the game starts or 15 minutes after the game starts. On average you're right. You're right on time but you're completely unreliable so the coach kicks you off the team. You know, there are very few examples where that's true but, you know, that – I feel like if you're accurate and unreliable you – what then is the point of your measurement?

So to me precision thing comes first. The reliability comes first. You have to think about am I able to reproduce it. If you reproducibly wrong then that doesn't help you either. So in terms of sequencing I don't feel a strong case for where to put risk adjustment but I think it does warrant its own separate – it's what you were saying (Karen) so maybe separate section so it's separate section right now. It's under validity but to me it's less of concern where it goes and that it goes but putting validity first, it's the scratchy record thing for me.

Man: All right, probably time for a break (unintelligible) out there who are not (unintelligible). I'm not feeling – I'm just an obvious consensus that (unintelligible) so maybe a break. Think about it during the break, experiences that even if we say it's a 10 minute break it never will be. It can't possibly do a 10 minute break. Fifteen is minimum so I think we're talking about ten minutes to the hour reconvene.

Woman: Thank you.

((Crosstalk))

Man: Okay if we could take our seats, reconvene. It'll take a minute but see if we can get going again soon. (Michael Abram) said to me at the break I didn't know quantitative people could talk so much.

If I could just follow up a little bit on new lead discussion. There are a lot of side discussions that I'm not sure are going to take us immediately and directly to a consensus on the issue of sequencing and labeling. (Karen) reminds me though that if we're going to make any significant change in the forms and instructions to developers (unintelligible) you'd be too late to

implement that before the next upcoming cycle which means have a little time to kick this around.

So we may continue to talk during watch and there's just some difficult issues again which may reflect our different academic tradition. Some folks just can't imagine discussing validity before…before reliability. It seems like they have to establish one before the other. There's the discussion interesting about well, you do it one sequence for the data elements but then you might do it in a different sequence for the measures scored and both cases recognizing the prominence of risk adjustment.

So I just don't think speaking of revenues that we're going to get this sorted out but (Karen) points out we don't need to right now. If we can get there fine and maybe somebody will have an idea at lunch but I think we're going to try to (unintelligible) just a little bit here.

Woman: Yes we have the lost the (unintelligible) that we've been playing around with. It seems like that we're really behind but because we only got through one of our reliability sites. Actually I think we're doing okay because all these things kind of inverse with that so some of these slides we can just ignore, not so much this one.

I think reliability not meaning (unintelligible) possibly the measures what (Sherri) mentioned earlier so we totally agree with that. And there's a few other things in here. There's always going to be some error and reliability is not an all or none property. These are just our usual assumptions so I'm going to just skip over this unless somebody feels like they don't agree with these assumptions. I'll give you just a second to look at this slide. Hopefully these are all correct.

Man: I don't want to (unintelligible) the color style. I like those. I just wondered if that's consistent than with how we (unintelligible) is operated. Somebody doesn't test the liability of certain time periods. Well show us a number. We say okay, what's a good number and see things reliable, (unintelligible). So I agree with the concept. I'm not sure we act on it.

Woman: It really should be because we often sin and say that this measure is reliable and we're not supposed to say that, right? But maybe we can think about we can more appropriately use our language to reflect this idea because I think we all understand it and we all agree with it but we're lazy in our language and (unintelligible). So that's just something that we need to pay attention to. if anybody has ideas on, you know, should we always stay in this particular context or some…you know, is there some uses loaded here because we always think about programmatic use but yes.

Okay I want to skip the terminology. We've already talked about data elements and scores. Just a little bit how much do you guys (unintelligible). You guys wanted to see talking about data element reliability because most of our discussion thus far has been about four level and kind of intersection of our terminology. Did we want to delve into data element reliability or did we want to finish kind of our score level discussion in whatever time we have left with the data elements?

Man: I think we can defer it to later but the one thing that I think we wanted to be sure to talk about was single – single item data elements and you know, obviously the internal consistency is not relevant in what should be required or expected.

Woman: Okay. Why don't we do that at the end because we're turning it on its head, finish our score level and then do the data because that's that one, okay. We

talked about this already in terms of definitions and let me be clear. When I said that validity is a bit of a rabbit hole I didn't mean the concept. I meant the word, that word, you know. So let's just ignore that last bullet there and we talked about decisions and whether we're going to concede network or whether we're not.

I think we covered that in our discussion (unintelligible). Okay. So some of our common approaches, this is getting to (J.C's) idea of definition first then method and rational later, all pretty much there. All the jack wants to have a little bit of that. So this is getting to our method so there's things and differences in demonstrating classification between providers (unintelligible) most of common approaches are. And I think one of the questions that I would have, I don't know if made it to this slide, not quite.

Just in general when I talk about reliability I used to always say it has to do…we already covered this a little bit. It has to do with me able to – can you distinguish between (unintelligible) in differences which in my mind was that decision to. But I might need to be changing my language in talking about risk of misqualification. So (unintelligible) your advice on just how we describe what we mean by the liability. That didn't quite make that side event. It's okay.

So we have this idea of the differences between providers with half, with ICC, with half versus Pierson, (Larry) did. We get resistance (unintelligible). That was great. Each guy was paying attention to ER back and forth. (Larry) has already given us some good information there. So when we say split half it's this idea of splitting a sample that you have and scheme agreement between the two samples.

Some people we do see use ICC. Other people use some kind of other correlation and our question was of all valid -- do I dare use that word -- or should be insisting on one versus the other. I think even that, that's very specific advice we can provide (unintelligible).

Man: So if I could I didn't play around with the file that you sent (Larry) but it was persuaded by your email. I'm that one. I could regret but I could put one decision or recommendation the table which is ICC. (Larry) emailed it to those of you who read it demonstrated that you can be misled by simple correlations. So that's something we could agree to go into the paper.

Woman: (Unintelligible).

Man: Any disagreement? Wow. Okay. We may want to just trade the other but I – turning to (Larry) a little bit there are also some things that mathematically look like and might actually be equivalent to ICC but they're called something else like IUR or other things and I think one of the things that would be really nice to establish, work but also get developed is these. But things are truly different, even the concept or (unintelligible) and once things (unintelligible) the same thing.

And I think we – in the email exchange our most recent article said IURs expectedly a (unintelligible) to see although in the article there's just a post related. Now there's the post of okay well, what does that mean?

Man: Internal things.

Man: Yes. I mean, 2.71 just like another or (unintelligible). I don't have a problem with this suggestion about us stepping up and saying ICC but was that recommendation all bringing along with it IUR, something, something,

something? In fact, you know, what I've seen occasionally is (unintelligible) is something that comes to your sensory label signaling (unintelligible) so what's that? Is that just another name for the same thing. This to me is the hardest hard of the better (unintelligible).

Man: (Larry)?

(Larry): This is a 30 second comment. I think that the goal of – of coming together is again hooking me up with a list of best practices and saying this is what we want. And that will encourage the systems (unintelligible) for them first and if we'll incur consistency in how we evaluate the measures and I think – for so example when you're evaluating split half reliability, if you look at the literature, if you look at atoms and (unintelligible) and all the other (unintelligible) the medical literature on this pretty much invariably they use the signal noticed by (unintelligible).

So let's use that as one way of assessing reliability. And then others, when they do the split half reliability test – sorry I misspoke. When they do the split half reliability test you use the ICC. You don't really see the IUR mentioned very much, at least not in mine using the literature. So rather than using things we can just say for spilt half reliability (unintelligible).

You don't have to cover – I don't think we need to cover all the different possibilities (unintelligible). You can just keep (unintelligible). I would think that (unintelligible) evaluating (unintelligible). They're not looking to – they want to know what (unintelligible) give them that template (unintelligible).

Man: IUR, (unintelligible) state right? For the (unintelligible) base they kind of operate closed (unintelligible). You can use in the same data assessed. (Unintelligible) so I always (unintelligible) putting as much (unintelligible) so we know that was one of the (unintelligible).

Woman: This is one of my three for the hour but I agree that we should give some guidance but I'm a little nervous constraining it to (unintelligible) because one half reliability, is nothing magical about that. That comes from the whole testing thing where you have 500 – 1000 math questions and you can't give them all to everybody because (unintelligible).

So you split the saying and 100 questions. Split it in half and give random half of (Kristie) to one group and a random half of (Kristie) to the other group and you get more or less (unintelligible). But what's wrong with doing that 100, not just simulating but doing 100 random samples of X numbers of hospital – X number of samples of that hospitalization.

This (unintelligible) would give you a similar kind of answer and I would be – like to see or say we have to do (unintelligible) because (unintelligible) get that same level of (unintelligible) so there are at least answer the (unintelligible) questions so (unintelligible) offline instead of names and recommendations about other kinds of things and then (unintelligible) consideration.

Man: I think that (Larry's) email is persuasive but if I understand it it really relates to risk adjustment outcome measures. I have no idea whether or not this idea slide says things like (unintelligible) flu vaccine at a certain time. It was kind of simple process measures like that or measures based on people respond to a survey with the most (unintelligible).

Woman: (Unintelligible) things (Larry)? Sorry.

Man: I'm not even sure what it means when you have a simple proportion. You know, (unintelligible) form. I'm not sure how you would do these

calculations. Maybe it's not. I think we need to kind of think it through before we make a blanket statement (unintelligible) this whole time. We should think through how it applies in the times and measures that typically come for this many.

I'm not sure we can argue from first principles here. The ICC is the gold standard. Everything else sort of box mates it and (VBH) determines. But my suggestion that that might not be correct and I think in the measures I was looking at this last cycle Acumen who some had lost some but consistently did I think a nice job in presenting material.

Present – gave us the opportunity to look comparatively at a number of different measures because they calculated to single to noise, they calculated Pierson split sample correlations. They did some other stuff but I'd like to see before we make a final decision on preferred methods, is to actually see in several different data sets with several different kinds of measures the – the – the – the performance across them and with enough detail in the analysis to understand which ones we think are performing well and which we think are more questionable.

I'm inclined to – to give credence to its ICC but I don't think – I'd rather – I'm a (unintelligible). I would rather see some comparative statistics across the similar data sets so we can understand how they perform differently and why they perform differently so we can make a more informed direction.

Man: May thank you with that. Thanks for actually putting that together. So I don't know. Is it possible to put the email that you sent out with (unintelligible)? So one of the statements I had, there is basically the first (unintelligible) how to make the list, the (unintelligible) email. But I can look at the (unintelligible) and I use data quite differently and then (unintelligible)

sharing that work and I can apply (unintelligible) this morning (unintelligible) simplest thing, there is no systematic (unintelligible) difference between ISMR-1 (unintelligible) ISMR2 and then we go to second one so I guess that's my question. (Unintelligible) transfer between the first (unintelligible). The only difference, the way it's getting (unintelligible). You can go to that poll. So I think that the way you created ISMR2 is basically (unintelligible). That's random (unintelligible).

And then we are saying that there is no difference between (unintelligible) and the second one is simply (unintelligible) by our (unintelligible). So I guess the only difference there is (unintelligible) in the first one, in the first difference. It is the (unintelligible) ISMR by one and then the second is since we are (unintelligible) ISMR1 by two and I guess my question there is why are we saying the (unintelligible) that we don't have any (unintelligible) because the way we see it ISM2 (unintelligible) what I'm saying?

Man: So the staff – the staff is finding email. How about if we come back to it? Maybe (DQ) has a comment. We'll come back to it.

Man: Used the correlation, I think one situation, I hope (unintelligible). You compare the streamline so 0.9. You should compare the line to the line of equality. You make them only the line close from line of equality because you can get 0.9 or even 1 very flat lined. Need two measure, one is very different from the other one. So I think the correlation, maybe you should treat it well. Make sure the (unintelligible) is square because it really matter. Make (unintelligible) square and then you compare the line to line of 45 degree (unintelligible). The correlation (unintelligible) 45 degree nominated (unintelligible).

Man: Let me go back to the first thing first. I think lots of (unintelligible).

Man: Go ahead and then (Paul).

Man: So…sorry. So if you look at this fund, okay, basically again this is split half sampling, okay? And what you're looking at is you're looking at some rate of the adjusted, with the unadjusted (unintelligible) difference. And they essentially rely on that 45 degree (unintelligible) and so when you calculate the Pierson correlation (unintelligible) and you calculate the Spearman correlation, core version and the ICC they're all very close to the fund (unintelligible), next slide.

Here what I did is I did a slow – a change in the slope, okay? So what you see now is one – one hospital has – in one sample the rates are twice as high as (unintelligible) who were changing slow and they lie – they lie on a very close, on a very tight line. So the two correlation provisions are again very close to one but the intercept correlation coefficient is way less than the market, okay? So if you would just use the plastic correlation (unintelligible) you would then think that these – this measure is – is very reliable and the ICC shows that it's not, next slide.

So here what I did is I just did an intercept shift. So I said that the rates in one – in one sample are the same as in the other sample plus some fixed number and I think I used the number 2. Again when you calculate the correlation coefficients, Spearman and Pierson are very, very close to one but the ICC is low, okay?

And then in the third example I did both a change in the intercept and a change in the slope. I don't know if you can advance the slide. There it is and again the two correlation coefficients Pierson and Spearman are very close to one. again the ICC is very low. So this is by no means a – this is not ready

for – this is…I was just trying to do a quick little demonstration. I just spent 20 minutes on my computer to show this but it shows, essentially at least…but it shows essentially is that the correlation coefficients are probably not the way to go if you're evaluating (unintelligible), if you're doing split sample liability testing.

You do want to use the past correlation. This is just a simple and pure sample. I'm not giving the statistical background for this and I was looking at (DQ) to do that. But I just wanted to demonstrate that the ICC is a better way to do it and this is in fact what people do (unintelligible).

Man: So (Gene) did you have a question then you want to add?

(Gene): Yes. I think my (unintelligible) to get to your point. And I think that beyond the (unintelligible) getting the (unintelligible). One is (unintelligible), you know, but that's the (unintelligible) but then from that point on (unintelligible) again that's the ideal case. That's why (unintelligible) simulation was starting with one but I think in the real world we're going to be getting something (unintelligible) from one and we don't know in that (unintelligible) what to do that or where's (unintelligible). That's what we're recommending.

Man: Okay. We have (Paul) and then (Ron) on the phone and then (Gene).

(Paul): So I – it's preview a little bit from outer space but the one thing that I very much like the idea of being a little bit prescriptive as not techniques that are acceptable. On the other hand I'm a little bit concerned about different technologies that are coming down the road in terms of artificial intelligence which is frequently a black box but yet may produce very reliable results.

And so I'm not sure that we – I think we have to keep kind of a little bit of an open mind as to what are the criteria by which we are judging these techniques. In other words if we could define what it is that we like about ICC what we like about the strapping and then we more sort of potentially leave the door open for something else that we haven't thought of or, you know, will come down the road that would fulfill those criteria. It might put us in a little bit more circumspect position.

Man: So (Ron) on the phone and then (Dave). (Gene) you change your mind? Okay. (Ron)? Maybe you're on mute (Ron). We can't hear you. (Ron) are you – if you're there we can't hear you.

(Ron): Sorry. I hit the speaker instead of mute. The subsequent not ready for publication presentation kind of took care of my point but I did wonder if – because I couldn't remember we seen any measures submitted that had examples of exactly of what he was talking about, examples where people did both in what the results were. I don't remember any…

Man: Does anyone remember any? I don't, seems like people pick their poison and then go with it. (DQ)?

(DQ): I think once (Larry) demonstrated I think. It's very convincing and you…I mean, it's pretty clear.

Man: I have a sort of a two part question with a little devil's advocate into it. I understand mathematically what you did. I'm not sure I can think of a practical situation in the real world where the shift or the intercept change would occur, if the slope change would occur so just a little illustration of that. And I guess then I also observe that in all of our panels you know, there's a

very tight distribution along this line and the highs and the middle are middles and the lows are lows.

And it seems like if that's constantly true and the two forms of correlation are always high but the ICC is moving all over the place, you know, the ICC that gets at the fine level truth there's the correlation that gets the funding level truth. So I just need to know more.

Man: So if you look – for example if you look at this particular graphic, okay, that red line is – it doesn't seem to be but it's the identity line, okay? And so if a measure reliable, meaning if it's reproducible you should be getting the same number in the first half that you're getting in the second half, okay? And what this is showing you is that all those two numbers are correlated, very highly correlated.

They certainly don't agree and the correlation coefficients are saying that they agree if you believe it, right? If you were using the Pierson correlation coefficients as a measure of split sample reliability testing you would get 0.99 here and so you would walk away saying that based on that correlation coefficient that the measure is reliable, okay? But clearly what this shows you is that the measure is not reliable. They do not agree with one another.

(Jeff): On the point…

Man: Go ahead (Jeff) and then (Susan).

(Jeff): Okay. So (Larry) I appreciate what you've done here but you've done is to systematically shift each of the points. And when we're talking about split sample testing it's a random split so the question I would ask is what – how likely we are to see the – the distribution in a random split deviated as far as

you have presented it from – from the – from the identity line, from the, you know…and that is potentially empirical.

I do agree if there are some deviations of course two splits in terms of either the slope or intercept then we're going to see the ICC and the Pierson correlations diverge. But I don't think it's going to be this much and the example we've got up on the screen, in the first sample the range is zero to two and in the second sample it's two to six and in the split sample of the same data there's no way the two distributions cannot – are not going to have some overlap.

So that's, you know, so yes. I mean, since we're dealing with random and not systematic selection in the two halves I think it's less of a problem than you presented. I'm not going to say it's not a problem but it's less of a problem than I think is suggested by the graphs and the calculations you did.

(Susan): So I just want to remind all of us and have taught statistics at some point in our career and I still do occasionally, that we ought to look at graph before we belive the statistics because if we had a quadratic function of the two it would show zero correlation but they would be perfect view, right? So we all know that classical example and I – I agree with you that – that this is a – we're talking about a systemic calibration problem, right?

So should I go to your scale? And it's like everybody's scale was – had the 0 set at 10 and that's, you know, that's what I believe, is that my scale's off by that much too. But I think we – I think in most of these measurement problems, since we have a random split that we should have a certain amount of random variation and it's unlikely they'd be functions of each other.

I agree (unintelligible) sort of perfect textbook situations that would happen. I also question that the likelihood of this kind of perfectly modeled bias or even close to perfectly modeled bias might be kind of unlikely.

Man: So…is it fair to say I knew that it was too simple just to get agreement? But it seems to me that it's fair to say that the ICC provides an additional level of comfort that is exaggerated in these illustrations. Maybe to (Sherri's) point seeing what (unintelligible) for a second, a second turn.

The text of the paper can allow for other approaches but – but I don't think there's a downside to the ICC. It adds a level of comfort, additional protection against slope and intercept bias or difference that there was exaggerated to make a point but I don't think it's – I don't see the downside. It's like getting purified water without using plastic or something.

Man: Okay? We'll move on. Does that sound good?

Woman: Okay. Let me go back to this first bullet, the distinguishes difference, classification. We kind of use this as a – I have then go through saying this is kind of a signal to noise analysis. The other day on the phone days and it sounds like we've – that's what I heard this morning. I wanted to see if I'm understanding correctly this idea of doing like the atoms data binomial signal to noise or some other. There's other ways of doing a signal to noise that doing a split half is kind of doing the same thing.

Is that correct or are they entered in two different questions? I think that's the first thing I want to understand. Are they pretty much answering the same question or are they answering something different or telling us something different?

Man: Well I…we'll put this way. So we keep moving. I think it's pretty much the same. If anyone in the room can see a difference please speak up.

Woman: (unintelligible).

Man: (unintelligible).

Woman: Use your mic too.

Woman: Can you restate the question please?

Woman: Sure. In the last year or so I was really kind of convinced that doing a "signal" to noise analysis, whatever flavor of that you do is something as kind of a different thing than doing split half, that they were telling me two different types of things, now I'm not so sure and matter of fact (Dave) you kind of convinced me on the phone the other day that maybe telling me in different words the same kind of concept.

So do we agree with that or are they telling us two different things? because where I want to go with this is if they're telling us two different things do we want people to give us both types of analyses or are we happy with one or the other or if they're telling us the same thing then it doesn't matter which flavor they get. That's what I want to know.

Man: (Sherri) was up first and I didn't see over here who was first. You guys can fight it out. Okay.

Man: So I'm probably not the best trained economiatrician or statistician in the room but they're trying to get the same issues. So if we go back to the issue we're trying to deal with it's the precision of the estimate for any individual

unit in here relative to the – the variant – the inherent variance and the phenomenon we're studying. So it's a matter of elements and precision.

The signals to noise uses all the data to make that calculation. The split sample by its very definition is using all the data but it's calculating the number on half the data twice for each of the entities. So you got to but in both cases it's trying to look at the degree of precision or imprecision and the estimates for an individual entity. And the split signal in the split sample, it says let's compare one estimate to another using half the data and how close are they and so same intent, different methods and therefore not necessarily completely correlated results.

Man: Sorry.

Woman: Are we – argue that you can't do this independent of, you know, sort of in general, that if you're looking at can you reproduce within a hospital the same score across two subgroups of that hospital's population? That's equivalent of split half reliability or boot strapping or whatever you use for that. If you're looking at the between with variance, between unit variance divided by between plus within unit variance then you're looking in your new class correlation. It may tell you something that addresses reliability but it's a different method of addressing reliability and it doesn't quite give you the same answer because you're…so I would say no.

Woman: And I would clarify I know it won't be the same answer but are we at least answering a similar question just in different ways?

Woman: No I don't think you are. I think you're giving – you're giving – you're addressing them out of variation that's attributable to between versus between plus the air term and it's giving you a different answer. It's giving you

something else about reliability so you're still guessing reliability but it's not using the same technique so in that way it's giving you a different answer.

Man: Thank you (Larry). (Paul) and (Mike) or (Mike), (Paul) or does it matter? Okay (DQ), (Larry), (Paul), (Mike).

(DQ): So if you look at a atom's formula there's nothing new. It is a ratio between variance divided by the (unintelligible) error. Even for a straight (unintelligible) ICC if you look at the formula it's also a ratio between – over between (unintelligible). We are not that different.

Man: (OFAR), two yes, one no, one in between. (Larry)?

(Larry): So I think they get at the same thing, okay? There are two very different approaches and from our discussions that we've had the – the split half reliability testing is a more conservative approach to estimating reliability which means that it's going to be – it's going to appear a little less reliable. How much? I don't know if you're using split half test reliability testing as opposed to signal or noise ratio. But at the end of the day they're both trying to get at the same concept and I think that in this case it's probably acceptable to allow measured developers to decide which approach they're going to use. But I don't think we need to have them give us both – both numbers.

Man: (Paul)?

(Paul): In the under informed group it seems to me that both are – rely on measurement error. And so if you could help me out to just understand what's the case where one would be different than the other, in other words what's the situation? And the reason why I'm asking this is because maybe there are certain situations where one's more appropriate than the other or if we

determine that well, really you did this but you should've done that. So is there – are there specific cases or situations that would help to illustrate how one is different than the other rather than on just theoretical grounds?

Man: (Andrew) is going to speak to that. (Mike) will get back to you.

(Andrew): Sort of, part of a question I guess. I'm far, far from a methodologist so I kind of hesitate to even speak up. But it seems to me like this would have your – you're regressing, sort of that precision question you're getting for an individual hospital. And you're sort of getting the repeat ability thing, you know. You do this twice or you're getting the same answer and it seems like the signal to noise, you're talking a lot about more about a relative discrimination ability, again the ability to distinguish those differences and it's much more dependent on their being differences.

And I guess to repeat your point the example I'm sort of thinking of is where they're all tightly bunched where there's not a lot of differences in performance. And if you did a split half, if you're, you know, if you're getting a precise and accurate measurement your split half I would think would give you a high – you make it a high score with a split half. But if you're doing a signal to noise, if you're tightly bunched it seems like you may be low. You may not be able to discriminate. I don't know if I'm right about that but it seems like it may…

Man: I have (unintelligible) example in front of me.

