SCIENTIFIC METHODS PANEL SUBGROUP 4

Moderator: Sheila Crawford
March 21, 2019
11:45 am CT

Woman1: Hi everyone. This is (Unintelligible) from NQF. We'll be starting shortly. Just confirming, (Andrew), are you on the phone, (Andrew) from NQF?

(Andrew): Yes, I am. I'm right here.

Woman1: Okay, thank you, (Andrew). (Andrew), did you want to go ahead and start or give it a couple more minutes in case people are still joining?

(Andrew): Yes, maybe let's give it – wait a few more minutes.

Woman1: Okay, thank you so much. So, we'll be starting shortly and please bear with us in meantime. Thank you, everyone.

(Andrew): I think we could probably start it.

Woman1: Okay, did you want me to ahead and do the intro and then the roll call and then I'll give it to you?

(Andrew): Sure.

Woman1: So, thank you, everyone, for joining the call and thank you for those who joined early and your patience as we get today's call started. This is the follow up call to subgroup number 4.

The first call was on Tuesday. We were able to get through the majority of the measures that needed to be discussed. We had one more measure, 3516, that still needs to be discussed and that will (unintelligible).

Just a reminder to our method panel members, three measures did pass review and they were not discussed on the last call and this is (unintelligible) unless – however, if you would like to pull them up to review more details, we can.

With that said, we did send you a survey link on Tuesday to the link to the Survey Monkey which you should be using to cast your votes.

Please let us know if you don't have it, but if you can fill that out as we go through, that's very useful. Staff will prompt you to cast your votes when timing is appropriate.

We only have one measure to discuss today but, you know, we do want to be conscious of everybody's time. So, let's try to be concise and hopefully we can get through this measure effectively.

And then I just wanted to remind everyone that this call is a public call. There will be developers on the line to answer any questions, some staff and panel members.

However, there is no opportunity for public comment. For recordkeeping, this call is being recorded. We ask that you please state your name every time you make a remark just so it's easier for us to keep track of that. So, with that, I'll do quick roll call. Is (Larry Vance) on the phone?

Man: Who was that again?

Woman1: I'm sorry. (Larry), is that you?

Man: (Unintelligible).

Woman1: Thank you. Sorry, I'll try to speak a little louder. (Lacy Fabian)?

(Lacy Fabian): Here. Thank you.

Woman1: (Glen) (Unintelligible)? (Matt Austin)?

(Matt Austin): Here.

Woman1: (Mike Soto).

(Mike Soto): I'm here too.

Woman1: Perfect. Thank you. With that, I'll give it to (Andrew).

(Andrew): Okay, thanks, (Unintelligible). I heard a couple more beeps here. Like, I'm just going to call out one more time. Did we have (Larry Glance) on? How about (Dean Nuzio)? All right, well, we can go ahead and talk about the measure.

This is – so, we're going to – last measure in our set. It is number 3516, percent of patients or residents experiencing one or more falls with major injuries.

It's a bit of an unusual (unintelligible) in some ways. It applies to three separate care settings, sort of three different measures. It's also a little unusual in that it is measuring very low frequency events which sort of changes some of the expectations to some degree for some of these testing results.

I know that some of our reviewers have some concerns about the low reliability of results. And maybe we can start there.

If anybody wants to talk a little bit about what their concerns are, I was thinking since we have a little extra time on this call - we don't want to take too much time, but maybe we could have our developers, if they're on the line, kind of talk a little bit about their rationale for this measure and why they got the low liability results that they got and how they interpret those results.

So, maybe we should start with our reviewers first and see if you have any comments or thoughts about reliability of this measure.

(Matt Austin): So, this is (Matt Austin). I'll speak up first. I mean, I'd look to others to help educate me. I guess I just was concerned with the appropriateness of using signals of noise where (events). But it's more maybe a point of ignorance than necessarily thinking it's incorrect.

(Mike): This is (Mike). I'm not so sure what it's incorrect. I guess I think that because the numbers are so small, they really are – this really is a lot of noise.

So the signal – it's really hard to find a signal. And the fact they said, well, we expect it to be low and, therefore, it is low, doesn't make – give me a lot of confidence in the reliability of the measure.

I'm not sure what method might be better but I think that what they're really finding, as you might expect, because the (unintelligible) is low, that they really have a lot of – even a probably with reliability.

Man: But maybe we could – with that, do we have our developers online?

Woman1: Yes.

((Crosstalk))

(Molly Vaughn): This is (Molly Vaughn).

Man: …talk a little bit about (unintelligible) results, why you used the method you did.

(Molly Vaughn): Right. Yes. So, first off, yes, you're correct, this is a measure that we would expect there to be a low incident of falls with major injury because it is a never event and does not occur frequently.

