

National Quality Forum

Moderator: Helen Burstin
September 21, 2015
1:00 p.m. ET

OPERATOR: This is Conference #: 28280920

Operator: Welcome, everyone. The webcast is about to begin. Please note today's call is being recorded. Please standby.

(Off-mike)

Helen Burstin: Hi, everybody. It's Helen Burstin, Chief Scientific Officer here at NQF. Welcome, and thank you for joining us today in our Members Only Town Hall to talk about an important new report we've put out for your comment. As you know, this report is still open for comment but we thought it might be helpful to give you some context and overview of the recommendations and engage in a dialogue with our members today as we work through this somewhat complex issue.

So, with that, I'm joined on the call today by Taroon Amin, who's our consultant to NQF, long engaged with us on some or more complex work like readmissions and cost (research) use. Taroon will kick off the discussion and then I will walk you through the themes and recommendations.

So with that, I'll turn it over to Taroon.

Taroon Amin: Thank you, Helen.

So, we're going to walk through a quick agenda here. We're going to walk you through some of the background, the panel's charge, and some of the key

themes and recommendations that emerged from the advisory panel recommendation.

I'd like to first begin by introducing the panel members and a esteemed group of advisors that help NQF walk through the panel charge, this included Helen Darling, the chair of the NQF Board of Directors, Lee Fleisher, our CSAC, Consensus Standards Approval Committee chair, and also Cristie Travis, our immediate past chair of the NQF CSAC, along with a group of measure developers and measure scientists, measure performance scientists, along with users (inaudible).

So with that, I'd actually welcome any introductory comments from any of the panel members particularly Lee, if you'd like to just say anything just to introduce yourselves and any opening comments that you have for the town hall.

Lee Fleisher: Sure, this is Lee Fleisher, I'm an anesthesiologist at Penn and chair of the CSAC. And greatly appreciate NQF's taking up this issue which has – excuse me, CSAC has been wrestling with for a while and the panel with – being led by Helen and Taroon at NQF and Helen Darling and Cristie and others were quite thoughtful.

As Taroon will mention, the charge really was to develop the framework as well as the potential benefits as well as unintended consequences of going in this direction of looking at different ways of intended use and how we might change the approval process. Clearly, the CSAC is ready once this has been considered by the membership and CSAC report to work with NQF staff to how to implement this and should that be the decision, and how to go forward with the different levels of approval. And then, of course, the committee framed, I think, importantly, the – how different levels may map to how they could be used for accountability and I think the report nicely outlines that. But really left that to the next step of – and where – who should be making those decisions really is sort of the phase two. This really – the committee focused on whether it's needed and what are the potential benefits of going in this direction.

With that, I'll turn it back and listen.

Taroon Amin: Greatly. Thank you very much.

I just want to confirm that if there's anyone else who wanted to make any introductory comments in the advisory panel who's on the call, we'd certainly welcome that.

And we'd certainly welcome that conversation as we get to some of the Q&A at the end of this call.

So, thank you, Lee.

As Lee pointed out at the – in some of his comments, NQF continues to strive to improve its processes to reflect the changing needs of the measurement system. Particularly over the last several years, stakeholder groups have questions whether the endorsement process should consider the specific intended or actual use of a measure or more broadly whether the specific use of a measure should be included in the measure endorsement criteria.

And so, very simply, whether NQF should move away from a binary yes or no, endorsement decision to something a little bit more nuance.

This effort by the NQF Intended Use Advisory Panel seeks to consider the merit and the various approaches of considering the measures intended or actual use as part of the measure endorsement process. So, as Lee pointed out, this was intended to develop a sort of a framework or an initial proposal about how a measure's intended use should be considered in the endorsement process. And the specifics of how the, you know, specific (ways), if you will, will be determined if they'll, you know, talking to these, CSAC will continue to iron out and, you know, based on the approval of this actual recommendation.

So moving to the next slide, this effort came – was primarily came out of efforts and all the way back to 2012, where the NQF Board of Directors impaneled what was considered the Consensus Task Force, that recommended a few directions for the advisory panel, and the two potential directions of the

Consensus Task Force asked the advisory panel to consider, was whether endorsement of measures should be for specific or actual use, so, for instance, whether endorsement for particular purposes, for example, internal quality improvement, public reporting or payment. With the consideration that measures that are useful for one of these applications may not be suitable for all of those potential uses.

The second potential direction that was considered by the Consensus Task Force was moving forward with levels and endorsement by using measure grading. The levels of endorsement would allow NQF to apply the same criteria to all measures, but to grade endorsed measures differently and makes the ability of how well a measure met the NQF endorsement criteria transparency users.

So, in the next slide, we – I just want to – we wanted to walk you through what the Intended Use Advisory Panel's charge was by the NQF board. The charge was to consider several critical topic areas, including identifying the various used cases for NQF-endorsed measures. Secondly, really discussing and distinguishing among the various used cases and identifying the need, if any, if different attributes would be required, depending on the different intended use of the measures.

