NATIONAL QUALITY FORUM

Moderator: N/A
March 19, 2019
3:27 pm CT

Miranda Kuwahara: Good afternoon and thank you all for joining the Scientific Methods Panel Subgroup Number 4 (unintelligible) Measure Evaluation conference call. My name is Miranda Kuwahara with NQF and we'll begin to poll with some brief housekeeping remarks before diving into the measure discussion.

So to begin, a discussion guide was sent to subgroup members on Monday morning. This document will guide save measure discussions and (unintelligible) for the order presented on that document. Consensus was not reached for the first nine measures and we will be focusing our time on those measures during our call today.

All other measures will not be (discussed) unless a member of the subgroup would like to pull a measure. If the subgroup chooses not to discuss any additional measures, the decision from your preliminary analysis will be final and at the end of today's call, we will ask subgroup members if they would like to pull a measure for discussion.

In that same e-mail that went-out Monday morning containing the discussions was also a (unintelligible) Survey Monkey. We asked our subgroup members

to please pull-up (unintelligible) and capture both as we talk through each measure and staff will prompt you to cast your votes when (unintelligible).

The same vein (in priming) it is limited on today's call and due to the number of measures we have slated for discussion, it's likely that we'll continue the discussion during the follow-up meeting scheduled for this Thursday, March 21st from 3:00 to 5:00 pm Eastern Time.

And finally we do want to note that this is a public call. We have a developer representative (unintelligible) they will answer questions from staff or from panel members; however, there is no opportunity for public comment.

For recordkeeping purposes, we ask that subgroup members and developers please state your name each time before you provide any remarks and with that I will turn it over to my colleague Andrew Lyzenga to conduct roll call and disclosures of interest.

Andrew Lyzenga: Thanks, Miranda. Hello, everybody and welcome. We'll combine our roll call here with a disclosure of interest. I'll just make a few remarks here and try to move through it fairly quickly.

You received a disclosure of interest form from us (unintelligible) or names to the committee and annually thereafter and you received a measure-specific disclosure of interest with each measure review cycle and that asks too about your relationship with any measures under review or any related or competing measures that we've identified.

Between these two forms we ask you a number of questions about your professional activities and the degree of your involvement with any measures under review. In the interest of transparency, today we'll ask you to orally

disclose any information you provided on those forms that you believe is relevant to this committee and specifically the measures you reviewed as a member of this subgroup and any related or competing measures.

We don't need a summary of your resume. We're really just interested in especially interested in grants research or consulting, measure development activities related to the measures under review by this (unintelligible). You reviewed 12 measures this cycle. I'll try to quickly go through those just to remind you.

Those measures are 0138 an HSN catheter-associated urinary tract infection measure, 0139 an HSN central line-associated bloodstream infection outcome measure, 0141 patient fall rate, 0202 falls within (unintelligible), 3501 hospital harm opioid-related adverse events, 3502 hybrid hospital-wide risk standardized mortality measure, 3503 hospital harm severe hypoglycemia, 3498E hospital harm pressure injury, 3516% of patients or residents experiencing one or more falls with major injury, 3504 claims-only hospital-wide risk standardized and mortality measure, 3493 risk standardized complication rate (unintelligible) elective primary total hip arthroplasty and/or total knee arthroplasty for (mips) eligible commissions and eligible commission groups, and finally 3494 hospital 90-day all-cause risk standardized mortality rate following (cavitch) surgery.

You've heard this before but I have just a few reminders. You sit on this group as an individual. You don't represent the interests of your employer or anyone who may have nominated you for this committee. We're interested in your disclosures of both paid and unpaid activities that are relevant to the work in front of you.

Finally just because you've disclosed does not mean that you have a conflict of interest. We do oral disclosures in the spirit of openness and transparency so with that I will call each of our subgroup members' names. Please state your name and if you have anything to disclose. Matt Austin?

Matt Austin: Yes, good afternoon, this is Matt Austin. The only disclosure I would offer is I am part of a broad technical advisory group that the hospital harm measure developers have reached-out to from time to time for input and feedback.

Andrew Lyzenga: Okay, thank you. Lacy Fabian?

Lacy Fabian: I like to say the end so I had missed the 3516% of patients or residents experiencing one or more falls with major injury on the disclosure but went back and gave the disclosure for that measure so I'm not providing any review for that one given my work at (Mitre) we had provided consulting support with that group within CMS who ultimately did that measure and worked with measure development.

Andrew Lyzenga: Okay, thank you. (Larry Glans)? (Larry) are you on? He may be on mute if you are. Go to Gene Nuccio?

Eugene Nuccio: Yes, hi, I'm Eugene Nuccio. I formerly was part of the home health quality reporting program measure development and maintenance system and I have no direct involvement in any of the measures here; however, I was aware of the work done for the impact measure from the staff and LTC and (earth) program but I was not part of that effort.

Andrew Lyzenga: Okay, thank you. ((Michael Soto))?

(Michael Soto): Hi, everyone, (Michael Soto) here. I don't have any relationship with any of the measures.

Andrew Lyzenga: Okay. Thank you. This one is tech again, (Larry) are you on? I thought I might have heard (Larry) earlier on the call but maybe I was mistaken. Is that (Larry Glans) who just joined us by any chance? (Unintelligible). Well, thank you all. I'd like to remind you that if you believe that you might have a conflict of interest at any time during this meeting, please do speak up.

You can do so in real time during this call or you can send a message via chat or e-mail to anyone on the NQF staff. If you believe that a fellow member may have a conflict of interest or is behaving in a biased manner, you can also point this out during the meeting or send a message to NQF staff.

Do you have any questions or anything you'd like to discuss based on the disclosures made today? All right, hearing none, thank you all for your cooperation with that. We'll go ahead and get started with the meeting. As Miranda mentioned we'll move through these in the order they are listed in the discussion guide so we will start with Measure 0138, it's just NHS and (howdy) measure.

This is a maintenance measure, for standardized infection ratio of healthcare-associated, catheter-associated urinary tract infections. (Some) patient care locations except Level 2 or Level 3 NICUs. We had consensus not reached decision in our preliminary ratings on both reliability and validity. We did have data element validity testing conducted for the measure.

It is NQF policy to allow data element validity testing to serve as a demonstration of data element reliability. I think there may still have been some confusion on this point so I just sort of wanted to reiterate that so if that

was a concern of any of the members that data element validity had been used for purposes of demonstrating reliability, that's something that we have provided in our guidance to developers that they are allowed to do.

I will let's see, so with that said were there any other concerns about reliability from our panel members, again we had in the submission form that's out there's the first validity testing section and we had and they provided some data element validity testing from states that have implemented this measure.

Let's see, just taking a look, they provided sensitivity, specificity, TPV and NPV (unintelligible) data elements, provided the validations indicated a (pool me) sensitivity of roughly 88%, specificity of 99%, positive predicted value of 96.9% and negative predictive value of 96.9%.

Do you want to talk about reliability? Again we had consensus not reached on this criterion so was the concern, sorry, go ahead.

Eugene Nuccio: Andrew, this is Gene Nuccio. I did rate this as low and my objection was that I didn't see the reliability data provided; however, given the position of NQF with validity, demonstration for data elements is sufficient and I'd be happy to change my vote to moderate on this.

Andrew Lyzenga: Okay. Great, thank you so much. Any other thoughts or comments on reliability? I think we will still have to revote to reach, you know, try to reach consensus or get a new decision on that. Is everybody comfortable taking a revote on reliability or does anybody want to talk about anything else?

(Ashley): Hi, Andrew, this is (Ashley). Just I just had a suggestion that maybe having a discussion about validity since that is the component that would kind of count

for reliability to make sure that there's kind of general agreement that that validity assessment is adequate. That might help from having to kind of go back.

Andrew Lyzenga: Yes, I think that's a fair point so yes, let's go ahead and move to the validity testing then so or validity overall rather. We can start with testing for this is the data element validity testing again can apply to both reliability and validity at the data element level.

There was some concern that the developer lift the data coming from all 50 states but the testing results were only provided for five states. I think that was just, you know, the data does come from all 50 states in the NHSN as I understand it but the testing conducted by five states so that you know, that's acceptable from our perspective so I just wanted to sort of clarify that.

There was also a note that the data element and validity testing was performed only for the numerator data elements, not for factors used (unintelligible) model. I would also note on this one that the risk adjustment variables are actually not at the patient level for this measure.

