NATIONAL QUALITY FORUM

Moderator: Measure Developer Maintenance
December 12, 2016
12:00 p.m. ET

Operator: This is Conference #32498691.

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

(Shan): Good afternoon, everyone. Welcome to the call. Please note, all committee members' lines will be opened for the duration of our time together today. So please be sure to use your mute button when not speaking or presenting to reduce background noise. Please keep your computer speakers turned off so we do not have any feedback. And please, do not place the call on hold at any time.

We will be voting. Committee members only will be voting today. During that time, we will have additional instructions and information. And then now, it is my pleasure to welcome our NQF staff. Katie, let's get started.

Katie Streeter: Thanks, (Shan). Hi, everyone, this is Katie Streeter. I'm a Senior Project Manager here at NQF. I'd like to thank you all for joining us today for the Musculoskeletal Standing Committee Webinar.

I'm also joined here with Karen Johnson, who is senior director at NQF, and Ann Hammersmith who is our general counsel here at NQF.

OK. So, today, we'll start off by doing Standing Committee introductions. We'll be speaking to the off-cycle work that this committee will be doing as well as the roles of the Standing Committee.

I'll give you a brief overview of the measure evaluation process, a little context and background of why we're here today. And then we'll start our discussion of consideration of candidate measures, we'll be reviewing measure 005 -- or 0514 and 0052.

We will be discussing harmonization of these two measures and then we'll close with public comment and next steps.

So now for Standing Committee introductions, I will turn it over to Ann Hammersmith.

Ann Hammersmith: Hello, everyone. We'll combine the introduction and the disclosures of interest because it's a little bit quicker that way. Before we start with the disclosures, I'm just going to give you a little fund thumbnail sketch of what we would look for you to disclose.

First, you are all sitting as individuals on this committee. You don't represent your employer, you don't represent anyone who may have nominated you to serve on this committee. You're on the committee because you're a subject matter expert and you serve as an individual.

For disclosures, we are interested in your disclosure only of things that are directly related to the committee's work. We're typically interested in consulting research support grants and speaking.

Also want to remind you that disclosures extend beyond financial disclosures, so you may have served on a committee or something like that as a volunteer, but that would still be relevant.

Just because you disclose does not mean that you have a conflict of interest. Part of the reason we do this is for transparency and openness so that everyone understands the committee members' background.

So with that, I'll ask you to identify yourself, tell us who you're with and if you have anything you wish to disclose. And I will call on each member of the committee. Roger Chou.

Roger Chou: Hi. I -- my disclosures are that I work with the American College of Physicians and American Pain Society to develop guidelines and standards for use of imaging for low back pain.

I mean, my stuff is (stated) in some of the measures, so I think that's obvious and I receive some funding to do so from those groups.

Ann Hammersmith: OK. Thank you. Kim Templeton.

Kimberly Templeton: I'm Kim Templeton. I'm at the University of Kansas in Kansas City and I have nothing to disclose.

Ann Hammersmith: Thank you. Thiru Annaswamy?

Thiru Annaswamy: Thiru Annaswamy, Physical Medicine & Rehab of V.A. North Texas Health Care System. I work on a couple of committees, one with the academy, I'm the chair of the Evidence Quality and Performance Committee.

And I'm in the North American Spine Society's Evidence-Based Guidelines Committee. As a part of those committee roles, I'm involved in developing, overseeing quality measures and guidelines, et cetera.

And I'm also being funded by research supported by PCORI and AHRQ in the past that may have looked into topics relevant to imaging in lumbar spine disorders.

Ann Hammersmith: OK. Thank you. Carlos Bagley. Is Carlos Bagley on the line? Steve Brotman.

Steve Brotman: Steve Brotman, AdvaMed, nothing to disclose.

Ann Hammersmith: Sean Bryan.

Sean Bryan: Sean Bryan, Greenville Health System, Greenville South Carolina and the American Medical Society for Sports Medicine. I'm a former board member. I serve on the Health Care Transformation and Quality Measures Committee, but have received no funding and have not worked with any measure developers.

Ann Hammersmith: OK. Thank you. Craig Butler.

Craig Butler: I'm Craig Butler. I'm currently working with UnitedHealthcare and also worked on -- with the Academy of Orthopaedic Surgeons on the workgroup on (indiscernible) by its function and pain measures they're adapting, but I really don't have any outputs.

Ann Hammersmith: OK. Thank you. Kelly Clayton.

(Off mic)

Kelly Clayton: Yes. I have nothing to disclose.

Ann Hammersmith: OK. Thank you. James Daniels.