Man: You want to put it on the screen? Want me to go to (Mike) and maybe you can get…is it possible to plug his?

Man: Actually I was going to ask for something just like that. My understanding is exactly as you said, that imagine if you did an analysis with a certain data set and then you then brought in data with providers who work very different in terms of their – their performance. The signal to noise ratio would look better. I'm not clear that split half would but I think that's what we want to have but I don't know. I don't know that for sure. I think that I'd like to see some empirical examples to lay that out.

Man: (DQ) and then maybe the examples coming through an email.

(DQ): So one distinction is when you calculate it ICC for a split half you get one score for the measure. If you're using the formula in the errors to (unintelligible) the two types of implementation. One is data binomial. One is (unintelligible) model. To do that calculation you get one score for each measure entity. So you have 4000 entities, you have 4000 score. Now we think – people I think reporting the (unintelligible) of the reliability (unintelligible) entities. So that's a huge distinction. One score for the measure, one you get a score for each measure entity and then you report a (unintelligible).

Man: Still waiting for the email so (Jeff).

(Jeff): So if it turns out to be true that this is part of the distinction between the two methods is there any chance that we could sort of move tingle to noise and some more usability realm because it seems to sort of answer the question about by the measure. Can I use this measure as a consumer to choose A versus B or is it payer kind of uses measure to pay A versus B?

And if the answer to that is no because there's too much noise then the solution is we'll make the measure more reliable. So the team's like it'd be

sort of more helpful. And then it deals with situational reliability, sort of concerns that we're saying this is the scenario that we're postulating and in this scenario the measure is usable because it has the right empirical property.

Man: Interesting suggestion. Still waiting so (Paul)?

(Paul): In the course of discussion it occurs to me what the difference – and it actually came up. If you do signal to noise ratio, if all of the providers have extremely little variance then, you know, if it's A over A + B and A and A are very, very close then the component of B is going to be very small. And so therefore the – the reliability will start to go down because the part that B plays in that equation will start to increase as A – as the difference in A shrinks.

So in that situation you might have a very good rating by ICC but a very poor rating by signal to noise ratio because you are – as your signal is becoming more and more uniform in reality it's becoming more and more uniform. The component that noise plays is going to be bigger and so that's a situation where ICC might be a much better measure under the reliability of the measure.

Man: Or not depending on what you're using it for, if you're using it to distinguish right and then (unintelligible) ICC.

Man: Yes. Maybe it's to follow up on – on this line. I'm not sure where we're going to end. I can call one of the measures that came through. This was one about a patient rating of coordination of care or shared decision making. And as part of their analysis on what I think (unintelligible), they created artificially different scenarios. They said we're going to make scenarios that are lower, in the middle and they got, you know, got ratings of these scenarios using the scale and then they were able to show that yes, the ones that they

had selected themselves to be low – low ratings and they were on high, so on and so forth and again that – in that particular setting they put it forward as evidence of validity and it was okay.

They did some other things as well but in this discussion about reliability I'm just wondering are we going to see a situation sometime where somebody goes out and gathers us info of entities that are selected specifically to span a wide range of performance and they're done on that basis. And then they go ahead and they gather the data and they calculate the statistics and they show us a big number, a high and then they say yes but that's not the real world. The real world is not as different in its performance. That is spread to the distribution to say (unintelligible).

What are going to do? Do we accept their number and say it's okay? Do we reject it because it's not the sample that's represented in the real world? My other inclination is to do the latter but I – we did have this one example. It's not just like the data (unintelligible) where it happened. While we're waiting for this to come on (Sherri) and (Paul)?

(Sherri): I think there's – it's really hard for me not to talk about purposes measurement and the type of measure that we're talking about because for example in these composite patient reported measures PROs you also in the ICC term and I am a big advocate of ICC. Don't misunderstand me. It's just in that denominator is also the within patient across the items error terms.

They got two error terms and between score variation. In readmission so that – ICCs are going to look miserable when you're comparing position level performance there and they're going to look miserable for hospital. (Ron Hayes) had an article a long time ago that actually looked at chrome box

Alpha versus the ICC. At the patient level it's great. At the provider level it stinks.

So, you know, there are – we're going to end up okay, it's reliable at the patient level. It's not reliable at the doctor level. Is it going to be used at the provider or performance measurement capex? We have that problem. I'm on division compared technical advisory panel. There's trouble and you're using it to compensate physicians' performance yikes. So I think, you know, without talking about what type of measure are we talking about, that's why maybe we're stuck with a grid for this type of measure.

Here's the menu of options, you know. And these are the assumptions we're making. `Also you can't do this in a big group but maybe that's the strategy for kind of moving this forward on this issue. I would just hate to lock us in to something that narrows options and then the developers are dealing with a different type of measure that isn't responsive to that kind of treatment.

Man: I…I…you want to respond to that?

Man: When you're saying that the ICC would be good at the physician level but bad at the patient level, how do you do a…

Woman: No stop.

Man: I don't understand that at all.

Woman: The opposite. If you get a chrome doc to Alpha where across the items for the patient it's…

Man: You're looking at survey items.

Woman: It's good…yes for a composite level.

Man: Okay.

Woman: And then now you're looking at within a doctor a cost patient's of that doctor. That's another error term and then you've got the between doctor variation whereas if you're looking at readmission band is so small and variance is so small, the perturbation in that kind of measure. You need real confidence that those perturbations are not noise.

Man: So while we're getting this ready I don't know. It's almost there. I personally don't – you said we have to settle for a grid. I don't think a grid is a bad idea as long as have good text to wrap around it and – and provide, you know, a finite set of options with considerations as to, you know, to (Paul's) earlier point when one is more appropriate than another or when two are equally regarded, getting at the same thing but two different options. Anyone object to a menu?

It's not going to be a…the term these days but a vast diverse menu. It's going to be a short menu but it will be a menu. We're ready to go now (Jack)? Go ahead (Mike).

(Mike): I think in principal I like that idea a lot. I think though that it really opens up a whole other question because we tend to say that well, what we tentatively say, this measure is endorsed by NQF. Well stop, not for this purpose. We, you know, except to say it was – we test under such and such condition but not – we don't really say that it's endorsed for this purpose and not for that purpose and in a way we're opening that up for that purpose.

Man: Probably want to steer away from the term endorsed. Maybe it's recommended by this committee or suggested. We'll get the right language from this tab no doubt. Okay. So as I said one of the things that I deeply appreciated in Acumen's presentation and materials is they provided multiple measures for the look and reliability.

And this is one of the measures that I was not happy with the level of reliability on and what you've got here, these are physician level measures. So the 10 is at the group level and the 10 and TI is at the individual physician level and as (Jang Qu) pointed out reliability is calculated for each individual entity. And then what they – what we've seen presented on the signal to noise is the mean reliability, the mean across all those individual entities.

We see at the group level we have multiple doctors and a larger sample. The mean reliability here is 0.7 and at the individual physician level the mean reliability is 0.5 in terms of the signal to noise ratio. They also did split sample testing. That's on the right side here, did the Pierson correlation of – of – for both of them and while you see what I consider a reverably big difference in mean reliability on the signal to noise the two correlation coefficients of 0.48, 0.42 and I was not happy with either of those.

They also show a quintile a quintile mapping for the – for the two splits on the half, two splits. So you can hear that (TO) for what that 0.48, 0.42 correlation means and what it means is in – let's take the 10 example where you had a mean reliability 0.7 Pierson correlation of 0.48. In the first quartile only 40% of the folks who were in the first quintile and half the sample wound up in the first quintile in the second half. And among those at the highest quintile the fifth quintile is 10…you guys on the other end but it's also 41% being in the fifth quintile and both and everybody else being distributed.

So you've got four different ways of presenting the data here to assess how precise is this measure based upon signal to noise or based upon split sample. How consistent is the ranking? Not going to say stable. How consistent is the ranking or the location and the distribution, of course the two split samples and the correlation is basically recording that as a summary statistic.

So here is signal to noise and split sample testing. And if you accept 0.7 as the reliability standard you would accept 10 but not NTI. If you accept 0.7 or 0.8 on a split sample you would accept neither.

Man: Let me just point out over on the left hand side (unintelligible) there so (unintelligible) to apply.

Man: We got one with 0.4 and they said according to literature 0.4 is considered by CMS as reliable. No it's not.

Woman: So I guess this is a great example. I think (Mary) had said earlier that she thinks the split half is the more conservative number and that would be playing out here. But it also brings up kind of a different way of presenting information, that quintile thing which we don't usually see but it's quite persuasive no matter what the other numbers are.

Man: Again one of the things I hope will come out of this meeting or continue discussions with some sense of the standard, when I looked at the quintile thing, you know, 60% of the sample not being in the highest or lowest quintile bothered me a lot as far as the samples. I sort of settled around 20 to 25% as an acceptable shift of 70 to 75% of the sample staying in the first top or bottom quintile but that was mine and I would really like some guidance on what – is there a consensus on how much in precision we are willing to accept in an endorsed measure.

Man: So if you're looking at – I think this example is really very helpful. If you're looking at changes and classification then you could do cap analysis and there are – there is a scale, a landis scale that you could use for the (unintelligible) to see what's acceptable and what's not acceptable rather than say 20 or 25%. You could use the landis scale to accomplish that.

You know, I don't think at the end of the day we're going to come up with and I think (Sherri) made this point before. I think that you know, I don't think we're necessarily going to come up with a single threshold but I would suggest that we do as a group and whether we do it today or offline is we may have to do an offline, is we should say okay. Above a certain threshold yes we're really comfortable with that, okay? And then below a certain threshold this is okay, okay?

And then there's going to be a gray zone where really I don't know that we're ever going to be able to necessarily come up with a consensus. But I think if we could at least say below a certain number is really unacceptable I think that would be a very solid accomplishment.

Man: I do (unintelligible) thinking point 0.4 to 0.7 as the gray zone.

Man: Yes pretty much. I think below 0.4, I think most of those should be (unintelligible), 0.7, that's sort of what people publish in the literature and between 0.4 and 0.7, a little bit of a crap shoot.

Man: When the pressure is on the metric case. (Karen) do you feel like you got a good solid answer to your question?

(Karen): No. I think I'm hearing that – and tell me if I'm wrong. I think I'm hearing that these are acceptable ways of getting at a similar idea that different ways obviously will give you different numbers, pretty obvious, right? Maybe I'm shifting my answer to yes.

I think I'm also hearing that you guys aren't necessarily saying we would prefer to see one or the other or both. And then I guess – I guess is there any way to give guidance, maybe just general guidance if like you said you look at the…in this example. This has happened to me in an example where your average reliability group level is at the 0.7.

Most – you know, maybe that's the other things kind of tilted you in the other direction. What if they hadn't given us these other numbers? All they had given us was the 0.7. We probably maybe would have come to a different conclusion and I think that's what I'm not sure what to do with. Should we ask for all three of these kinds of things?

Man: That's a good – I was wondering the same thing. Maybe (Jack) could – so had you only seen the left side would you have said insufficient or would you have been more favorable?

(Jack): I certainly would have said insufficient on the NTI. On the 10 level we're right at the border there. One of the things that I think we don't know and the Landis scale, I'm not quite sure where those decisions came from. They're not too inconsistent with what we're seeing here but I think the advantage of having multiple and looking at one of the reasons why it calls for it to get some data, just wanted to see how stable the relationships are between these things and whether the calculation method and the specific method really produces – can produce very divergent answers.

If they all produce the same answer then it's just a matter of understanding what you're seeing when you look at one if signal to noise and either split sample or Monte Carlo simulations as samples from things produced – can produce very different answers then I think we've got an issue of do we accept the most conservative, do we accept the least conservative? But until we know whether they're all actually measuring the same thing pretty much on the same, you know, highly correlated way it's hard to make – it's hard to answer your question (Dave).

Man: Well maybe…to get back to (Karen's) question maybe it's a matter of saying all of these approaches are getting at more or less the same thing. If you show us one you're at a higher risk than if you show us two or three and we've…

Man: That's true.

Man: (Jack) just to clarify what particular test led to the numbers in the top left table?

(Jack): There's an application of the Adam signal to noise model. So basically I assume they ran a (unintelligible). I would have to go back and double check but I think they ran a hierarchal (unintelligible) what you're seeing in the distribution there is the signal to noise ratio for each of the calculated – for each of the entities in the calculation and the mean in the left hand margin. But it's basically the Adam hierarch signal to noise which is usually implemented as a hierarchal model.

Man: So it's not quite exactly like this illustration but it's just interesting to notice that when we do split half and when we run the Pierson correlation we get bolder numbers and so they fall under the 0.7 magic number. So we just need

to remember that a Pierson (unintelligible) doesn't automatically give you (unintelligible) compared to something else, if something else is not (unintelligible) ICC but it's (unintelligible).

Man: Yes. But if you go back to (Adam's) paper where he said 0.7, that 0.7 reliability was the single test. It was a two group high-low kind of comparison in this classification. At 0.7 we're about 25% of this sample.

Man: Exactly.

Man: (Sherri)? (Mike) and (Jack), (Larry)?

(Sherri): I'm back to (unintelligible) from the type of measure the options kind of grid menu strategy because I think that the type of measurements are used different because of its dichotomous and you have less variance to work with. Continuously you have more variance to work with and then the sources of variation, we haven't even talked about that and that's the design we can actually estimate components of variation and which belongs to the hospital, which belongs to the clinician, which belongs to the patient and so on especially in risk adjustment. That gets to be at least immediate strategy (unintelligible).

I'm struck by it's kind of hard to give a single answer (Karen). You know, it depends and it depends and I feel like my husband the (unintelligible). It depends on a whole bunch of different things so it depends on the type of measure you're dealing with and I'm still nervous about the consequences of misclassification.

I think those are different for different purposes measurement. I appreciate (Mike's) issue but I think the consequences they've got in this classification, you know, adjust what that threshold looks like.

Man: Two things that strike me is that we have to keep on paying attention to, is how much of the results depend on how much actual variation there is among providers in the Pep 10? And the other thing which we haven't spoken too much about but also there is how many, you know, providers do we have in each unit or how many patients have we served and so on? that's the difference in the (unintelligible), that you know, how reliable however we measure it depends on the data in which it's been tested.

And I think that – that may or may not relate to the circumstances where it would be used in the future. And being more clear about how to match those things up I think is something I think we need to pay attention to more over the line.

Man: So I think again this is a great example. I think one other strength is this example and this is something that our (unintelligible) had discussions about, is that when you're estimating the – when you're getting at reliability by looking at the signal to noise ratio as was brought up you're calculating that for individual providers and then if you look at the distribution, and then what this particular presentation does is it says let's look at the signal to noise ratio as a function of volume.

So 10th percentile, 25th, 50th, 75 and 90th percentile, it's very clearly showing you that say for ten this measure is extraordinarily reliable for the highest volume providers. So that's telling you hey, you definitely use it for those. And then for the lowest volume providers, if the 10th percentile is 0.5 and that's a little bit of (unintelligible) but then if you go down to the – at the

NPI level then at 10th percentile it's 0.39 and it very sort of seems like it's outside of the range where you would feel comfortable using it.

And this gets to the point of when – and maybe this is a little bit too – too much in the weeds but when you're going back and you're evaluating the reliability of the measure you can say, you know, based on the data that you're presenting us based on the signal to noise ratio and this is a strength to signal to noise ratio. We agree that this measure is reliable as long as your case volumes are greater than (unintelligible) because maybe in the case of the NPI greater than the 25th percentile.

So that gives you a little bit more information to go with. So there are definitely – there's some pluses and minuses to both approaches. If you're looking at the split half reliability testing you really can't determine using that particular testing (unintelligible) what the minimum case volume is going to be or when you're doing a signal noise ratio you can't so that's a strength.

Woman: (unintelligible).

Man: Okay, okay. Okay thanks. So if it was – but that's driving the variability, right and the variance between providers but I get – I get what you're saying so you can – so when you're specifying what the measure could be as reliable it's a different (unintelligible) same thing.

Woman: (unintelligible).

Woman: I was going to say use your mic if you can.

Man: So that becomes a bit of a problem then. I was thinking in the context of when you have – typically I would think that when they're doing this kind of

analysis they would stratify the providers by case quality because…because then you could sort of say look, for the 20th percentile if you're not 20th percentile case (unintelligible) it's just not going to be reliable measure and you should be using it. So you know, currently CMS says we're not going to report on you if you have fewer than 25 cases per reporting period per paper – for – for quality reporting, public reporting.

And this is a little bit more impure in terms of saying you know, if you have less than a certain number of cases the measure is not reliable. You shouldn't be using it for reporting purposes. You shouldn't be using it to pay for performance and that's the strength of the signal to noise ratio (unintelligible).

Man: (Andrew) and then (DQ). You were going to say something?

(Andrew): Yes, just briefly I was just wondering if we might be able to as sort of an intermediate step or something, you know, because the process, you know, recommended methods we could make some guidance on reporting results and ask for some level of granularity. We don't own. You gift this kind of thing. It's just the – what they put in the text there. Pierson correlation of 0.48 and 0.42 for, you know, signal to noise that's 0.726, you know, mean they don't give us always broken up by percentiles or quintiles and maybe that's something we can – I don't know if we have to, you know, say report it in quintile or report it in quartiles or percentiles or whatever but is there something we can recommend in terms of reporting results in different circumstances?

Man: I was afraid someone was going to ask that. (DQ) you want to…?

(DQ): (unintelligible) solution of the reliability so it's not by measure score or by volume but to get to (unintelligible) point it is the function of (unintelligible)

only not the (unintelligible). It's (unintelligible) what volume that you (unintelligible).

Woman: And to (Andrew's) point I think we had discussed earlier that at a minimum when we get the signal to noise it would be nice to see some kind of distribution since we know every provider gets to realize early that we'd like to see that and there were some equivocation obviously with complete consensus on whether we would want to see that filled out by campus guys or not.

I think the discussion of that point was that will tell you, you know, the story for that particular context but if you change context you might end up with a different minimum. So that was kind of the iffyness there but at least if we…you know, it would tell us something about that particular context. So I think that is something that we would probably unless somebody objects to it, probably not along with (Sherri)'s grid, calling it (Sherri's) grid.

Man: So we have 26 minutes before public comment. Should we rule (unintelligible) and next slide? Now we're on Slide 3.

Woman: I'm okay with leaving off the bootstrapping question for now. we can come back. I think that might be just another flavor of the (unintelligible).

Man: But I do think I didn't hear a lot of objection to including it in the grid, in (Sherri's) grid.

Woman: Yes and then say…okay. This question actually, data element liability cannot (unintelligible), this is…what – I want to leave this to a little bit later. This is a criteria question so apologies here. What I did want to talk – and I think we talked about a little bit in terms of where there's tight, you know, narrow

variation in things that sometimes we get very rare events which is another way to get tight variation.

So I wanted to get a flavor where we have rare events or very tight scores. And I don't want to get into thresholds and maybe this is a little bit beyond where we need to be but should we be expecting different reliability numbers? So…yes it's a mortality measure. You know that, you know, it's not a 0 to 100 score so your variation is – is fairly small. Would we expect a pretty high variation or wouldn't we? What do you guys think about that?

Man: If I could just say a little bit of detail I don't know if (Jack) and I may have been in the same session last week at Academy Health but there was a presentation about – I don't know if it was, it was some kind of touching rate mortality (unintelligible). And they show the descriptive statistics so if you want national averages some fraction of a percent and then the – what was just some smaller fraction with percent in them to max at some higher fractional new percent. And it's not as interesting when we show the ICC statistics of (unintelligible) liability at 0.95, 0.96. I think really how can that be? I mean, how can you – there would just soon to be by their definition and by their data.

So little between variability. Now I guess I accept that them, if at the same time it's within variability it's trivial not to and maybe you could get that and maybe that's what it is (unintelligible) over here. But that just – is there any special consideration that we should have for us developers to have about low rates or if you have high rate measures and maybe4 there are none. Maybe it just all runs the same and it's all fine.

Man: So (Mike) I s next but (Sherri) seems to want to say something on this point. (Mike) are you on this point or do you want to…

Woman: I already (unintelligible).

Man: (Mike)?

Woman: Two in the business, there's sort of reconsidered two types of options, a bunch of different options but those criterion reference testing which means you get a threshold impact and there's norm reference testing. And norm reference testing just does it by the distribution so if your distribution does it from 100% the 95%, somebody's going to fail. And so you know, the idea is that a very narrow band for variation.

If you introduce norm reference testing somebody is going to fail in that distribution. So I have a question for you (Karen). What are we talking about? Are we talking…again it goes back to how it's being used. Are you talking about norm reference testing (unintelligible)where you would just decide in a distribution of score the (unintelligible) or criteria reference testing. And if there's very little variation in measurement we consider that not – you don't measure things that don't vary. Don't measure percent of people get their blood pressure taken and (unintelligible). It doesn't vary.

(Karen): I don't think I can quite answer your question. I think it's getting back in my mind to one of the questions that we started out with which is when we are talking about reliability in the context we talked about today are we really interested in distinguishing differences or are we trying to quantify the risks of misclassification. So I don't know if that's answering your question or…it's not exactly but that's where I was going with this because when we have really narrow (unintelligible) it might be – we might have a low risk of misclassification but we might be able to tell the difference between providers and what are we supposed to do with that?

And then kind of the core (Larry) is if we have, you know, that tight variation would I ever expect see a 0.9? I don't know. I'm kind of lost by it.

(Sherri): So maternal mortality should happen zero times. If it happens once it's bad. So it depends on the measure you're looking at and what to make of it. You know, what's the interpretation in terms of quality so I, you know, and how the precision of an estimate of rare events become in a way (unintelligible) when the variation is so – it's so bad that if you miss – identify that individual's side, you know, you're really doing a disservice to the whole measurement that's sensible in that case.

But it kind of depends on what you make of the variation because variation should vary and there should be very little, you know, then in a way the reliability issue comes back to how well you measured maternal mortality. So if, you know, that's why I'm back to the grid thing because I can't do this without some understanding of what you're making. Which type of measure do you use? What are they being used for and how much of the variation should you expect to see in the measure?

Man: (Mike) and then (Gene).

(Mike): I think I agree that how you're going to use it matters and this is a great example why. Just no doubt that rare events have low precision reliability, whatever you want to call it. This is – it's a fact if the goal is to say using that as a measure of the quality if it's here provided. So I think we shouldn't give, you know, to discount that. That's just an important feature of the measure.

It doesn't mean that healthcare providers can't say hey we need to look into everyone's cases but that doesn't mean – but we shouldn't say this is a good

measure by normal CQF standards for being able to measure the quality of the care providers.

(Gene): In addition to the issue of rare events -- things that could never happen -- the – I found in reading through some of the measures that there was a desire to sort of restrict the number of eligible cases that led you to the calculation. And the way to deal them with the developer then dealt with the problem of so few events in the sample by extending the period of time with which come the data where collected and I think Acumen is actually one of the people in this sort of case.

My concern is that if you extend the period of time with which the data are collected, say, three years there's a lot in the world that changes in three years, quite notably the way clinicians deal with medical problems. And so, I find it like trying to drive down a hill on a very, you know, just given (unintelligible) okay, you want to, you know, (unintelligible) by looking in the rearview mirror.

It's not very effective and so I think that we need to think about the period of time where data should be selected but that then go into computing the measure to know whether or not, you know, what it's doing, what the data means because so much has changed over that period of time.

Man: (Jeff) and then (Larry).

(Jeff): This is sort of peak hyperbole to make the distinction between reliability and use so this is a perfect example. The measure was reliable and I conclude at 0.95 path on reliability. And the question is well, what can we use this for and probably not selecting providers but there's now a lot of variability. It's not going to make a huge difference in my likelihood as (unintelligible) events but

it might be useful for – for telling providers where they need to allocate some resources. I don't know if you have some of these events. Maybe you need to do some quality improvement.