There's the, you know, the risk adjustment is based on the type of unit that the patient is in so that I think maybe the reason that there was not testing done to those data elements. Do we have additional thoughts or concerns or comments on validity testing?

Matt Austin: Yes, so this is Matt Austin. The concern I had was NQF's guidance is that it's the measure is being re-endorsed that there needs to be empirical validity testing of the measure's score and I did not see that nor was a rationale offered for why that wasn't done.

Andrew Lyzenga: I believe actually we don't require measure score level testing necessary. We do require empirical validity testing of some sort but that can be done at the data element level. For a maintenance measure we require empirical testing either at the data element level or at the measured score level over face validity.

We don't require face validity for maintenance measure unless the developer has provided a justification or rationale for that that is acceptable to the committee but data element validity testing will suffice, you know, for you know, our requirements for a maintenance measure.

Lacy Fabian: This is Lacy Fabian. Matt Austin that was my impression as well so (unintelligible) not the case or requirement then I can change my assessment.

Matt Austin: Okay.

Andrew Lyzenga: There were a couple of other concerns I think raised beyond just the testing so I can maybe go over those quickly. Sorry, I lost my place here so the risk adjustment there was some concern of the, you know, the lack of social risk factors included in the risk adjustment model but the reviewers did note that since patient level factors aren't collected, it's unclear how the developer could have done that but they couldn't have really tested that either.

There was some concern about no statistical results like a CC statistic or a, you know, a model power were reported for the risk adjustment model and some other concerns about the testing done for the risk adjustment model. I don't know if we want to talk about that a little bit about the information provided in support of the risk adjustment.

So that may have been (Larry) that brought that up and I don't know if (Larry) has joined us yet. (Larry Glans) are you on? Doesn't sound like it. Does anybody else have any concerns about risk adjustment they want to discuss? All right, well the final item, point of concern was sorry, did somebody have something to say about that?

Woman: I didn't know if you were asking for comment from a developer or from the committee.

Andrew Lyzenga: I was looking for committee comments. Maybe we can talk about the last concern and then, you know, come back if you have any comments you want to make about that. (Unintelligible) I think the last, sorry, go ahead.

(Gene Pomment): This is (Gene Pomment) commenting. Traditionally the risk adjustment is applied to the score rather than for the provider and given that there's a kind of a lack of information about how that score differs amongst different providers I'm a little uncomfortable trying to make a comment about how well the risk model works (if we don't have) a lot of information about how the score works.

Andrew Lyzenga: Well, maybe this would be a good time to get a bit of clarification from the developer. Did you have some remarks you'd like to make about the risk adjustment approach and clarifications for our panel members?

((Crosstalk))

(Michael Soto): This is (Michael Soto) maybe while they're getting ready to do that, I just want to add a little bit that a standardized incidence (rated) the kind of risk adjustment so it's really different from a lot of the ones that we typically see for outcome measures. They do provide the guidance from CDC about how to

do SIRs and they have followed those methods (I think) well so it may be that it's kind of a different approach than we're used to.

Andrew Lyzenga: Do we still have the developer on?

(Michael Soto): He may be on mute. You know, somebody just spoke-up, did we lose them? All right, well I think the final concern I think was about the meaningfulness of differences identified by the measure. There was some note that a fairly small proportion of facilities have an opportunity for improvement roughly 9% the implication being that around 91% of facilities might not have an opportunity for improvement. Any comments about that from the committee?

Matt Austin: Yes, I mean, this is Matt Austin. I mean, I personally thought that there is some variation of performance, you know, 13.66% versus statistically significant less than 1.0 and 8.68% were statistically significant, greater than 1.0 so at least 22-ish percent of hospitals are different in the mean.

(Michael Soto): Okay.

Matt Austin: I don't know if there's a magic threshold for what is a meaningful difference.

(Michael Soto): Yes, right.

Matt Austin: But for me 22% of hospitals being identified as good or for performance is something.

Andrew Lyzenga: Any other thoughts from the committee on that? Are we ready to move on? So just to sort of read back what we just talked about, it sounds like there is some comfort with the testing being provided at the data element level and for that testing information to be used to demonstrate reliability of the data

element level as well so is everybody comfortable taking a revote on reliability and validity at this point?

Matt Austin: So yes, Andrew this is Matt Austin, another further question so they and maybe I need to just pull-up the algorithm for validity but if one agrees that the data element validity testing is sufficient, what would one rate validity?

Andrew Lyzenga: So yes, I was actually just about to say that, that moderate is actually the highest eligible rating because they only provided data element validity testing so that is I should have mentioned that before, the measure would only be eligible for a high rating if they had in fact provided the score level validity testing and the same with reliability. It's only eligible for a moderate for reliability. Hearing that said, should we go ahead and vote on that?

Matt Austin: That's fine.

Andrew Lyzenga: Okay, while you're doing that, the next measure is very similar. This is the (cloud-V) measure outcome measure from the NHSN and I think that we have the same concerns pretty much identical to the previous one.

So I don't know if maybe we could kind of skip over this discussion if everybody's comfortable with your votes on reliability and validity for 0138, we can just go ahead and take a revote on 0139 as well unless there is anything that anybody does want to discuss about 0139.

Matt Austin: No, that's fine, go ahead.

Andrew Lyzenga: Okay. Time to revote.

Man: Yes.

Andrew Lyzenga: All right, let's do a revote on that one too as well. All right, we can move on then to the next measure.

((Crosstalk))

Woman: Hi, this is I'm sorry to interrupt but earlier you mentioned that you wouldn't be discussing a few measures in this call. Would you be able to let us know which of those measures that would be?

Andrew Lyzenga: Sure, so it's 3504 claims only hospital-wide risk and standardized mortality measure. That received the high rating for both reliability and validity. 3493 risk standardized complication rate following elective primary PHA or PKA for (mips) eligible clinicians, that received a high for reliability and a moderate for validity.

And then 3494 hospital 90-day all-cause risk standardized mortality rate following (unintelligible) received a high rating for both reliability and validity so those three measures, 3504, 3493 and 3494 will not be discussed on the call.

Woman: Thank you so much.

Andrew Lyzenga: You're welcome.

Man: What about 3516, that was the first one on the agenda list?

Andrew Lyzenga: Actually it's last on my list here so we are planning on talking about it but it's not until the end of our agenda here. Can we move ahead to the next measure then? All right, let's do that. The next measure is Measure 0141 patient fall

rate. This is all documented falls with or without injury, experienced (citations) on eligible unit types in a calendar quarter, that's recorded as total falls per 1000 patient days.

We had a high/moderate rating for reliability on this one so I don't think we need to discuss reliability but for validity, we did have a consensus not reached decision. We had some concern about the testing. Reviewers were not entirely confident that the analysis provided by the developers was a demonstration of measured validity.

They're thinking it may have it may demonstrate something more like reliability against pulling the information up for this one. Did anybody have any comments while I do that?

Matt Austin: ETS, so this is Matt Austin. I may have been the one to make that comment and partly I guess in looking to my colleagues on the subcommittee to help educate me or I'd love to hear their thoughts on whether they thought that the testing that was provided did demonstrate validity.

(Michael Soto): Matt, this is (Michael Soto). I'm totally with you on that. I didn't understand how we worked that either.

Matt Austin: Okay.

Eugene Nuccio: And Matt, this is Gene. I would concur. I'm a little suspect of the bootstrapping method in that depending on how the bootstrapping is done, you could be testing the quality of the bootstrapping rather than the validity of the measure itself.

Andrew Lyzenga: So do we have the developers on the line for this measure?

Emily Cramer: Yes, this is Emily Cramer with the University of Kansas. We are the measure developer and the American Nurses Association is the steward.

Andrew Lyzenga: Okay, great. Could you maybe give us a little bit of explanation for your method of assessing validity particularly the bootstrapping approach, maybe a little bit more detail on that for our panel members?

Emily Cramer: Sure, so we tested it several different ways in the past, the last two re-endorsements we've used this particular method and the reason for using the bootstrapping is really to create this sort of empirically-based distribution that we can use to sort of approximate the true fall rate because we can't know the true fall rate.

And so in order to determine whether or not the reported fall rate is nearer what would be a true fall rate, we've used the empirically-derived distribution from a bootstrapping technique so we create sort of this empirical distribution and then rank the hospitals based on that and then the comparison is how close the rank of the hospital on their fall rate actually matches the bootstrapped.