James Daniels: Yes, it's (Jady) Daniels here. I work at Southern Illinois University. I run Sports Medicine Fellowship. I'm on the task force of the American Academy of Orthopaedic Surgeons and also American College of Occupational and Environmental Medicine for similar topics of this. I've written a couple of book chapters and articles, also on hospital task force. That's it.

Ann Hammersmith: Thank you. Christian Dodge. Is Christian Dodge on the line? Katherine Gray.

Katherine Gray: Yes. I am the president of SAGE Health Management Solutions which is a clinical decision support technology. And, also, I'm the executive director of a non-profited SAGE Evidence-Based Medicine & Practice Institute that is developing clinical content.

And, I may from time to time serve on one of their panels which could, you know, touch on one of these topics as part of the great review.

Ann Hammersmith: OK. Thanks. Marcie Harris Hayes.

Marcie Harris Hayes: I'm at Washington University and also on the task force to develop clinical practice guidelines for hip disorders for the American Physical Therapy Association, and I have no disclosures.

Ann Hammersmith: OK. Thanks. Mark Jarrett.

Mark Jarrett: Hi. I'm the chief quality officer at the Northwell Health and I'm a rheumatologist and I have no disclosures.

Ann Hammersmith: Thank you. Puja Khanna. Is Puja Khanna on the line? Wendy Marinkovich.

Wendy Marinkovich: Hi, this is Wendy Marinkovich. I'm with Blue Cross, Blue Shield Association, and I have no disclosures.

Ann Hammersmith: Thank you. Jason Matuszak.

Jason Matuszak: Hi, I'm chief of Sports Medicine and Excelsior Orthopaedics in Buffalo, New York. And I have no relevant disclosures.

Ann Hammersmith: OK. Thank you. Catherine Roberts.

Catherine Roberts: Hi, Catherine Roberts, Mayo Clinic, Phoenix, Arizona. I have no involvement in the measures being discussed. I also have no consulting research grant or speaking support.

I do, however, write and review measures both for the American Academy of Orthopaedic Surgeons, their appropriate used criteria and the American College of Radiology Appropriateness Criteria where I am a muscular skeletal panel chair.

Ann Hammersmith: OK. Thanks. Arthur Schuna.

Arthur Schuna: Hi. I'm Arthur Schuna from the V.A. Madison, Wisconsin where I worked for 41 years. Currently retired and I have no disclosures.

Ann Hammersmith: Thank you. John Ventura.

John Ventura: Hi, John Ventura, an owner of a consulting company, Spine Care Partners. And, I do sit on the technical expert panel for the Physician Quality Reporting System for the pain and function measures.

Ann Hammersmith: OK. Thank you. Christopher Visco.

Christopher Visco: Hi, Christopher Visco here from Columbia University and the residency program director and Sports Medicine Fellowship director. And I also sit in the board for the Association of Academic Physiatrists, but I have no relative disclosures.

Ann Hammersmith: OK, thank you.

Has any else joined who I missed?

OK. Just a few reminders, again, we said it's an individual as a subject matter expert on this committee. If during the course of the meeting you think that you have a conflict of interest, you think that a fellow committee member has a conflict of interest, or someone is behaving in a biased manner, please speak

up in real time. You're welcome to bring it up during the meeting. You can contact your co-chairs or you can contact NQF staff and we will work to resolve any issue.

Any questions from any committee members?

Christian Dodge: I just wanted to...

Ann Hammersmith: Yes.

Christian Dodge: Hello? I just wanted to say, I'm Christian Dodge, I was having microphone issues earlier.

Ann Hammersmith: Oh, OK.

Christian Dodge: ...Association of Naturopathic Physicians and no disclosures. Sorry about that.

Ann Hammersmith: OK. Thank you very much.

Anyone else? OK, thank you.

Katie Streeter: Thanks, Ann. This is Katie again and I'll be moving on and talking to you about roles of the Standing Committee. We bring together this group of experts to evaluate the measures in depth, and to make recommendations to NQF membership for endorsement. And then the membership will vote on the measures.

As a reminder, the role of the co-chair in today's meeting will be to co-facilitate the meeting, keep the committee on track to meet the goals of the project without hindering critical discussion or input, and to participate as a standing committee member.

Roles of the Standing Committee specific to measure evaluation, and also, I'd like to point out that we are fortunate to have the measure developers present at our meeting. They will be available to introduce the measures, to respond

to questions from the committee and to correct any misunderstandings about their measures during our discussion.

During the measure evaluation, committee members often offer suggestions for improvement to the measures. These suggestions can be considered by the developer for future improvements. However, the committee is expected to evaluate and make recommendations on the measures as submitted.