Secondly, to examine whether the NQF endorsement criteria requires updating and internally, to propose a path forward on whether and if so to incorporate the specific use of measures and the NQF endorsement process.

And so, with that, I'll turn it over to our chief scientific officer, Dr. Helen Burstin, to walk us through the key themes that emerged from our deliberations and the key recommendations that emerged, all of which we would be interested in having this conversation about today. And we'd certainly welcome additional public comments from you during the public comment period.

So with that, I'll turn it over to Helen.

Helen Burstin: Thanks, Taroon. And thanks to everybody for joining us today.

It's actually quite a crowd, so I'm glad to see this stimulated as much interest for you as it did for us.

So, I'll dive deep into each of these themes and then the recommendations but please let us know, please keep track of your questions and we'll have plenty of time for Q&A as we wrap up the overall discussion.

So, first here, so the first issue we really face, so we didn't actually – go back please, sorry. Where we didn't – slipping it back again. Sorry, go back to first slide.

(Off-mike)

Helen Burstin: OK. We found overall limited consensus among the advisory panel members, of whether there was a clear hierarchy among various accountability application.

As we began this work, we heard from some that there was a sense that, for example, some of the higher stakes financial penalty programs, for example, might be viewed as being the top of the hierarchy. Others viewed public reporting as being at the top of the hierarchy, but ultimately, there was a sense among the panel, and I'll get into a bit more detail on each of these, that there really wasn't a hierarchy that we could work from.

Second, although one area we did find very much agreement upon on the advisory panel was that there is really a qualitative difference between measures used for quality improvement only and those used in accountability applications. And we'll come back to the implications of this for NQF and our processes.

And finally, there was a sense that field experience to be defined, as we'll talk about, somehow a sense of the measure having been used by end users in the field before being put into some of these accountability applications, would be a useful addition in the evaluation of measures for accountability applications.

All right, next slide. Thank you.

Theme one. So, a little bit more about the limited consensus.

So certainly, we heard some folks, particularly from the provider clinician community, who thought that a hierarchy could be based on what the risk would be of potentially misclassifying providers in the program, and then potentially that subsequent financial impact to providers. There was a sense that hierarchy of financial impact would be difficult to determine from both accountability applications, because most accountability applications may have different financial impact on providers.

And lastly, a sense that a hierarchy based on financial trick to providers only could be misguided, and that a hierarchy is really inherently dependent on the perspective of the user and we heard this. For example, why is it any more important that a provider necessarily be misclassified than a patient potentially picking the wrong provider as part of their selection effort? So that was really what we heard as part of theme one, really unlimited consensus overall as to whether or not there was a hierarchy.

Theme two, really emphasizes question and some of you may know for a very long time, NQF endorsement has always been intended for a quality improvement and accountability, and it's that end that has always been a bit in play here.

And we really had a very thoughtful discussion as part of the advisory panel about measures that were really intended only for Q.I. initially, as they're being used by providers to improve performance overtime, often not used publicly in any way, off in the back-of-the-envelope kind of work that drives improvement on the ground.

There was a sense that many of those measures really required a different standard for evidence and testing. It was fine to have many of those measures perhaps not be reliable and then the real shape because they are being used to drive improvement as long as they are moving you forward on your trajectory, that might be fine. Excuse me.

And lastly, there was a sense there might need to be different criteria for this Q.I. only measures to help create an environment where organizations could

more easily use and share measures. So as we come back to this towards the end around the recommendations, one of the key questions would be, what is the role for NQF in these Q.I. only measures, and some these about making it easy to see these measures, share these measures, approve on this – improve on these measures on the ground, but perhaps not necessarily do, as rigorous, an assessment.

Nest slide, please.

So, theme three was an especially interesting one, because we thought it was really important, but frankly, we had some overtime defining it. That they sound strange but more on that to follow. But essentially, what we heard was that, somehow information from the field, i.e. those on the ground, using measures, being measured, trying to use measures will be a really important input into the evaluation of measures.

As some of you know, we are often times evaluating a brand new measure that has never yet been used doesn't have experience. And there was a sense that field experience data could provide really important information on the gap identified from how the measure is specified. And some of you know, we currently get gaps from what the literature would suggest, as well as an assessment of how well the measure identified variation and meaningful differences of performance across providers, the other side of our required gap assessment for the evidence criteria.

But, which we'll come back to and we'd very much welcome this in your comments. There was not clear consensus exactly what field experience would include or wouldn't include.

And there was a sense that we needed to – it was important to create a mechanism, we've been referring to this for a very long time as sort of feedback loops of how a measure is performing once it's implemented really having that monitoring mechanism in place so we can incentivize developers to provide this data overtime for review. So, much more on that to follow.

So, those are the three themes.

Essentially, what emerged from the advisory panel deliberation was, as I mentioned, since there was not a clear sense of hierarchy going forward or the ability to clearly tie this measure is OK for payments but not for public reporting, for example, a recommendation was to move forward with some sort of grading of measures with transparency based on the NQF endorsement criteria. And I'll walk through in detail each of these recommendations.