I mean, we would hope it does not occur frequently in a setting. So, when you're interpreting the lower liability, the (signal) to noise ratio is the measure of how well, you know, how well a measure can detect differences across facilities.

So, for quality measures in long term care in (unintelligible) care settings, we're – you typically see values that are .1 or lower.

And as one of the reviewers mentioned on here, we did expect a low amount of cross-provider variation because this measure is a never event as in it is largely preventable and should never occur.

That being said, we feel that measuring falls with major injury is very important from a quality of care perspective across post-acute care settings as it is associated with very poor health outcomes and poor care across facilities, so.

(Mike): Well, this is (Mike) again. I mean, I certainly understand the desire to have a measure of this. But we have to look at the reliability as it is.

And I think that, as you would expect, because the numbers are so small, this kind of measure inherently doesn't have much reliability. As much as we'd like to have a measure of it, it just doesn't get me over the (unintelligible) hump.

Woman: I guess – so for – because better quality for this measure is associated with a lower score, and you know, you wouldn't expect falls with major injury to occur at a high rate because it's, you know, I mean, we're looking at, you know, comas and a concussion and very, very poor – I mean, very, very bad falls here.

So, I guess, from (unintelligible) perspective, what kind of reliability would you expect for a measure that is measuring such low incidence?

Man: I mean, the problem is that most of these units will have zero or one event in most periods. And, you know, so they're – and the measure basically just has a lot of randomness in it simply because of that.

I don't see how there's any way that you can deal with such small numbers and come up with a reliable measure that's going to be able to separate the units that are doing better than the others because there's just simply so much noise inherently with such small counts.

Woman: I'm going to turn it over – go ahead.

(Dan Barts): Hi. This is (Dan Barts) with RTI. So, the observations are non-random. They're actually – the problem with the reliability analyses that we have is that they're based on an entirely different field of study where, you know, you would – it comes from the field of psychometrics where you would design a test so that you don't have ceiling or floor effects.

This is quite different because we really want the floor effect. Like, as a society, we want people not to fall with major injury. So, like, a floor effect is actually quite desirable.

So, and when we have, you know, such large – so, I mean, if our measure mean across (unintelligible) is usually about like half percent which means, like, for every one incident, you have 199 non-incidents.

So, actually if you think of it that way, the data are quite consistent. It's just that the measures that we have were developed for things like IQ tests, which are, you know, just a totally different thing.

So, I would, you know, argue against the idea that there are – that the data are randomly scattered or that there's a lot of noise.

(Mike): Well, this is (Mike) again. I don't want to keep on repeating myself, but I think that, whether you think of it as the top or the bottom, the fact is that it's just distinguishes one unit from the next is whether or not they had a fall.

And that has a lot more randomness than it has signal inherently because we're looking at these rare events.

(Dan Barts): That's – no, I understand where you're coming from, but we've also demonstrated that there is a significant effect of facility. You know, we can run (a nova) test and reject the hypothesis if there's no effect of facility membership.

So, there are significant differences between facilities. (The effect) size is not large, but of course, we wouldn't expect it to be.

(Mike): Yes, well – yes, I suppose I could see how that would be true. But the real thing is, if you report this measure, you know, and one of the nursing homes, one of the SNFs happens to have a fall one day, one of the measuring periods, it's going to look like it has a bad outcome.

And then next year or the next period, it could have zero and there's really nothing changed in the quality. It was not this rare event happened. And that's just simply not a reliable measure.

((Crosstalk))

Man: That's understandable but we also have a sample size restriction which means that the proportions are more stable. You know, so if we have a SNF with, you know, one person in and then that person happens to fall, you know, they would have the score of 100%. That doesn't happen. We have a minimal sample size…

((Crosstalk))

Man: …the issue is the size of the numerator.

Man: Yes. Well, we're not talking about falls that can be random. We're talking about falls with major injury which is (unintelligible).

Man: Okay. Well, again, I don't have anything more to say. I don't think that this analysis demonstrates reliability and I don't think that – I wouldn't have expected it to because I think that – because it's based on such small numbers, it just can't have it.

Man: Okay.

(Matt Austin): So, (Mike), this is (Matt Austin). I'm just asking a question. I mean, would sort of consistency over time be a stronger way of looking at reliability?

(Mike): Well, I think that – if they found consistency over time, that would be evidence of it, but I think that they won't. We don't know. Of course, they didn't try.

(Matt Austin): Right. But I would think that might demonstrate a stronger quality (signal). If I'm a facility that has had a fall every year, (in five) years, that would seem to indicate that there really is a quality problem as opposed to the (unintelligible).