We wanted to point out a few changes to our processes here at NQF. These include off-cycle opportunities for the Standing Committee, which is why we are here today. Modifications to the CDP process change and emphasis when evaluating maintenance measures and additional staff guidance. And this includes the preliminary analysis and ratings that we provided you with.

So to highlight the one major change to our NQF Consensus Development Process here, I'm sorry -- the board ratification should actually be crossed out. We no longer go through that step in our process.

So, the way we do things now is that the measures go through Consensus Standards Approval Committee, CSAC decision. And then from there, we go right into appeals. We no longer have the board ratification.

So for the evaluation process, to assist the committee evaluation of each measure against the criteria, NQF staff prepared preliminary analysis of the measure submission. This will be used as a starting point for the committee discussion and evaluation. If you were assigned on a measure, you will lead the discussion for each criteria. And we did assign three committee members to each measure to act as lead discussants.

During today's meeting, the committee will discuss and rate each measure against the criteria and make recommendations for endorsement.

NQF has recently streamlined a maintenance process. In the maintenance measure forms, you'll see that any new information is in red and all the information is in black. The intent was to decrease the developer and

committee workload particularly when there were no updates to the measures. During the meeting, if there are no updates to the criteria and the committee agrees, then we will not vote on these -- on those criteria.

So here's a refresher on NQF endorsement criteria. The criteria are in the specific order and there is (hierarchy). There is logic to looking at them in the specific order.

The first one is importance to measure and report followed by reliability and validity, scientific acceptability of the measure properties. I'd like to point out that criteria one or two -- criteria one and two are must-pass criteria.

I'd also like to note that we'll be discussing harmonization a little bit later after we get through the first set of criteria.

So this slide here, we wanted to highlight the changes in our maintenance process. As I mentioned earlier, we have streamlined our process for maintenance review.

And so from our perspective of, do we need to keep redoing and rediscussing the same old same old. We ask the developers to tell whether the evidence has changed or not. And then we'll ask the committee if they agree. But if the evidence was followed three or four years ago and nothing has changed, this probably is still pretty solid now.

So we do provide a summary of the prior review. And then the question puts you as the committee is, here's what the evidence was, are you aware of any changes? And if not, can you just accept this measure has passed the criterions and we'll just move on? Clearly, this is a standing committee decision.

There is, however, an increased emphasis on data for current performance gaps in cares and variation.

Criteria number two asks you to take a look at the scientific acceptability of the measure properties. What we focus on here is the reliability and validity of the measure.

At the end of the day, we're looking for those measures that can provide reliable and valid results, so that we can make judgments about the provider who's being measured in terms of the quality of care they provide.

Measure specifications are addressed under reliability and validity. Specifications are the instructions for calculating the measure. We ask you to really pay close attention to specifications of the measure as (your assess) as reliability and validity. Precise specifications are the foundation for reliability. And specifications consistent with evidence are the foundation for validity.

Again, we wanted to quickly show you the change in emphasis for the maintenance measures. We're really looking for a much greater focus on measure use and usefulness for this measure, including both impact and intended -- unintended consequences.

Feasibility tends to be mostly about the data source and the burden of collecting, submitting and calculating the data. And we're looking to (see) measures that are endorsed by NQF be publicly reported and will be used in accountability program. And so we'll be sharing with you any of the details that we know about how a measure is being used. And so that will be a part of your evaluation on usability and use.

So the process for our measure discussions today will start off by asking each measure developer to briefly introduce their measure. They will be given two to three minutes for this introduction. We'll then ask the lead discussants to begin the committee discussion by providing a summary of the pre-meeting evaluation comments. And actually, I updated the comments this morning. We actually did not receive any pre-meeting evaluation comments.

Developers will be available to respond to questions at the discretion of the committee. And the committee will vote on all subcriteria -- criteria and subcriteria.

As a reminder, a process for achieving consensus, we do need 66 percent of the committee to be present in to be voting, which we do have. To be recommended, measures must have greater than 60 percent of the committee voting, yes, which is high and/or moderate meeting. Consensus is not reached if we are between 40 percent and 60 percent. And, consensus not reached measures will move forward to comment and the committee will revote on these measures after the public commenting period, any ideas to gather all of our -- all input from stakeholders across the board.

So, a reminder of why we are here today, we are here to review and vote on measures 0052 and 0514.

As you may recall, these measures were submitted for maintenance review in 2014. The committee did not recommend the measures for continued endorsement. There were several recommendations need for each measures and the developers agreed to defer the measures and make changes based on the recommendations.