So it just seems like if we have a metric like ICC that gives us good consistent data that's not dependent on how much provider variation there is and you know, then that's their liability specific. And then we do this other kind of analysis, you know, which really tells us more about what we can use this for and how effective it will be if you use it for various things. It could be like, signal to noise.

Man: (Marianne) then (Paul).

Man: So I'd like to push back a little bit. I think that we should have one set of standards for most measures. I think you can't even – we're having a hard time coming up with any kind of thresholds let alone for – for different types of measures. I think we're really (unintelligible) going to come up with nothing and some (unintelligible) things.

Second thing is that, you know, if an outcome which is ordinarily uncommon say for maternal mortality. We cannot create (unintelligible) because if you look at hospital mortality for hospital maternal mortality the vast majority of hospitals will have zero (unintelligible). I don't know if that's a really very good example. Going back to the Academy Health example you brought up, you know, even if it's a very uncommon outcome so it's 1 in 1000 and if you have some hospitals that are 1 in 1000 and ones that are 5 or 6 in 1000 and if they have a lot of patients that they're looking at this in their 30,000 at certain locations you could have a reliable measure.

But again it goes back to the point that I was trying to make is, is that I don't think we should come up with different thresholds for different resources. I respect what you're saying (Sherri) and I understand where you're coming from. And I'm not applying this to survey measures but I know nothing about surveys but in terms of binary outcome measures (unintelligible) section Yes, No complication measure, yes, no. Maybe we could have one set of criteria.

I think (unintelligible) you come up (unintelligible) that we all agree on, let alone what different sets we (unintelligible).

Man: Just to clarify my question that's a perfect answer because I was just wondering is it even mathematically possible to have an ICC with that, with rates, not only that (unintelligible) but with that (unintelligible) variation. We have to see if it's mathematically possible. Okay.

Man: I just – I was – I was having trouble imagining it could be so but I…

Man: (unintelligible).

Man: Something like that. It could be big numbers. I'd have to go back and look it through.

Man: Okay (Paul) and then (Matt).

(Paul): There are two issues that have been raised that are potentially very troublesome but at least (unintelligible). One is (Gene) brought up about – he brought up the stability issue because time is another factor. And I think if there is some reason to believe that things are changing so rapidly that the measure may not be "reliable" then that's a consideration in of itself. In other words there's no sense to a measure which is reliable based on past data when

we have a reason to believe that it will not be predictable future data because the whole point of having the measure was it's going to be used over some period of time.

Unless they have some sort of dynamic which can, you know, using artificial intelligence or something that's automatically refreshed the measure, you know, real time then it's something that's changing. So if there's a reason to believe that's something's changing so rapidly then I think that impacts our perspective on reliability. The other is just the possibility of talking about very low incidence (unintelligible).

It gets very challenging because the small hospital may have zero events and a zero sample size and yet you don't really believe that they're rated zero. This is, you know, what hierarchal regression tries to deal with, you know, the zero doesn't really equal zero but that…you know, that's a particular challenge. I think that if you have very low incidence events it affects – you have to very careful how you assess for liability in a measure because zero may not be zero.

Man: So in the (unintelligible) our subgroup actually reviewed a measure that I would consider to be a measure of a rare event to never event. It falls with major injury and a preponderance of the facilities have been rated here. This measure developer used a signal to noise analysis for the reliability analysis. That's the variability testing.

I guess what was going through my head was maybe more than setting thresholds. Should we be providing some guidance around what are the appropriate methods to use for testing rare events and how they're different than measures where events are more frequent and not strong enough in

statistics to know that there is a difference or there is a difference to that answer. That's just something to maybe process.

Man: You're still up. Okay (unintelligible).

Man: I can just respond to that. I think the point that (Paul) made was a very good one. You can use hierarchal modeling and get shrink using shrinkage estimators to come up with risk adjusted rates for those low volume providers. The problem is that the nature of the shrinkage is that basically what we're doing is you're taking hospitals that have very few cases and we're essentially making them out to look average.

So you're getting a number and it's relatively stable but it's probably not very accurate or more to the point. It's probably pretty meaningless. There are other approaches to doing this and this involves using something called shrinkage target. I don't want to take up the time right now to discuss that but what it does is 15 seconds if you know that you have a particular condition that has a strong volume outcome associated, so for example aortic valve replacements.

Well clearly we have empiric evidence that higher volume centers do way better than lower volume centers. If you use conventional hierarchal modeling we have low volume centers, you'll shrink their performance down to average and that's clearly a case of misclassification. If on the other hand you use this technique called shrinkage targets it's also a type of hierarchal modeling.

What you're doing is you're effectively shrinking the performance of these low volume hospitals towards the performance of other low volume hospitals.

So you're giving a more accurate…more accurate representation of the performance of those hospitals that have very, very (unintelligible).

Woman: Okay. So now I'm into my afternoon three where I only did two this afternoon but you know there's another – this is really (unintelligible) going to kill me but (Sharon), (Lisa), (Norman) and I in 2006 were (Dwight) (unintelligible) composite and for somebody like me we ended up with the rare events but it represents the construct that you're trying to represent like quality maternal mortality would represent quality of care for (unintelligible) but it's rare.

You put it with other things that you know also represent quality of care. And you know, in mortality we all cause mortality. Our readmission doesn't go with mortality as we know. Empirically it isn't correlated so maybe that's the wrong thing to put it with but you would create a composite out of a certain number of things that you know about that possible as we wait to see a decision.

So you take rare events and you add rare events that should, you know, conceptually go together to create a composite out of it. Then you also don't need as many samples. So it is a way around this rare event (unintelligible).

Man: So I will clarify that this measure that we did review actually was not risk adjusted. I think the expectation with it being solved was that there is evidence based practices that nursing homes and hospitals should have in place to prevent these events from happening. I guess I would just push back and say with your guidance (Larry) the different frame non-risk adjusted measure than a risk adjusted measure.

And I actually did look at your comments for this particular measure because we're running them separate but I think that was one of your criticisms of the measure, was that it was not risk adjustment.

Man: And to really respond I thought it should've been risk adjusted. I think that in any situation where there are best practices that can avoid that outcome you still have different list levels. So if you have one facility that has 30% of the patients that had a stroke they clearly (unintelligible) their risk to fall with another story that has a much healthier patient population despite the fact that you have (unintelligible) that are available.

I would still think…you can use hierarchal modeling and not put patient level (unintelligible) they list that so – and if you wouldn't do that and you have a lot of low volume facilities again you would shrink them back (unintelligible). That would still be…I don't remember that particular method very well. I don't know if that's what they did or not.

Man: So let's…thank you. Let's use the last five minutes or so to talk about data element liability. We imagine it might be quick. No priors. (Karen) you want to present this?

(Karen): So sorry. This is a…a bit of a continuation of the session that we had in of our monthly calls that the idea for ECB instrument based measures I think correct me if I'm wrong (Dave). When there's multiple items going into one performance measure we typically see it's from (unintelligible), nothing like being the (unintelligible) we often see but then, you know, what happens and what should we do about when the performance measure is based on a single item.

For example would you, you know, recommend this whatever to your client as an example. Often we can think about a test/retest option that may or may not be a possibility. So I think and then in the (unintelligible) I'm sorry. I didn't – I'm not quite sure. You may decide them in whether that's the right thing but maybe (Gene) you can help me out (unintelligible).

Man: Let's just…yes let's just talk about Alpha and test/retest. So when I come back to Alpha I would say that there are a number of example of getting back to the (Sherri) grid idea that yes, it's an easy thing. All the computer programs spit it out but, you know, somebody else could do a test of a, you know, (unintelligible) analysis, give a dimensional model and that would actually maybe even be superior.

So there should be other options given but I guess the general search for agreement if you will is that there should be with multi-item scales or multi-item data elements that there be some evidence of the internal consistency or even dimensionality of that. That's one.

And then the test/retest is, you know, called better for some percent of the kind of reliability that relates to reproducibility. But I guess the discussion point that where we left off as I recall is okay. With the single item that has come up with caps. For example as long as you don't get internal consistency as one item so what do you look at for reliability. Some of us said test/retest.

I think there was some pushback from the developer that said that's not appropriate. So let's see what the committee feels in the next three to four minutes. That is to say you have a situation where you're not going to get internal consistency because one question or one item, one piece of information in the data element should there be test/retest or reproducibility testing, (DQ) and then (Sherri).

(DQ): For the single item question there made (unintelligible) item that had been validated I'm not troubled by that but there are measure that may create a single item. And then we just assume the (unintelligible) validity and they go further and they call that – that's called (unintelligible) subscale so kind of circular like…so you want to run it and create this item everything is taken care of. That's one item that troubles me.

(Sherri): So I think the – the concern I have up here is that you're talking about different levels of analysis. So it's the patient level analysis that makes sense to cross multi-item scale to do chrome box Alpha for example. But now you're talking about using the average of the patients scores across the patient and physician. That's a different analysis. This is the data element, okay. Then the concern I have is the test/retest (unintelligible) because there's so many (unintelligible).

They use tests a person functional limitation (unintelligible) and they're great. They've been (unintelligible). Then they've got limit across the street and get hit by a bus so within two weeks you have two full variation. Now they're a wreck but it's actually because something changed their health date. The two things in the interval where the only thing that should've happened was variation (unintelligible) is very true. And for that reason I don't like test/retest of liability (unintelligible).

Man: I think you just created another (Larry) example of an extreme case. Yes.

Man: (Sherri) let's imagine that it's relative to (unintelligible) test/retest where the number is actually really high so yes it goes back to what happened and it's still (unintelligible).

(Sherri): It based yours on an interval where you probably shouldn't see anything else that has (unintelligible) to somebody that would – you change their score or improve their score, you know. They would (unintelligible) because they were in a terrible mood when they answered the question questions and now they're in a great mood and as so instability of the future of the standard scores versus measurement error, okay.

The true score variation is (unintelligible) and I – you know, again I would accept that but boy, you'd have to – the rationale for choosing test/retest reliability should have to include something about why we don't think there is actually going to be two score variation interval (unintelligible).

Man: How about it's 12:30? We have the public comment? How about we do (Gene), (Laci) and (Karen) but then we go to public comment? Does that work for the protocol?

(Gene): Real quick we did but in post acute care world oftentimes the clinician comes in and assesses the patient that sort of here at the end of the 60 days or whatever. What we found was the discipline of the – of the assessor makes a huge difference in how the (unintelligible) patient is assessed even if it's within 24 hours of assessment. You get a bit of a problem of change of characteristic.

So if you want to improve your scores in post acute care you have – you have the PT come in first who assess everything very low and then at the end of care you – you have a nurse come in who assesses everything high and all your patients improve.

(Laci): Building on (Sherri's) point I think it goes back also to maybe it's not such an extreme example either. To (Paul's) comment earlier about teams of artificial

intelligence and things rapidly evolving and progressing is that issue of the (unintelligible) comes into play. So it becomes a question of well, do we do it the next day, the test/retest? Do we wait six months? We brought in (unintelligible). Now we spot not such an extreme example.

(Karen): The two scenarios that I've seen are people doing (unintelligible) of when you have multiple items such as kind of not doing anything at all when you have single items. So that's kind of one thing, is that acceptable or should we – is that something? The other example that we've seen is – and I'll just give you the measure.

It's an experience measure for family members that (unintelligible). So they didn't want to be test/retest because they felt like typically it wasn't kosher to, you know, ask them questions even, you know, even in a really short timeframe. So if you know, test/retest may be acceptable but what is a – they don't do anything or what if they don't feel like test/retest is appropriate? Is there any other option that we could offer people to do that would give us reliability for this (unintelligible)?

Man: I mean, (Sherri) mentioned earlier if it's multi-item you can do a split half analysis of the items within the scale. So that would be – if ethically I can understand how you would not want to re-ask the question and if somebody was grieving but you could look at this with what you have. Reliability was ten questions by (unintelligible).

Woman: (unintelligible).

Man: Well yes. Well you ask for an exception.

Woman: No you could do ICC (unintelligible). You know, if unit of analysis is not the patient, it's anything else but the patient. I guess that proves the data element vision.

((Crosstalk))

Man: Yes. So thank you. I think now things clear as can be. Let's go to public comment.

Woman: Hi, if you're on the phone and you would like to comment feel free to raise your hand or put a comment in the comment box. If you are not on the webinar and you'd like to make a comment you can. Just speak up now. We did not receive any comments, but if you just didn't have a chance to, feel free to chat us and we will make sure to get those stated before toward the next session. Thank you.

Man: So if I could just - I guess we're going to break the lunch a little bit early, but just say by way of - based on what I do with most of my time, I mean I heard a little – fair amount of worrying about test, retest reliability. But if .7 is the standard, it's, really with data elements that come from patient report at least it's really not hard to get .7 or above. And if you can't, there's probably a problem.

Even if some people get hit by a bus, I mean, but it's - I think to me it was more a matter of, you know, are we going to - are we going to ask developers to try and get this information? I personally am in favor of it, but I do think that we need to make a decision on that. But it's unusual for me, at least in the work we do, to see the coefficient below .7. So it's not like they're going to lose on it in most cases.

I think we need a lunch break. Let's start back though promptly at 1:15. Like try for - at 1: to remind yourself to get to your seats. How's that?

Man: Okay folks, afternoon session time. See if we can get everybody back. First of all, just a brief thank you for the excellent discussion and comments during the morning time. I know we left a few things hanging and that maybe it's a little understatement, but gave us a number of things that I think we can work with, say in the context of a draft paper. I know you can't see this, but I took many, many points that we can follow up on as opposed to unresolvable questions.

So I think we can - all that scribbling is actually informative. It was something important that somebody said or some other people said. So without any further ado, then I think we will tee up the validity session and then on we go.

Woman: Just real quick. Just the - more of a kind of logistics check. How many of you guys, show of hands, are going to have to take off a little bit early before let's say, should have that time my head, before three? At three, at three, do I see (Larry), (Zikew), and (Jean), you guys all have to leave by three? Okay. So we're just trying to get a sense of who we might lose and you know, where we - how we might need to adjust the agenda to make sure we get the high priority items and stuff. But I think we'll be in good shape if we kind of stick to the agenda for now and get through validity by three and then we'll continue. Okay. Thanks.

(Ron Goldman): This is (Ron Goldman) (unintelligible) when you sign out of the webex.

Woman: Oh, okay, thanks (Ron).

Woman: Okay, let's get started on the validity. I think, well I'm not going to change. Validity, this is just reminding us of the sub-criteria that we have under validity and we want to know whether the results from measures produce valid results about the quality of health care. Can we, are we making the correct interpretation about quality when we look at those results?

And then underneath we always think about testing, but then exclusions, risk adjustment, differences, comparability with the data all we consider kind of grid, we call it threats to validity.

There we go. So again, the correctness of measurement, the extent to which which (unintelligible) draw conclusions. And the one that we're most familiar with, does the measure assess (unintelligible)? Definitely think it does

Context matters. (David), is this when you wanted to add in?

(David): Well this is - I'm not sure we need to get hung up on this, because it may take us down one of the rabbit holes. This is - I just came across this interesting article that was way long and way technical. But it basically makes the point that – ends up in the last sentence if you just want to jump to there, essentially raises the question of whether you can ever declare a test to be valid as opposed to a test and a context and a use and what not.

So this is now in the context of validity, some issues that we talked about in terms of reliability. I don't know where it takes us. So this - back of your minds, so let's keep going.

Woman: Okay. So doing kind of a similar thing, we think about data element validity and score level of validity. One is about the correctness of the data elements as compared to an authoritative source. So at the patient level data, how

accurate are they? And then at the score level, we've traditionally talked about it as making the correct conclusions about quality based on the (unintelligible).

So that's how we have traditionally thought about validity and influence. So we might want to just open it up briefly to see if - the way we did before. I think we had similar two by two table before for reliability and I think we were coalescing around the ideas of maybe there's Xs in all of those boxes. I don't know if we can come to the same kind of conclusion here about validity, but maybe for a couple of minutes chat and it looks like Mike is ready to go. Mike?

(Mike): So one thing I like about the major score validity is it actually makes reference to the quality competence. And I think that that was missing from the first statement about validity needed, and we used the correctness of the measure. It wasn't that this measure really tracks the quality concept that it purports to do. And I think that could be clear.

But I think that when it says correctness of conclusions, that may go a little bit further than appropriate here. Again, I think that the issue here is whether or not this measure really relates to and tracks the equality concept better (unintelligible).

Man: Just a quick response. I'm glad you made the comment, someone thought on the prior slide there was no mention of quality per se, which suggests that a measure could be valid sort of somehow floating out in air. This (unintelligible) quality I guess in my own mind is (unintelligible) correctness of conclusions might (unintelligible) has a higher score and quality measure and results in higher quality.

Man: That we all have (unintelligible).

Man: And - we'll get into this later. But we have examples of measure squares going forward with the whole discussion about validity that to my mind doesn't touch on that issue. It says, you know, this correlates with that, but it doesn't argue that either one of them (unintelligible) quality. So I think it's going to end up a meaningful (unintelligible).

Woman: And definitely a great segue to what will be the meat of our discussion. Does anybody have any other reflections or (unintelligible) or definitions (unintelligible). (Sherry)'s going to use her first of three, right? Sure.

(Sherry): So there's different kinds of validity. So there's content validity. Does it reflect the construct (unintelligible) standard that we were looking at. So predictive validity and discriminant validity. (Unintelligible) strike me as relevant because this, you know, this kind of exercise.

What, is there any concern at NQF about whether - what kinds of validation we're going to ask the developers to do?

Woman: It's a great question. When we thought about data element validity, we are thinking about what is public health (unintelligible) validity. So we assume that there's a gold standard – there is truth somewhere that you can compare those data elements to? So from that perspective on the data element side, we're looking for criteria validity. If it would be kind of fair to kind of call it that.

We had not specified that it has to be at the score level, that we are looking for criteria versus construct versus discriminate (unintelligible). So to date we have not had a requirement along those lines. And I think that some of you

would really like to have some predictive validity or something along those lines. We have not - we don't go that far thus far in our requirements.

Woman: A follow-up question? This is, so what you get - you don't really look at criteria validity in some of these measures, you're more looking at content. And especially in different levels of stages of development of measures because some measures are like brand new and then you would evaluate them differently from well, like paths and stuff that you, you know, they've been around and (unintelligible) for a long time. So is there any cheering up of what you're going to ask of measures at different levels of development

(Sherry): In the past there hasn't been. I think everything of those - that nature is on the table if you are interested in thinking along those lines. I will tell you that sometimes people don't always get the white labels on what they're doing, So, you know, so different people will call things, for example, they'll call it predictive, but it's not really, or they're using predictive in a different way.

So, for the last few years here at NQF we've just been calling everything kind of construct, realizing that the construct may be a criterion and it might be going one way versus the other, but we're just kind of not worrying too much about the label, but that is part of what we want to discuss today. Because, you know, what really should, you know, the question is really what comparators, right? And I think that is something that we really need to work on.

Man: (Sherry) has a really important point in that, does - the basic question to answer is does the measure actually measure what it purports to? In other words, if it measures what it purported to, that would be valid measure of quality. But that's a separate question from does the measure actually measure that. And is it a valid measure of that metric?

And I think both questions really have to be answered affirmatively, and there's slightly different – or they're actually technically different questions.

Woman: Okay. So thinking about data element validity, again we've thought about - just thinking about data elements there are extracted in some way using measurement, and how do they compare to what we think is true, often in a medical record that you know, could be something else. And so, that's how we thought about data element testing. What was it affecting?

And we - the consideration – the accuracy of the right word, truth to the concept. I'm not sure how useful this conversation is. Do you feel like we want to go into words right now? I usually use accuracy. Is there any disagreement with the word accuracy when we're talking about data elements and comparing the gold standard?

Man: Let me just rephrase and ask if it picks up what you mean by accuracy. Let's say you've got a measure of say functional status, outcome. So that's the one we were talking about here. And it seems like to me one of the challenges is to show that that number moves with the patient and go up, down with some dimension of quality, measure of quality.

Now if somebody showed me that information, I'd be inclined to say yes, it's a valid measure because the data. So is that what you're capturing by accuracy? So what does accuracy mean?

(Sherry): Let me make sure, let me say - basically what we're saying is – let's say somebody abstracted a diagnosis of, we want to find the diabetic patients. Are you really pulling diabetic (unintelligible) or are you pulling diabetic men with heart disease? You know that's what we're getting to in this piece.

So are we (unintelligible).

Man: That actually sounds more like accuracy, but I'm trying to raise the question of do we need to know that if you're pulling diabetics you're actually pulling truly diabetics, or – if, and, but that's actually one kind of data that's really (unintelligible) it's the numerators that we're – in either case, are you asking if (unintelligible) high number of (unintelligible).

In this case, neither one would (unintelligible) just saying, did I get the denominator right, did I get the numerator right. And (unintelligible) measures.

But I can imagine some data elements for which you could at least show data to say that (unintelligible).

(Sherry): Can you give me an example? That would help me.

Man: Well, the example, a couple of examples, like functional status Maybe it's some kind of context, that is independent quality. And you need to show that it (unintelligible) quality. Or I'll take readmission. I know (unintelligible) quality, but the empirical evidence is they're only something like 5, 6% of the variation that can be attributed to measurable features of quality of care.

So yeah, you can get the admission collected (unintelligible) meaning you get the admissions that exist, but at (unintelligible). I'm just raising, what are we fundamentally asking for and then what (unintelligible). I may be pushing it too far and certainly I think we should (unintelligible) typically could we expect, and maybe that's wrong. It's just (unintelligible). Yeah, (unintelligible).

(Larry): I'm afraid I raised my hand before I heard what you were talking about. So I'm not sure that I'm going to be addressing the great point. But going back to data element validity, and the question about whether we should be using the term accuracy. So when I think about this, I think you're basically - let's take the example of coding for patient-level risk factors.

So, most, many, many of the measures we use are based on administrative data. So what we would like to do is look at the validity of the administrative data for picking up a clinical condition. So you have a patient sample, you know what the ICD codes are, you're using some kind of coding algorithm to map those ICD codes, different diagnostic conditions.

So for example, you know, whole bunch of ICD codes are mapped to congestive heart failure. And then you're going to have an abstractor who's gonna go and look at the medical record, the authoritative source, and pick up diagnoses in the medical record of whether the patient has congestive heart failure or not.

And so to me, data validity is really agreement between the - your data elements as specified. So you've got ICD codes mapping into a diagnosis for congestive heart failure, so the agreement between that and the actual medical record. And - so not so much accuracy but agreement.

We actually don't really know what the true gold standard is. We don't even know if the medical record is accurate or not. That's a whole ’nother discussion. But, so it’s not necessarily accurate, but just a remit with what we think is the best source. Does that make sense?

Man: Yeah, I think the discussion of the last few minutes highlights the multiple levels at which we're being asked to assess validity, including the question I

would have. So if you think about the root source of the data, in many cases it is for all intents and purposes, medical record, and that gets abstracted somewhere. So what we're seeing is an abstracted version of the data that's part of the denominator, the numerator. And one of the validity questions is have they got those codes right?

But the other question is the one that (Larry) raised, which is, did whoever was abstracting it and produced those codes get the - get it right from the original source of the data.

So I think one of the questions we have is, we expect validity testing and checking to go all the way back to the original sources of data. Do we want to - are we being asked to check the appropriateness of the decisions that have been made about abstracting the data from the sources, these diagnosis codes are in, these diagnosis codes are not in, that's the second level.

And then (David), you raised a third level, which is, is this really a quality measure? And that's to be blunt something I've been hunting. Somebody says, I've got an admission from the community measure, I'm not quite sure how it's going to be used. I'm not second guessing whether this is a quality measure that's a usability - from my perspective that’s a usability question. I'll kick that one to the steering committee.