The second was to incorporate this concept of field experience, user experience into the measure endorsement process. Third, that we should consider how rapidly evolving measures are being used, and consider a way if we have grades of endorsement to allow for grade advancement and incentivize that grade advancement through grader use and testing.

Next.

As I mentioned in the themes, consider a different standard for Q.I. measures if that remains part of NQF purview. But this to think about and encourage the MAP, the Measures Application Partnership, who just met for their first meeting just this past week, to consider how these measure grades could help you use in the selection of measures as they've been doing that work for the federal program, more than 20 federal programs each year as part of our pre-rulemaking work for HHS.

And then finally, thinking about what would be the future work that would enable us to perhaps get closure to what drove a lot of these discussions upfront, and think about what is the interaction between program attributes, meaning a particular payment program, a public reporting program, et cetera, or even, you know, (tiering) of health plans and selection of providers into those programs, and how that may relate to individual measure attributes, and that seems like kind of a significant core in our measurement science knowledge.

So, next please.

So, a bit more detail on recommendation one. My first request here is we would very much welcome your thoughts about better names for the grades.

For now, at least, we're (inaudible) as a triple A double layer A, or if somebody on our standing committee advisory panel said, "Really, that's better and good."

Good, better and best, I'm sorry. So, best, the AAA measures would be the highest grade possible on all of our must-pass endorsement criteria, including some of these key issues around higher moderate on feasibility, usability and use. And in addition here, that's currently not part of our process, would be field experience for at least one year.

The better measures, the high-grade measures and really important here that all of these would be endorsed, that's just gradations within endorsement, but at higher moderate ratings for all NQF criteria, there may not be much like field experience yet.

And then finally, the moderate grade, the good measures that would meet the bar would really be moderate levels of criteria and limited field experience or performance data.

And, I'll tell you that our group of senior staff met last week for a couple of hours and tried to, in fact, take each of the criteria and put them into each of these boxes and that will be a fair amount of work to finalize.

But very interesting, if we began really thinking about where does transparency fit in here, where does field experience fit in here, what is, for example, measure score level testing versus data element level testing and where would that all fit in. So, much more of that good stuff to follow with CSAC in a couple of (months).

Next slide.

The second recommendation is about this question of, how do we incorporate field experience into the measure endorsement process, and although there were some struggle with defining it exactly, the goal essentially was to have the ability to review performance data on the performance measure prior to widespread implementation. And I think that secondary, that you – this also included, at least in our deliberation this last week, that those being measured

would have an opportunity to see those results prior to widespread implementation.

And that analysis could include, for example, the bullet listed here, what's the gap and measure performance, how well it identifies variations or really meaningful differences of performance across providers. And thinking about those benefits and risks associated with the measure by end users. As you're looking at these measure results, what kind of feedback would you provide? How might that measure be improved? And certainly, the requirements for field experience would likely evolve as the maturity of the measure evolve as well.

The third recommendation was really to think about how we would consider how rapidly evolving the use of measures are in this day and age, and think about how we can incentivize developers to want to bring the measures back to move forward and move up in grade through additional use and testing. So we talked, for example, about the importance of looking at both the intended and unintended effects of a measure when it's in used, recognize how difficult it is to access those data by measure developers, but recognizing how important that is.

And secondly, the recommendation that is part of our measure maintenance process, we could potentially look as an opportunity to think about how we could incentivize raising the measure's grade through additional use and testing, and make that a positive incentive to move up the scale. And some of these are probably seen in the materials we've put forward over the last six to nine months or so, where we've talked about streamlining measure maintenance and putting much more of an emphasis, in fact, on the gap and performance and use and usability over evidence and repeat testing. So, in some ways, this is a nice logical coming together potentially of the work of the Intended Use Panel and the work specifically of measure maintenance streamlining.

Next, please.

The fourth recommendation comes back to this issue of different standards for quality improvement. We've talked about encouraging potentially a different process that would allow this Q.I. only measures to be evaluated differently than other measures and potentially improved.

So, for example, we know most of the clinical measures that fail, usually fail on evidence and particularly this linkage between process and outcome, might that be something you would eventually assess and gather data on as your measure is being approved for Q.I. and then bring it back forward to move to a different level.

While Q.I. measures could demonstrate opportunity for improvement of variation of class providers, there was a sense that overall, you would probably have lower requirements for our current must-pass criteria as I mentioned for evidence, reliability and validity compared to endorsement.

And this is – would be a new process for NQF. And in some ways, somewhat analogous to what we brought forward over the last year about approval for trial use for new eMeasures. We want to – we would want to specifically see comment on your sense of how useful an approach that would be.

Is there a logical reason to bring forward Q.I. measures, and we've currently heard on the call with our advisory panel that some members of the advisory panel specifically thought even in the Q.I. environment, they would still want some way of having a sense that had gone through a process to assure they were using good Q.I. measures.