((Crosstalk))

(Mike): Right. For sure, if they found that that – I would regard that as evidence, but that's not what they did and I guess they don't expect to find that simply because of the nature of the small numbers of events.

Man: But is it not still an important thing to measure?

Man: Not if you're measuring noise. I mean, this could put some of these (unintelligible) in a lot of trouble just because they had bad luck in one period. They had one fall (down) period.

For something – whether it's important to measure it and whether it's reliable are two different things. You need both and it's not our job to assess the measurement.

I mean, (it is) important on this committee. Our job at this moment is assess the reliability. I simply don't think that it's been demonstrated.

Man: Any other thoughts from our committee members, or viewers, rather? (Lacey), any comments from your end?

(Lacey): Sorry, this is the measure that I recused myself on, but…

((Crosstalk))

Man: Ah, I see. I see. Sorry.

(Lacey): No worries.

Man: All right, well, it did receive a low or insufficient rating in (this back here). So, we actually don’t necessarily need to vote on that. I thought it would be worth just bringing up again to kind of talk through the issues, see if anybody potentially had a change of mind.

Maybe we could still do a revote on it just to ensure that we’ve got a reliable vote and see what the result is there. And I think since we have a little, you know, extra time, we can go ahead and talk about validity as well.

We had a consensus not reached on validity that the (unintelligible) did a test at the score level which included a confidence interval analysis and variation by discharge that there was, again, some concern about whether these were appropriate methods to test validity, whether (unintelligible). Do we have any thoughts or discussion on that?

(Matt): This is (Matt). I actually personally voted moderate. I actually – well, the confidence interval analysis didn’t quite pan out the way they were hoping, I though their analysis, looking at the discharge destinations did align with their hypothesis.

And so, a falls with a serious injury are more likely to be discharged to a hospital.

Man: Any other thoughts on that?

Man: I struggled with that point when I was reading it the first time, but I do feel a little bit better about it now. But I guess I’m – so is this – we’re talking about the – basically the relationship between the rate of falls and what happened to them after they get out of this facility?

So, for instance, I mean, what I'd written in my notes here is that residents who had a fall that resulted in a major injury were discharged to acute (segs) at a higher rate than residents who experienced a fall that resulted in a less severe injury.

Well, yes, that's what you would expect, that if they had a fall with injury, they would more likely to go a hospital to take care of it, but I'm not sure why that is a measure of validity as opposed to…

Man: I guess maybe it's assessing whether the fall did lead – did, indeed, lead to a – what was a major injury. I don't know. Maybe the developer can…

Man: Yes, I guess, maybe – I'll say one more thing. I guess I think there may be a confusion here between what happens to the individuals and a measure of quality of the setting.

Man: (Mike), this (unintelligible) clarifying. You're saying that perhaps the validity was looking at an individual resident level as opposed to a validity level.

((Crosstalk))

(Mike): Right. Suppose that you had people from (Nips) that were all exactly the same with respect to their quality. But then you looked at the residents who had a major injury and you found out they were more likely to go to a hospital.

Okay, yes, that's because they had a major injury. It doesn't tell me anything about the difference from quality between the SNFs.

(Molly Vaughn): This is the measure developer, (Molly Vaughn). Is it okay if I jump in or?

Man: Yes, please do.

(Molly Vaughn): Sure. So, this analysis is a variation by discharge estimation. This is looking at the item level validity of this measure. So, we wanted to make sure that the item that is measuring falls with major injury is, you know, is in fact, you know, if someone indicates that they had a fall with major injury there, (in fact), these same patients discharged to acute care settings or having worse healthcare outcomes.

So, I mean, it's showing the construct validity basically of the measure. So hospitalizations are associated with more adverse health outcomes and higher cost in healthcare utilization.

So, that's kind of just to give you some insight into why we were looking at that in addition to the confidence in our interval analysis that we (unintelligible). So it's just one piece of validity testing.

Man: Maybe like some of the other measures we viewed in the last call, it might be better kind as data element and validity testing which, again, does meet our standards.

Man: Yes, that makes a little more sense to me.

(Matt): However, I guess – this is (Matt) – it did indicate that a performance measure score testing, and I think maybe the concern is the measure specified at the facility level of, you know, and the state level.

And I guess that's where I was getting a little stuck, too, is did they need to demonstrate validity testing for both levels because the testing they provided feels more stay level than facility level.

Man: Well, again, you know, it might be better to describe – again, when it looks like we're looking at the, you know, something more like data element validity testing.

I mean, I don't know if we want to proclaim that, but that's kind of how we were looking at some of the earlier measures. Maybe just sort of a difference in terminology here.