But do they get the measure right? Are they actually measuring what they say, is to me the issue. And with some of these measures we've got this issue of how we check the validity, and we've been using concurrent validity and it was correlated with something. And the rules say can't do face validity after the first - but for some of these measures, what's being measured is straightforward enough in my way that the face validity is the measure.

Yes, you either got admitted from the community or you didn't, that's a face validity issue. I think that was the measure where they were correlating with something else and were getting something like .3 correlations because there was nothing else in the data set that actually tracked what they were doing for the whole concept of, you know - correlational validity was irrelevant to that measure, but the face validity sort of held up.

So I think one of the issues is whether we should be looking at this issue. We got to think about all the different levels. And I think (Larry) got that right. We often need to, I think - one of the things I'd like to see on our agenda is whether for some measures face validity can remain the basis for defining the validity of the measure. It's measuring what they said they measured, whether we think that's a useful thing to measure or not.

Man: I just - I didn't want to get into the measure level validity yet, but I did want to reply about the data element validity. So we didn't really talk about this, but there's two pieces, data reliability and data validity. So to my - in my way of thinking data reliability means that you have the first abstractor looks at the medical records and authoritative source and codes the ICD.

And then you have a second independent abstractor who codes the very same ICD codes. And then you look at the – that's data level reliability. In terms of data validity - and this is what I - the comment that I made, there what you have is you have some kind of mapping algorithm that takes the ICD codes and maps them to congestive heart failure.

And so then you - you already have the administrative data and then you have an abstractor who goes to the medical record and looks for congestive heart failure and sees whether those two agree or not. So they're two different types

of agreement. One is actually data reliability, data element reliability. The second point is data element validity. Those are two very different concepts.

(Jen): This is (Jen) from the phone. Can I jump in? I feel like Jenny from the block. We're touching on - or kind of in this conversation is administrative data, Medicare claims data. And I - for me this is another place where maybe we can have some grids or guidance because we see lots and lots of measures coming from claims data.

And I'll give an example from my space, which is episodes of care. Pneumonia is often coded NOS. So there's a very specific, a nonspecific coding, 80% of the time in fact pneumonia is NOS. And the rest of the time it's more specific like bacterial, fungal. So you have a lot more specificity in terms of figuring out what kind of pneumonia.

So when we build an episode of care, we're thinking a lot about how specific can we get in terms of what was really happening for that patient. Because we know that big NOS bucket has got fungal and bacterial and all the other detail - that just not coded for those other cases.

So we have focused on a lot of time thinking about the fact that Medicare claims are for payment, but we want to use them for clinical information. In fact, is that a valid way to use those data. But my question for this group is, a lot of measure developers are facing that same set of questions and it can be very expensive to, you know, abstract the medical record to figure out if the claims data match.

And one of my sort of issues in thinking about all of these topics is accessibility. Do we make the bar so high and so costly that very few folks

can be measure developers. So anyway, in this whole nexus, I just want to share all those thoughts at once, 'cause I'll go back on mute.

Man: Yeah. So actually my - that was really good. Thanks (Jen). It's kind of similar to what I was - where I was going. But I think first to your point, (David), that when we think about data element validity, it is - we're not worried about how they're using it, that we're worried about is it accurate way to measure, the way we want to measure? And I'll go back to (Larry)'s CHF example and it hits a little bit on the claims data to0. So there's probably - for CHF, there's probably now in I 10, there's probably what, 150 codes for it.

I made that up, but it's a lot. You know, I don't really care. If I want CHF patients, I don't necessarily care if it's left astolic, diastolic – and all MDs in the room, I'm really sorry. But, so what I care about is it some flavor of CHF that I need? And so am I getting that subset right?

And so - so I think in terms of data element validity, it's not about the purpose. I think that's the score. Are we measuring quality, right? Can I get you in a readmission bucket yes/no accurately by looking at the claims. That's data element validity.

The other thing I was going to say and (Jen) your stuff, your comment leads right into this. I think we need to decide if claims - if administrative data, namely claims data, is a valid data set to rely on for measurement. Because I don't think that, you know, the RTIs, the Acumens, the Yales of the world should all have to go out and abstract records and figure out if these are accurate, right?

The – there's plenty of compliance people keeping an eye on that, I know in our shop and I'm sure in every shop. So I think we - something very useful we

could do for the measurement developers is to come out with some position on the validity of that data score. Sorry, I combined like eight things. So I don't have to go three.

Woman: Yeah. I still get confused by data element. So, to me this - if you're asking is this the right data source for this piece of information that we think reflects quality? Then back to (Paul)'s point about, well what do you do with PFI? For example, physical function. Do you say, is this the right data source? Because we can - then we can associate it with gait and if we get (unintelligible) we get different answers. Who's right?

Or pain for example. That's a great one. If the patient reports pain on the questionnaire and then they report - somebody writes pain level in the chart and there's a disagreement, who's right? So where's the right source for which data, which kinds of data is back to the (unintelligible). Are there different right data sources? I mean ask patients, when were you in the hospital the last time? Give me the date. That's dumb.

Because patients do, they do all kinds of different things. So that's the wrong data source for that piece of information. I'm still back to, it just depends on what you - which piece of information that reflects measure that reflects quality (unintelligible).

Man: So I think ultimately what we want to know is does the data element reflect what actually is going with that patient? I mean, does the patient have bacterial pneumonia? When was that patient admitted to the hospital? And I think that's depending on the question that we're asking - medical records could be better or worse, but ultimately, what we really want to know is does the thing that we use that's going to go into a score reflect the condition of the patient and the care provided to that patient?

Man: (Jeff), then (Christie).

(Jeff): (Unintelligible) use the word data element, it's kind of confusing. We're not really talking about - so like for claims data for example, we could just say, okay, we believe the primary diagnosis code, you know, we think there's enough auditing processes in place that we believe that. We believe age, we believe data depth. So we're not going to concern ourselves with the validity of those data elements.

What we're really interested in is the validity of the measure components. So our measure is made up of constructs, you know, diabetes as a construct or age of the construct. And what we're interested in is the validity of the constructs that we build from data elements – and so maybe we shouldn't call it data element validity, sort of measure of construct. That's even confusing 'cause we talk about constructs, but quality (unintelligible), it's like the component of the measures.

Man: Yeah. Interesting. I'm just trying to think of the implications of that, that it's going to have to soak in.

(Christie): And sometimes we just use our handle in the measure themselves swhen we measure patients, right? So you might not trust that one diagnosis of diabetes is diabetes, but they require two diagnoses within 30 days - with a minimum of 30 days apart and then they might even require that (unintelligible) his medications or if it's an infection, you know, you're not going to trust the infection element. But if you require maybe some treatment or some follow up or something else.

So you can handle a lot of that. Or some of that, you know, in the way you construct a measure because you know those individual data elements aren't always reliable. Right? But yeah, I totally agree that, I mean that we've got classification systems that have been well developed and funded by government and the CPF system that people use and agree is accurate. Identify CPF patients and wave - there are many ways to do it, but we can't expect every single measure developer to do that on their own. I don't see why we would want to do that.

And the same with the claims data, yeah. They're constantly being reviewed and audited and et cetera.

Man: Can I come back? So, you know, ideally we'd have like a library, right, of data elements that's known for the (unintelligible) and it would only be if they draw a novel construct from those data elements and pass them to demonstrate the validity of those novel constructs. But if somebody else, you know, created the same construct and already demonstrated the validity of that, they could, you know, they could use that. Sort of redo it.

Woman: And since then again, just a careful bait and switch in the claims data space. When CMS is auditing claims, again the purpose is payment. So they're auditing the accuracy of the billing. They are not auditing the clinical accuracy. And so again, just to be very careful, and we've talked a lot about these episode measures from Accumen. They reference CMS's audit and compliance work to kind of get at this issue, but in fact that auditing is for a totally different purpose. So just wanted to kind of point that out.

Man: That's good. Thanks. Okay. I've got (John), and then (Matt) and then (Joe).

(John): Yeah, and especially in regard to claims data, on the evaluation forms we review it's not necessarily clear what data elements we're expecting them to report back on. (Jeff) hinted at, and to take it out a bit further, are we expecting them to tell us about the accuracy of those codes used for the numerator, the denominator, exclusions, risk adjustment? If somebody wants to say, well, whatever is the most important, they're all extremely important.

I generally thought the most important is, are they capturing those numerator events correctly? But when (Larry) started talking about the diagnosis of heart failure, et cetera, I mean if you are, you're missing 50%, 20, 30, 40% of the people who hit the denominator that's, you know, obviously pretty significant too.

One time I was working with a measure that will go unannounced and from one quarter to the next, they had accidentally dropped one code for - on their risk adjustment, and all of a sudden we had 33% of the hospitals that were worse than expected and nobody was better than expected.

So like to get clear on, you know, which data elements within the claim, because you know, I think as (Jeff) was hinting at, we can't say that, oh, the accuracy of claims, that is X. I mean, like (Jeff) is saying, we're probably pretty confident age is correct. And if they died is incorrect, is typically correct, but in working with - developing a readmissions measure, we, you know, one example is we saw the admission source and where the person was transferred to - extremely inaccurate, at least the time period I was looking at. So we need to get clear on, not just claims but what types of data elements we're talking about.

(Matt): Yeah, I was just going to agree with (Jeff). I think if there would be the ability to build some sort of library of - specifically maybe ICD-10 diagnosis and

procedure codes, I think that'd be incredibly helpful for measure developers. I'm not sure there's any great value in having developer after developer go back and validate those same codes. Our team submitted a stroke misdiagnosis measure in the last cycle and we fortunately were able to look to the literature where someone had already done the validation of the stroke diagnosis codes, and we were able to reference that as part of our application.

(Joe): So this is where I really struggle, because I come at this from a different perspective. A lot of times when I'm reading these - especially one that has to do with claims code, because you know, I'm in a more operational world where we're abstracting for quality, whether directly from the EHR or having somebody manually review it.

And you know, the caution I always give to measure developers is make sure you understand what that data is used for, why it's being entered and actually who's even entering it because we have a lot of back and forth with our coding department, which they have different rules, different specifications and way they code out principal diagnosis. And then everything that follows after it.

And I could give you examples of specific measures that, you know, I - when I read it said this isn't going to capture the right population. But when I'm reviewing these, having that knowledge, it's like I can say, well, you're capturing accurately what you're trying to, but that's not the right code or the right - it shouldn't be principal diagnosis. It should be any diagnosis.

And then, speaking to your comment, (David) about the, are you measuring quality? Now you're talking about even another thing. So, I was trying to find the first set of measures we went through where I was saying, while you are measuring a difference between these hospitals, but you can't say that you measure in quality because you're not giving me any material to say its

measurement shows better outcomes, you know, ED throughput's always been kind of a good example of that. Just the median time from door to admit or discharge and there's not a lot of evidence that that improves outcomes.

So it might actually have the unintended consequence of discharging patients before you've done a thorough examination because you know you're under that constraint and actually do harm to the patient. So, you know, maybe this committee can kind of give some guidance around that. When you're doing this evaluation, do you remove yourself knowing that knowledge had not factored into your evaluation? Just say, yeah, you're measuring this but it's not the right thing to be measuring.

Man: And certainly there – we're touching on this distinction that we've made, I think with reasonable success. Certain things are in our purview about validity, and certain things (unintelligible) purview. And I think we've indicated a couple of them. And again, in my initial comment, I wasn't advocating that we take another step or try to clear up some of that territory. It was just sort of putting a range of things out there and saying which ones do we want to specifically say we're not in the business of, but okay. (Christa)?

(Christa): So this goes back to the question of which data elements are – so we are sure are valid, and this brings me back to my social determinants of health, risk adjustment, you know, efforts that I've reviewed where that, you know, lots of times they are testing but the data they're using is not really very valid. It's not measuring social determinants for that person, because you know at the block level or the zip code level where, you know there are only 220,000 geographic American community survey block areas.

So it's going to be a disparate measure and it's not going to capture fully the FDH for that person, and then they don't risk adjust for social determinants

when, you know, the debate is pretty much - the evidence is pretty clear that they do have some impact on some of these outcomes, but we're not finding it and then they.

So you almost have to look at as a step before to make sure the risk adjustment is accurate, which (unintelligible).

Man: We've come to a pause and (unintelligible) I guess I'll go back. Given this whole range of discussion, I raised the question is accuracy (unintelligible) move on, (unintelligible). So I'm actually okay with a set of questions. (Unintelligible) one is them (unintelligible) provide some sort of updated inventory of sort of data elements with established validity. (Unintelligible), or along with that, (unintelligible) evidence relevant to a particular say (unintelligible) of data, to say that the data elements (unintelligible).

(Unintelligible) primary data analysis by the developer if there's reasonable published evidence that (unintelligible). So there a couple things I think have come out of this, we can follow up on.

Woman: And just to be clear. I'm pretty sure I'm working with now, but I'm pretty sure that we have written, if not, I'll make sure that it's been there, that we do have kind of reviews as (Matt) said. So somebody has already validated the particular data elements in Medicare claims, for example. Please use it, don't go out and do it yourself, you don't need to do that. So that is definitely kind of a short (unintelligible) that we can offer (unintelligible) I think it's already, that should already be documented as something that we (unintelligible).

Man: This one is in relation to that. Just quickly. So if we had a data element already validated in a previous and endorsed NQF measure, is that acceptable, do you think? I don't know how I feel about that (unintelligible).

Woman: I think it would be, I don't see why not. I don't know that we've seen that before. Usually people go to the published literature. Depending on what they are. Stroke diagnosis, I know there are several papers where people have, back in ICD 9 land have done…

Man: I was just thinking, like if you're submitting a measure and you've already said, yeah, it's okay to use, I would never say (unintelligible) source, it's okay to use (unintelligible) source (unintelligible) We've already, it might, you know it's a measure we've done, you've already, we've already vetted it. Would they have to go through those hoops again. It seems like (unintelligible) system, but maybe have them disclose the measurement?

Woman: What we would ask is that they would just copy and paste what they did before. Because, yeah. So we want everything to be in one packet. So we don't want to say, go look at and measure whatever and look at what we did. We don't want them to do that. Yeah. But they wouldn't have to do it again.

Man: Okay. I had two up that are gone. I assume you know what you're doing, you didn't just tip it over on accident.

Man: So if I could maybe just - I think maybe to sort of chime in on (Susan)'s point. I mean I think for our measure we were able to go to the published literature, but I think that's actually the exception more than the rule. And so to me it's just, could we create some way of, even if it's not in the published literature, it still facilitates that knowledge or learning, and I don't know if that's NQF's role or - and maybe we do wanted to sort of go back to the published source. It just seems like - I'm not sure most measure developers, if they've confirmed administrative codes are going to publish on that, I think is the limitation.

Man: (Unintelligible) there is not, there might be some valid measures, you know validity might be (unintelligible) and others where the measure is completely (unintelligible) validated. So in those cases I wouldn't want to be asking the developer to submit or sort of approve that measure that they are proposing is valid or (unintelligible) and then (unintelligible) whatever you were saying, you know, (unintelligible) validated that to (unintelligible).

So what is our role, for example, what is it that we'd be looking for in a situation like that, where the measure is – they are proposing a measure for which the validity has not been (unintelligible).

Man: Just quickly, the validity of the data elements?

Man: Yeah, yeah.

Man: (unintelligible). But I guess seemingly, that's essentially (unintelligible) of the discussion. That measures coming through for the first time, there is a body of knowledge someplace (unintelligible) but otherwise they have to show it. Simple answer, but is that a sufficient insufficient answer?

Man: But again, I think going back to (Denny)'s point, it would be (unintelligible) to do (unintelligible) again, that's a (unintelligible) they have to do it both, they are proposing (unintelligible) now, that particular (unintelligible) has not been validated earlier then (unintelligible) they have to be validated with, you know, going back to some other (unintelligible).

So I think we could do it but again, then (unintelligible) going back to what (Jenny) was saying earlier, then (unintelligible). Again, I don't know how you are going to handle it.

Man: Let me just push a little again, and then we do have to move along here for (unintelligible). If we think about the (unintelligible) data elements, it seems to me that for example if you're talking about the (unintelligible) surveys, either those are already validated by some psychometrics or the person who developed the measure (unintelligible) actually starts using them, (unintelligible) and you've got clinical data (unintelligible) claims, you know, that - those three things, does that cover most of the waterfront?

I guess the point would be that in all three of those domains, they're probably, you know, highlight is that there's some existing validation data to refer to, and then if there's not, I guess it's on the developer, he's got to show they know what they're using are valid. Like somehow they have to convince us. Either by working back to something that exists or bringing something new. But I guess (unintelligible) for a whole lot of things, there's a body of prior knowledge.

Woman: But just one reminder and we're actually stepping a little bit into - you guys have recommendations for changes to that criteria. For some kinds of measures, just a regular run of the mill measure, not at (unintelligible) measure, for example. Right now our criteria say you can tell us about data element validity or you can tell us about score level validity. We don't actually at this moment require both.

And one of the questions that we were going to ask you is, should we be requiring both, and there are resource implications to these type of thing, but again, most measure types that have come in the door we don't actually require data element validation currently.

Man: I would just briefly comment on that last comment that you made. And I think in a perfect world, the building blocks should be valid and that's what we'd

like to see. But the reality is that the resources are just not available to do validation of data elements. I mean, imagine taking measure developers and asking me them to actually look at both data element and (unintelligible).

Man: Music from God. That was really good. One thing we can weave into the upcoming discussion would be, you know, can one mathematically or logically draw the conclusion that if at the measure square level it's valid, does that imply or will guarantee that at the data element level. That would at least provide a basis for this either/or rule, but okay, let's just, let's let that evolve. (Jeff)?

(Jeff): Just a quick comment on using other sources or previous sources to do the data validity. You know, I think you also have to consider the context of how that data was used and what you're citing. So to me, if you're citing, you know, claims and it reflects that it's appropriate for this population, but I'm using it this way, which is different. It's not, you can't use that source as your point of validity.

Woman: I forget what you call homogenization, what do you guys call that? Harmonization. Sorry. To me, you put it in a blender and it comes out, you know? So, the thing about different sources - and we haven't considered – it came up I think once in some committee I was on, whether the two sources of pain, inflammation, for example after a total knee replacement, one was from the patient questionnaire and the other was from the medical record. Can you put those together? And what does that do to data element validity?

Man: What does put them together mean?

Woman: Can you harm, I don't know what harmonize means.

Man: Harmonize often means pick one or the other, does not necessarily mean put them together.

Woman: Oh, I thought it meant you can kind of create, you can equate them, is what I thought it meant.

Man: It can mean a lot of different things.

Woman: Oh, well in that case, I don't know what that means either, but is the question still relevant? Can you put two different data sources together (unintelligible)for the same quality measures? Is that ever allowed?

Woman: I think we have seen that. (Christie), you might know more. Sometimes we get like diagnoses – you mentioned this before, sometimes diagnoses out of the claims or medical record and/or Part D, we know a drug was dispensed. So that's a little bit I think maybe of what you're getting at. Not exactly the same. I don't know that I've seen a patient report and, you know, either the patient reports or somebody else has written down something.

I don't know that I've seen a measure like that, but I don't see why we couldn't have a measure like that if somebody wanted to develop it.

Woman: Yeah, not really seen alternative sources of data like that where you'd combine them because if they disagree, (unintelligible), how would you do it. But one way, I mean, CMS wouldn't require you to validate the pain scale for example, because they say they made a statement that pain is what I say it is. If I say I have a 10, it's a 10. It doesn't matter. I don't care what the nurse says, I don't care what anybody else, right?

And of course that's going to change on a daily or even during the day, but you know, there are those kinds of items too that, you know, it is what I say, period.

Man: I mean I think if you have two different sources of data for an element, it's a simple regression question for you to estimate one from the other. You can do that, but then that, the error in that estimation should end up in the denominator when it gets to the performance measure squad. Right? But then you're okay, then you're all right.

Woman: Yes. I guess what I was getting at is if they're - I mean there's error in every measure and so you really are not necessarily kind of combining sources of error. But if the nurse writes it down, that's the nurse's report about the patient. So then the medical record is the data source. But if you're inputting data directly from the patient, as some systems are now starting to do, you've got two different sources of the things (unintelligible).

Woman: Well and then, you know, the MDS assessment (unintelligible) nursing homes, they have to stay with the residents. They can't say - it's not the nurse doing the assessment anymore, it's the patient voice, right? So they've changed the orientation of a lot of those questions and there's a lot of them like that, you know, the depression scale, the pain scale, the behavioral scale, there's a bunch of those where it's all about what resident says, not what the nurse says anymore.

Man: (unintelligible) but there are different recommendations, some that (Eugene) mentioned, some are therapy, some are nurse, some are physician. So, and the measure even mention that they tend to grade patients differently. So it's the same outcome (unintelligible).

Woman: Okay. Okay. I'll put that one out. Let's talk a little bit about measure score. For many measures we, new measures particularly, we do allow face validity. It is a determination by experts that measures reflect care quality, differentiate good from poor quality. We're kind of, this is one place where we are fairly definite at least in what we're looking in terms of face validity. And we really want people to think about the score, not whether, you know, not how valid the development process was and whether we liked the things that go into the measure, but actually the actual score and whether it's reflecting quality or not.

And then we have empirical testing. And you see here the way that we've written it (unintelligible) relationship and the measure results of some of the concepts, trying to get to whether or not our conclusions are correct. Again, that's us kind of using - the idea of construct validation is the umbrella term, that knowing that the hypothesized relationship maybe to some criterion, maybe to some predictive outcome later on will not be (unintelligible).

This is just kind of common approaches. Again, construct validity. People typically do correlations. We can do other things to some measure, we can do (unintelligible) if you have some kind of external way of splitting up providers. We can see if a particular measure kind of discriminates in that way. Those are the typical ones that we see. Okay.

So part of what really came up, I think it's come up before, but I really noticed it in this last cycle is, you know, what is the actual, you know what comparator should be used, right? So if you're saying we're doing score level validation, we're correlating or doing some other analytic thing with another measure or another set of measures, what really is a reasonable thing to accept.

So, we've put up here comparing past measures to themselves. That's pretty typical really of any instrument-based measure where there are maybe multiple performance measures calculated out in the same instrument. So for example, hospital perhaps, I think there's actually 11 different performance measures calculated in text, and that's again pretty typical for other types of measures.

So they would do a concept validity and say, okay, if the hospital is doing well on the respect domain, that one they probably, you know, we think that they're probably doing well on the communication, right? That's the hypothesis and they just kind of compare between themselves. We've allowed that kind of thing to go through, right? Because you're comparing the scores probably in some kind of correlation way of doing it.

For cost measures - actually you may have to help me out in this one. Comparing other claims based measures with the same data elements.

Woman: Yeah. So I think (Jack) and (Jen) can certainly help expand on thus, but there was a concern particularly with the cost measures that came through one of the subgroups, which were, and I think (Susan) was a part of that group too - that the same methodology essentially for validating the measure at the score level was used where – it was a claims-based cost measure. And for construct validity they essentially compared another claims-based measure with the same data elements to show a kind of directionality.

So they used a claims-based cost measure that included readmissions or like the costs or the claims associated with readmissions and they compared it to another claims-based readmissions measure and both measures shared the same data elements. So I think there was definitely some disagreement among the subgroup on whether that second measure was actually independent for

kind of external measures of - for which to kind of compare that and demonstrate validity.

So I – (Jack) please jump in and rescue me and make sure I'm accurately describing that.

(Jack): So what I recall, which – this basis for making any judgements. But what I recall is there were two kinds of comparisons that we saw being made. One was a correlational comparison of a measure with another measure that some of the same components we would expect to be similar. So if there's readmissions for a specific condition that was being correlated with the overall cause of readmission measure as an illustration that they were both getting at the same thing.