So it's sort of a way for us to simulate a true fall rate and then look at the comparison of the actual fall rate to the true fall rate or the estimated true fall rate and so that's why we just classify it as a validity if you wanted to look at the observed compared to the true fall rates.

(Michael Soto): So this is (Michael Soto) and I mean, it seems to me like that's an interesting way to look at reliability but it really doesn't get at validity for which you need something external to the measure itself and you just this to me sounds like a way that says that the measure is well (really get) that validity (that strikes me).

You either have to say this can predict something or is related to something that we think it should have predicted or it's related to other aspects of this concept, things like that.

Woman: So we have, I mean, we have in the past used some of those methods and (unintelligible) those have come-back with I think good scores in the past and so we just changed to this method recently so I think there's evidence from other sources that we could potentially provide too if that helped strengthen this measure (unintelligible) documentation yet.

Andrew Lyzenga: I'm not sure if we have much opportunity to give more information at this point because they'd have to take submission (unintelligible) provided.

Matt Austin: Would there be any possibility to have the person who's coughing mute themselves if they could? Thank you.

Andrew Lyzenga: Thanks, Matt. Any other thoughts on this method of validity testing from our other panel members?

Eugene Nuccio: I'm sorry, I muted myself, this is Gene Nuccio real quick. The stratification that you used on I think I talked about nursing homes and hospitals but then you had multiple other groups. Could you describe how all that stuff works together and I'm sorry that I'm not making my question very clear? You said there's 9 subtypes, unit types for stratification and then your testing was done at the nursing unit and the hospital unit.

Are you reporting this at the nursing and hospital unit where the analysis was done or are you reporting this at the 9 subunit level?

Woman: So it's been there's actually two versions of the measure. There's one at the unit level which is we definitely tested on the unit type measure and that that's where the stratification really happened and the reason for that is because it's all on in-patient nursing unit and so for example critical care units are expected to have a much lower fall rate than for example med-surg units because the patients are just less mobile on the critical care unit.

So that stratification is meant to make the comparisons within units more valid because that we compared two units of the same type. The (hotta) level measure actually involves an aggregate measure of all the units that are submitting data so that the data we used to test this is from the national database of nursing quality indicators.

It's also made at the unit level so in that hospital-level measure in order to again level the playing field, we create a strategy, we used that same stratification and weighed it based on standardized scores and weighted by the number of the patient volume in each of those unit types.

So a hospital with for example six critical care units isn't being compared to a hospital with one critical care unit because their falls wouldn't necessarily be lower if they had fewer units that were submitting on the med-surg units or rehab units where the fall rate is much higher so the stratification is used for comparison at that unit level and then also it's kind of the methodology we used to roll it up to that hospital level.

The measures are reported through NDHQI at both the unit level and the hospital level using that roll-up technique, that aggregation technique.

Andrew Lyzenga: And does the stratification take all that into consideration in terms of I mean, you used the term inappropriately but conceptually I think correct, the case mix of the hospital.

Woman: Yes, so it's not true case mix but …

Andrew Lyzenga: Yes, I recognized that …

((Crosstalk))

Andrew Lyzenga: … nine units.

Woman: … yes, so it would so that we created standardized scores for unit type so across the entire sample, the standardized score and so we created these scores for each unit type and then within that we would within each hospital we weight their hospital score by each unit that submitted data, that unit type score plus to times patient volume in that unit type so we'd be proportionate to that number of type of patients that are seen at that hospital.

Andrew Lyzenga: Okay, thank you.

(Michael Soto): This is (Michael Soto) again. Do you report the results by across the different strategies? It seems to me that what you said there about how you expect to find greater fall rates in some kind of units than in others is actually a measure of validity that I think is missing.

Woman: So we do report at the unit level and within NDHQI there's benchmarks (unintelligible) by that unit type and so and the hospitals get each individual unit a unit type and then a hospital level …

((Crosstalk))

(Michael Soto): I mean, is the data in the submission?

Woman: Yes.

(Michael Soto): Where would I find that?

Woman: Oh, you mean, the scores in the submission?

(Michael Soto): You said earlier that you expected that in certain kind of units, the rates would be higher in those because of the nature of the unit.

Woman: Yes.

(Michael Soto): And that I think it's that you found that but I don't see where that is actually recorded in the material that you cited.

Woman: Oh yes, that's a good point so we do include that in the NQF submissions. They previously had been included as part of the measure information form. We included to show just differences across the units in that the main measure information form hasn't included it in the measure testing form here so yes, you wouldn't be able to do that right (unintelligible) based on this document.

(Michael Soto): Okay.

Matt Austin: This is Matt Austin. I had two other concerns with the validity testing. One was at least from what I could see, it looks like the validity testing was just done at the hospital level and yet the measure is specified at both the hospital level and the unit level. Am I misunderstanding that?

Woman: The validity testing was done at (rows) although I'm looking at this and it looks like it those are not included here so no, I don't think you're missing anything.

Matt Austin: The other piece that I was a bit concerned about was in (clue) of the narrative there was some discussion about risk factors, gauge history of falls but then there's no risk adjustment in the measure so it's maybe trying to understand why the decision not to risk adjust given that there seems to be some risk factors for falling?

Woman: So we one of the reasons we don't collect data on those risk factors currently so that's one of the reasons we haven't done that the risk adjustment and we also haven't tested a model for risk adjusting based on falls. There's a number of I think there's a number of models out there.

There's also a lot of discussion about risk factors for falling need to be included and should be included and what ones aren't and so we haven't come-up with a reliable risk adjustment model yet so those are the two reasons we haven't tested it here.

Matt Austin: Okay, thank you very much.

Woman: Sure.

Andrew Lyzenga: Okay, any other thoughts on that subject? I think there was a little bit of there was also some lack of clarity on the method of assessing meaningful differences. Let's see, so the see what they did, the treating of ranked hospital scores and unbiased estimators is ranked true fall rate. I'm assuming the hospital's percentile ranked scores are independent with (galcean) or uniform

distribution to hospitals with true fall rates at the 42.5th and 57.5th percentiles that have hospital scores differing by at least 15 percentile ranks in the same direction in 50% of repeated samples.

That sort of I'm not sure if that was the result that was that came-out of the testing. Could we get a little bit of clarity on that from the developer as well and how your method of testing meaningful differences in the results?

Woman: Sure, let me (unintelligible).

Matt Austin: It's Section 2.B4.

Andrew Lyzenga: 2.B4, I mean, 1.2 over.

Woman: Oh yes, so the way that we're treating the ranks would show that dependent on their percentile you would actually see significant differences in hospital scores. They would actually differ within that percentile rank, they would actually have quite a bit of difference and they would differ by that and it will be half of the sample.

So we would see it over and over and over that they have that difference of at least 15 percentile rank, the hospitals that are different actually show-up as different. Is that addressing the question?

Andrew Lyzenga: That's satisfactory for our panel members?

Matt Austin: Yes, that's helpful, thank you.

Andrew Lyzenga: All right, well are we ready to revote on validity for this one as well?

Matt Austin: Andrew, this is Matt, just to clarify, do we need to vote on reliability as well or just …

Andrew Lyzenga: No, we had a passing vote on reliability so there's no need to vote on that one again, just validity. All right, without hearing no objections, we'll ask you to go ahead to vote on reliability, or sorry, validity as your Year 141 and we'll go ahead and move to measure 0202 which is falls with injury.

Had the same kinds of results in terms of our preliminary ratings. We got a high moderate result for reliability so no need to discuss that where we vote on it. For validity we again got a consensus not reached decision. I believe we have the same kind of concerns here again, some lack of clarity and concern about the method for use for assessing validity.

Have we talked about that enough from the previous measure? Do we want to have any more discussion about the testing method?

(Michael Soto): This is Mike. I had pretty much the same concerns about this one as before but …

Matt Austin: This is Matt, I have the exact same concerns as the previous measures so I don't need any (unintelligible).

Eugene Nuccio Yes, and this is Gene, same thing. Just a quick question not related perhaps to your what we've been chatting about but how do you see and this is a question to the developer, how would you see this measure about falls and reps, falls with injury being serious falls being related to other measures that are being developed in the post-acute environment to an acute care environment for the impact measures?