Others wondered whether this was a space, in fact, we needed to be in. Or thirdly, you've heard a lot of concerns about the fact that there really isn't a library of these Q.I. measures and use. And that some of these, more of a library function of at least gathering all of the current measures that are in use to avoid people recreating those Q.I. measures in every health system, clinic and practice in America.

So, with that, next recommendation, please.

Recommendation five, specifically links back to the MAP at how might they use those measure grades to help with the selection of measures for specific programs.

So, for example, if the MAP determined that a particular program based on its programmatic needs, really required AAA measures while others may only need AA measures that might be useful information. And I think part of it is also that in addition to the actual grade itself, there would be a great deal of transparency of the evaluation and what was rated positively and not, for example, was their field experience yes, no, was also, I think, be very useful to the MAP process as well.

Next.

And as mentioned earlier, there was certainly a sense that as we're increasingly looking towards doing more and more work in this area of measurement science, that we should consider what might be a set of measure attributes that we want to begin to really understand how they could line up with a set of program attributes so that we better matching going forward.

So, for example, measure attributes around potential from its classification, the precision of risk adjustment modeling, or program attributes like the nature of the financial incentive, for example. It would be the kinds of things that clearly will need more work and would line up very nicely with some of the work NQF has been doing back to the SES risk adjustment work and now carrying through to new work we've just been funded around attribution and variation and comparability of measures.

So with that, I'm going to turn it back to Taroon to walk you through the final steps of the process and then we will have plenty of time for questions.

Taroon?

Taroon Amin: Thank you very much, Helen.

So, to walk you through kind of where we are in this process, the comment period for the advisory panel themes and recommendations are open until

September 30th at 6:00 p.m. The advisory panel (speaks) public comments on the key themes and recommendations that we just outlined, and we will – the panel will consider these comments during an upcoming comment call on October 28th.

Following this call, the recommendations will be finalized – presented and then finalized to the CSAC and to the NQF Board of Directors.

So to walk you through just the visual on the timeline, we are currently in the public comment period. And we were having this all-member or member exclusive, I should say, town hall webinar today and we're going through – going to – next to the (full) comment call and then CSAC and then the executive committee.

So with that, I will open it up to see if there are – just to see what comments we have from members, NQF members on this really important topic.

So with that, if we can open the lines or if there are any other comments from the advisory panel, we'd also take those as well.

Helen Burstin: Yes. And we've already got two in the chapter in, so maybe we could start with those, let people queue up with the operator.

So first is from our good friend, Patrick Romano. Patrick, I don't know if you want to do it verbally, but I'll kick it off. He specifically says, (our ties) maintain a library of Q.I. measures, the National Quality Measures Clearinghouse, was there's some concern about the adequacy of that library, and if so, should those concerns be addressed to the NQMC process?

I think that's a great question, Patrick, and I know that the NQMC still has a form that has to be completed, there is still a bit of a burden to submission and it's not clear how much of it truly on the ground.

We've used this Q.I. system only in our – this Q.I. measure only in our health system. And it's really moved the needle for us and others who are out there should try this. It's not clear to us whether those are actually part of that.

We do have a close (directly) relationship with (Equery) and the folks there. So, we'll certainly follow up with them. But Patrick, do you want to say anything else from that? Do you want to open up – if you want to just queue up, we'll properly take your comment, Patrick?

Operator: Yes. To make a comment, please press star then the number one.

Helen Burstin: All right. While we're waiting, there is another question here from (Richard Morrison) with the AAA, AA and A rankings include a clear statement, how accurate and stable the measure is or the degree of statistical validity.

Taroon, do you want to take that one?

Taroon Amin: Yes, absolutely.

So, the panel discussed this at length in the sense that the current NQF endorsement criteria allow for, obviously, review for statistical reliability and then also allow for face validity assessment. But, the panel did say that the – you know, as we move forward to AA and AAA, the rigor and the results of the statistical reliability testing would be evaluated more closely. And also that face validity would not be sufficient off in the AA and AAA categories.

So they're absolutely would be (that) to consideration, quite frankly, that would be a major differentiator between the AA and AAA rating against the A rating, which is essentially the current NQF endorsement criteria.

Helen Burstin: Operator, are there any calls on the phone? I mean, any questions on phone?

Operator: Yes, ma'am. You do have a question from Akin Demehin.

Helen Burstin: Great. Hi, Akin.

Akin Demehin: Hi Helen. Good afternoon. This is Akin Demehin from the American Hospital Association.

First of all, just wanted to thank the NQF for exploring this issue and taking it on, I know it has multiple facets and multiple different viewpoints. From our perspective, the notion of an endorsement that carries within a degree of

precision about how the measure could be used would be something that's very attractive from our perspective.

In reading through the draft report, we had a couple of kind of initial reaction to it, and possibly a question that I'll get to at the end.