The other kind of measure - the kind of comparison that I pull in this set of measures was for the cost measures. The key thing about the cost measures is they both have standardized pricing. So whatever your DRG is, that's the price. The standard payment from the DRG is what you're going to get paid. And then you've got other sources of the 30-day measures, cost measures, got other sources of additional costs.

You've got post-acute institutionalization, the skilled nursing home or home health or physician services, drugs not included. And what they were saying is - and readmissions to the hospital. And what they were saying is, we expect this measure to go up as additional post-acute stuff happens.

So we could show that if you had an admission to a skilled nursing facility or rehab facility, your costs were higher than for those who didn't. Oor if you had a readmission, your costs were higher than those who didn't. And that was the argument saying we're measuring costs, variations in cost associated with

services over the 30-day period. And this is our validation that the costs are being measured because the costs are higher when you have more services.

Man: One of the things I'd probably say is positive, but one of the things that made it convincing when I read these proposals, these applications, is when the developers articulate a kind of a theory of the, you know, for process measures (unintelligible) outcome measures, you know, like, you know, people admitted to the hospital, area measures, a little bit of diabetes emergency department.

Why is that a measure of quality? Whose quality is that? You have to articulate that and then you have to say, you know, what kind of data can you bring to bear to see whether or not that theory actually holds up?. So I think that the extent that the developers can articulate a theory, and then use that to guide their empirical testing, that's (unintelligible).

Man: Thanks, and I was thinking, someone's thinking (unintelligible). Given all the logic (unintelligible) they can actually show a valid measure of cost, or is it a measure of quality, and that's the logical approach from cost to quality. So I think you have to have a theory of that.

Also the cast, those of us in the group who (unintelligible) it's common to say, you know, here's a set of questions that are to do with communication that will produce (unintelligible) different things (unintelligible) that would produce a score. The things are correlated therefore they validate each other, basically, or there's a global question, would you recommend this hospital and so they would say (unintelligible).

My question then and now, does any of that formulation suggest that these are good measures of quality and we know patients are happy even though they

said they had a good experience, but it's any of them (unintelligible). The answer may be, well yes it does obviously because it can't possibly be anything else, which is often the rationale that's in it.

(John): That's obviously a part of the theory, is how do these things relate to quality.

Man: Can I comment on that (David) because the cost and resource use committee, which I sat on, would be the first to say we're not measuring quality. We measuring - we're trying to measure cost and resource use. And the issue is do you get - how does quality track with the cost? So there was a lot of interest in measures of value and a real feeling that we were not there in figuring out the - how to integrate the measures of quality and the measures of cost to get measures of value.

So we were - the committee was saying we're still at an early stage of seeing if we can even get the cost and resource use measures right. We'll deal with value down the road, and value as far as integrating with the quality measures, but it's not the same as quality measures.

Man: That's an important point. I thank you for the reminder. Not everything in our full scope is going to link to quality. I mean that's when you mention quality forum, but, and cost measures are a good example. So sorry, I probably set the bar a little too high, although I think it's in other measures, it's a challenge to say that the variation in score was indeed (unintelligible).

I think we had …

(Jack): So, the way I think of measure of validity, I'll capsulate it, I think that for a lot of clinical outcomes, mortality, complications, for those types of things, I don't think you need to make a really strong argument that that's the measure

of quality. I mean if you have fewer deaths, if you have fewer complications, fewer infections. I think that that in itself says something. Okay.

But in terms of the conceptual piece, I think for some other measures, I think then I think what you're saying becomes much more important. Some of the very simple clinical measures we're looking at, it's pretty self-evident.

I think of measure validity along three different lines. The first one is face validity and I know that it's not a critically important piece for the NQF, or the NQF is trying to kind of downsize that a little bit. I do think face validity is very important for the people down where actually - at the sharp end, who are the clinicians, the nurses, the physicians, everybody was taking care of these cases.

Because if you don't have face validity on a measure, the people who are working with the patients are going to say, come on, really? And the whole process starts to fall apart. You've got to be able to convince clinicians that these measures have face validity. So I think you need to keep that.

The second piece is construct. This is the empirical testing piece. I see the value in that. On the other hand, I think that there are some real limitations there because what you're essentially you're saying is we have this really - this new measure, and we're going to compare it against other measures that are quote, unquote credible measures. So what makes them credible?

Well, precedent, we've already endorsed them. They don't really know if they're any better. And so if we have four construct validity with those credible measures, it may not be because our new measure is bad. It may be because our old measures were bad.

So what I think is actually really pertinent and maybe most important is predictive validity. And this is, and I'm limiting myself here to clinical outcome measures, risk adjusted clinical outcome measures. So, suppose in a perfect world that you have this very, very good risk adjustment model. And for each patient you're able to plug in their clinical risk factors and then based on that predict what tape of mortality, what their predicted probability of death is and say that model is really, really good. So then you can take all the patients that a particular provider cared for them, whether that provider is a hospital or a physician or a network, whatever.

Let's say it's a hospital. We can take all of their patients and you can come up with a predicted probability of death and then average them together and you get their - and I'm really simplifying this, I know there's a lot of other ways of doing this - but you can come up with their expected mortality rate. And then you know what the actual outcomes were.

So now you have the observed mortality there. So you have the expected and you have the observed and again, assuming that your risk adjustment model is really good, okay, then you essentially have a gold standard, right? You can compare the observed to the expected and if the observed is significantly worse than the expected then you have low quality, and if the observed is significantly less than expected, you have high quality.

So, and the way I think about this is when you're looking at measure level validity, once you get past face validity, which I think is really important, I think that the next piece and what's critically important is looking at the predictive validity of the risk adjustment model.

And so then you start thinking about how do you look at that? And depending on what model – what kind of model it is, you can, you know, for many, many

of the outcomes that we look at, which are binary, where you're looking at logistic regression and we look at discrimination and we look at calibration. But, so again, to kind of recap, I think face validity is very important because if a model is – doesn't have face validity - if a measure doesn't have face validity, it's not, you know, your clinicians are not going to buy into it and hopefully your policymakers will not buy into it.

But once you get past that point, although I see some value in empirical testing in terms of a construct validity, I don't think that's that important and I think we should be downplaying that. I think the really important piece is the predictive validity of the risk adjustment model. Now I'm kind of confining my discussion here to the clinical outcome measures that - where there's a basis for risk adjustment.

Man: Just quick observation and we'll move on. I don't disagree with any of that except often when you're using the term predictability.

(Jack): I'm using it in a different sense.

Man: Understand, I'm just – so everybody hears the difference. Okay, (Sherry)? And then (Christa).

(Sherry): I don't want to get into a (unintelligible) differential, you know, the frequent dispasian argument here, but (unintelligible) is an MD. But so we did publish in Medical Care in 2009 an article where we actually looked at the position effect on hemoglobin A1C, for example, on LDL measures, on clinical measures. And it turns out that if you test it empirically, round about eight is where the hemoglobin A1C value becomes - looks like it becomes a patient effect, not a physician effect.

So I don't think any of that testing has been done empirically or very little of it has been done empirically on clinical measures to look at the differentiation of what portion of the variance belongs at the physician level and what portion is attributable to the patient level. And that brings up attribution, which I was really hoping to avoid, but because all the testing is done at the patient level, it goes on to predict retinopathy or cardiovascular illness.

And then you say, okay, that's a good measure of physician quality. Nuh-uh. Not unless you actually test the empirical part of the variation in that measure that belongs to the physician. And that would be fair to compare physicians.

So I'm thinking, you know, we didn't, we haven't asked for enough of that kind of stuff even for the clinical measures. And then when you get to the patient experience measures, once again we have the error that you have to consider, you know, what portion of the correlation should belong to the – it's tested at the patient level, not at the physician level.

And that's where the interclass correlation coefficients go to 0.05. Well is that a really good measure of the physician performance then? You know. So I think that that's really a question mark in my mind and it certainly brings up the issue of - face validity is critical to the clinical audience. Is it sufficient? It's necessary, but it isn't sufficient.

So I would just say I endorse the first part of your statement and not the second, which means for me hearing, because some of these measures are brand new and in pediatrics especially you get brand new measures out there that are needed, but they haven't gone through iterations of the kinds of testing that, for example, caps and other measures have gone through.

So in fairness to the provider community, are we, is it a good recommendation to NQF, start tearing up these measures and promoting measure development in areas where there's gaps where we don't know or, you know, and are we accepting the unit of analysis, validity, path thing? And by the way, this is just a - endogeneity of yeah, you use the overall rating to test how, you know, correlated it is with their ratings of courtesy and respect or how clean the hospital is, hello, and you've got an endogeneity problem, because another pet peeves of mine.

So a recommendation, I think it's still cheering and validity should follow where the measure is in the course of its development.

(Larry): Sure. I'm just going to push back ever so gently. So when you're looking at what portion of the outcomes is attributable to the provider versus what portion is attributable to patient, you can do that. In fact you can do that using hierarchical modeling and looking at the inter-class correlation coefficient. And we've done that for cardiac surgery, for cardiac surgeons. I've done it for cardiac anesthesiologists and it turns out that the cardiac surgeons, it's about 4% for cardiac anesthesiologists it's 1%.

Now the problem with that is that it used to have the same issue of shrinkage, the interclass correlation coefficient. Basically if you have a lot of low volume providers, that interclass correlation coefficient will go down. So it's really hard to empirically establish how much of it is patient versus how much of it is providers. I would venture to say that very few of us believe that we would have – that cardiac surgeons have nothing to do with our out, even though the model says - or the ICC says it's only about 4% of the variation.

You know that - I think surgeon quality or hospital quality is probably pretty important. So I don't know that we need to empirically show that the ICC and

the provider portion is that high because I don't think we can. The reason that we can't is because those volumes (unintelligible). I think if you had much, much higher volumes you would probably see a bigger effect.

But regardless, I think that - I don't think we can rest this on face validity or construct validity. I think face validity is important but it's not sufficient. I think construct validity has a lot of limitations and I think that the predictive validity of the risk adjustment model is absolutely critical because again, it allows you to have some objective benchmark. If you can predict with a reasonable level of certainty what the patient outcomes should be and then you know what their actual outcomes are, then you have a way of benchmarking performance and benchmarking quality.

(Sherry): Can I just push back on that for one quick second. I don't want to get into this cross-table debate, but I think if your system has a place in this, I really disagree that you should stop with (unintelligible). No, but I think you can do it, (Larry). And I think there are ways to do it that haven't - we haven't explored thoroughly enough yet and we haven’t even tried in a lot of areas.

I think the surgeon effect in your cardiac rehab example, the cardiac rehab unit might have as much to do with the outcome of the cardiac surgery. Then you know, then the physician clearly has a point of some contribution to make. But is that a really - a system issue, a clinician issue, a patient issue, especially for poor and underserved patients. So I do think that craft empiricism has a place in this and I don't think we're there to kind of drop the curtain on it.

Man: Well (Sherry) and (Larry) are talking to one another. I also think they’re talking past one another, and it's around the issue of attribution, and I heard two different things. One is, and I do think risk adjustment is a vehicle for

trying to attribute, because you say certain amount of variants in outcomes here, some of those attributable to who the patient is, and we're trying to capture that in a good risk adjustment model and we will assert that any variation above the risk adjustment model is - we're going to attribute to the providers.

That in essence is what's going on in a risk adjusted model. (Sherry) in talking about the diabetes thing suggested - and to some extent the cardio rehab thing - suggested we need to disaggregate this. That if you're looking at control of hemoglobin A1C, she said down to the level, you know, seven is our control target.

So if you're using a 01 variable that you get to control, you wind up with one answer, but what she said is their analysis said getting down to eight is heavily influenced by the physician and the physician. And the care team's interaction would go beyond the immediate physician. Care team's interaction between eight and seven, it's heavily influenced by the patient. And that's a different attribution because it's cutting the data differently than the risk adjustment model would.

And what I heard (Sherry) saying is, how should we be cutting the data to get the attribution right? And sometimes the attribution is driven by the data. So I remember when the readmission - the original readmission, unnecessary readmission measure came out, we were going to pick the diagnoses for which we should not see a readmission. And CMS looked at that and got enough feedback and the data set, you can't do that.

So they shifted to a risk-adjusted, observed-to-expected model for the readmission measure, which is, you know, and it was simply because they didn't think the data they had enabled them to do the kind of cutting that

(Sherry) was talking about doing for the diabetes control measure. But observed-to-expected is not an ideal way to measure readmission risk and attribute all of the, observed above expected or below expected to the institution.

And I think that's what we're talking about. It is an attribution issue. Risk adjustment is one way to get at attribution. And the question is, is that the only way or is it even the most appropriate way for some kinds of measures to understand what's being contributed by the providers?

(Christa): I forgot what I was going to say, but I knew all this literature kind of pronouncing 70% of health outcomes are due to the patient and behaviors and social risk factors along with 10% clinical (unintelligible). So what always struck me about evaluating sets of measures like the Medicare Advantage, five star measures right, there are 55 of them, almost none of them are correlated.

I don't think you're gonna get a lot out of this correlation issue, right? Even ones that you expect to be correlated, like HBA1C testing and HBA1C level. They're not correlated because everybody does HBA1C testing, or what about medication adherence to diabetes medications? Is that related to HPA1C level? No. No.

So is it a problem of risk adjustment and the lack thereof or not properly risk-adjusted? And the same is true with the readmission measures. I mean the all-cause, 30-day readmission measure is adjusted for some stuff. Age, you know, gender, and chronic conditions but more likely the back of the hospital. Same with these new potentially avoidable hospitalization measures, right? And they're supposed to be for, you know, ambulatory-sensitive conditions that you're really not supposed to be going back to the hospital for.

So there's a whole lot of disagreement about that, but those measures are not related to this, the all-cause readmission measures. A plan who does good or a provider who does good or hospital who does good on the all-cause readmission measures is stuck on the potentially avoidable readmissions measures. Right?

So, it doesn't work. And is it because those measures don't include some key stuff which are related to social risk factors that really affect the likelihood of going back to the hospital because they're not adjusted for that yet. They're not even adjusted for dual status yet. And in fact a recent round that came around, they said, yeah, (unintelligible) but we're not going to adjust to that either, low income.

So there's problems I guess in the measures that have already been approved and out there pretty clearly, I think, when we try to apply some of these criteria.

(Michael): Okay. So, the discussion took a little bit of a different course than I thought when I first put my card up, but what I was, the point I wanted to make goes back to a point that (Mike Stoto) made about trying to encourage, facilitate, help, guide developers proffer a theory behind the correlation. So I'm not at all uninterested or trying to dis (Larry) for his suggestion that, you know, risk adjusted, you know, observed versus expected, that kinds of thing may be the highest type of validity that we would aspire towards and things that are connected most closely to quality.

But stepping back and using the third bullet example that's on the slide. It's been, it's been my experience, especially in the behavioral health area that folks look at just simpler correlations with process measures. And let me just describe to you guys a scenario that I saw that actually was good and the only

thing I was frustrated about is that the developer didn't take it farther and I thought it might've been good - a demonstration of validity, a construct of validity and you guys can react to it and tell us if you think indeed that was the case.

So the measure was a screening measure. Okay. No - and just whether or not screening for substance use disorders. Okay. And there is no strong evidence one way or another that, you know, you need to risk adjust this. We should be screening everybody. You could argue, okay, so let's assume that risk adjustment isn't a big issue here.

What they did for validity is they said, okay, we're going to look at the same providers and see how often they screen for two other things. So they did two other validity checks. One was screening for depression and one was screening for infectious disease. Okay. Again, all just screening, not results, not more nuanced than that.

And then they wrote this up very briefly and then they said, hey, you know, we saw correlations and I'll even give you, I remember the numbers - bizarre that I do, but it was - .61 was a correlation between depression and substance use screening, and .31 quite a bit less close, significant effects, big numbers, was the infectious disease. And then they said, hey, we demonstrated validity, period. That was it.

I wasn't very pleased with that. I wanted at least a bit more interpretation. I think they actually had a chance and I think in the preliminary analysis of the comments back to them, I suggested this. They actually demonstrated a bit more nuance. I'm not surprised, I don't think, perhaps many of you are, that depression and substance use screening might go more strongly together. This,

by the way, that was a Pearson correlation, not an inter-classic correlation. We didn't expect perfect agreement.

And then infectious disease. That goes with (unintelligible) substance use disorder but perhaps not quite as tightly coupled to behavioral health providers. So they could have, with a couple of simple literature citations, talked about how they expected those things to cross-correlate, thus those things to validate. They had two experiments. They could have demonstrated that. I know it's purely construct validity, but had they done that in a persuasive way I think it would have passed our current criteria.

So question for you all. Do you agree? Is that good? Should we keep that or is it still too loose, too disconnected from quality? Which by the way, I should add, you know, something that comes up in our, another part of the application we don't ask you all to review, which is the evidence piece at the very beginning. You know, what is the connection between the process and the quality measures?

We asked the standing committee to worry about that. We asked you guys to worry about scientific acceptability, right? In this case we're talking validity. So again, I'll pose the question. The scenario I just described, have they told a nice story about those two correlations? Would you be persuaded past on validity? Is that good enough, or do they need to go the extra step, do something more involved connecting it to the more distal and true outcomes?

Man: Just a quick process thing. I'm also looking at the clock, got a break scheduled (unintelligible) at three o'clock and I'm turning in my labs. Are there specific issues on which you definitely want input/decision or consensus? Because although you've put a question to us (unintelligible), what – we've got too

many things going on with not enough time to do them. So, I don't want us to lose this, but….

(Karen): Well, I think (Michael)'s question is, you know, it's kind of the same question we have here, what should be an appropriate comparator? So we have allowed pretty much anybody to pull any measure and correlate it. And sometimes there's kind of a trivial one, but they technically do what we ask, right? So you kind of have to say yes.

But you know, should we get a little more strict and say, you know, we're not going to accept certain types of things, or should we leave it as is and just let you decide, you know, tell the story and tell us why you think it's a good comparator and go from there. I think is pretty much the question that I'm looking for. And I think that's the same thing that (Michael) mentioned.

So in other words, it might be really hard to say you have to go that extra step and do the outcome, right? That's not going to be possible for everybody. Maybe the best they can do is some kind of process measure that sort of makes sense or doesn't. So how do we deal with that? Because sometimes people, sometimes we really do see something that's very, very weak, but it checks our box. So how do we kind of go, how do we get better without going overboard? That's what I'm trying to get to.

(Christie), you're nodding your head so I hope you understand…

Man: First of all, we had, (Karen) what I didn't hear that you wanted to turn us away to some topic other than the (unintelligible). And I'm happy to let it run, it's fine. Okay.

(Karen): I think this is the main question. If you get to the face validity question, that's fine. If not, we'll come back and - it could be related. Yeah.

Man: Okay (unintelligible) you've been up for a long time.

Woman: I just had a quick response to that, and I don't know why you'd apply a less standard to that than you do to picking, you know, risk factors like social risk factors. You require both a conceptual justification and an empirical justification. I think you need to do the same. That's all I wanted.

Man: I just want to come out on the, I mean look at the predictability of risk adjustment model. And we need to keep in mind that this is in the context of developing quality measure. So the goal is not always maximize the performance of the model, and certain variable because we care quality of (unintelligible) we don't want to account for that. Even though you put that model into much better model performance.

That's obvious, right, when you try to include (unintelligible) better obtain the first (unintelligible) you don't want to do that same data because (unintelligible). Just keep in mind, it may not be the best prediction-wise, but we are doing this for evaluating quality of care.

Man: It feels to me like (Michael)'s example of screening for substance use disorder is a perfect example of whether, you know, when do we need the correlational validation of the measure, as opposed to this speaks for itself. If you think substance use disorder screening is - should be done and its presence or absence is a measure of quality for substance use disorder screening, there's no need for a correlation.

And if you don't have any exceptions to the list, which (Michael) said, everybody should be screened for it. It's a yes/no, every patient, that's the, then the question is, are you measuring that accurately? Why are we looking for correlation? In this case where face validity would seem to be suspicion for justifying the validity of the measure. Do you want to argue that this screening is part of a general measure of screening as a quality measure?

So it's standing in for a general sense, does this provider do adequate screening? That's a different issue. And there the question is does it correlate with other screening measures make sense? But if you think screening for substance use disorder is inherently a measure of quality of care and that's what we want to measure, on its face you're measuring it. Why look for a correlation?

So I think it's the context. Is this the quality you want to check or is it in some sense a stand-in or a surrogate or a marker for the quality you want to check. And if it's a marker then you've got to compare it to others. But if it's what you want to do, it's done.

Man: Thanks. I have my card up, I was going to make the exact same point the exact same way. So I'll just second it, thanks.

(Jen): Well, I was just going to say, I think we asked for that extra level, just to make sure – even, it's really a question of showing this really work in the wild, right? It may be valid to do it and you think you're doing it right that testing may show something that you thought you had to be told that you didn’t. So that is - it's just that extra level to go and yeah, you just want to validate that what you expect to see is, you know, what you're actually going to see. Not just take it for granted that screening is good so let's not look any further.

I'm not being very articulate here, but the idea is that things could go wrong in the implementation or in the - some definition or some calculation. My God, you could, you know, have a typo in your SAS code or something like that and it goes wrong. And that testing might actually point that out to you. So that's why we asked for that.

Man: (unintelligible) you have two different measures for screening for different things, they're sort of correlated. (unintelligible).

(Jen): It doesn't, the only thing I think personally, what I think it would do is if you chose a poor correlation, it probably shows that you've either had a bad hypothesis. You know, your narrative wasn't so good. Or something's gone wrong somewhere. It doesn't tell you which one is that wrong one, but they would make you go back and look. So it doesn't, it would only prove the negative. That's what I'm trying to say.

You're not proving, you know, you're trying to show that something isn't going wrong and it's only a really little kind of minor way of doing it. And what we'd love to see I think is people continually doing additional testing with different variables, et cetera, different datasets, whatever. We haven't enforced that. (Sherry) to your point there's resource requirements and try not to be too onerous. But you know, in the perfect world, people would continually (unintelligible) and continually make the case that, hey, nothing's going wrong, at least from my viewpoint.

Man: So when I put, when I put my card up, I was gonna say, I thought the same thing that (Jack) said. So thank you. I wouldn't, I was going to say exactly the same thing because you said it better than I would have said it, but it does. It does allow me to maybe turn this a little bit to the difference between our committee's role and (unintelligible) committee's role. Because I think what

you presented was a judgment. You're asking us to make a clinical judgment, is depression screening a sufficient proxy for this substance abuse. I would say clinically no.

So if I was in the standing committee I wouldn't accept it. And my question to this community is, and maybe this is where you're coming from (Karen), we are sometimes asked to allow proxies or estimates of the real quality measure. And I think sometimes we do say, well if the coefficient's there, that's okay. Even if it might not have, you know, clear conceptual - sufficient conceptual similarity for standing committees.

So I guess my question to all of us is how do we decide in a given review whether we should consider, you know, the face validity of the situation or the conceptual similarity?

Man: So let me make one important comment, okay, with regard to this. Depression, and substance use disorder are related strongly related antecedent. Or, can be causal, right? You've taken a lot of alcohol, you'll manifest depression symptoms. So standing committee is certainly going to be aware of that and is going to be persuaded by that. Similarly, but less strongly infectious disease is related to substance abuse. Let me give you the vector, the pathway. Okay?

People who use alcohol or have alcohol addiction, let's say, or substance use - have a higher risk for other addictions including IV drug use that puts them at higher risk for HIV and hepatitis C, et cetera. Okay, so you want to be doing screenings. So they're related. That was the point I was trying to make, that I was persuaded.