Woman: So we have a few years ago did a harmonization study on our falls compared to other fall measures in NQF and I think there's been a few more potentially developed and then one of the key differences is in just the denominator. A lot of the numerator definitions are pretty similar and a lot of the data elements are similar but the key difference is this denominator.

So these are for in-patient fall rates and we've used hospital days because that's the most relevant exposure variable for fall rates. Some post-acute care settings, that's also appropriate. Others I think they've used things like total patients, patient admissions rather than patient days. In outpatient settings a lot of times they use patient visits so I think the primary difference is in that denominator definition. Does that answer the question?

Eugene Nuccio: Okay, thank you.

Andrew Lyzenga: Any other discussion or can we go ahead and vote on this one as well? Hearing no objection, go ahead and vote for Measure 0202 on validity and we will move now to Measure 3501E which is hospital harm opioid-related adverse events. This is an ECQM an electronic clinical quality measure. We had a consensus not reached decision in our preliminary rating on both reliability and validity.

We had let's see, validity or sorry, reliability tested at the element and score level. There was some concern about the sample size that used in the testing, some lack of clarity around the methodology used. With respect to the data element testing, some concern about the sensitivity of the opioid administered with date and time data elements.

Regarding the score level testing, again some concern about the sample size which was fairly limited, only five hospitals and let's see, and I think some it

was a question I had is I might describe what they described as score-level testing as more of a data element testing. I think that may have been raised by our reviewers as well. Any discussion about the reliability on this measure?

Oh, I'm sorry, that was mistaken. I was talking about validity testing. For reliability they did conduct a signal-to-noise but again we have some concern about the size of the testing sample. There's a pretty high median reliability score, .95 with a range of .86 to .96 and limited sample size there.

Eugene Nuccio: Sorry, this is Gene again and I have to ask this to my fellow committee members or panel members, the measures limited to the first 24 hours of hospitalization and I'm not sure how that does or does not perhaps overlap with ED admissions or time. Is there any concern that the 24 hours element is problematic in terms of the scoring? I don't know how this activity happens in hospitals so I'm just looking for your guidance on that matter. Mike, do you have any thoughts?

Matt Austin: This is Matt. Your question again is how does the measure incorporate the ED?

Eugene Nuccio Well, not only the ED but it says the measure's limited to the first 24 hours of hospitalization.

Matt Austin: Uh huh.

Eugene Nuccio: And so I guess I wasn't clear about whether the time clock starts with ED time or if it's admission to hospital time and if 24 hours is sufficient time to get a representative piece of information about the use of (naroxylen) or naloxone.

Matt Austin: So this is Matt. I interpreted it as the clock started when the patient was admitted to the hospital so I changed to inpatient status.

(Karen Dorsey): I just want to remind you all that we're the developers, this is (Karen Dorsey) from EL Corps. We're on the line and happy to clarify if you all would like us to.

Matt Austin: Yes, I think that would be appropriate, could you clarify that point about the start of that 24-hour period or does that (need) the emergency department?

(Karen Dorsey): So the measure captures all administration of naloxone that happens during the entire period of an inpatient admission including care that was provided in the emergency department or an observation status if that was then like converted to an inpatient admission so naloxone is counted everywhere continuous outpatient and inpatient locations that are all consumed in an episode of an inpatient admission.

The 24-hour piece comes-in only in the case that if naloxone is given in the first 24 hours, the specifications require that there was an opioid administered by the hospital prior to the naloxone administration to avoid counting as harm the naloxone that's given to reverse overdoses that happened in the community.

Andrew Lyzenga: Thank you for that clarification.

(Michael Soto): This is (Michael Soto). Can we come back to the issue about score validity versus data element validity? I'm looking at it again and I realized that I think although I maybe didn't pick-up at the time that they do seem to be both like data elements rather than score even though they're distinguished in the measuring in the testing documentation. Can you explain why the second part

of that is a score of validity test or maybe what you did and why you just score validity?

(Karen Dorsey): Sorry, that was a question for the developer?

(Michael Soto): Yes.

(Karen Dorsey): Right, so what we did was perform adjudication in the medical records to confirm that a harm actually occurred, that naloxone was administered to reverse the effects of an opioid that had been administered in the hospital and so we reconsidered that like the gold standard, right, so that the medical adjudication of a harm in terms of thinking about patient safety measures that …

((Crosstalk))

(Michael Soto): But why is it an updated element as opposed to score validity?

(Karen Dorsey): … because we were adjudicating the occurrence of the harm so we think of data element validity as can you show the naloxone date and time that was extracted from the EHR is the same date and time that you find when you look at the clinical interface?

Can you confirm that the opioid medication was administration that extracted from the EHR is the same one you find when you look at the clinical interface but this is a clinical judgment that's made to actually confirm that a harm event occurred, that the intent of the measure which was to capture naloxone to reverse opioid administration by hospital staff was actually met by the electronic specifications.

So we have a clinical adjudicator looking at the entire medical record, all the information as they're able to look at and making a determination that yes, the harm is you have to find it did actually occur, not just that the data elements were present and accurate. Does that make sense?

(Michael Soto): I understand what you did and I actually think that that's a strong thing to do but I wouldn't call that data element validity and it's possible to approve it at the moderate level but it just what I would …

((Crosstalk))

(Karen Dorsey): So I'm just curious in that context what you guys now just for our learning what you all would consider a better test of the validity of …

((Crosstalk))

(Michael Soto): Well, generally the thing for validity is that you have to compare the score across a range of units to some other external measure that you think should be related. That's an (unintelligible) maybe the other committee members have a different view.

((Crosstalk))

Andrew Lyzenga: This is Andrew from NQF. That's typically our understanding as well when you're talking about score level validity testing. You want to see some demonstration that that score that's, you know, rather than at the patient level, you know, at the facility level or whatever level of analysis you're talking about, that's sort of rolled-up and calculated score is an accurate reflection of quality that's been provided at that facility.

It's usually again entails comparison or correlation of that score with some external measure of you know, that may be related or you know, that you would hypothesize would be related in some way if that makes sense.

(Michael Soto): Maybe other kinds of injuries for instance.

Andrew Lyzenga: So if you were using that as an example you would want to see that measures that scored well on this measure also scored well you know, on a rate of other kinds of injuries or what is it like?

(Karen Dorsey): Okay.

Andrew Lyzenga: But note, sorry, go ahead.

(Karen Dorsey): No, I was going to say okay, thanks, we appreciate the feedback.

Andrew Lyzenga: I should note that data element validity testing again is acceptable by our requirements so if the panel does either way if the panel does accept the results of the data element validity testing, that's sufficient to pass the measure on to the standing committee although if we did judge it to have only data but with element validity testing, the highest eligible rating would be moderate.

Eugene Nuccio: This is Gene Nuccio calling, speaking again. Your approach does not use any kind of risk adjustment nor any sort of stratification by (unintelligible) by the location of the type of populations served by the hospitals.

Could you discuss that again and tell us why you chose not to do any kind of risk adjustment because it would seem to me that, you know, the likelihood of hospitals the prevalence of this would vary quite widely depending on where

the hospital is located and this type of typical patient population that's served especially if you're including any kind of ED administration of the drug.

(Karen Dorsey): Sure, so I think we did fill-out our rationale on the forms but I'll just say briefly that for this measure in particular because we require administration of an opioid in the first 24 hours to count the administration of naloxone as a harm, we do not imagine any need to think about the differences in populations coming through the door.

So we basically do not count in the numerator as a harm anyone who's getting naloxone to reverse an opioid that they received or took in the community outside of the hospital, only those administered in the hospital by hospital staff.

And so in that instance where we're only talking about naloxone reversal in reaction to over-administration of opioids in the hospital, there really is no clinical rationale for why some populations would be at greater risk of that harm than others since it's entirely in the control of the hospital both the dosing and whether they're administering opioids.

But also the degree to which they're monitoring for adverse events before someone gets so severe that they would require naloxone to reverse those effects. Does that make sense?

Eugene Nuccio: Yes, and I'm sorry, I probably missed that detail on the explanation that community-induced opioid use that's been reversed by the (Laxon) is not - they're excluded in your calculations?

(Karen Dorsey): That's correct and we assessed the accuracy of that in our adjudication, so we confirm by adjudication that the specifications were effectively doing that.

Eugene Nuccio: Yes.