The first is, I think there is – there are differences in how measures behave in specific context. And so, in reading through the level of grades that you came up with is kind of a straw man here, it struck us that it didn't seem particularly tightly linked to a particular application for a measure and sometimes, there is no way of knowing whether a measure retains its accuracy and validity when it gets applied in a different context. I think a lot of that has been part of the debate about the PSI-90 measure which was initially tested on an all-payer claims database and yet is used on Medicare claims only.

And so, I think exploring the linkage between that a little more explicitly would probably be helpful.

The second thought that we had is – certainly appreciate the notion of measures having an opportunity to increase their grade asset for.

I don't know that I would lay it out as quite so much as such a determinants like path, that is, I don't know that every single measure necessarily needs to get on a path towards being endorsed for use and, for example, a paper performance program. I think there are some measures that are just never going to really be oriented towards that particular application and that's fine, they can still add value, but maybe not in that particular context.

The other reason why we think that (membership) is so important is, as you know, we've really been pushing for greater parsimony and focus of measurement efforts nationally. And so, setting up a system that kind of almost passively encourages measures to step up to the plate of the publicly reported measures, I think, makes it a little more difficult to hone – (for us) all to hone in on the most important measures that are really going to drive performance forward.

So, I guess the last comment I have is really more of a question about process. So, certainly, understand that the panels come together to provide some recommendation, how are those recommendations be used by the CSAC in the board? In reading through the report, it didn't feel like it would be kind of a straight up, we'll adapt the recommendations or no. It's not kind of a binary decision.

It feels like it's being set up as providing thought for additional direction by the CSAC and the board. But, if you could describe that a little more, that'd be helpful to us. Thanks so much.

Helen Burstin: Thanks, Akin ...

Taroon Amin: So Helen ...

Helen Burstin: ...(inaudible). Go ahead.

Taroon Amin: So, Helen, I could start on – I mean, those are very thoughtful comments. And then if you want to just add on or however on each of them, I think that they're all sort of interesting.

Helen Burstin: Yes.

Taroon Amin: So, the first one, Akin, on the issue of how measures behave in different context just to sort of paraphrase a little bit, you know, this was a significant part of the conversations among the advisory panel and sort of reflects a bit of what you see here in terms of recommendation number six.

There was no question that the group tells that program implementers – particularly, there was element of the way that program was being designed and implemented, that would require different attributes of measures. You know, some of the examples that Helen pointed out was, you know, how the comparison groups are created, the measure score threshold, if you will. You know, whether there was statistical test that were used to distinguish between performance categories.

There were a number of these sorts of program attributes that the panel identified and really, you know, basically agreed with this, a notion that individual measure attributes sort of interact with this program attributes and there needs to be further examination in terms of how we evaluate this interaction. Mainly because it didn't seem like this was really a measure property following the jurisdiction of the NQF endorsement property – process. But still regardless, it represents a concern or emerging concern among those that are being measured around, well, we may have a great measure that performs well in terms of testing, but actually when it gets implemented in the program, folks have some concerns and there doesn't seem to be an appropriate venue for (those) concerns to be vetted since those are really program attributes.

And so, there was definitely an acknowledgement of this being an issue and certainly recommended additional work in this area.

The second comment, I think, is – and also on – is very important comment as well around not all the measures need to beyond this pathway toward AAA.

And I think one interesting notion, I guess, was that the way that the panel sort of characterized this, and I think we will need to think more thoughtfully about the comment that you're raising, is that there – you know, there – the field is looking for a – meaning the community, I should say, the – those folks that represent NQF membership, are looking for additional evolution of measures as they're being implemented and creating a process in which that information is getting back to stakeholders on how well a measure is performing once it's actually endorsed.

And I think from that perspective that the idea will, you know, let's try to create some incentive to move along the pathway, not necessarily that the measures need to "grow" in the level of maybe sophistication or additional testing.

But your point is well not necessarily that we want to move, you know, or quite frankly they can move to a AAA rating.

The third point is also sort of an interesting one, and one that Lee sort of pointed out at the beginning of this (patient) which is, you know, the – first of all, you know, how does this advisory panel, you know, what was the level of the recognitions and, you know, some of the challenges that we've already kind of heard already in terms of how will this, you know, how will this guidance and recognitions be implemented.

And I think that that's sort of an important point that I'd like the staff and reflect on for a moment, which is to say that the advisory panel was given a very broad charge by the NQF board. They – and the NQF board was really looking – and the CSAC was looking for a strategic guidance on, you know, should we actually include, you know, the specific program in the measure endorsement criteria and require that every measure state that A priority. And this question of whether there was a hierarchy was really took almost three hours of the call of the advisory panel because there's been a lot of debate on this, as Helen pointed out in her comments.

And so, this advisory panel, really, is sort of setting up a strategic direction since it, obviously, has a huge implication to the NQF endorsement process and its meaning. The advisory panel came up with sort of a strategic vision for where the NQF endorsement process should go, in which the CSAC will consider this sort of vision, the straw man and the AAA grading methodology. And then, the CSAC will actually try to do the hard work of mapping along with staff (inaudible), actually did the hard work of mapping the criteria, you know, more specifically to each of the grades.