What I was disappointed about by the developer is that they didn't convey that to the audience that was reading it. You guys especially, right, not knowing

that they should cite and they can, there's boatloads of empirical data that supports those correlations I just described to you. Moreover, they support the idea that one correlation, depression to substance use disorder, is going to be stronger than infectious disease screening for substance use disorder. And you can sort of logically think about that too in terms of the comfort level of say the screeners and what they're comfortable - what kinds of questions that they're comfortable asking.

So that was my point, as opposed to suggesting just that this is a measure that was self-evident of quality measure. I really wasn't trying to make that point. I was trying to say that it was persuasive and mostly the presentation of the developer that didn't make it clear to somebody who then would be evaluating the quantitative connotation behind demonstrating that to you all. So.

Man: Well time management. I got four cards up. If you can be brief, let's say its those four and then break. Otherwise we're (unintelligible). (Jeff), (Mike), (Sherry), (Lacy). So (Lacy) gets last word.

(Jeff): Respectfully I think that the idea of thinking of validity in terms of how one responds to a measure, how one sort of interprets the measure, helpful if in the depression screening - that my response to that is, you know, to get rid of all my problematic cases, then even though conceptually it's a good thing to do, our response to that was exactly opposite of what I wanted. So that makes it an invalid measure and that's why you do the empirical testing.

(Mike): I think that (Michael)'s example was really a good example of what I had in mind with the theory of the measure. It was expected to be correlated with these two and more with this one than that one. I think that the challenge that we have is we look at this checklist one by one and it's hard to kind of articulate a theory that runs all the way through that.

And I think maybe that's a way we can - if we can find a way to get the developers to articulate that and sort of carry that through that narrative, I think that would help.

(Sherry): So I think back to (Michael)'s example and (Larry)'s example, when you validate something at the patient level. So patients who are screened are more likely to have certain kinds of issues, et cetera. For patients who have certain conditions go on to have – to die earlier than everybody. When it's evaluated at the patient level, that's a different issue. And then you go on to make it an attribute of the physician or an attribute of the screener.

So if you're looking at screening for behavioral issues, say in the behavioral zone but keep it at the physician level, not at the patient level. So then you don't get back to correlating things at the patient level and using it to validate the physician-level performance. I think that's a big, you know, keeping that unit of analysis in mind, helps you actually choose (unintelligible).

If physicians are good at this, what else should they be good at in terms of screening? Then the infectious disease thing maybe transitions over into yeah, it's great at the patient level, but it doesn't quite work with - I'm just making that up, because I don't know this area well.

(Lacy): Yeah. So, but sometimes I think that when we're parsing everything out, sometimes we start to miss the forest. We're down in the weeds and I wonder if it goes back to (Karen)'s comment about how do we also think about raising the bar. The other thing is if we've got, you know, and the screening measure might be a good example here. What are we saying about value and quality? If we're looking screening measures? Is there follow up? Did they do anything

after they did the screening, like are we at the point to ask that question of what's next.

So is there an opportunity if you've got first measures to come in right now we have the face validity but you don't have to do the empirical testing. But if these measures come back, so if the screening measure comes back in the four years from now, is that the time to say, right, but what are you predicting? What value are you adding, you know, above and beyond the other kinds of things that you've done in the past. Is that a way to start imaging things to a different direction.

Man: We are now at least by that clock just a bit less than three o'clock. We can come back, it'll be 3:15. (Karen), last call, is there anything we must, must get from folks or can we just catch them privately? Okay. If I wasn't clear, I try to be clear. We are now on break.

Man: All right, here we go. The final stretch.

Man: Okay, everybody back? Off we go.

Woman: All right. We wanted to talk - we did want to talk a little bit about recommendations or potential recommendations for changes to criteria, but before we do that, we wanted to talk just real briefly about validity. Going back to this idea of the methods panel being the gatekeeper for measures that fail on the methods panel side.

We are getting pushback on that. So we wanted to get a flavor from you guys. Do you think it would be a wise move for us to consider, either you guys not being a gatekeeper at all or maybe being…

Man: For what?

Woman: For reliability and validity or, you know, is there some merit in thinking about whether you guys would provide more advisory but not an up or down maybe on validity? Or maybe another flavor of that would be kind of like we ask you to tell us your concerns about factors in risk adjustment that don't fail it because of that. You know, are there other things like that that we should be doing within validity but still allow you to take something down if the stats are really kind of poor, the method's kind of poor?

And actually maybe you can rephrase my wording to make it clear.

Ashlie: I'm not sure I would rephrase. I think that's - you stated the problem. I think the other - maybe just to pile on a bit- is, I think it's a bit of a, maybe a bit of a process question, but also, thinking about - I think some of the propositions we've had on the table is that, you know, if maybe because we feel that reliability, for example, is very largely kind of based empirically unless there's kind of a clinical judgment and with reliability, right?

They do an ICC or the split half or whatever test that they choose, and then there's less kind of quibbling that a standing committee may have about whether or not something was clinically appropriate. And that for us lies a lot more squarely in with the methods panel. And I think from a staff perspective, we feel like that should really stay under the purview solely of the methods panel.

I think the question of validity, as you look at all the sub criteria and some of our discussion earlier, I think that gets a little bit muddier, right? Because there are some clinical aspects to evaluating you might say all of the sub-criteria to some degree. And I think the way we've structured the process right

now with there being a gatekeeper function as (Karen) described is that we have kind of implicitly or more explicitly, some might argue, weighted the front portion of our process to say if there are any methodological concerns with the liability or validity that that kind of outweighs or should kind of pre, you know, sort those measures by not passing them onto the committee because those methodological or statistical concerns kind of outweigh anything the clinical committee might have to say or add on to that evaluation of those criteria's.

So I think we're just testing our hypothesis to see whether or not that resonates with you guys and how we kind of implement the process. And so we're looking for your perspective on where you kind of think the SMP and the standing committee's concentric circle, like, where does the overlap lie and where should the stop gap be for consideration of the measure. So hopefully that was helpful.

Jack: A quick reaction to this. If you think about the way in which validity is being established in some of these measures, we see some which are based upon expert judgment so we want to exclude this case? Do we want to include this case? And for some of it it's data driven.

And I think this committee can do a very good job of evaluating whether the data driven decisions meet the standards that we have (unintelligible) driven decisions.

With respect to some of the clinical judgment elements, the expert opinion elements, I think my tendency is to put those down the road to the standing committee with a sort of flag that my approval is conditional on this.

But the other thing that I noticed is when I talked about their expert panels and reviewed by the technical expert panel and, you know, it did this, what we're not seeing is the degree of consensus in those groups.

So I don't know whether this was a contentious point and it squeaked through on a 59/49 vote or whether everybody said no, this is clearly the way we ought to do it.

So I'd like to see some more inclusion of the sense of where they're using opinion and an expert panel how much consensus there was about the specific decision, not just in general but the specific decisions with the same (unintelligible) we make. And I think that would help both us and the standing committee.

(David): All right. Thank you. (Paul) next and then Ron on the phone and then (John).

(Paul): Basically it's a lot of what Jack said. I think the only point - I think we have methodological concerns that we made clear, I think that should be veto power.

However if our concerns are not clearly methodologically appearing as (unintelligible), which we were discussing about space validity and whatnot, then I think we have to make clear what our concerns are because they can certainly be open to debate.

But I very much like the idea of having the (unintelligible) federally profit inside of the decision process for the methodological expert panel consensus.

(David): Ron. I'm not getting you, Ron.

Ron: Oh, okay. Can you hear me now?

(David): Good. Yes.

Ron: Yes. I think, and this is the question that's kind of been haunting us all morning is I've been in the midst of one of those stop gap decisions. And it was not popular and became escalated to a higher level.

And I don't think it benefits anybody to have that happen. But on the other hand we all can look back on measures that we've seen some through that have either been for a lot of things we're talking about today weak on their reliability testing or weak on their validity testing or sometimes even both.

And the purpose of the committee was to improve those. So I think as we talked this morning about the 40/70 rule and have had some stalwarts trying to head us in that direction, we have not been definitively as we could have been retrospectively nor have we issued kind of the standards we needed to in the past -- we're getting there -- to inform the committees of the kinds of things of what we really do - and any developers, I'm sorry, what we really do expect to see.

And I have to admit I'm a little more biased towards reliability than validity, but I'm having an increasing amount of distrust in face validity especially with some examples we saw during our last session and a couple that were alluded to today.

So I still believe in the mission of the committee. I think we continue to have to work on the guardrails of what ones we're absolutely going to adamantly be against moving forward.

And the ones where we think we do want to give some advice - on the last slide before we took a break there was minimum standards for validity. And we probably should make sure we communicate to the standing committees that, yes, it meets minimum standards but it is just minimum standards and we have higher expectations than that over time as we talked about.

So I'm kind of dancing with that issue of whether we should see a stop gap group of not. I guess the answer might be in some circumstances, absolutely yes and in other words we are truly advisory based on expert analysis of reliability and validity that they don't have necessarily in the standing committees. That's it.

(David): (John), your card is down. Are you in or out?

(John): Yes, well, I, you know, get my comment out first just because, (Karen), I thought you said and correct me if I - help me out if I missed some of what you said. I thought you introduced this topic by saying we're getting some pushback on our role or something like that.

Can you say a little bit more about that because that might help him. You know, what are we trying to fix here? What are the issues?

(Karen): Sure. In all fairness even back when we first did our redesign (unintelligible) kind of have to do with the methods for (improvement) measures. And if they say no, we're not going to push them on standing committees even if at that time they said, I don't like that. I really want multiple people to prove to me to evaluate measures and (recap them).

And that hasn't changed. We still have some very - there are those people who really - and let me just be blunt. I can see you as having too much power. And they would prefer the standing committees to be able to make the final call.

And, again, they make the final call on ones that you currently - they make the final call on the ones that you guys push forward to them. The ones you guys say no to, they don't think. That's how we set it up (unintelligible). And (Elisa) can say these things much more delicately than I can. I don't know if you want to add to that.

(Elisa): No, I think you captured it. I think, you know, part of this is a role that we've created as well. I think it's questions for us about the process and whether or not the standing committee should look at all of the measures that are put forward.

We were trying to solve for making all of the committees more consistent and recognizing they didn't have the expertise that the scientific methods panel has to look at reliability and validity. But I think it does create this black box that they don't see everything.

And so we are trying to, you know, determine whether or not it's evaluating everything or making sure this SMP's role and review is transparent. And so, you know, it may be sufficient for standing committees to just see the results of your review, whether or not it passed or not.

(John): So I'm going to speak in favor of continuing to have all that power. But I will say if at the end of the day we become advisory, we would either have to revisit the value of how deeply we go into the reviews.

I think maybe we just switch to having review and submit reports to the standing committee. But I'm not in favor of that. I mean, I think it's better to have this set up the way it is.

And I paid particular attention to one of your first slides, (Karen), where you showed that when you add up the consensus passed and consensus not reached, it gets to about 75%, so that's not that bad.

And, you know, I think in the back of our minds, at least, our goal was to get to that more like, 80% or 90% over time, as measured developers become more - you know, we're clearer about what we expect.

We become better at providing what we expect. And, you know, if 9 out 10 make it to the standing committee, that's a pretty good rate. It's not like we're really wielding all that power.

Having said that, I have no objection to sending our rejections to the standing committee and letting them overrule it if they choose to overrule it.

(Ashlie): So I think we certainly considered that. And my question, and I think the piece that we struggled with was obviously we created the methods panel because you guys are methodological experts.

And what you guys can help us with, what rationale - or given that the expertise on the standing committee is not necessarily of methodologists, what rationale potentially would they give to counter a recommendation that you give that's kind of methodologically and specifically be.

And I think that's where we're having, the trouble is, like, okay, if we do that then what are we saying? So, yes, I'm kind of struggling out loud.

But I think the question is, like, so if the standing committee could overturn it, what's the rationale for that if that's not their kind of expertise. Yes, I'm just...

(John): So, if you don't mind, at least the two or three committees that I have been on have a few methodologists on them. It could be those people that say, hey, I took a look at this and I think they're wrong as one example.

On the other hand, it could be that there's some overwhelming face validity that they're just going to - I mean, I don't think it would happen that much. But I think part of it might be that they're not happy that we're not seeing some of these things.

And I don't really believe that there would be a lot of overruling. But it could be that some of the methodologists on the standing committees could exercise.

(David): I've got three cards up - but let me just answer, (John), your background question. I don't know if this is on some slides that we've seen. It seems to be reviewing some of these results from stakeholder interviews and feedback.

And my recollection is that the most favorable comments about us came from the standing committee. That they like what we're doing. They're happy about what we're doing. And I think that speaks to this issue of developers that they seem, at least perhaps in relative terms, to enjoy the fact that we strip off the (unintelligible). They won't have to look at. Okay.

Now the group on the other hand that's negative is the developers who say you guys are just one of the staff and you're making our lives too hard and you're (agreeing) with the science group.

So I think if we get into this question of do we want to shift the process of what we're doing. It's based on attention to the positive comments from the standing committee that seem to like what's going on. Okay. You're nodding. So we have...

(Ashlie): That's a very accurate characterization. And we do have some slides later on about the process that we're going over. And the standing committee members who responded to our survey, again, I don't remember the end. It's in the slides.

But I think it was, like, 86% really like the role that the methods panel sees in kind of stopping measures that failed reliability and validity from going forward to them.

It's like, you know, why should we spend our time reviewing these if, in fact, those two criteria have not been met. And so I think, yes, you're right.

(David): (Paul) and then Christie and (Sherry).

(Paul): So just in the way of processing communication, as it is now, does the standing committee have access to our reviews? I think they should. And I think also the white papers that we're putting out also will probably help.

But I think part of it might just be a product of communication that might be facilitated. I certainly, you know, for (David)'s issues, but (unintelligible) and most people (unintelligible) very pleased to have that headache off the table and say, okay. Well, you know, this is methodologically sound. Now let me figure out whether or not we're going to do this and this is related to quality.

(Christie): Yes. I was going to say I'm a little disappointed we're even having this conversation. I mean, you guys are the endorsements people. And you make this decision that you need this committee for the reasons you need it.

It's sort of like I don't know anything about plant science and I just say all of you guys are climate scientists but I don't believe it so, you know. You know, who am I to say, right?

God bless them if they want to review all of that stuff. But you guys set up this committee for a very valid reason. And if it doesn't pass the scientific reliability, it doesn't pass.

I mean, unless you know better than the six people who evaluated that measure, you know, you might be a couple on there but not six. I mean, we don't always agree, right? And, you know, yes, I would just say no.

But, you know, what I was also going to say about even changing the process is I feel like after doing this for two years now, I need to see both the reliability and validity data.

Have you all gone back and changed some of your answers on the top part after you reviewed the bottom part? I mean, it does influence even, I think, the scientific acceptability to see all of this stuff.

I mean, earlier, somebody said, oh, they have all this background. But you guys don't read that. I do. I feel like I have to have that knowledge to even do this. So, yes, I would not change the process.

(Sherry): When this committee was first explained to me, I thought the goals of it were twofold. One to raise the bar and improve the standards. Now that has to be

done after a certain amount of time because it's not fair to hold people accountable for standards that weren't someplace in the measures.

And the second one was to review complicated measures. So I'm assuming that that the NQF staff goes through that process of reviewing things that should just go straight to the standing committee and then we get what you all have considered to be yikes.

And so to me this is, like, it's got two functions. And if the goal is to raise the bar and have these discussions about what constitutes an adequate reliability assessment under which kinds of circumstances, that's one purpose.

And then are we now evaluating based on older standards, existing measures, but then we can crab about well, I don't think they did that right. But, you know, that's what's happened to me several times.

Yes, I don't think they did that in certainly an optimal way, but maybe not even in a sort of standard way. And I'm a little confused about, okay, is this still our role here to up the bar and help you all create new standards but then you communicate back to the measures developer or are we now the screen through which you're going to filter the measures and is that a different function or are we on the same page?

(Karen): I don't know that initially I would have said you guys would raise the bar, at least not initially. I think you will help us decide how and when we might want to raise the bar. So we're kind of getting there.

I think definitely we had anticipated and received that we're getting more consistency. And in that way we're raising the bar for NQF so there's that.

Another kind of the corollary of that is knowing that so many standing committee members don't have that expertise, making it more of a comfortable experience for them. Mike thought that taking this stuff off the table for them.

So that was another kind of third point of view that I mentioned that was in there. Did I miss one of the things that you asked?

(Sherry): I think the issue of sort of raising the bar came up in the context of well, wait a minute, you know, is that really - and it came up today, thresholds. I mean that's different from your prior guidance.

And that algorithm that you put together has been evolving and that's been changing and some of it has been a function, I think, of what this committee has helped to do.

But then the question is, you know, are we primarily serving them as the screener function or this now still a dual process kind of committee.

(Karen): I would say both, yes, it's both. You're helping us screen so, you know, you guys are saying right now these don't make - you know, so our standing committee members don't have to worry about how we set it up.

But you are also helping us raise the bar, helping us give better guidance to the field, all those things. Including another way of raising the bar without - when you say raising the bar, I'm thinking making things harder. And that's not fair of me to think about it like that.

(Sherry): Just one more follow-up point that raising the bar does mean improving the way we evaluate quality. And if that's the goal is to improve, then giving

feedback on every measure theoretically should help raise the bar, you know, even for people who didn't - why didn't you get?

You know, on what criteria did you fall down and how could you improve the measure next time so when it comes back around, if it comes back around, you need to address the following points like feedback, think sheets, for example. I think that would help.

(Karen): And to be clear the ones that you guys have said no to, we do give that kind of feedback to the developers. To date we have not shared all of that kind of ins and outs of dirty laundry, if you will. We have not shared that with any committee.

We are figuring that we do need to do that because they want to know kind of the ins and outs. And maybe it's okay just to tell them that without allowing them to overturn what you guys have said. So that's what we're talking about.

(John): And just to kind of (Karen)'s last point, could it be more of a remand than an overrule? So provide transparency to the steering committee about what the results were and if they have concerns they just send it back. And that doesn't sort of challenge the competence of either body, you know, but it retains our role as the arbiter of fact.

(Andrew): Yes. Maybe along those lines could it even be sent back to a different subcommittee to be reviewed? Maybe even without the knowledge that it's a measure that's been sent back. So it's sort of that they're getting two independent review of the measure. Just one idea.

(Sherry): You mean like a different subgroup?

(Andrew): Yes, a different subgroup, exactly, yes, yes, right. So some sort of almost an appeal mechanism without us being maybe super formal. Also I guess I'm trying to clarify because I've heard maybe a little bit of inconsistency.

What I've heard from some of the results was that the standing committees liked having the methods panel. So is the concern is that they're not finding out the details why we didn't pass a measure on validity or reliability or is the issue from the measure developer side and their frustration? I've heard a little bit of mixed.

(Ashlie): I think it's both. We've heard both. And I think, you know, different stakeholder groups have expressed different concerns. And so we're trying to take all that in and then, you know, weigh and balance and figure out where the action really is.

And I think where we've landed so far is transparency really is a first step and let's see how that goes. And then if there's still concerns then we could maybe revisit it. But I think transparency, I think, is certainly a first step that we'll be working on.

Jack: Yes, I want to go back to Christie's comment about the process. This was identified as a key element in a proven process. And I think the experience with it was are we reinforcing in the sense that it's accomplishing that?

I've been on the standing committee. I'm not the only person in this room who has been on the standing committee. But the consideration of validity and reliability has been uneven in the standing committee.

Sometimes it is the single thing that the committee is focused on. And other times it's quickly glossed over and we're focusing on other aspects, including usability and so forth.

And the couple of methodologists that are on the standing committees are great. But what we've seen in the process here is you put together five people and there's not a consensus in half the time before those five people get a chance to discuss things.

And the five people, that 50% lack of consensus, has gone down to 10%. So I think the process here is working. And the other conclusion from looking at that experience is that assessing validity and reliability is hard. And it requires a fair number of eyeballs and heads engaged in the process more than any standing committee is likely to have.

So I think the experience justifies, reinforces that the goals and the rationale for it have, in fact, been realized in practice.

(Andrew): I just kind of wanted to play devil's advocate just because there hadn't been much of a pushback. And I don't actually (get branded) by this argument personally. But I've heard some suggestions - you know, as we've said this is hard. We don't always have agreement. And there are always errors in measurement.

Measures are always imperfect and there are always tradeoffs. And we do hear from, you know, some of our, I don't know, say, more aggressive consumer or purchaser representatives sometimes that even in perfect measures drive better behavior among providers.

And, you know, in those instances where there is, you know, maybe the reliability has been not great. And this town might vote it down. You know, there might be some on a committee that would say it's imperfect but it's worth it. We really need a measure in this area. We need to push behavior.

Again, I don't, you know, necessarily buy that. I like the gatekeeper role. But just again to play devil's advocate and put that other position out there for consideration.

(Paul): Just to bring an analogy from a slightly different situation. I guess a couple years now, I became a statistical editor for the Journal of Thoracic and Cardiovascular Surgery. It's one of the leading journals in cardiac surgery.

And that was a part of a push on the part of the journal to improve the statistical methodology for those that were being submitted. And the process has grown quite a bit.

But there was tremendous pushback from many ticked off surgeons and authors threatening that they were no longer submit papers and this said and the other thing. But the end result is that the numbers of papers submitted to the Journal has escalated. The impact factor has escalated.

And so if we are actually committed to quality, then we're committed to quality. And if the measure doesn't pass methodologic muster then applying it just in the hope that it will drive quality, it's more of a wish on very shaky ground. It's a crap shoot.

And I think if there are genuine methodologic concerns, I think that we should do our work and find out.

(David): There's just one quick response, Andrew, and then you can jump in too. I've also been in a number of different meetings over many, many years now. And we've that (purchaser) views have been expressed about inaccurate measures as being accurate and reliable. And then we push people in the right direction.

I guess my response to that would be if that's really your goal, it shakes the whole foundation of NQF, I think. It doesn't mean NQF wouldn't exist anymore but it would just be a fundamentally different operation.

But the mechanisms and the endorsement throughout the committee have been about the reliability, validity, accuracy, certainness and informative value of measures.

If you just throw all of that overboard and say, well, we just want measures that push the field forward in certain ways. We don't care if they're good measures or not.

I mean, without that, the board wouldn't have to do it. But then it would seem like they would just take down all kinds of things that currently exist that I personally am in favor of.

And I understand a good measurement. And I believe good measurement does drive the field. I believe the add measurement results in resentment and hostility and complaints. And so, you know, I absolutely do not share that feeling. I'm not saying anything. We're here. I've said it in other public forums. But I just observed that.

I think we exist and then there are other functions that exist to promote high quality, accurate, fair good measurement and it shouldn't be about pushing bad measurements just because its - without something in place.

(John): Your journal example would be the only journal example like that. I mean, many examples, like Journal of Clinical Oncology, for example, that have been way up in its quality because it required statistical review and treatment cooperative groups. And it required statistic, you know, centers, you know have improved. I mean, there are a lot of examples out there.

And so I do think that's worth maybe keeping in mind along with - and I'm going to return to this, the fact that, you know, it's not - we're not that hard. I mean, we're rigorous. But 75% at this stage compared to NIH W section, it's like everyone would be applying for grants at 75%.

So it's, I think, maybe the elephant in the room. If I'm hearing the elephant correctly, the recent elephant, is also the payer of our meetings, which is kind of ironic, right?

I mean, so I'm sure we referenced. You know, we'll have to see if CMS is going to pay for this meeting or how many meetings they're going to - so, you know, this seems to be a discussion that needs to happen with CMS frankly.

And maybe reassure them that if we could have maybe some data, like you showed us some data, like, it's actually getting better and not just in terms of the synergies of, you know, things that have gone through but what it's led to and the quality of the measures.

So if we could compare the measures coming this year to five years ago in some sort of way and show that, I think that might, you know, help them justify, you know, continuing.

(David): Mike then (Sherry).