Man: All right. Any more discussion on that point? I know we kind of skipped over reliability. There were again some concerns about the sample size used, but the method was appropriate I believe and the results provided were fairly strong. Did anybody have any concerns - remaining concerns about reliability? Are we ready to vote on both reliability and validity then?

Matt Austin: This is Matt Austin, if I could maybe just bring up an issue on validity.

Man: Sure, of course.

Matt Austin: Maybe just whereas it's getting stuck in my head was the low sensitivity of the opioid asset administration data element. In terms of how well that is actually being captured in the electronic health record. And if we're not getting that right, how do we know that we're actually, you know, capturing a harm event that is an actual harm event?

(Karen Dorsey): So that problem was a problem of not having an exhaustive list of medications in the adjudication tool. And the naming convention is differing with respect to the adjudicator looking for the exact match for the medication name. And so that was not a reflection of the absence of an opioid in the medical record. It was literally because we created a tool that had a dropdown menu and we were missing some names.

Matt Austin: So we really I guess think of that as a lower or a slower, but we don't necessarily know where it actually is, is that…

(Karen Dorsey): Right. It's higher than that.

Matt Austin: Okay.

Man: All right, thank you. All right, any objecting - objection to just going ahead and voting and reliability and validity for this measure? All right, hearing none. Please go to the SurveyMonkey and vote 3501E, provide your ratings for both reliability and validity.

And our next measure is number 3502. This is the Hybrid Hospital-Wide All-Condition, All-Procedure Risk-Standardized Mortality Measure. I know we have - I believe (Michael Abrams) maybe on the call. I think he was going to walk us through this measure. (Michael), are you on?

(Michael Abrams): Yes, I'm here. Can you hear me okay?

Man: We can.

Matt Austin: Can you repeat the number again please?

(Michael Abrams): So this is measure 3502. You'll find it on page 14 of your PDF. The title is Hybrid Hospital-Wide All-Condition, All-Procedure Risk-Standardized Mortality Measure, so again on page 14 of your guide for the committee.

This is a new measure. What I'm going to do is actually try to go through all of the specifications and the reliability and validity points directly and then at the end - save at the end some questions and concerns brought up by you all before moving on to voting. So I'll try to quickly review things and try to anchor you to page numbers on the PDF as well as the points with the embosomed titles there.

So starting with the fact at the very top, this is a new measure and a brief description of it is that it estimates hospital level 30-day risk standardized mortality rate defined as a death from any cause within 30 days after what is referred to as an indexed - an index admission date for patients who are between the age of 50 to 94.

They're referring to this as a hybrid measure. This is the third bullet point down below there. A hybrid measure related to another measure which reviewed the cycle and you came to consensus on in past, measure 3504, the difference between this current measure 3502 and that other measure is that this uses electronic health records rather than claims thus I guess the reason that they use the term "hybrid", because it actually does use some claim based type information as well. But then gathers lab values and other things from the electronic health record and couples those together does creating a hybrid.

It's also the current measure is different than the previous one, because it expands the age down to the age of 50. And finally, one point that they do make which will come up in our discussion is that no empirical of validity testing that actually happens with the current measure. Instead they rely on validity testing that was done on the all-claims that is the non-hybrid measure.

This information (for you) here about the numerator and the denominator, just to remind you if you want to refer back to that. This measure is referred to and then jumping down a bullet - a couple of bullets there, this measure is an outcomes measure of course, but also the developer has referred to it as an e-measure. This measure could be problematic with regard to the testing they provided. I'll make those points later. But they are calling it an e-measure principally because it's based on electronic health records.

Under the data source bullet there, you can see that we're talking about Kaiser from an anti-health records in particular across 21 hospitals referring to something on the order of 340,000 index admissions that they're referring to in their testing.

Next bullet down there, exclusions. Just remind you, they do have a number of exclusions, discharge against medical advice being the top bullet there under exclusions and other represented. And you can see and their argument is that these exclusions were rationale, but more over quite rare, so it shouldn't disturb their results and their analysis that much.

The level of analysis now, skipping down to the next bullet is that the hospital based level. And then just after that there is a long description that I provided for you and indeed for myself to remember about the rich risk adjustment models that they proffer there using this Bayesian Markov Monte Carlo simulation model. And running this model by the way across 15 different chronic conditions or what they referred to as service lines or what you might think of as clusters of chronic conditions where they fit their regression models in order to derive their coefficients to then ultimately calculate an expected rate, a risk adjusted rate for each condition that they would then employ for their risk-adjustment model.

I'm not going to spend too much time on that other than to say that they do spend a good deal of time describing the details of their sort of case mixed adjustment that's largely driven by diagnosis. And diagnosis I might add that happen as you would hope and expect at the beginning of the admission as oppose to at the discharge, right, so that they're sort of coming uplift or quite explicitly coming up with some quantization for the risk in each of the - across each of the 21 hospitals that they will be looking at. Or in case it seems they

actually - that they actually deployed to their national claim sample only in order to do risk adjustment.

And they also took a look at simultaneously and separately, now I'm well down on to the top of page 15. They took a look at the social risk factors of dual that is Medicare and Medicaid eligibility status presumably is a proxy for disability. And then AHRQ and SES, the short story there was that although they saw significant effects, they weren't as strong as their case model. So that involved diagnosis, so they really relied on that.

And - so they proper this case mix model which they employ and then you can see in the middle of page 15, the diagnosis as well as age as a risk factor that they employed for their main risk model. And then further down in a bulleted list here, the bottom of page 15 I should add that they additionally in their risk model employed the electronic health record information they got from Kaiser. The Kaiser EHR record which added these sort of clinical variables, things like heart rate and oxygen saturation, et cetera. So they did use those.

As perhaps you've inferred from my description, there is some confusion potentially about where they deployed these different risk adjusters. They generally argue that they couldn't do full testing in the EHR record, because it wasn't nationally representative. And that's why they deferred to the claims based measure in order to do much of their testing, but that was a little confusing of course, because they certainly tested their risk model using EHR data elements. So perhaps the developer can clarify that at some point.

Moving on to page - the top of page 16, just above the bullets that entitled ratings for reliability. You can see there is a description of their bootstrapping samples and again a reminder about the different chronic conditions that they looked at individually to (define) their coefficients.

But then the bullet below that indicates that when they did employ the electronic health records, they did an imputation approach using some normalized variables randomly selected or normalized values for like, you know, blood pressure and white blood cell count and so forth. That seemed reasonable although they did disclose that for some of the variables there was as many as 50% missing laboratory values. So that's something to keep in mind about this measure.

Getting now to reliability and validity directly, you all gave reliability a moderate rating. There was no element level testing that presumably could be okay as we discussed earlier because this is a new measure.

And you heard instead to split half with adjustment they say and looked at an Interclass Correlation Coefficient then at the score level and that ICC was 0.683, pretty good on adjust - you know, without another adjustment that they say that they have employed. And then when they do an adjusted ICC and it's a little unclear what they mean by that. There is something about the full year sampling versus the 15-month sampling presumably. They also see a pretty reasonable ICC, even a bit higher at 0.77. So that's the reliability presentation.

I'm going to move on now since you all passed that to the validity point, so we're at the middle of page 16. You all did not pass on that, that's why we're having this discussion now. No element level testing was done and score level testing was done with the - only the claims data and for the measure that - the other measure that you reviewed, 3504. In fact, it's pretty much that other validity testing that they referred to.

And so as an example and a reminder to you, they compared the score level testing that they were seeing nationally with their general algorithm for - or cause for fatality to nurse the bed ratios and the hospital star ratings and notice the positive or negative correlation that was, you know, consistent with what you might expect in those regards. And they also then referred back to their TEP related to the claims evaluation. So there was sort of a general validity approach that was applied to this measure as well as looking at it again in a claims-based framework from previous data.

And then finally under validity of course they give us the information from the claims data again only related to the different levels that hospitals were performing at. That's near the bottom of page 16 and just a reminder to you, the worst hospital was at 3.95% rate and the desperate forming hospital at 8.7% and then they give inter-quartile ranges as well.

We talked about missing data already, so let's move on then to key items of concern related to this measure. And one comes actually from staff, because there is some confusion about whether this is an ECQM measure or not. It has been proffered as such. And so under these items of concern on the bottom of page 16, there is a quote directly from our guidance which I will direct at the developer. Is the developer on the line for this measure I should ask?