And the CSAC will have to do some of the work related to how this gets implemented, once the strategic plan has sort of identified and vetted through the NQF stakeholders, particularly NQF members.

So I'll stop with that and I'd welcome any comments from Helen on this topic as well.

Helen Burstin: Yes, thanks, Taroon and thanks, Akin, those are great questions. I think the one thing I would just say at the very end just to be very – sort of pointed in your response, yes, the CSAC will be asked to be making an (up or down)

recommendation about the grading and, in fact, they will be at our November meeting at least based on our initial work, be going through a fairly detail way of actually expressing if this moves forward, the grades by each endorsement criteria and sub-criteria.

So, more on that certainly to follow, with an expectation that it would go to the board to follow. So that's a definite yes.

So, lots of good questions there. We've already got a couple more in the chat. So I just want to quickly pull up those including a very thoughtful comment from and Dr. (Hogenmiller) at Homewatch just making the point that, you know, so much of the success of really thinking about this question of field experience will really be determined overtime as the processes are developed and the outcome unfolds overtime. So certainly, we agree this is an important step forward, but no quick answers here.

(Cindy Heltad) put forward a question, asked about what examples, for example, measures for Q.I. that only a health system would be expected to submit for NQF endorsement with this new grading system. (So that we see) so many one-off measures with no intention of seeking endorsement (is stuff) so I can tell.

I'm not sure the new grading system will cause health systems to harmonize their similar untested measures. Has the committee discussed how the grading process would affect the QCDR measure testing and application for measure endorsement?

So, certainly no, there was no discussion of the QCDR piece of this. And it's interesting questions, (Cindy), I think our feeling was more so that not so much that the health systems or really anyone would be – we'd have to think this through, but the Q.I. only pathway, it may be that they're just coming forward really for the sake of sharing and learning from others.

So, a classic example we've talked about for many years is how many hospitals have slightly different definitions of what's unexpected return to the (O.R.), I mean, because some of this be an opportunity to have among NQF members a chance to iterate to say, what have you figured out to do that works

better, so that it isn't always the sense of kind of reinventing each time somebody tries to move it forward, but we would welcome your thoughts specifically about the latter issue as well.

Sam Tierney, hi, Sam, who's only on the web today from AMA's PCPI, asked if there had been any thought or consideration given to the additional resources required to develop a (deniable) measure that could earn a AAA.

As the measure develops, we've got to take a significant time, amount of time and resources to develop a measure that would meet the criteria as they stand now and adding this layer of complexity would seem to require even greater time and resources, certainly for measure development.

There was not a lot of time to (certainly) discuss among the advisory panel, there were several – a couple of measure developers on the panel including Elizabeth Drye from Yale as an example.

So there was a sense that the AAAs would be a higher bar, but there wasn't necessarily an expectation that all of that needed to be available immediately. You know, there are measures, often times, not necessarily being used for some applications, where perhaps, there is that opportunity to gain more insight. And, certainly as we've talked about with PCPI, we know one of the really difficult issues is the fact that it's very difficult for you to, in fact, get and other developers actual on the ground results of what's happening.

So, perhaps, some of this could also be thinking through with the developing community, we did have a call this week with – last week with our measure developer advisory panel to talk this through and they had some good suggestions here.

And, think about what would be potentially additional resources overtime that might be brought forward. CMS was part of (this kickoff) which is part of these discussions as well. So, at least for those federally funded projects, this was certainly something we talked about in terms of the ability to get to that level of rigor.

All right. Couple more in the chat. Operator, please let me know if there's other people in the queue, we can alternate back and forth between chat and live calls. Is there anybody waiting currently, operator?

Operator: Yes. We do have one from John Shaw.

Helen Burstin: Hi, John.

John Shaw: Hi. Hi, Helen, how are you?

Just wanted to say to you and the committee, I really like the way everything was presented and outlined and wanted to just highlight my favorite. And, that's the – including the field experience.

As a lot of the discussion has already mentioned, this is a complex situation. And we have a lot of things moving rapidly. And the feedback loop for field experience, really, gives us an opportunity to move forward without having as much as of a risk at misclassification.

So, in general, it's impossible, I would think, to be perfect out of the gate for any of the measures. And so, some type of fine tuning is going to be necessary.

From what I recall, a driver for the intended use examination were some of the committee members that had voted an endorsement and when the measures came up again and the MAP said, "Oh, if knew this is how it was going to be used, I might have voted differently during endorsement." And, others had shared that we're really looking at the same measures for a lot of the same things multiple times and that's inefficient.

What incorporating the field experience indicator gives us is, basically, the underpinnings of a continuous quality improvement.

So, some of the issues and concerns that people had were, we really wanted to reduce the potential harm of misclassification, and I agree that it should be for all stakeholders, not just providers ,but also the patients and plans and even communities.