Mike: Just a quick point about the goal function. I was really pleased to see today how people referred to those specific examples and also a general impression that they had to do in the review?

So think we are probably doing a better of job of thinking through the standards because we have this experience of having reputably difficult cases than we otherwise would have.

(Sherry): Yes. I think you can get methodological rigor mortis where you end up with, you know, absolutely paralyzed folks who think just because they don't want to fail they don't do anything.

But CMS upped the bar. You know, we didn't do that. They put these measures to different purpose. When you're starting to adjust somebody's compensation based on a measure of their performance, it raises the standards that you need to apply before you declare something good for that purpose.

And, you know, we didn't invent that. That came from outside. And I think that has changed the character and nature of these discussions more than anything else.

It isn't that, oh, yes, if you publish a bunch of measures, people will do better. Well, you better make them, you know, resonate with the clinicians and the other people who want the same value.

And I think teachers rejected performance evaluations and compensation awhile back because they couldn't get what part of the variance in the classroom belong to the teacher and what belonged to the student.

And I think as we learn from other industries, we may decide, that, yes, certain measures aren't appropriate for that purpose. I know I keep harping on this but it is important to kind of keep that in mind as these changes go on.

Christie: Yes, I second what (Sherry) said. I mean, these measures are now worth billions of dollars to the health plans and to the providers. And we do see a lot of these measures being used, say, at the provider level.

I mean, I tell the health plans, you can't use this at the provider level because there's only four people in the denominator and it doesn't mean anything. But they say, but it will help them improve - you know, it will help them move - okay, fine. But that's all internal and I don't really care, right? I mean, that's fine.

But I also want to go back to where the - and it may not have been 100% of where it originated, but on the standing committee on disparities, and, you know, one of the recommendations from that committee, you know, how the heck do you get people to start to implement, you know, risk adjustments and then it got around the whole conversations about the appropriateness of the risk adjusted models and the fact that that wasn't really evaluated at all in the past.

And so it came from the disparities committee and we still haven't touched that nut. Wait until we start rejecting things based on, right? So I think we still have a long ways to go to meet the original sort of origin of why this committee was even put together.

Yes, I think I heard consensus for sure on this one. So, yes...

((Crosstalk))

Ron: This is Ron. Well, you think we have a long way to go. We haven't even scratched the surface on patient reported outcomes, shared decision-making outcomes and a whole plethora of disparate data sources out there, which are really going to scare up discussions about reliability and validity again.

(John): Thank you, Ron.

(Karen): We're debating kind of your mental state right now and how many brain cells you may or may not have.

(David): Do you think it sounds questionable?

(Karen): I'm thinking about mine and I'm wondering because what we wanted to do was revisit a conversation that we started back in January where we gave you some ideas about potential recommendations that you may want to make in terms of our criteria.

And at the time you thought it was a little too early to make those recommendations. You wanted to way for the papers, at least, you know, (Larry)'s paper was a little further along. So we wanted to tee that up for you today. And this is our current guidance.

So do you guys feel like you want to delve into the potential recommendations for criteria or would you rather talk a little bit more about process changes that we might make, which might be a little less onerous. And maybe some informational updates that we can have.

And let me just show you to remind you and this is actually a great chart, whoever came up with this, just a reminder that depending on the type of

measure we have different requirements for reliability and there's kind of a similar chart for validity.

And one of the questions has to do with those either/ors, you know, and should we be requiring both levels of testing for both reliability and validity or one or the other, you know, some of those kinds of recommendations.

So it's a little bit of a thorny issue. We hit it a little bit earlier when we talked about data element validation and whether, you know, is that reasonable to ask for all measures or is that just too much to ask?

But, you know, we could also ask like right now, all other measures except instrument based. And composite measures, for example, we don't insist on score level testing, right? And a lot people don't like that, right? Let me stop there.

Do you want to delve into these questions now or do you want to table this and we'll try it on a phone call in the next month or two? What do you think? It's entirely up to you and how much energy you have.

(David): Let me just ask at a level beyond that. If this is going to be a deep complicated discussion with all sorts of differences of opinion, then I can see deferring.

It had seemed to me, but again, it just may own bias I was hearing, but there was actually a fairly strong consensus of opinion that this either/or was a problem. And people raised questions about why did that ever happen and how can you possibly evaluate measures without measure score level or reliability?

Now if that continues to be a consensus or if I'm not imaging that, presumably this is something we can deal with fairly quickly and where it's not past our mental capability to do it.

So, Christie, if I can just go back. Are you wishing to defer this because you think it's going to be a long two hour discussion?

Christie: I think that either or might not be as simple as you think. But maybe I'm making it more complicated. But then...

(David): So if it's not sort of as simple as, I think, sort of in terms of how we get to some endpoint then I'll happily defer it. Because it is late in the afternoon and we have a few other things.

So I don't have a problem with that. I'm just exploring just in case if it is something we could nail down we perhaps ought to do it. But, okay. (Sherry)?

(Sherry): I don't feel like this is a very thorny issue. First of all, I think you are missing a row there because there are instrument-based composites. So there are both instrument-based and composite measures. So I think you've got to have at least one more row in this.

But otherwise, I think we were all checking all the boxes before. And unless people have changed their minds a lot, I don't think - (David), what's the big issue?

(Andrew): So maybe just to clarify, would the conversation just be about all other measures and that either/or decision or do you want us to also revisit instrument-based composites in ECQMs and whether those checkmarks are in the right place?

(Karen): I think we would want to start with the either/ors and maybe come back to the ECQMs. ECQMs are going to be harder probably because you've got to have - I mean, usually to do score level testing you need a fair amount of data. And that's in the kind of limiting factor for ECQMs. So that might make us lean towards maybe not putting an X in that box.

Kind of on that same idea, you know, if you've got a claim space measure or a registry based measure or something like that, most people do score level reliability anyway, right? They do it.

But if you have something different than that then it becomes more difficult to do that score level testing. And, you know, back in the day, we had a lot more of those kind of measures than we do nowadays.

And it definitely made sense back in the day to have the either or. So, I mean, we can certainly talk about it if you guys want to.

(David): I think we're beginning to talk about it, but we should establish our whether we can talk about it. So I got (Lacy), Mike, (Sherry), but let's try to keep on at least this process question. Are we even going to get into this?

(Lacy): So, right. So I think my observation is that there are a lot of panel members missing. I do recall that I thought it seemed like there was consensus about the either/or that we would want to check the boxes.

So unless we thought we had that consensus, we could say tentatively that's the consensus and then bring it back when we have the panel members all together to raise if there's, no, we're not going to go with that since we lost a bunch of folks or just defer the whole thing.

(David): Just to rephrase, so I think I would agree with that to say that we could feel in a group right now there probably is consensus. But we could make that a phone meeting agenda item and sort of say can we confirm this? But then it's open for discussion if people don't feel that way. Mike?

Mike: I guess I don't feel this consensus because I'm not even sure what it is. And I guess I was troubled by what (Karen) said about how measures for reliability can't be required for ECQM because it's too hard to get. Well, if it's needed, it's needed.

So I guess I don't understand what's at stake here and I would vote for having a more extended discussion at a later date.

(David): And that's okay. That's just what I'm trying to establish right now. To the extent I was talking about consensus it was that in that lower row, there should be boxes on both left and right. But if that's not clear in consent, then obviously we have to carry it further.

Mike: It's not clear to me what the consensus is.

(David): The consensus I thought we had heard was yes, boxes both places not either/or. That's okay. (Sherry).

(Sherry): I have a recommendation. Why not check all the boxes and put a provisional recommendation and then flag unless otherwise justified.

Ron: I like that. I really do. I mean, that's exactly what - this is Ron. It's exactly what was going through my mind, too, is to set that expectation and then ask for in the submission of why something doesn't need to be done. And it's

going to be paralleled in the validity discussion we have, too, because we talked about claims earlier. But, yes, why not have all the boxes checked?

(David): (John), you had your card up.

(John): Yes. As, (Marybeth), noted, we're missing probably roughly half the people. And if we're talking about something that's sort of establishing the rules of the road for us it would be nice to get most people to agree on what kind of rules of the road we're developing.

So I would suggest to defer this to a call. And I think one time before we noted - maybe we should talk about meeting every other month and for longer periods of time because when these topics come up it seems like given A, B and C is already on the agenda then, whammo, you only have, like, 30 minutes to talk about something that could get pretty needy. And then we end up, you know, nothing making a decision. So I would defer it to a call, I would suggest.

(David): Yes. And a friendly note to that, I think looping a couple things together we can, indeed, push this out to a call. But what we might do is put something like (Sherry) said, so specifically out there as an option to be considered as opposed to leaving it totally open ended.

And these people can say do I like that or don't I like it and why? And then if there's an Option B, let's put that out there and say do I like that? Or do I like it? And I think we'll just do that more efficiently than if it's open ended.

(Karen): Perfect. Okay. Let's go and (Ashlie), do you - I think we would have time - we're trying to decide if we want to do the toolkit. I think let's table that toolkit unless, (David), maybe we do want to think about are there some additional

white papers that we want to put on our list for future development? Or do we want to - do you want to have that conversation now or do you want to table that one, too?

(David): Well, I think we should do it quickly in the sense that after today's discussion and in other times when we've been doing this work and thinking about it, if there's interest in a different new other paper then this would be a time to suggest it.

I think generally, you know, we've been open if anybody sees the opportunity and can claim the boundaries of it, and particularly wants to lead it, we could put it on a list. I don't think at least to this point we've been harsh or restrictive on saying no you can't do that or we won't do that.

But this would be our opportunity. So this is an opportunity and feels the energy and commitment to do it, let's get at them.

(Sherry): I think the attribution issue is a real mess. And I think that the provider community really is looking for help on, you know, what part of this belongs to - what can I do to influence this measure that you just are going to hold me accountable for?

So I think the issue of attribution and, you know, maybe we can go back and forth about is this an empirical problem. Is it, you know, a tiered up problem really on a measured sort of, what part of it belongs to where and are there some guidance issues, some guidance we can give or reassurance that we're not that we can give to the provider community about attribution would be helpful.

Jen: This is Jen. I don't know if you guys can hear me. There is an NQF attribution committee. And I absolutely agree. I think attribution has reliability and validity implications so playing those out would be valuable.

But I don't know how it dovetails with some of the other attribution work at NQF.

(Ashlie): So this is (Ashlie). Thanks, Jen, for chiming in. Actually I led that work along with Erin O'Rourke and (Sirnamine). And so we have done a couple of papers already on attribution.

They did not focus on reliability and validity evaluation. We were really just fundamentally just asking a question, like, how do you define attribution? I think, just how people, like, talk about it is very different. There's different definitions.

And so we are kind of really looking at established kind of common definitions for, like, what an attribution model includes? What kind of decisions are you making?

Is it at the measure level, the program level, the interaction of how a measure may have attribution baked in, interact with the program but then also have attribution, you know, build into the program, and, you know, kind of considerations for that.

So we've done that to some degree already. They're very I would say foundational papers kind of orienting people to the challenges and some of the issues.

But we did not delve into the reliability and validity issues so I think there's certainly space for that. But I'm happy to send those to you just as a little bit of a kind of an orientation of what we've done to see whether those might be helpful to kind of build upon in some way.

Woman: (Unintelligible).

(Ashlie): Oh, okay.

(David): Oh, I was starting to write that down.

(John): Just on the outcomes paper, I must have missed the meeting where I was named as the lead? Am I (unintelligible)?

((Crosstalk))

(John): All right. Maybe I just forgot. I mean, well, we did write that book which is kind of half the story. And then I think you ended up getting a report from RPI on the other half? Maybe not. Remember there were actually two - or maybe we could stitch them together and update them. Is there a committee of people?

(Karen): I don't think I have one.

(John): All right. So we need volunteers to join the writing committee. We can't be that big so not everyone.

Jen: This is patient reported outcomes, is that the topic?

(John): Yes. So we have Dave Nerenz, Dave Cella so far.

(Lacy): So have you already shared other (unintelligible) - could you share the other papers you were talking about putting together to see what the (unintelligible) was with those?

(John): Yes, we can do that. Is the RTI one out?

(Karen): Yes.

(John): Yes, I mean, I think they turned it in but I don't know that ...

(Karen): So did NQF actually have...

(John): Yes, they didn't actually publish it.

(Karen): I don't think they published it. But it's definitely on our Web site. We can definitely provide that, yes.

(John): Okay. All right. Well, we may detail a few of you on the side if you don't raise your hands. You can also raise your hands in the next, you know, week or so and we'll form a committee.

(Karen): And maybe some of the trick would be even scoping out what that does. Because, like you said, there's early then...

(Karen): Why don't we send out the two pieces and then, you know, sure, yes.

(Sherry): So the shared decision-making measures are going to present a real interesting problem for attribution and also for reliability and validity. So then what's the care look like when decision is shared?

And what are - you know, where is the attribution going there and, you know, how do we - you know, accountability, how do you, you know, change the nature of, you know, things like lawsuits, you know, accountability when things go south. Where does the responsibility lie? There are a lot of things that are going on with shared decision-making.

And full disclosure, I sit on the shared decision-making technical advisory panel for the ARC grants I have now. But a lot of these questions are coming up and there aren't any good answers.

So the kinds of, you know, scientific issues that are involved in shared decision-making that are unique to that might be an important - and once again, I am not volunteering to write it. I'm just putting out a topic that might be useful to consider.

(David): Your experience with this, we're talking about measures plural or are there really, say just one or two on this topic?

(Sherry): Well, that's an interesting question because I think the shared decision-making, there are various different applications of it that go on in surgery are different than those that go on in primary care, that are different than go on in prevention and screening.

I think the topic changes the character and the nature of the discussion. And, you know, I think that those are interesting in terms of - especially if the fields and the measures start to evolve, I don't think one size fits all is going - one size is going to end up fitting none, but I think it might actually be an evolving question.

(David): I'm just thinking there was one specific measure that came through actually in the last two cycles. It got tabled or rejected and came back again. But that's just one measure.

(Sherry): No. There are multiple different measures that are going on right now, shared decision making that are not - that are tailored to specific applications.

(David): Any other suggestions? And obviously at any time if somebody gets a brilliant idea in the middle of the night and wants to share it, we're up. We might even be reading emails overnight, you never know.

(Karen): Okay. I'm going to hand it over, I think, to (Ashlie). To just give you some - a little bit more - we've referred to several of these things throughout the day, but this is a little bit more formal look at what we've done and what we've learned so far.

And so let's go through that and we'll spend maybe 15 or 20 minutes just (unintelligible).

(Ashlie): Sure. So I may breeze through these a bit. I think we've actually already talked about this a fair amount already. But just to kind of point out that we've been trying to do these incremental improvement cycles and laid out a couple here.

Again, as (Karen) mentioned this morning, we started out with the idea that this group would be kind of independent reviewers, kind of sitting, you know, alone submitting your evaluations.

And we would collate those and we would over time transition to subgroup calls and hashing out the measure discussions for those measures that didn't reach consensus.

We have increased the degree to which developers can interact with the methods panel during their discussions. We've developed this discussion guide - initially we developed this doc to help us, like, facilitate the discussion protocols for those measures that didn't reach consensus for once we started subgroup calls.

We shared that with you guys to help you kind of get a sense of where the discussion should lie in terms of where the areas of disagreement were with the preliminary evaluations.

Obviously the consensus not reached measures no longer go to the co-chairs for adjudication. We, you know, kind of served that to the subgroup calls. Or some of the kind of incremental improvements we've made and, again, what we discussed earlier, we do feel like the changes that have been have been fairly successful in helping to not only reduce the burden of the committee but for the co-chairs as well and reduce the number of - since it's not (REACH) measures that go to the committee, ultimately we'd love to have kind of a very definitive statement from the SMP on where the votes or the evaluation of the reliability and validity are from the methods panel.

So, again, we're working on transparency and improving the efficiency of our processes in terms of re-work or having to re-evaluate measures as well. And so we did recently meet internally to review our processes and determine how we might work on some of these issues.

We've talked a bit already. This is just a bit, kind of an excerpt from our internal kind of chart for what we were focused on. The lack of transparency, how the intended submit period in general for that period of time when you do your evaluation internally, projects that are also working on those measures

that did not go through method's panel and how that process is kind of paralleling your evaluation, we are looking at that as well.

And then this idea that developers are really seeking for more opportunity to interact with you guys to clarify and kind of provide, I guess, rebuttal for concerns that come up during your evaluation. So, again, we want to increase transparency and implement any process improvements for the upcoming cycle.

So we did surveys. We did process mapping. We were looking to eliminate waste and address our problem statements. Again, we sent out surveys to six of our major stakeholders groups, including you guys, our (CPAC). Those who comment on measures or wish to comment on measures, standing committee members and measure developers.

So this is just a quick summary of the number of respondents that we have. We didn't do a response rate. I don't have the exact numbers of the total. But I think we have close to, like, 400 or so individual developers, maybe 110 or so developer organizations. Obviously we have 22 methods panel members. A number of you guys did your surveys. Thank you.

We probably have approximately 15 or so standing committees close to 300 standing committee members. We have about 22 (CSAC) members, 17 (CSAC) members. Commenters, I think our distribution list was around 1,000 or so and then our staff. That number is out of about, I would say, 30.

So relatively okay response rates, but we did get some feedback. So what we're sharing here is some of the survey results that were specifically asked related to the methods panel kind of processes and your activity.

So then this is again around the preliminary results around the perception of the methods panel value. We could see obviously that staff, we really, you know, between staff, the commenters and the standing committee members obviously we all are in the SMP camp and developers certainly are feeling the pressure for getting through the process. So that's a relatively significant difference there, not statistically significant but significant.

This is one of the stats that I referred to earlier where we asked the standing committee members about the gatekeeping role and whether or not they think we should change it. Whether, you know, measures should be filtered at the methods panel.

And so 86% we can see of those 36 folks that we should keep it as and that the standing committee should not be evaluating measures that don't pass reliability (unintelligible). So, again, support for our current process.

So these are just some of the comments that we pulled out, kind of the qualitative comments that we pulled out from the survey responses and from staff.

Some of the comments that we heard about this gatekeeper role was that (BAC) really believes that it is protecting the rigor of the criteria. That it is being applied more accurately.

That, again, (BAC) believes the standing committees don't really want to evaluate the measures that aren't what they call ready or, you know, haven't passed the reliability and validity criteria, haven't met that bar.

And, again, that there would be more work for standing committees if that filtering was not done at the method panel.

Again, very similar sentiments from standing committee members that, you know, if it's failing reliability and validity, why should we be reviewing it? Again, these issues with the second to the last bullet around the transparency that they want kind of more information about why measures are failing and, again, this idea that it is saving time and effort from the standing committee to not evaluate measures within that.

Man: (Unintelligible) standing committee, right? I mean, you cannot (unintelligible) this.

(Ashlie): Right. I think it's the - and, again, this is coming from staff. So I think probably some of staff's interaction with some of the standing committee members. Maybe the 8% on one of those previous slides, the folks who feel like that they should get...

Man: (Unintelligible).

(Ashlie): Right. Who should see all the measures and they want to know kind of what's going on. Again, I think increasing transparency...

Man: Transparency (unintelligible).

(Ashlie): Yes, exactly, exactly. I think that will go a long way.

(Karen): And not really to scale. But just at the (unintelligible).

(Ashlie): Yes. So the failed stuff is - everything else, the passing measures are public but the failed measures do not get publicized in that way.

These are, again, more comments from the developers and from those who comment on our measures around the value of the methods panel. Again, from the developers this is a reflection of those.

I think the 78% who don't know that this panel is of value. Again, it looks like another hurdle, an extra level of review, more requirements. I think, again, this comment on whether or not there's expertise review, enough expertise review. But base measures was in question.

Some commenters, I think, again, that the perception is that there may be a lot of effort to get through the methods panel, but the benefit maybe hasn't fully been realized or that it isn't some incremental benefit potentially for having the methods panel in place.

That this idea that, you know, measures that don't go through the SMP that are reviewed by staff internally that aren't complex and then go to the committee aren't exposed to the same level of rigor in terms of applying the criteria by staff potentially then that they would if they came to the methods panel. So that's certainly one perspective to consider.

I'll keep moving here. Again, standing committee members, some comments that the methods panel, you know, really is doing the deep dive that's really needed on the reliability and validity.

Let's see, yes, I think from a patient perspective there is a comment here that, you know, it is really helpful to have the methods panel's perspective to help them not to feel the responsibility of having to provide input on some elements that they may not be trained to provide.

Man: (Unintelligible).

(Ashlie) Yes. So the next few slides are a bit of the summary of where staff landed at the end of our internal improvement effort. And some of what we're proposing to implement actually with the fall cycle we wanted to run by you certainly.

These aren't final. They're all proposed at this point. So if there's any input on how we might, you know, change or, you know, shift some of these items, we're certainly open to that.

But some of the improvements that we identified so you will see the current process of what we currently do and then what we're recommending on the right.

So currently what we do when measures are submitted we do sort of an internal review to try to - we were doing this from the perspective of we were trying to help developers try to improve their submissions and really put their best foot forward.

We do a really quick kind of review of their measures to make sure they've, you know, been as responsive as possible to the question. Do you need to provide more description on one of your approaches?

You know, we also are doing checks to make sure there aren't any what we would call fatal flaws, like, if it's an administrative claims based measure at this point do they test using IP sim codes?

If it's an instrument based measure, did they do data element and measures for reliability testing? So we're doing all those checks to make sure they've met the minimum criteria, but then we would also do checks to make sure that they are putting forward the best submission possible.

Because of the limited time that we have in the intended submit period, that turnaround for us to do the check would probably have been nice to have, was a very quick turnaround not only for us but to give the developers an opportunity to react to that they only had two days, 48 hours.

And for them even though we thought we were helping them, they were, like, that's too much. It's too fast. We can't turn it around that quickly. It's not fair. And so it was a heavy lift for us and it's a heavy lift for them. And they were unhappy with it.

So we've actually decided to eliminate that step. There was also some issue with whether or not we were, you know, developers weren't really to address everything in the 48 hours from an efficiency standpoint of whether or not the value-added steps, we really just felt like it wasn't really a value add. We removed that step.

What we will continue to do is to make sure that measures that come in the door do meet the minimum criteria. So we'll be making sure that the appropriate type of testing was provided based on the type of measure, that they've responded to every question that they needed to respond to.

And we will filter out those measures that don't meet that kind of minimum criteria and only those measures that do will be forwarded out and distributed to the methods panel.

Any questions about that. Sure. (Sherry)?

(Sherry): I mean, one of the things that's hardest to do when you review anything, proposals or anything else, is figure out what's not there.

And so the idea that, you know, this is just going to be really radical but, some of the formatting and the structure of the application itself gets redundant and confusing.

You're trying to figure out where on earth this is. And you can't find it there. Did I just miss where it's supposed to be in the section it's supposed to be in or is it really not there?

And I think maybe some of that could be the way the application form is now, the submission form is now, isn't streamlining that. So look here for this. And then if it's missing, you know, you can figure out that it's not there.

I think it would make a review of these things in general a lot easier.

Christie: Yes. I'll second that. You see a lot of texts repeated. I'm going to read that again. I'm have to start reading it. I go I read this already. And it would save all of us time.

(Ashlie): Yes. I think we get a little - we tend to not want to change the measure developer's submission, right? So if they submit something, we either have them change it because we don't want them to say well that's not what we - you know, you removed this. We didn't want you to remove that. That's not what we wanted removed.

So we don't mess with their packet. Like once they submit their packet if there's something, like, egregious that needs to be changed obviously we would kind of take if offline if it's - but things like that we try not to mess with so much.

But we can maybe look at maybe the testing form or maybe the way it's organized about how we can be more instructive to developers about what information to put there.

But I think we certainly don't want to be kind of tinkering with their permission after this.

(Sherry): I was actually talking about taking the template.