During the CSAC review, the committee or the CSAC noted concerns about the committee's interpretation of NQF criteria related to measure exclusions. And, as I mentioned, the developers made changes based on those recommendations and the measures were submitted to NQF in October of this year.

So with that, I'd like to move on to discussing the first measure, 0052. I also like to see -- make sure we do have the measure developers available on the phone?

Jenna Williams-Bader: Hi, this is Jenna Williams-Bader from NCQA. Can you hear me?

Katie Streeter: Yes, I can, thank you.

And before you begin with introducing the measure, I'd like to ask if any committee members had any questions about our process for today or anything that was just mentioned.

Roger Chou: Yes, this is Roger. Can you clarify what you mean by the CSAC has concerns about how the committee was interpreting exclusions, the NQF exclusion criteria?

Karen Johnson: So this is Karen. In rating the report, and apologies, I wasn't staffing that project in 2014, so all I have to go by is what was in the report.

But I think the difficulty that the CSAC must have had, there are a lot of red flag options, if you will, from the appropriate use criteria, and not all of those were included as part of the measure specifications. And my understanding was that was the difficulty that the committee had with those measures.

I think when the CSAC looked at it, I think they weren't sure, and again, this is my interpretation, but I don't think that they were sure that that was appropriate to penalize the measure for not having those exclusions, probably because many of those exclusions are fairly -- they're not -- they don't exclude a lot of people. So in other words, they're low frequency exclusions. I think that was probably part of the difficulty.

But since the developer (worry) at least amenable to reconsidering some of those exclusions, they ended up doing a deferral rather than going with the decision to drop endorsement. Our developers are actually on the line. If they remember something different than what I'm kind of figuring out based on the reports, feel free to jump in and explain a little bit. But that was my understanding of what happened.

And maybe other committee members might remember, too. I -- it has been couple of years at least.

Roger Chou: Yes. So there's some data in the current measure that talks about the percentage of people with potential exclusion, which is very low, but that's very inconsistent with the data which actually -- there's evidence showing that if you look at populations you present with low back pain, up to a third or half will have a red flag condition.

So, I don't understand where the CSAC's criticism or critique is coming from. Is that -- was that a database critique or is that just a, oh, we don't think that these -- I mean, I don't get that.

Karen Johnson: You know, I am not sure, I might be able, in the next little while, to find our transcript.

Jenna Williams-Bader: Hey, Karen this is Jenna from NCQA. Perhaps, I could say something about that?

Karen Johnson: That is great. Yes, if you remember, Jenna, that'd be great.

Jenna Williams-Bader: Yes. So on the -- what we (thought) as far as the exclusions were exclusions that we tested. I mean, there are two types of testing, we looked at claims for a larger number of patients and then we looked at those claims in the medical record for a smaller number of patients to see what might be appearing in the medical record that's not appearing in claims.

My understanding is that one of the exclusions that's included in studies that were referenced -- the staff referenced are people over the age of 50 or just older adults. And our measure actually only goes up to age 50. So, that might actually be one of the reasons why we're seeing (relative) exclusions here because if you do think of age as an exclusion criterion, that is obviously going to exclude quite a large number of people where some of these exclusions are much rare.

Karen Johnson: OK, so that might have been (it) if they were looking at the low. I do remember that in the report, there was discussion of low prevalence of -- from the exclusion. So that could have been part of it. I don't think that would

answer the CMS measure because that one doesn't have the age exclusion, but that might have been part of it.

Does anybody else have a memory of anything?

Kimberly Templeton: This is Kim. I don't -- I can't (say) anything from that standpoint, but I would say, I would share Roger's concerns that the red flags that were brought up by the committee or discussion the last time were not included. I'm also concerned that there's not an exclusion for people with prior spine surgery.

Karen Johnson: OK.

Jenna Williams-Bader: So, again, I was going to cover this in the opening remarks, but since it was brought up, our measure does have what's called a negative diagnosis history. So, patients who had a claim for low back pain in the six months prior to what we call our index episode start date are excluded, and that would also exclude patients who'd had surgery in the prior six months for low back pain.

Roger Chou: Yes, this is helpful. I mean, I didn't want to get into the details about the exclusion criteria. I was trying to get at what we were being criticized for because it sounded like we are being critiqued for not understanding the exclusion things. But, I don't think we -- I think we did understand what the exclusion, we didn't have the data at that time about what the proportion of the, you know, the exclusions would be. So, I don't see how we could have been critiqued for that. We were concerned that there were red flag conditions that weren't specified in the criteria.