But to do that, even if we're wrong out of the gate, if you have a process that very quickly brings in the field experience ...

Helen Burstin: Yes.

John Shaw: ...it allows you to modify it quickly. So thanks ...

Helen Burstin: Exactly our thoughts, John, yes, very helpful, appreciate it.

We'll take one from the chat, we'll go back to the phone line, so thank you for all of your good comments.

(Karin Revin) for the AMA asked us to explain the process of how the reports recommendations would be operationalized, making (it probably) be critical this information be released to the – for comment part, any approval of any changes, then just point this report is actually out for comment right now. So, please submit your comments.

I think the one thing we will see going forward is we'll put forward to CSAC at least a straw man proposal of how the current criteria and the inclusion of field experience would fit in there.

(Karin) also asked if, you know, if the changes are not defined further, will there at least be preliminary guidance and how different scenario should be handled. And we did talk about potentially putting forward some case examples for clarity as we're moving this forward.

She had concerns, but otherwise, there was a risk for this process and then she has an even greater level of the inconsistency, especially since the process is subjective as it is now.

And, you know, for example, one of the implications of a group adapt a grading measure for public reporting is not a bad thing. And thank you, (Karin). It's interesting – again, I think it's important to note that the group wasn't going to go down the pathway, they didn't feel like there was actually enough science or grounding there to, in fact, say that specific grades of

measures link to specific uses. So there would not necessarily be any implications in how it could be used certainly by, you know, different groups.

But, one of the things we've discovered is you at least have gone through the initial (staff) process of trying to sort out the criteria into the various buckets, is how much we are relying upon the current algorithm so you already have for evidence and testing, and clearly, pointed out the need for similar algorithms for our other criteria.

And we think the algorithms have certainly made it less subjective, but we would certainly welcome thoughts in that going forward.

There's another one in the chat, but operator, is there somebody else queued up?

Operator: There are no comments from the phone lines at this time.

Helen Burstin: Super, OK.

So, the next comment from the chat box is from (Elisa Crawford). Hi, (Elisa). And it said specifically made the kind of – more aspirational measure design to move the needle broader or less defined, but nonetheless extremely important aspects of care, are likely be rated lower than measures that are very focused on specific procedures or outcomes that won't have as big an impact on quality or cost of care.

So, could the grading system, for example, encourage uptake of measures that although evidence base, reliable, usable, et cetera, are not as capable of moving the things that really matter.

And we actually had this discussion, (Elisa), it would have – helpful to get some of your folks listening in on this, were in that certainly to follow when we present this to CSAC in November.

So we specifically had this pretty extensive discussion about how to handle evidence for some of the really important new emerging areas where it's often harder to get it, like care coordination.

And the last thing we intend to do with this is make it harder to bring forward the measures that matter. So we actually had some thoughts that we might put forward, to think about how we handle that. We're also hopeful of the fact that at least our current criterion requires only a justification for the relationship between the outcome – or rationale between the justification for the outcome and its relationship with the potential (prophecies) or structures. At least what we've seen to date is that that does allow significant more flexibility for those more newer and innovative measures where you certainly may not yet have the extensive evidence that this particular measure has had outcome yet, although everyone agrees it's really important to do.

Some more on that to follow, thanks.

Anyone else on the line, operator?

Operator: No, ma'am. There's no one on the line.

Helen Burstin: All right. So, one more comment here from Akin, he raised one more issue that I think is important, so I'm going to read it again, Akin.

One of the thought is how to frame the hierarchy of grades, and although he agrees, the question shouldn't necessarily turn on how much (by now the risks is displayed), it is important to think about it in the context of reporting and pay for performance the extent to which a measure really differentiate the performance of providers. Is it accurate or reliable, in that that provider A provides better care than provider B. Then that's very much a very useful framing, it isn't about that. And in fact, you could easily and we heard this quite (impassionately) from Cristie Travis, our former CSAC chair who was on the panel who made the case that, for example, you wanted to be able to differentiate performance as a consumer or a purchaser as much as you wanted to certainly for providers to feel that comfort as well.

So very useful, very thoughtful comments.

Let me see if there's anything else that has come in.

So, one more coming in, (I hope).

In the interim, feel free to continue to queue up if you'd like.

(Karin) is back.

Let's see. So, (Karin) makes the point that the current NQF submission format developers provide information on opportunity for improvement, identify variation and meaningful differences. Is there more that should be asked or a new criteria needed? This needs to be clarified, so technically, it's already addressed but not always producing the desirable results.

You know, it's an excellent point and we talked a fair amount about that as well of what kind of information was available and actually, you know, just last weekend in our (C.V.) committee, a couple of measures were put forward, truly without any gap. There was no information yet available from the measure in PQRS, and the committee was actually frankly uncomfortable about not yet having that information as they are making that decision.