(Ashlie): Oh, the template, okay, okay.

(Karen): It's a little ironic because a few years ago we did it split out much more into some data elements on the sample. We did have it split out a lot more. And people wanted it more simplified in the way that we had it now and one of the big problems then was people were putting things in the totally wrong place.

So, you know, they were putting their score level stuff way over here. So it had a different set of problems. So we can certainly revisit it. It makes me laugh because we did used to have it split up and we changed it. Maybe it's time to go back and try again.

(Ashlie): Yes. I'll keep moving here. So a couple of improvements on the structure and transparency. And, again, to my point in the very first session this morning on our updates, again, we're looking to expand the size of the methods panel. Again, this move was primarily to address kind of the burden and the workload of the methods panel.

To (Karen)'s point earlier, the number of measures that are being reviewed by each subgroup is much higher than we initially expected when the group was

first formed. We certainly recognize that you guys are volunteers and you have day jobs.

And so we're trying to figure out, you know, how to best distribute that workload across - is it, you know, more people? The time frame that we have to do the work we're kind of stuck in that time frame. So how do we find other ways to distribute the work? And so expanding the work is one of the solutions that we've come up.

But, again, certainly there are things that we give up with that. We don't have kind of the smaller, maybe more cohesive group. But, I mean, there are tradeoffs, right? So we're certainly open to more feedback on that. But that is one of the approaches we're considering.

Again, this idea that we would rather than convening over conference calls for the measure reviews that convene in person twice a year to review the measures and during that in-person in an effort to allow developers to interact more they would be allowed to come to the meeting, you know, ask questions, very much how we do with standing committee meetings.

For those of you who are members are standing committees, would we allow the developers to come to the table with the committee? During the evaluation of their measures they can respond to questions and so forth.

So trying to kind of recreate that same interaction with developers that we have for evaluation measures by the standing committee.

We'll also be opening up, and again, the last two bullets are around kind of being more transparent about the processes. Currently we only post the agenda

given the quick turnaround time. Often there isn't time to make sure that all of the meeting materials are posted.

Our discussion guide and so forth, we'll be working to make sure that that is more transparent going forward as well as allowing public commenting on the methods panel's conference calls or in-person meetings, however we move forward with that.

So, again, I mentioned earlier about right now developers can only respond to questions and concerns. On the conference calls that we have we do not allow them to submit additional documentation after their initial submission is completed and submitted into the process.

And, again, we really did this to kind of balance, again, we give you four weeks to review a set of measures. Then, you know, the developer may get on the call and then they want to submit all of this additional documentation and we're kind of like, whoa, you know, our volunteers spent time already reviewing this and now they're supposed to kind of turn on a dime and, you know, consider what you have just submitted, you know, trying to balance your time that you've spent.

Also encouraging developers to take advantage of technical assistance that we offer at the beginning of the process where some of these issues we might be able - we'll be able to catch some of these issues where things might be missing.

But also keeping in mind, I think, what Dave Cella mentioned this morning about, you know, sometimes there's this idea or a feeling that the developers have this information. And if they just had the opportunity to submit in writing or give you this table that you're looking for that, you know, the

question could be resolved relatively quickly and a measure could move forward.

And I think the dilemma that we have is we kind of have to do it for everyone or do it for no one because those decisions about who gets to do that and who doesn't get to do that can sometimes be construed as, you know, not being unfair to one party or another.

So we're trying to figure out if we can do a (unintelligible) one and that's really where we led it. And so our improvement would be that we've looked at our timeline and if we're able to have an in-person meeting, we figured out that there would be an opportunity of about a one week period for us to provide developers with a combined document with all of your preliminary evaluations.

They would have one week to look at that and provide any additional kind of written documentation or rebuttal to your concerns. And that would be appended to your meeting materials.

And you would have a week to then review that in preparation for an in-person meeting where all the final discussions on all of the measures and voting would occur.

So, again, trying to build an opportunity for them to provide additional documentation. You would have an opportunity to review that after your initial review and then everyone comes together at the end at an in-person meeting to vote.

So I did want to just pause there and see if there's any kind of initial reactions to having a process like that.

(David): A couple of quick thoughts and one you certainly hit it at. I'm trying to envision this room, like twice as big with twice as many people in it. For the in-person meeting, it's just really going to be cumbersome.

You know, once in a while I've been in meetings that big. I just was in a CMS meeting. And the room was enormous. You couldn't see people. People didn't get a chance to talk. So I really have some concerns about just the mechanics of that.

Now I should say I deeply appreciate your concern for our welfare and I can see how in terms of the workload in just reviewing the body of measures, particularly if that body is going to grow. You know, I appreciate the concern. I'm just not sure that's the right method.

Now there's a little bit of a pie in the sky hope that our workload required to review each measure may get a little better as we drive more consistency, I would say, about what comes in, how is it presented, what does it look like?

I know I find myself getting better over the cycles. It's just trying to develop the thing. I get familiar with the forms. I know how people deal with them. You know, I can scan six paragraphs of essentially not helpful stuff and go right to the key paragraphs.

So I don't know if I can count on that. But at least when we review (unintelligible). So I am a little concern about the expansion part.

And also just one thought about sort of this perceived unfairness. I understand the point and I can see how that would happen. But also I think an argument or the sense against that charge would be one, if all of the developers had the

same right to listen in on phone calls and if they all have the same right to participate, two things happen that could be very - one is that in some cases there is actually a statement made by one of us that said, gee, if we just had the correlation between X and Y this issue would be settled.

And in fact they've got that in the dataset, I bet they can do it in two hours. And that came up in one of the measures I was on. I felt really badly that they couldn't just take the two hours and get it done.

But that is specifically a measure. And if somebody else is not in that situation and therefore they're not invited to show us an X, Y correlation, it's not unfair. It's relevant to (unintelligible).

So I do think the situations in which we could perhaps open the door a little more widely to some kind of quick response without being accused much or not being accused accurately of being unfair. It's driven by circumstances. And, in fact, it may be driven our request or the developer's initiatives saying, okay, we can do that. And others saying, no, we can't do that. Okay. Fair enough. Do it if you can.

(Ashlie): No. I think that's really helpful. I think the other point of that in our current structure with the eight conference calls is there really wasn't time. So giving them time to do that, resubmit it, sending it to you guys to review and then getting everyone's vote again.

You know, I think given the time that we have, by the time we finish the conference calls, we're really looking at just a very short period of time for us to summarize all the information, to get it to the standing committees, to give to the developers.

So we really didn't have that extra time to have that back and forth or even for that process to occur for you to review, revote and to get everyone's votes and collate.

So it's a timing issue as well as a fairness issue. But I think perhaps with the new structure, even without the expansion, if we were to have an in-person meeting, that does buy us some time. It gives us about a week.

So I'm curious about your thoughts about the in-person meeting versus doing conference calls.

Christie: Can I ask a question about the in-person meeting?

(Ashlie): Sure.

Christie: So it's a lot of people but since we're doing the final votes, will it be subgroups? We'll meet in subgroups? I mean, everybody is not going to do the final vote, right?

(Ashlie): So I think we're still - we still have not figured that out actually. And also there was a thought about whether we would divide the groups, right? So there would be 40 but then 20 do an evaluation this cycle and 20 do an evaluation this cycle.

Or, like, how, you know, just so it's not - I think there's a lot of permutations about how we could use the 40 people. But I think all of that is still on the table.

(David): And something like that would certainly address the issue just within meeting mechanics. I mean, basically each thing would look just like this except there would be two with different names and faces around.

I think it's still a challenge just to have a common set of criteria, a common understanding of the terminology, form and culture. I can see that's on the table.

I think that's really one of the most important things we have to try to create. You know, a couple of these negative comments about us, you know, we don't always agree with each other or, we don't - well I think we really need to drive that through a meeting.

There will be circumstances where we don't agree with each other no matter what, but I do think we ought to try to get to a point where we think largely along the same lines or we speak in the same kind of voice or use the same words. And either/or we split the meetings with 40 people, it's just harder. It's got to be harder.

(Sherry), and let's see, (Marybeth), (Sherry) and (John). I'm not sure who was first.

(John): So, well, she answered one of my questions. But, I thought, correct me if I'm wrong, I thought earlier in the day when we talked about the function of getting together for in-person meeting, I thought somebody said or the slide said it was for the purpose to review measures where we did not reach consensus.

But then maybe I misheard that because within the last five minutes I thought I heard you say you say you envision the purpose when we get together in

person would be like the standing committee where you're generally reviewing the pot of measures that you have, which obviously then is not limited to the measures where we even reached consensus.

So what's the thought on what is the function of that in-person meeting?

(Ashlie): Right. We would still focus on, I think, my analogy to the standing committee was just in the way that they convene and allow developers to come to the stable to participate in the discussion.

But the agenda would still be focused on consensus has not been reached on.

(Marybeth): Correct me if I'm wrong, (Ashlie) and (Karen) because my CDC process is a little bit antiquated after you guys changed some of it. When do you evaluate for importance?

And is importance still the number one criteria, the first criteria? If it is, so when do you evaluate for that? Do we do the measures first and then they go to the steering committee and then they evaluate for it?

Might it be easier to cut down on the workload if the importance of the measure goes first? Looking at that first and then eliminating those that they don't think are important at that point? That's just an out of the box thought.

(Ashlie): That's actually interesting.

(Lacy): How much does it happen, I guess would be my question?

(Karen): We do have measures that are inevitable. So it could be that you guys spend a lot of time evaluating the liability and validity and three months down the

road it goes on evidence. So it's an interesting idea. We kind of assumed the - yes, it's interesting. We'll have to think about that.

(Sherry): I think we blurred two functions here. And we spent a lot of time on, you know, evolving methods, you know, discussing methodology and a lot of stuff that would be included in the process of how we think about before we even measure an evaluation.

And this is weird, but if you separated that function from the screening of measures, and you had a subgroup that was more interested in improving the, you know, the evaluation in pertinent measures and in methods and stuff like that, and that was their focus versus screening the measures, would it fit more like a study section?

If you're going to have 40 people, then I would argue that you have to retreat the - I know you don't want this. But the NIH kind of way of doing primary associating and then have the entire group vote and not discuss.

You know, so they could discuss a few little things in the issue but have the primary subgroup discuss the measure and then have everybody ask a few questions and then vote.

Because otherwise, I agree with (David). You get 40 people in the room, 40 people like us - you get three more like me and Jack, you know, you'll never get anywhere.

So I do think that that, you know, might be one of considerations you make is removing the methodologic improvements kind of issue from the actual discussion of the measure.

(Marybeth): I just had a quick sort of question/comment on this improvement. So in the initial review these would just be measures where there wasn't consensus? They would go back to the developers so that they could write a response to us. Is that one that would happen?

(Ashlie): Right. It's just the consensus not reached. Because essentially we would keep our current process for measures that passed based on your preliminary evaluations, which we kind of automatically move forward to the committee.

We generally don't bring those up for discussion on the call unless we think...

(Marybeth): So unanimous passing or.

(Ashlie): It's not unanimous but it's clear majority. So it has to be, like, you can't be, like, in the gray. It can't be, like, a split. Like, if there's six people on the subgroup it can't be a, you know, six moderate, six low. That would be discussed on the call.

(Marybeth): So what I worry about is if you have - so I know when we've done our call sometimes I'll go into thinking, no way. And then someone will bring up a good point or the opposite.

And so I wonder what's the possibility, I'm just throwing it out there, that we might create extra work for the developer that might go away after our discussion. Does that make sense at all?

(Ashlie): No, it's true. It's true. It was the only. That's a good point. It was the only time frame we could find for them to provide input.

(Marybeth): Yes. It is probably the only timeframe. If we do it, we'll have to keep an eye on it and make sure we're not making work for them because that won't change their opinion of us probably.

(Ashlie): Yes. It's a good point. It's a good point.

(Marybeth): (Unintelligible).

(Lacy): Yes. I have two points. One might be a clarification related to the process for the transparency. So if it's only - so I'm thinking about the ones who just fail outright.

They don't ever have an opportunity for the - you know, like if you fail, if we all agree they failed outright because they missed their correlation tables. They don't ever get a chance to give their correlation table.

But then the ones where somebody - half of us said we wanted the correlation table and half of us said, oh, we didn't know we needed it. Only those developers are going to get the chance to give us the correlation tables.

(Ashlie): Yes. That's a good point.

(Karen): It's an interesting problem.

(Lacy): I just wanted to clarify that, yes, where the buckets were falling for giving extra feedback. And then I just wanted to second (Marybeth)'s idea about, I feel like if there's any kind of primary criteria, if there's anyway that can happen in the process first. You know, if it's the do not pass go if there's an important - you know.

(Karen): It's an interesting idea. The only problem with doing it that way was we would still want the standing committees to have the option of doing additional content clinical discussions after you guys do your validity things.

So you kind of need to go first for sure on validity. You see what I'm saying? But we'll think about it and see if there's some, you know, yes. We'll think about it.

(Andrew): Hopefully suggested it probably less frequent that an outcome measure doesn't pass on evidence because it's sort of a lower bar and those are the measures we would clinically review.

(Karen): That's very true. Yes. Usually the measures - you're right, (Andrew). Usually the measures that you guys are seeing, they're typically not going to fail on evidence. It's more the process measures or the intermediate clinical outcomes. But you guys do see some of these you don't see a whole lot of them.

Ron: So it's Ron.

(Ashlie): Go ahead, Ron.

Ron: Well I'd like to give you something to consider too. And I'm mostly dealing with our 80% disapproval group, which were all on the previous slide.

I wondered if you would give some thought to developer Webex's, maybe three a year, maybe more. The papers, I hope, will take care of many of things I'm going to mention.

And I recognize that our developers basically fall into three categories. Those that think they know more than anybody on this panel anyway. Those that probably just need little tweaks as far as some of what we talked about today.

Here's the information we're specifically looking for. Please make sure you provide it to us. And then the ones that are at the very low end of the spectrum that need a lot of help in submitting measures.

But until our formal guidance comes out in the form of papers or this interaction with all of them, there can't be too many developers. I mean, just take the number of measures and you know who develops most of them.

So, I mean, we could probably organize two or three Webex's to answer questions, give direction, talk about the direction things are going. Talk about how some things have changed and what they used to be, a lot of the topics that we talked about today about reliability and validity.

And that may well help also help also drive some improvement in the process. So just give some consideration to Webex because we couldn't possibly do it in person.

(Ashlie): Thanks, Ron. So we actually did a little bit of that in our developer workshop yesterday. And we do offer - most months we offer a measure developer webinar.

So we're doing some of that and I think we probably could, as we have more guidance to share, could use those opportunities to share more guidance with that group because we do have fairly good participation on that from our developers who submit measures.

So it's certainly on our radar. So thank you for that, Ron.

So we've already been over this slide so I'm not going to spend a lot of time on it. But, again, definitely thinking about what (Lacy) recommended about the measures that fail at preliminary analysis. But essentially our current process is that both consensus is not reached and measures that pass are going to the committee.

Measures that do not pass do not go to the committee. And we provide a short summary of what we will be doing going forward is that committee members will basically get all of the information that the methods panel got about the measure as well as a detailed summary of your evaluation. And they will have an opportunity to discuss that measure but will not be allowed to revote.

So that's basically the end. I think we've talked through a lot of this already. But if there are any other issues or concerns about the process as we've been implementing it the last couple of years, if you have any, you know, recommendations on what's been working well, what hasn't been working well, certainly we're interested in hearing that.

And I think you guys have already had a lot of really great recommendations. So if there are others, we're really interested in hearing them.

(David): I just have one remaining challenge to mention, but certainly at quarter, ten to five we're not going to deal with it. But I'm not sure how we do. When we've gone through all of our individual and group discussion about reliability and validity and we've looked at all the data and we've got tables. We got everything looked at, you asked us to call something high, medium, low or insufficient.

I struggle on the boundaries between those categories. When do I call it high? When do I call it moderate? And that actually it doesn't matter much because they both pass. And usually I end up calling it moderate just because I'm usually not very impressed, but okay.

But, boy, that cutoff between moderate and low that really matters. And it's right back to our threshold stuff. Somebody shows me some kind of statistic is .58.

Well, it's not high. But boy it's a tough call. And it's tough for me as an individual. And then we get the five of us together and we kick it around. And we can't reach consensus because I don't know how you do reach consensus.

So I don't know what to do that one. And I would have suggested it long ago if I thought I had an answer. That is tough.

(Ashlie): So we do have algorithms for both reliability and validity that gives some direction on what constitutes a high, what constitutes a moderate. So, for example, if a measure has both a data element and measures for testing, that is appropriate and, you know, that kind of meets the score, that would be a high, right? Versus if a measure only - well, it depends.

((Crosstalk))

(Ashlie): Yes, right. So, again, things are still relative. But we do have a little bit of guidance on that. And I'm not sure how frequently you guys refer that or if you found that helpful at all during those periods. Do you ever refer to that?

(David): The way you're describing it's okay. But it's kind of these threshold things that if you only have this one thing, you can only give it high. Okay. I get that. But

it's a judgment call. You've got both things and they're both looking kind of squishy. What do you call it? And eventually it has to pass/fail.

(Karen): Oh, yes. Thank you.

Woman: I've got an appointment.

(Karen): Oh, good.

Woman: I can't miss that.

(Karen): Okay, good. Thank you for everything today.

Woman: (Unintelligible).

(Sherry): I want to second (David)'s point because the distinction between high and moderate is meaningless and the distinction between low and insufficient is meaningless unless you're giving it back to the developer.

If that distinction is meaningful to the developer, I will struggle with those things. But distinguishing high from moderate to me is meaningless. It's pass/fail. And if it failed, was it failing due to, you know, it was crummy what you did or you didn't do some things or what you did didn't give us enough.

So, you know, I mean, having these distinctions be meaningful to this group, anyway, I think is a heck of a lot better than having these categorizations that don't mean anything.

(Karen): We should definitely think about that internally. At one point back in 2011, I guess, when we went to the high, moderate, low, insufficient, the hope was to drive additional testing.

So at that time, if you only did one level you would only get moderate. And if you did both you would get high. So it was trying to, you know, give a carrot for doing more testing. We actually changed that.

(David): It's a really little carrot.

(Karen): It is a little carrot. And then we even changed that later because we said, well, hey, it's score levels that we're really, really interested because we do things, you know, based on - we make decisions based on the scores. So, again, it was another little carrot.

I do see, personally, I have less trouble between low versus insufficient and I find that that is an easier call to make. But I struggle when we do our staff ones, you know, is it really high or is it moderate? And at the end of the day it doesn't really matter. So we will think about that.

Mike: I'm not so troubled about the fact that we sometimes disagree. And I think that, you know, in some cases it's easy and then the algorithm is linked to the right (unintelligible) that we all agree with.

But some of these things really are hard. And when the data aren't there that's why you need a committee like this. And the fact that we don't always agree is a sign we're struggling with the hard cases and it justifies our existence.

Jack: So several years ago I think you experimented with like a two stage process? Remember that? I think the developers liked that. I mean, I liked it because for

stages you got feedback on importance. And it was early enough feedback that you could actually do something about it, you know, before you submitted it.

And so the other thought is sort of around peer review, you know, like is there a way that measure developers who wanted it could get some peer review of their submissions, you know, before, you know, as a way of providing feedback, you know, before they actually go through an evaluation cycle, like another developer.

((Crosstalk))

(Andrew): Maybe they could get some...

((Crosstalk))

(Andrew): Yes.

((Crosstalk))

(Andrew): I understand. It wouldn't have to be required, you know, but if you wanted it. Plus if you served as a peer reviewer you could get, like, a higher position in the field.

((Crosstalk))

(Andrew): Little carrot.

(John): Yes, just, you call for a couple maybe improvement ideas or just wild ideas. One is, you know, we're going to be involving developers it sounds like more in our calls and meetings in the future.

And just first the observation that in the last couple of calls you get some of developers who are just kind of rambling, you know. They are just unsolicited. You know, they slip into pitching their measure like they've done a thousand times.

But just we, you know, need to have some really strong guidance for, you know, basically, I don't want to make this sound mean, but, you know, they speak when they're spoken to. You know, here's our question. Answer the question. Okay. Now that goes back to the group, you know, to set some boundaries because we only have so much time.

Just one other thought, you know, when we're really struggling, I think we'll continue to struggle for years, and that's fine, with really flushing out what we mean by validity and reliability.

I would suggest in the future to bring in an ad hoc person to help us struggle with that. You know, the obvious example is we referred to (John) Adams in virtually call we had and maybe because I know who he is and I know him a little bit.

It wouldn't hurt to approach (John) and ask him to sit on a meeting we're having. You know, we're constantly trying to, like, read his mind through the papers. I think it could help accelerate our conversation and contribute to it to bring in some outside eyes and ears to help us think about it in a meeting.

You know, we might have to pay him. But I think it could be of assistance.

(David): Yes. Time check observation. Before we do close, we do have to allow for public comment. Maybe now is the time. I don't see any other cards up.

Woman: Okay. If you would like to make a public comment, please raise your hand or make a comment in the comment level, commentary level box as I mentioned. (John Shaw), you had a comment?

(John Shaw): Hi. I typed it into the chat box just because I wasn't sure we'd get some time. But one thing that came to mind in listening to the discussion today is the importance of patient and caregiver voice.

There's a new NQF policy of having a minimum of two actual patients and caregivers on every NQF committee. And I would suggest that each subgroup have them on as well.

And particularly since they're the experts in a couple areas that seemed to be problematic during the discussions, looking at facility, looking at PROMs, social determinants of health, shared decision-making and so on. And so that would be a strong recommendation.

(David): Thank you.

Woman: And, Ron, has his hand up so he might want to just say something. Ron? Oh, it's down. He didn't mean it.

(Karen): Well, we took every minute of the day that we had. We got through probably, what 66% of our agenda. So we didn't get through everything. So I personally think I enjoyed the meeting. And I think we came to some agreement on some things and some agreement on where we're going to go next trying to get additional agreements.

So I just want to say thanks for coming in. Thanks for participating. We appreciate it very much. And I'll hand it over to Dave to kind of close this out.

(David): And without holding anybody back further, I really appreciate the work that we did today and the comments. I just think people were on point. People were either agreeing or disagreeing with each other. We're cordial. We're professional. We're respectful.

You don't see that in every meeting. It's a good feel in the room. I like how we work together. We did not reach consensus or have clear answers on all points. I wish we did but it's not because we were talking about baseball or the weather or politics or something. I think that we used our time very well. But I kind of think people did a nice job of having a flow of discussion around a topic.

And I think in places where we still have loose ends hanging, which are quite a few, some, for example, we can address some of the morning points on reliability we can take, as part of this paperwork.

You know, we can put out strong men or women hypotheses, however, you want to say it, and say let's get this out in front of people in writing with specific framing on it and then we'll get comments back. And it's just a way of pushing that forward.

We will have future phone calls. I don't think we have a distinct paper set up to pick up validity issues, but we still have ways of sort of framing specific alternatives.

We talked about a couple of those sort of midway afternoon saying, you know, we could do A. We could do B. Which one do you like? I think actually that was in the check box thing, but okay.

So I'm happy with how we spent the day. I was looking forward to it coming in. I know it has to end. Everybody is getting tired. I would find it kind of fun if we came tomorrow. But okay.

(Sherry): Thank you for your leadership, too, you and Dave did a very nice job of - it didn't turn into night of the long knives. So we're all (unintelligible) back here.

(Ashlie): We second that. Thanks to you and Dave.

(David): I appreciate that. I think we tried to keep a positive and professional podium and have people feel that their suggestions are always welcome and there's no such thing as a bad idea.

It may turn out technically there's a wrong idea. I'm guilty of that frequently. But, you know, we just point out that. Somebody else who knows better says something and then I'm good. So I think it's a good group. I found it enjoyable doing this.

END