(Karen Dorsey): Yes, we're still here. This is (Karen Dorsey from the LCOR). So this is a hybrid measure, not an e-measure. Let me just clarify that right off the bat. I don't know where the e-measure came from. I think it only came from the fact that we submit some information that's similar to what we submit for an e-measure for the data elements that are extracted from the EHR and used in risk adjustment. So it has some elements of an e-measure, but it's a hybrid measure much like the endorsed hybrid hospital-wide readmission measure that we also developed for CMS.

(Michael Abrams): Okay. And correct me if I'm wrong and for the benefit of the committee members too then this measure would not be just automatically deployed. It would be deployed with far more careful scrutiny much like you would do with the claims measure, claims in the data, et cetera thus it doesn't need to take the monitor at this point of an e-measure. Is that fair to say?

(Karen Dorsey): That is correct. Hybrid measures are all computed by CMS like claims measures are and have all the data processing that goes along with that.

(Michael Abrams): Very good. Thank you for clarifying that. So then let's move on to issues related to it, just starting with - so the reliability, again you all passed that. So does the committee have any additional things that they want to discuss at this point regarding the liability or perhaps we can, if not then we could move on and just discuss your concerns about validity.

All right, so hearing no apparent objections, maybe we will go on to the validity discourse. That will be - we're on page 17, now big bullet entitled validity. And that first bullet actually is to remain only to an e-measure, so we can skip over that. The second bullet there, score and element level testing is required. Score level testing in this case was not done. Again because of this absence of a nationally represented sample for the Kaiser EHR record. The committee want to discuss that in any way or does anyone on the committee want to get clarification from the developer about that point?

Matt Austin: So this is Matt Austin. My understanding is that they either needed to do score level validity testing or data element validity testing. I don't think both is required.

(Michael Abrams): That's correct from our perspective.

Woman: Yes, that's…

(Karen Dorsey): Sorry, this is (Karen) - this is the developer. I just have a point of clarification. So we do have face validity assessed for this measure that no longer a qualification sufficient to get your moderate vote or moderate rating.

(Michael Abrams): This is a new measure. Am I right about that?

(Karan Dorsey): Correct.

(Michael Abrams): So face validity can be acceptable for a new measure. For maintenance measures we do require empirical validity testing again unless there is some acceptable rationale or justification for only having people with validity. But for a new measure submission, we do still accept face validity. As we suggested would only be eligible for a moderate rating on validity, so.

(Michael Soto): So this is (Michael Soto). It seems to me that score levels on testing was done here. I'm not sure I understood though that the question that came up was about being for the old measure versus the new measure.

(Michael Abrams): Right. Score level testing wasn't - on validity was not done on the list of data from the Kaiser system. It was done with the previous claims approach and did not include the EHR adjustment for example. Am I correct about that? Does the developer want to clarify that?

(Karen Dorsey): We have face validity for both measures.

Woman: Measure scores.

(Karen Dorsey): Measure score face validity and we had empirical validity testing for the claims measure. I think for obvious reasons we could - we do not have the similar comparative measure that is specific to Kaiser that would lend itself to being able to do empiric score validity testing in such a small and particular sample.

But I will also say that the measures are identical except for the addition of the clinical data elements to the risk adjustment. And so we also don't see really a rationale reason why the validity testing that we did for the claims measure will not be wholly applicable for the hybrid measure which basically only has a slight enhancement in the risk model because of the presence of some clinical variables in addition to the risk model that's present for the claims measure.

(Michael Abrams): I understood that. Can you just clarify where these extra clinical variables deliberated by your TEP? Was that part of your…

(Karen Dorsey): Yes, absolutely. They were presented with the entire measure specification and I'll say that for the clinical data elements that are used in the hybrid measure, we also have data elements validity testing for those. So because they are the same clinical data elements that are used in the NQF endorsed hybrid hospital-wise readmission measure. So that's already been presented and adjudicated by NQF committees and they're the same identical data elements.

(Michael Abrams): Got it.

Woman: Yes. So (Michael), this is (Ashley). I think another question supposed to the pane will be is the developers representation of the face validity for the

measure, is that appropriate? And if the evaluation of the face validity is deemed appropriate then that rating should be moderate.

(Michael Soto): Yes, I'm not…

((Crosstalk))

(Michael Soto): This is (Mike) again. I was going to ask the same thing. So since this is a new measure, I understand that face validity if done appropriately can be good enough for a moderate score and then maybe the question is does the empirical testing at the score level that was done even though it isn't quite the new measure where that might actually - neither it maybe - I don't know whether that can even move it up to moderate to high or not, but maybe that's best that can be asked for. Is that?

Let me restate. My understanding is that since this is a new measure, face validity is enough to give it a moderate rate, that's true, right?

Woman: Yes. That is correct. This is (Ashley) from NQF. That is correct.

(Michael Soto): And then maybe we can say then that - could we say that even though the - if we regarded this score level empirical testing, even though it was not exactly on the new measure, that could be an enhancement. Could that move it up to high even if we agree that was an appropriate test or maybe we can't even get to high by the rules? I don't know.

Woman: Hi. This is (Ashley). (Andrew), correct me if I'm wrong. But in order for a high rating, there has to be both measure score and data element validity testing.

Man: That's my understanding, yes.

(Michael Soto): Okay. So we couldn't get to high anyway, this is the data element.

Man: Actually - no, I think actually you could have just score level testing and that will be enough to make you eligible for a high.

(Michael Soto): So then the question becomes do we regard the score level testing of a slightly different measure as close enough to potentially move it up from moderate to high I think, right?

(Michael Abrams): Yes, it's a fair question. It - this is (Michael) at NQF talking. So I think it depends on how close the claims are alone to the hybrid which, you know, as the developer just argue there is a lot of similarity and overlap if you are comfortable with that characterization. The more comfortable we are with that characterization, the more likely that I think you could use your discretion to decide this might exceed a moderate rating. But that's what it hinges upon. You know, so we're talking about comparing 21 hospitals and 340,000 people to, you know, millions of people across several, you know, 1,000 hospitals and what the differences might be there and how that might impact the scientific acceptability of the measure.

I want to point out just - I have only one last point that I think is a salient one for the committee to consider and that is the imputation point about adding the electronic health record. Remember some of the variables we're missing is many as 50% that is one half of them we're missing and they ransom imputation with the average values of blood levels or blood pressure levels that they had. Anybody concerned about that as something that might threaten validity in some way?

Matt Austin: So this is Matt Austin. I think I'm the one that made that comment and maybe others did as well. That was my biggest concern on the validity side was the - I mean what is the real impact of those missing values there as measure developers able to provide anything else around that topic.

Woman: Hi…

(Karen Dorsey): This is (Dorsey) speaking.

Woman: Sorry (Karen), go ahead.

(Karen Dorsey): No, please.

Woman: Sure. I just wanted to clarify where the data we're missing, so we have both surgical and non-surgical divisions. So the missingness of the 15% to 50% is limited to lab result values for only the surgical divisions, so there is much less missing data in the non-surgical divisions and for the vitals. So for the surgical divisions missing lab results, we did random imputation within using value in the normal range for that lab. And the reason why we couldn't do any kind of testing regarding the impact of the missing data is that we only have 22 hospitals in this dataset and we thought it would not be very, you know, illustrative as to what the impact will be.

Also let me just provide the rationale for why we use this imputation method. So these surgical patients that are missing initial lab data are more likely to be elective surgical admissions that has labs collected within 30 days prior to admission. So we were less concerned that the patient wasn't abnormally - with an abnormal lab value would undergo elective surgery without having labs checked again. So we felt it was, you know, a reasonable approach for this type of missing data.

(Karen Dorsey): Right. And I'll only add that those kinds of decisions were also vetted with our clinical experts and technical experts. So the validity that they put forward reflects that.

(Michael Abrams): Got it. And those points were made in the application, my recollection is - this is (Michael) speaking at NQF. Are there any other concerns or questions? I think that the point of action here is whether or not the committee wants to recast their vote on validity? It previously passed it on reliability, so if any of the final points or questions that the committee wants to post to the developer or otherwise?

Eugene Nuccio: This is (unintelligible). I have a quick question to the developer. You state that the measures restricted to hospitals was at least 100 admissions in that division. And while it doesn't specifically concerned our scientific methods panel I think, but it will be more in the math world. Are you concerned about reportability and excluding smaller or rural hospitals in the application of this measure?