They don't actually need to provide score level testing if they have provided data elements, so testing, you know, did you consider this to be, you know, some reasonable form of testing validity at, you know, one of those levels, you know, whatever they happen to call it?

You can still give it a moderate rating. It would only be eligible for a moderate rating if we considered data element validity testing, whereas, if it were score level, we could give it a high.

It sounds like (something we were) likely to do anyway. But, again, at the patient level, I don't know if it makes a difference whether you're rolling up at the state or the facility level, if you're getting element validity.

((Crosstalk))

Man: Oh, sorry, go ahead.

(Molly Vaughn): Oh, no, I…

Man: Go ahead. I'm sorry.

(Molly Vaughn): No, no, that's okay. This is (Molly Vaughn) again. I just wanted to clarify. The confidence interval of validity testing that we ran was at the facility level. So, you know, that score level, we're looking at, you know, variability in the score. So, I mean, we have both in our validity testing.

(Mike): Right. Yes, that's true. Thank you for that clarification. So, on the data element level, if we think about this second test as being a data element level, I mean, the fact that we – for the SNF setting, you know, 62% versus rates in the 20% for those with major injury versus less severe/no injury, I guess that doesn't tell me very much.

I think that you could be pretty far off on the data elements and still find the results like that. So, for instance, if somebody had a major injury, only less than 2/3 of them are going to an acute care setting?

You know, how major is that injury if they don't have to go someplace else to get it treated? So, I guess, it doesn't give me a lot of confidence even at the data element level.

(Matt): So, (Mike), this is (Matt). If I heard you correctly, your concern is that, like to the SNFs only 61% of patients who had a – or residents who had a fall resulting in major injury actually (unintelligible). You would expect it to be higher if it truly was a major injury.

(Mike): Yes. I mean, if the developer wanted to offer evidence that this is really picking up on major injuries, I guess I would've expected that to be higher.

You know, maybe it's okay. Maybe they can be treated where they were, but you know, if we were doing a comparison between what the medical record says and what the gold standard says, we'd be looking at lots of – we'd be looking for a lot higher rate of agreement than 61%.

Man: Right.

Man: Yes, that's sort of what (unintelligible), what is the gold standard, right? Data (stores) have been used for the measures of gold standard (unintelligible) gold standard.

(Dan): Hi. This is (Dan) again from the developer. So, with the data available to us, we didn't have what you would call, you know, necessarily a gold standard. But the intent of the analysis was to show that the different levels of the variable we're using predict different outcomes.

And, I mean, given that people are already in a care facility, I'm not (a clinician). I'm a statistician, but I would imagine that, you know, that for a lot of major injuries, maybe that, you know, the facility can provide care for people, you know, with certain injuries, you know, a broken hip, a subdermal hematoma.

But the importance was to show that the item is predicting different outcomes at the different levels. And, therefore, (validating) the item that this measure is based upon.

Man: So, I understand that, but what we don't have is a benchmark that would say, you know, how much higher you would expect it to be in the acute care – discharge at the acute care setting than the other ones.

I'm sorry. We don't have a benchmark to say how much higher you would expect it to be comparing the major injury to the other two categories. So, I don't know what to expect.

And so, when you talk about predicting, you know, that sounds to me more like a measure of convergence validity or predictive validity rather than data element validity.

So, you know, I think that to accept this, I would have to have a more well-developed presentation to make it clear exactly what's being done for what purpose and why these numbers are missing.

(Karen): This is (Karen) from the QS. This has been a really great discussion. I (unintelligible) criteria and I want to probably underscore what (Mike) just – validity testing demonstrates that the pure data elements are correct.

So, when we're thinking about data elements validation, that's what we're looking for – are the data elements, correct?

We go on to say that typically what we're looking at or looking for is (unintelligible) and another authoritative (unintelligible).

So, I think the question in front of you with this analysis is to (unintelligible) you feel like – feel confident that the data elements are (unintelligible) would be (unintelligible) against (unintelligible) and then go back against the medical records.

(You see) this kind of patient-level construct validation thing. (We feel) like that is (unintelligible) the data elements (unintelligible) measure.

Man: Thanks, (Karen). Any more comments or thoughts on this or questions for the developer? Hearing none, I think we can, I guess, go ahead and take another vote.

That is, I believe, everything we had to cover. I'm trying to think if there's anything else we should discuss here. I think that's it. Any final comments from our reviewers?

(Matt Austin): Nothing from me. Thank you very much. This is (Matt Austin).

Man: All set, too.

Man: All right, then we'll go ahead and ask you to – for your updated votes. Thank you for taking the time to participate. Thank you to the developers.

We'll reach out back to you with the results and some thoughts and feedback as well. So, again, thanks to everybody and we will be in touch again shortly.

Woman: Thank you.

END