So, I was just curious about why the CSAC said that we didn't understand what the exclusions criteria were supposed to be. But I think we can move forward and we can talk about the specific exclusion when we get into the criteria.

Karen Johnson: Yes, I think that's really good. And, if I can, I will try to locate the transcript from that meeting and to see if anything else jumps out. And I'll let you know as -- if I find it.

Katie Streeter: Thanks. So with that, I guess we can turn it over to NCQA to give their brief introduction of measure 0052.

Jenna Williams-Bader: Great, thank you so much.

So as I said, my name is Jenna Williams-Bader. I'm a director of Performance Measurement here at NCQA.

As a reminder, our measure is a health plan level measure that you did administrative claims to assess the percentage of patients with the primary diagnosis of low back pain who did not have an imaging study within 28 days of the diagnosis.

The measure has (been in use) by NCQA trade since 2005, and has been used in a number of public reporting accreditation and payment programs.

Now, we really took the feedback that we heard from this committee quite seriously the last time. And, so we decided to do a reevaluation of the measure in 2014 to make sure that we really were excluding the right patients from the measure.

So our first step was to review the importance, what we saw was that the guidelines continue to recommend against the use of imaging studies within the first six weeks unless there are indications of a serious underlying pathology.

In addition, there are now eight specialty societies who are recommending against imaging for low back -- uncomplicated low back pain in the Choosing Wisely initiative. So the number actually grows when you were doing the reevaluation. It was six, so there's even more specialty societies now for recommending against this imaging for uncomplicated low back pain.

We also see that there continued to be a performance gap. So it's between the highest and lowest performing plan, and the highest and lowest performing regions. So although the measure rate has been stable, we do see that there are plan to -- are continuing to achieve better performance on this measure which would indicate the other plans having opportunity to improve their performance.

The real (meet) of the work we did was to review the exclusions. So, first, we went back to the evidence to determine what evidence-based exclusions we might be missing.

As a reminder, what we had in the measure already was exclusions for patients over the age of 50, a history of or current diagnosis of cancer, IV drug abuse and neurologic impairment and trauma. We also, as I said, had a negative diagnosis history of six months. So, if patients were getting a surgery for low back pain in those six months, they'd also be excluded.

The one evidence-based exclusion we saw that was missing from the measure was a history of prolonged use of corticosteroids, so we decided to consider that exclusion. We also ask about what other exclusions we might want to consider during public comment. And we ask our expert panel, our measure specific expert panel, if there were any conditions that may reasonably indicate serious underlying pathology and that are significant enough to potentially have an impact on the measure rate, because again, this is a health plan level measure, we're not trying to exclude any indication of a serious underlying pathology. We really just want to be excluding those patients, or we want to have exclusions that are potentially going to have an impact on the rate across health plans.

So, out of those discussions with our experts, we decided to, in addition to excluding patients who had a history of prolonged corticosteroid use, we also decided to exclude patients with HIV, major organ transplant and spinal infection.

Another thing to note about the exclusions is that we look for a history (of them), they also go through 28 days after that initial diagnosis of back pain, so if an imaging study is done and cancer is found, as long as that cancer diagnosis is on a claim within 28 days of the low back pain diagnosis, the patient will be excluded.

We did not retest the validity of the measure because we had actually tested a number of the exclusions during our original measure testing, so we just provided that additional information in this recent submission. And as you'll see actually, the exclusions that are now in the measure did not have much of an impact on the rate when we did the original testing, but we thought it was important to exclude those patients due to face validity reasons.

As with our other measures, when we have a new measure, we use actual data from plans that get submitted to us to test reliability. And since we just released this updated specification this year, we won't have data for testing reliability. And so, we receive the data from plans that we weren't able to submit new reliability testing either.

To assess the face validity, we met with a measure specific measurement advisory panel, we posted the measure for public comment as I mentioned, and then we also reviewed with our committee on performance measurement who approved our changes and approved continued use of this measure in our programs.

That's the summary. Thank you very much again for giving us an opportunity to present that here.

Katie Streeter: Thanks, Jenna.

So we have three discussants that were assigned to this measure, Christopher Visco, Catherine Roberts and Carlos Bagley. And, we'd like to start our discussion on evidence. And, if I could ask if any of the discussants would like to summarize any thoughts about the evidence criteria and then we will proceed with voting.

Catherine Roberts: Sure, this is Catherine Roberts. Jenna, please pass along and (release) my personal thanks to the NCQA for their effort in evaluating our feedback, much appreciated.