So, in some ways, our hope is that some of this will spur more discussion and the broader measure development and measure use enterprise so that people get those data to submit, because frankly, it's really, really hard to get those data and it can sometimes take a fairly long time. And that's part of the reason why our requirement for more of the field experience to move up might incentivize end users and developers to work more collaboratively and get that information available.

OK. There is a hand raised. Yanling Yu, if you're on the – press one – star one for the operator to call on you.

Operator, do you have a comment incoming from Yanling Yu?

Operator: And at this time, please press star one.

Helen Burstin: Please press star one, Yanling Yu. We'll be happy to take your question. I don't see any other hands raised, although lots of you are on this call. All right.

Operator: And we do have Yanling online. Your line is open.

Helen Burstin: Hello, welcome. Yanling, are you with us?

Operator: Your line is open.

Yanling Yu: Can you hear me?

Helen Burstin: Yes, we can hear you now. Please.

Yanling Yu: OK, thank you.

I have a question about the recommendation for considered different endorsements vendor for Q.I. measures. I was just wondering whether the panel have considered the fact that sometime those – some (Q.R.) measures eventually involved with or, you know, some type of performance measure. They are not pure Q.I. eventually, you know, it would tie to public reporting or, you know, some type of a pay performance.

So, I wonder if a panel have thought about it. When you would set up a totally different endorsement standard for Q.I., what eventually a Q.I. measure involved in towards that direction, what the recommendation would be for this type of Q.I. measures?

Helen Burstin: Yes, thanks for that question. It's a great question and we did talk a lot about the evolution of measures, and that was part of the discussion, certainly we talked about it being able to go up in grade.

We didn't talk so much about kind of jumping pathways, which I think is what you're saying, from a Q.I. pathway to the more accountability pathway, but you're absolutely right, it's something further that we do need to discuss.

We have a lot of conversation and, you know, we had some really remarkable Q.I. insights from Don Goldmann, the chief scientific officer at IHI, who was on the panel as well. And, you know, one of the key questions is really whether the measures that are often used for Q.I. are really similar to the measures used for accountability or could some of those ideally be kind of the Q.I. measures that could almost, you know, emerge with additional testing and become accountability measures.

So, thank you for that thoughtful comment. It's something we'll need to consider going further.

Yanling Yu: Thank you.

Helen Burstin: Anything else in the chat? Anybody else on the phone, operator?

Operator: There are no comments at this time.

Helen Burstin: OK.

Do we have anything e-mailed to us that we want to – I know we had some folks e-mailing us questions earlier. Anything we'd like to – do you think we've covered everything? I think we have answered the questions that came in. You guys are all on the same page there.

Taroon, do you want to add anything?

Taroon Amin: No, I mean, this has been a very thoughtful conversation. I mean, I think maybe, you know, process might be another sort of interesting conversation. I think we've touched on this a bit, but where we're going from here, you know, the general strategy, overall recommendations have been sort of put forward by the advisory panel.

We'll take the feedback that we've gotten from this call. We've had the opportunity to have some discussions with several advisory panels, the NQF including our measure developer, advisory panel, and then our standing committee advisory panel which includes chairs and members of NQF measure endorsement standing committees on how this, you know, some questions and clarification that have come up to these conversations.

I will provide all of this feedback along with the comments that we received to the advisory panel.

As Helen pointed out, we will be sort of taking the description here in terms of the single, double, and triple A ratings and how they relate to the criteria.

We'll be having a more robust conversation around that specific topic with the Consensus Standards Approval Committee, who will consider whether or not, you know, essentially the main recommendation here around moving ahead with the grading methodology should be, you know, tested and implemented in upcoming projects.

Obviously, that will be considered by the NQF board as well in early December.

So again, we'd welcome any other questions or thoughts or comments, reflections on this conversation both again today if you have any. We welcome that conversation, but we'd also welcome that to the comment tool, and we will make sure to thoughtfully consider those and bring that forward in terms of our adjustments for this process.

Helen Burstin: Great. Thanks, Taroon.

And as you can see, it's not a very long report. We didn't want to drown you in this. It was really more, I think, a thought-provoking report of getting your ideas and census – NQF members of this being a viable direction. And you still have plenty of time, comment doesn't close until September 30th.

So, you know, at least a week or more to please take a look at the report based on what you've heard today. Feel free to make comments, I think we still have several, many. And then we'll have an opportunity to speak with the panel again towards the end of October and then, of course, we'll bring further deliberations to our public transparent CSAC, November 18th to 19th. Not sure exactly which day yet, but we'll post that agenda probably so you'll be able to see exactly when that will be addressed.

But, all of you are very busy, so if there aren't any other questions, I don't see anything else on the chat, or operator, anybody else queuing on the phone?

Operator: No ma'am, there are no comments.

Helen Burstin: Wonderful. All right.

Well, thank you for your attention. I'm glad we had this opportunity to talk through our members and get your insights into this important area of work. Thanks much, everybody. Bye.

Male: Thanks.

Operator: Thank you. This concludes today's conference call. You may now disconnect.

END