((Crosstalk))

Woman: Yes, definitely. Let me just pull it. I know it's here. So the restriction or the exclusion is limited to patients with or divisions, CCSs within divisions was less than 100 - hold on a second, so this is principle diagnosis within a CCS. So those - that's the grouper that we use was fewer than 100 admissions within the measurement year. And we don't think it's going to have a big impact. I don't know if anyone in the development team here for (unintelligible) has actually the figure for the number that were actually excluded?

I think actually, you know, in our form we have that. We have the total number and it's very, very small. It's like less than 1%, 0.7% or something like that as I recall correctly. I can pull it up. So based on that level of exclusion, we think that it's not going to have a significant impact on the measure or on the hospitals within - that have smaller numbers of patients that this particular exclusion isn't going to be the issue for smaller hospitals.

Eugene Nuccio: Great. And are you also applying this to Kaiser institutions only or are you suggesting that it could be used more widely?

Woman: No, that's an exclusion that's across both measures. The two measures are almost identical except for the risk adjustment that as the clinical data elements. So otherwise, the exclusion and inclusion criteria are all the same, so we apply this across the board.

Eugene Nuccio: Okay.

(Karen Dorsey): Right. And it's an exclusion of - that affects more patients included, but not the hospital's ability to receive a measure score, so that's the important thing. Very few hospitals with all of our exclusions that are identical for the claim's measure and the hybrid measure, very few hospitals fails against score even with the exclusion that (Doris) just explained in some detail.

Eugene Nuccio: Okay, thank you.

(Michael Abrams): Very good. So any other comments from or questions from the committee or can we have you move to a vote or revote on validity? Anything you want to add there, (Andrew), maybe I'll hand it back to you.

Matt Austin: So this is Matt Austin. Can I ask one more question, because I just want to I guess sort of make sure I vote in the way that is appropriate? Around the missing lab results for surgical patients and I don't know if I'm asking the right question here. But is there sort of a way to understand if you were to have imputed those values for patients for which you actually did have data, how close or how much do the scores change?

(Karen Dorsey): So I think - let me just say again that the way that we think about missing data with data that's extracted from the EHR is that almost no matter what kind of data element you're talking about and here we're talking about highly available data elements on the adult hospitalized population, but, you know, the reality is that there is always some missingness and there are a lot of ways to address missingness. And there are different impacts of different ways. This is the way that we addressed missingness for testing.

And I think (Doris) explained the rationale for that was based on the fact that the patient population for whom those values were missing were largely patients were getting elective surgery. And so that's the reason why for testing purposes based on the clinical rationale that was vetted with our experts, we basically put in normal values for those missing values.

You guys know, because you're experts. There is many, many ways that we could address that and perform sensitivity analysis for example for a preparation for implementation of such a measure, right. And I think there is a lot of room for stakeholder feedback that tends to happen after development and even through sort of the (unkept) process and public comment process. There are lot of different ways to deal with that, but we rested our initial decision in testing based on what we heard from our clinical experts as the most clinically rationale way to approach the missingness as it was specific to

that population for whom we did not have lab test which were the surgical population.

Certainly for implementation and as we get feedback on this measure, there is an opportunity to look at different mechanisms for doing this and pressure test down a little bit in terms of looking at the differences and score. I'll say that there is really no statistical rationale to think that we're going to see very, many differences. We can make much bigger changes in our risk model and see very insignificant bumps, right in performance and in hospital level scores. These are - this is a really sort of we were playing around the margins here when we talk about whether we're imputing this value or that value.

So I just want to make sure you guys sort of understand the full context of how we came for that decision and what we would recognize are the limitations of what we put forth, you know, for testing vis-à-vis the sort of full lifecycle of the measures. And these are issues that appropriately come back to NQF and endorsement maintenance, the more we learn about how to approach these things.

Matt Austin: Thank you so much. So I guess I have a question for the NQF staff which is when we endorse a measure, are we endorsing it with the current imputation approach or are we endorsing that with that the measure developer will explore different imputation approaches and we'll figure out which is the best one to use?

(Karen Dorsey): Okay, just hang on.

(Michael Abrams): Yes, so (Mike Abrams) here. You're endorsing it as specified which in this case will be listed as imputation approach in the hybrid measure.

Matt Austin: Okay, thank you.

(Karen Dorsey): Right. But just a point of clarification as the developer, this committee is not actually making an endorsement decision, right, only a decision about whether…

(Michael Abrams): That's correct.

(Karen Dorsey): …right, a decision about whether we met the criteria for feasibility such as they can go in front of the…

((Crosstalk))

(Michael Abrams): You're correct. This is a message panel, so they're approving the message for then to port into a committee that will decide on endorsement or not, that's correct.

Matt Austin: Decide on endorsement over the measure as specified?

(Michael Abrams): Yes. That is correct.

Man: Are we comfortable holding a vote at this point?

(Michael Abrams): I think so.

Man: All right. Hearing no objection, go ahead and enter your votes into the SurveyMonkey and we will move on to the next measure which is 3503, Hospital Harm - Severe Hypoglycemia. This actually - this has a new measure number, but it is actually a maintenance measure. We're trying to work out on the back end how to - we sort of establish that previous number.

This is an e-measure, an electronic clinical quality measure. It assess the proportion of inpatient admissions for patients 18 and older who received at least one anti-hyperglycemic medication during their hospitalization and who suffered a hypoglycemic event within 24 hours of the administration of an anti-hyperglycemic agent.

(Michael Soto): Can I just - on that last one, 3502, it was only for validity we were voting, is that correct?

Man: I believe so, yes.

(Michael Abrams): Yes, that's correct.

(Michael Soto): Thank you.

Man: So for 3503, we did get consensus not reached on both reliability and validity. So we'll need to vote on both of those. This I think it has pretty similar issues to when we just talked about a couple of measures ago, I think it was 3501. We got a score level reliability testing which, you know, came out with pretty good results, median reliability score of 0.89 roughly. Although there was again some concern about the sample size there, for validity similarly again we had data element testing and what the developer had called score level empirical testing and there is some question again whether that might better be classified as data element testing.

Maybe we could start with reliability. Again, some concern that only six hospitals were included which some thought might be an institution number to compute reliability. Any discussion about that? Give me a moment to pull up.

(Michael Soto): So this is (Michael Soto). I think that might have to do with the issue of how reliability is assessed empirically. So for instance, if you were doing kind of a retest, I think that six will not be anywhere near enough. But the beta by no meal really is looking at whether or not the differences among the units that you have are essentially different compared to the noise. So I think that the number of units is far less important for this beta by no meal approach if I understand that correctly. I wonder if those are great.

Man: Thanks Michael. Any other comments on that and do you agree with Michael about that?

Matt Austin: So this is Matt. I don't really have enough experience to agree or disagree.

(Michael Soto): You know, this is what the subgroup is writing in the paper about reliability assessment message and so on I think we'll be addressing, but haven't done yet, of course.

Man: With that said, we're comfortable taking a revote on reliability. Anybody have any additional thoughts or comments or questions for the developer?

I'm hearing none. Let's move on to validity. Again, here is some concern about the, you know, sample size used in testing and whether that was sufficient to determine validity. The - some concern about the developer's rationale for not risk adjusting and whether there might be a need for additional data to support this decision. Again, you have your sort of question about whether you would consider this to be score level testing or data element level validity testing which we'd only sort of implicate whether this is eligible for a moderate or a high, not a pass decision. Any…

Matt Austin: This is Matt.

Man: Yes, go ahead Matt.

Matt Austin: I don't know if this is the right place to provide this feedback. But is there somewhere documented - and nothing I saw - I guess it's not until this afternoon that I'm understanding that this is actually agent's measure.

Man: Yes. Apologies for that and sort of…

Matt Austin: Is that just we're only just a little bit as I know I'm trying to go back through my notes and figure out. It looks like they did do empirical validity testing of the score and the data elements. So I'm just trying to go back through that information.

Man: Again I think it's kind of similar to one we discussed a couple of measures ago where they did look at, you know, sensitivity specificity and PPV, NPV around that sort of core question of the adverse event and there is maybe some debate over whether that constitutes score level or data element level validity.

I think again at NQF, we would typically expect score level validity to be an assessment of that rolled up score at the level of that facility or whatever entity is being assessed.

Matt Austin: All right, thank you.