So, for the evidence, you may recall that the ICSI guideline was (graded) as the strong recommendation with a moderate evidence base, and we did suggest updating to the 2015 ACR Appropriateness Criteria, which was also supportive of the evidence that they had already given. But, I think that also strengthens their evidence by having more to back them up.

I think the question for committee was that, you know, they did update the evidence as we asked and it supported what they originally thought and we originally thought too, but do we actually need to repeat the discussion and vote if it's not really changing anything, they've just done what we asked.

Katie Streeter: That's correct, any other input from committee members?

Christopher Visco: Yes, Chris Visco here. I agree, this is an appropriate update to the evidence. I did not think that there's a need to repeat discussion and vote. I think this just supported the already present moderate recommendation.

Karen Johnson: Oh, great. So the way -- this is part of our new maintenance process. So, you know, this is just an update of evidence, seems to be in the same direction as before.

Does anybody disagree and would actually like to (vote)?

Hearing none, we will just take this as this measure passing the evidence subcriteria. So, thank you for that.

Katie Streeter: Great. So now, we'll move onto opportunity for improvement, gap in care.

Catherine Roberts: OK, I can keep going, Catherine Roberts.

So, basically, we have the data presented, the developer added data showing geographic variation and cited the study from the V.A. which found significantly higher rates of MRI in younger adults compared to older adults and lower rates in blacks compared to whites.

And, for the committee, our discussion questions are, is there a gap in care that warrants a national performance measure, and are you aware of other evidence that disparities exist in this area of health care?

Katie Streeter: Thanks, Catherine. We'd like to open it up for discussion. Any other comments from the committee about opportunity for improvement?

Roger Chou: This is Roger. I mean, the data at least from 2010 haven't shown any change in the, you know, in the rates of appropriate imaging. I just wonder if there -- were there any bump when this measure -- I think it was first adapted in 2004, was there -- has there been any bumps since it was first adapted, do we know?

Katie Streeter: Would the developers like to respond?

Jenna Williams-Bader: We would have to go back and take a look at that to see if it has changed.

We usually look at three to five year (then) as we're doing this kind of work. So, we could go back and look back, you know, 11 years.

Roger Chou: Yes, I mean, it's not critical, it's just be interesting because, like I said, there's no -- basically, the rates are completely stable in the last five years. But just, you know, wondering if the measure did have any impact or potential impact when it was first brought in.

Karen Johnson: Jenna, this is Karen from NQF. Could you explain too just so everybody is clear, 2014 is your most recent -- are you wait -- you're waiting on claims data to get more current data, I'm guessing?

Jenna Williams-Bader: It's not that we're waiting on -- we're waiting on the health plans that report on the hundreds of thousands of patients to give us their data in June that reflects -- they'll give it to us in June of 2017 that reflects practice in 2016, and that reflects the spec that we released last summer.

And so, we're -- when we get that, we'll have the nationwide snapshot of what this measure looks like with our updated specification.

Karen Johnson: Thank you.

Jenna Williams-Bader: (OK).

Roger Chou: Yes, and again, this is Roger. Just in terms of the, you know, the questions for the committee, I mean, you know, we reviewed this evidence several times for ACP and other groups. And, I think the data are consistent in showing that, you know, rates of inappropriate imaging continued to be relatively high in various studies and various health systems when people try to measure it.

The disparity stuff is interesting. I don't really know how to interpret that V.A. study. I mean, I don't know if it means that, you know, people are doing more less inappropriate imaging in black persons compared to white, or people just on imaging, black people in general are more often and, you know, those mean different things, I think.

So, I don't really know how to interpret any of the disparity stuff, but I think in terms of the, you know, higher rates of inappropriate imaging, I don't think there's any data recently to suggest that, you know, this isn't an issue.

Christopher Visco: Yes, Chris Visco here. I agree with that. And also supported by the huge variation in geographic performance rates swinging pretty widely -- or depending on (City).

Katie Streeter: OK, any other comments on opportunity for improvement before we vote?

OK. So, since this will be our -- the first vote we're taking on this Webinar, I'd like to just pause to give you instructions for the voting process. And I actually think (Shan) from CommPartners will help us walk through this.

(Shan): Absolutely, thank you so much, Katie.

In a moment, they will advance to the voting slide. When they do that, you will see the voting options to the side of your individual selections. Just simply click in that box next to the answer of your choice and your vote will be registered. Voting is open to committee members only and it will record your vote.

You do have the option to change your vote while voting is still active and open. If you've inadvertently clicked the wrong choice or wish to change your mind, simply click your second choice, it will pull your vote away from the third and register it with your final choice. Back to you, Katie.

Katie Streeter: Thanks, (Shan). So, voting is now open for importance to measure and report, (1B), performance gap for measure 0052.

Male: I can't see the voting buttons.

Female: Me neither.

Male: Nor can I.

Male: Yes, there's no box.

Female: There is no box.

Male: Do we need to use that (other line)?

(Shan): Katie, go ahead and advance to the next slide.

Katie Streeter: OK. Oh, OK.

Male: There it is.

Katie Streeter: Now, voting is open.

(Shan): There you go.

Katie Streeter: Thank you.

(Shan): My pleasure.

Male: Hope the Russians weren't involved.

Karen Johnson: And Kimberly, this is Karen, since you cannot vote through your system, would you mind verbally giving us a vote? Is that possible for you, and I can cast your vote for you?

Kimberly Templeton: Sure, no, I appreciate that, that's fine. And what are my options, yes or no?

Karen Johnson: One is high, two is moderate, three is low and four, insufficient.

Kimberly Templeton: I would say moderate.

Katie Streeter: Great. So, looks like we have 19 votes in, zero voted high, 16 moderate, three low and zero insufficient.

And now, we'll move onto scientific acceptability of measure property and we'll start our discussion on reliability.

Jenna Williams-Bader: Great. So, we're on page four of the document, and under reliability specifications, you'll see that the specifications have been updated. If you click on that, you can see how they've been updated. This also included some integration of telehealth visits, physical therapy. You'll -- I'll point out

that -- again, that the denominator includes patients from ages 18 only up to 50.

We have a value sets Excel attachment that shows what codes are used to identify the patients and the measure is not risk adjusted. So, we'll talk a little bit more about the integration of physical therapy and telehealth visits coming up. But the questions for us to talk about right now are, are the data elements clearly defined, did they include the right codes, do we think the logic and the calculation algorithm is clear, do -- how likely do we think it is that this can be consistently implemented, and are we happy with all the exclusions listed for this measure. I'll open that to discussion.

Mark Jarrett: This is Mark Jarrett. Just on the exclusions, you have prolonged use of steroids and major organ transplant cancer, but with the proliferation of, you know, biologicals and anti-TNFs and other drugs like that, either autoimmune disease or perhaps even just (perceptive) use of those drugs will lower the threshold or concern about infection in those patients.

Wendy Marinkovich: This is Wendy. I also have another question on that, the prolonged use of corticosteroids.

If -- I believe that some plans have pharmacy -- that don't have pharmacy data or they have delayed pharmacy data (is that to) separate -- sometimes the separate benefit. So, I might raise a little concern about being able to identify that when they're looking at the individual trying to get the denominator.

Jenna Williams-Bader: This is Jenna. Perhaps I could speak to that point. We did actually have a conversation about that whether or not we'd want to require a pharmacy benefit for plans reporting the measure because of that exclusion. And, on the one side, we knew that if we require the pharmacy benefit, it would reduce the number of plans who could report.

On the other hand, we have to think about how many patients might get missed because we weren't requiring the pharmacy benefit. And from our testing, at least we saw that the use of prolong -- or that the prolonged use of

corticosteroids was actually quite a low rate. So we -- again, if an exclusion plans don't have to chase the information down, so, we thought it would -- actually we would lose too many plans by requiring the pharmacy benefit. For an exclusion, we expect to have a very, very small impact if at all on their actual rates.

Kimberly Templeton: And this is Kim, if I could also bring up what I mentioned before on the exclusion criteria, the history of spinal surgery. I don't think (spinal surgery) (sponsors could have issued at any point). So I don't think that, at least from a clinical standpoint, that putting a limit on when that surgery occurred is going to be helpful for (this). I would say that we wanted anyone with prior spine surgery.

Male: I would agree with that.

Roger Chou: This is Roger. I just have a question. So, neurologic impairment is listed, which I think is, you know, appropriate. But, you know, a lot of people present with radicular symptoms. I don't have the Excel attachment, I don't know if that was sent with all the codes, but does that include people with radicular symptoms? I mean, it's much more common to have symptoms of radiculopathy than to have actual neurologic impairment.

And many guidelines and appropriateness criteria or whatever would say that, you know, having radiculopathy can be an indication for imaging, so I'm wondering how that was handled in the codes.

Jenna Williams-Bader: So, we actually did have discussions about that with our experts. The guidelines, both the one from ICSI and ACR if -- with just radiculopathy do not actually recommend imaging in the first six weeks.

Mary Barton: Right, that's what -- I think that's what Roger is saying.

Jenna Williams-Bader: OK.

Mary Barton: So -- yes, so, they're right, we -- our codes are limited to codes that would suggest neurologic impairment.