Eugene Nuccio: This is (Jamie). I'm a little concerned about the overall purpose of the measure. In their validity section they talk about testing six hospitals and there were - and there are two data tests. There were about 175 counters meaning that each of these six hospitals only had about 30 encounters each

over a period of I presume 12 months. I mean I understand these are clearly kind of low prevalence events. They certainly could be serious.

I'm wondering if these are more process measures than outcome measures that is we have a case of, you know, these six hospitals having 30 events and about a little more than half the time they're having events with harm and about half the time they're having events without harm.

Again, maybe I'm just thinking of - as also a MAP member, is this something that we should be concerned about measuring or is it, you know, just kind of a low incident kind of event that, you know, we understand the issue, but we're trying to capture something that that's a rare event.

Matt Austin: This is Matt Austin. That's an interesting thought, because as I reread this denominator and numerator statement, the denominator are basically patients who received an anti-hyperglycemic medication under in their hospital stay and the numerator are those who had a test for like glucose less than 40. So that's to me - well, to me that's not necessarily a process, because I'm not sure what the process is for measuring. To me the outcomes that's being measured is for patients who received a high anti-hyperglycemic agent to their blood glucose fall.

(Michael Soto): Yes, this is (Michael Soto). I agree that seems like an outcome to me too that the glucose level was low.

Eugene Nuccio: It did what it's supposed to do?

(Michael Soto): Right. You would only - you know, of course you will know that if you did the test, but of course that's from many outcomes.

(Lisa Tudor): This is (Lisa Tudor) from the measure developer. Would it be helpful to have some clarification? I don't want to speak if not helpful.

Man: Sure. I think that will do.

(Lisa Tudor): So certainly the capturing of hyperglycemia and this measure is actually paired with a hyperglycemia measure that is not in front of NQF at this time. That is dependent on the hospital capturing glucose. So there is a fundamental need to capture that although I would pass it that glycemic testing is probably the most commonly performed laboratory testing in hospitalized patients. And there is I think pretty universal testing and we focus on a denominator of patients that are at risk for these kinds of events and therefore they have diabetes or they are on diabetic hypoglycemic agents.

So we are focused on a population that is theoretically at risk. They are - you know, there is a possibility that there are patients that may have severe hypoglycemia that had no glucose level checked and we are missing those. Our clinical experts and technical expert panel felt that as well as some additional endocrinology experts felt that this was capturing the critical outcome of hypoglycemic events, severe hypoglycemic event (unintelligible) is much lower threshold for severe hypoglycemia.

But I think, you know, while there may be an extraordinarily rare case that is missed because the hospital is not monitoring a patient on hypoglycemic agents appropriately or clinicians felt that universally that this would not be a reason not to endorse the measure for face validity as we do have face validity although I know that's not adequate for maintenance face validity to add to whether we need to find it as a measure score or a data element of validity.

Man: Any comments from our panel members or discussion? Are we comfortable going ahead with a vote on this?

Matt Austin: I think so.

Man: All right. Let's go ahead and revote on reliability and validity? And, you know, one more measure I think we can at least start it, maybe we can get through. I'm sorry, we have more than one measure.

So the next one is hospital 3498E, Hospital Harm - Pressure Injury. We have a high rating for liability on this one. But a consensus is not reached rating for validity. Again, very similar analysis done here for validity and again depending on whether you request by what they did at score level or data element level maybe eligible for a high or moderate for validity.

But for this one, there was some concern among reviewers about weak validity resulting one of the tested datasets and reviewers suggested that inconsistent use of structured field in EHR does raise concern and it's about data quality and documentation practices, also some concern with the decision not to risk adjust or stratify reported results.

Any - I don't know if you've finished with your voting on the last one. Again, just to remind you, this is 3498E. This is an electronic clinical quality measure which is a proportion of inpatient admissions for patients 18 and older. We develop a new page 2, 3 or 4 pressure injury, deep pressure - deep tissue pressure injury or unstageable pressure injury during hospitalization.

Any discussion about the validity results here?

Matt Austin: Yes, this is Matt. I did vote low, because I did have some concerns with the inconsistent use of structured fields and identification of hospital acquired pressure ulcers. So back to my notes to exactly I couldn't find any more detail, but that was my biggest concern was if hospitals aren't actually using structured fields to document pressure ulcers, are we getting a false signal of what the problem is? And I think it's (unintelligible) five that brought my attention to it where they compare the electronic EHR extraction in the manual chart of distraction and had a PPV of 68.4% in one dataset and 44.9% in the other.

Man: Any other thoughts from our other panel members about that or I don't know if our developers have any comment on that, those low results in that one dataset?

Woman: This is (Nicole) from the measure development side from the (LCOR). I just wanted to get quick clarification that while you write the PPV is lower at some of that test rates, the hospitals were actually using the structured fields to capture pressure injuries. They just weren't doing it consistently within the first 24 hours which is why a lot of them show up as new pressure injuries in the EHR and the ones who validate it, it turns out they were actually there, but they were down in unstructured field. So eventually they were being documented in structured fields. It just wasn't within the first 24 hours with (unintelligible) capabilities.

Matt Austin: From my understanding correctly that they will be miscategorized then right by your measure, because it would show them as being hospital acquired when they weren't seen hospital acquired?

Woman: Correct.

Matt Austin: Okay. So we would potentially be overstating a hospital's performance in that area.

Woman: Correct, until the documentation fraction is changed.

((Crosstalk))

(Karen Dorsey): I just want to add that we actually discussed this pretty extensively with our technical expert panel and actually it was a discussion that came up at the MAP as well. So I think you are characterizing it absolutely correctly that if there is not adequate documentation that an injury was present on admission or present in those first 24 hours that a hospital that fail to document appropriately would be charged with a harm in that case.

But our technical expert panel and some members of the MAP Committee, you know, pretty consistently the feedback that we've gotten is that that kind of problematic documentation is legitimately part of the quality signal meaning that if hospitals are not accurately documenting these kinds of fundamental aspects of the presence of pressure injuries, they can't adequately track their progression and they can't communicate shift-to-shift among the nursing staff appropriately and that's being penalized for that kind of sloppy documentation is appropriate. And so that was explicitly discussed when the technical expert panel provided their vote on the face validity of this measure which they endorsed.

Matt Austin: Right. I guess to me - this is Matt Austin, it just becomes down to. I would argue that there is obviously quality issue that this measure is addressing, it's just - it doesn't really help identify the quality issues documentation or if the quality issue is management of patient's skin.

(Karen Dorsey): Right. And it's really both initially, right, because the documentation does come up as an indicated harm. And like most of the quality measures that we've developed for CMS, if we are providing hospitals with sort of valid information about quality and in this case whether that's, you know, purely about the acquisition of pressure injury or a mix of information about the acquisition of pressure injury and problems that they're having with an appropriate documentation that gets reflected in their rate that all of that is valuable and actionable information for the hospitals to - that they need and will benefit from in terms of focusing efforts on improving quality in that area. And I say that's what we heard from our technical experts.

I will say that it is true that, you know, a complete failure to document is an approach to game. I think that that's really an implementation issue for CMS to consider and address. So I would put that in the implementation bucket and something that certainly when the measures come back for maintenance and endorsement and the committee consider some of these implementation issues should be a matter for discussion.

But in the short term, you know, we think that that rate reflects important quality information for a hospital to have even though that's a mix of information about the quality if their documentation and the quality of skin care.

Matt Austin: I think that can be summarized by the method drives improvement, measure I mean.

Man: Any other discussion on validity or reliability? I think we have to take a vote on both.

Matt Austin: Turning to the notes that says that reliability is high, because there were three highs, one moderate and one low.

Man: I get that wrong.

((Crosstalk))

Matt Austin: 20, okay, it's just…

Man: Yes, page 20. Sorry, yes, you're right, sorry about that, so just validity. Have we had enough discussion that you feel comfortable to revoting on validity? Hearing no objections, I will ask you to go ahead and enter your vote, your rating on validity in the SurveyMonkey and I think that will probably wrap up our call for today. We're just a little bit past the hour.

As (Miranda) mentioned earlier, we have a follow-up call on Thursday, 3:00 pm Eastern Time. So we do have one more measure to address and that's hopefully will be a fairly brief call, but we'd ask you to attend that call as well. And if there is nothing else from any of our panel members, we'll let you go. Any final comments or thoughts?

Matt Austin: None.

Man: All right, thank you.

END