Jenna Williams-Bader: (Cut-off point on) specifically.

Kimberly Templeton: And in the last meeting, we did bring up other conditions that (illustrates) red flags like fever, et cetera, that could be an indication to spinal infection. I assume that either we have spinal infection (illustrates) and exclusion. However, frankly, the diagnosis is not made until you've had the MRI.

Jenna Williams-Bader: So -- and as I mentioned, we actually do take if there are claims for those exclusions in the 28 days after the low vaccine diagnosis, then the patients are also excluded. So if a patient came in, for example, they've had pain for seven days, and you had -- maybe they had a fever but indicated that they should have an imaging study, if you do the imaging study and find an infection and bring them in to treat them within that 28-day period, then the patient will actually be an exclusion.

But, I think for fever, it's a little -- fever can also be a symptom of many other things not necessarily a spinal infection. So, again, we were trying to think about specificity and sensitivity when choosing the exclusion.

Kimberly Templeton: No, appreciate that, but I think it's fine or the back pain in -- when you also have a fever, it'd be very concerning for an infection, but as long as you're being squeezed later on down the road, that's fine.

Thiru Annaswamy: This is Thiru Annaswamy. I have a question along the same lines that many time, the back pain, low back pain, uncomplicated low back pain may be the code used when initially the x-ray is ordered, and subsequently, this may had developed into a radiculopathy with neurological impairment or cancer, or other exclusions. But it may not necessarily happen in the 28-day period, which it's seemingly arbitrary of how the 28 days were -- was decided on.

But if -- and also, I wanted some clarification on that. And I'm also hearing some mixed messages on symptoms of radiculopathy, but no neurological impairment, would that be excluded or would those patients be included in this calculation?

Jenna Williams-Bader: So, if they have radiculopathy but no neurological signs, then they would be included in the measure population.

Mary Barton: Right.

Jenna Williams-Bader: If they have (color point off), then they are excluded.

Mary Barton: And in terms of a 28 days, this is Mary Barton, vice president of Performance Measurement here at NCQA. Sorry, I didn't introduce myself earlier.

With -- you know, for the -- for a measure, for any measure, there has to be a timeframe because you cannot ask health plans to all report something that will be comparable across them unless they're all (hoeing) to the same specifications.

And so we need to specify the time period to look in for an exclusion and a time period to look in for the numerator, et cetera. So the 28 days after the imaging is not -- I suppose you could say arbitrary, to me, I would say, you know, it's a measured decision.

We have to draw line somewhere. Most of those conditions that you're talking about are ones that would -- cancer and a (cold) spinal infection are things that you would hope that your clinicians would be vigorous about pursuing. You wouldn't expect those to be investigations that we're drawing out five months or six months.

So I'm a little surprised to hear the -- to hear your response to the 28-day time window. So I'm -- I'd be curious if you find that in your health systems, that's a -- that it's a common theme that the workup for, you know, (on a cold)

infection or cancer that's presenting as low back pain would usually take more than a month.

Thiru Annaswamy: No, I wouldn't, in response to that and fairness, yes, most of the times, a cancer or something dangerous would present itself fairly quickly you -- one would hope. But, obviously, if you're shooting for 100 percent on this measure, many times, you might lose some of those patients, who may be appropriately (immense), but somehow fell out the 28-day limit.

But if you are shooting for less than 100 percent, which is OK, and we can accept some of those, were genuinely (image), appropriately (image), but somehow they just feel outside the criteria, the timelines, then I guess it's OK.

Jenna Williams-Bader: And certainly ...

(Multiple Speakers)

Jenna Williams-Bader: Sorry. Go ahead.

Jason Matuszak: This is Jason Matuszak. Actually, I wouldn't be that surprise to see a lot of stuff falling outside of four weeks, just because we expect those clinicians that, you know, if we're going to give somebody a fair legitimate shot of conservative management for what we considered to be a mechanical low back pain, four weeks is certainly within a reasonable period of time before you're even seeing the person back for a second visit, and you see them back for a second visit and then you make a determination that you have to do more investigating.

So, do you guys have the data that shows that more (steroid) pathology is picked up within four weeks on a routine basis versus being out at six or eight weeks because that's my concern? And I understand this is a health plan measure, but health plan measures turn around and become provider measures when the health plan, you know, started to ask providers how we're managing these things.

So, that's the question I think that people have, is what data are you presenting showing that, you know, 98 percent of the time that it's something serious. It's within 28 days with that initial claim of low back pain.

Mary Barton: Thank you for your question, and I can certainly appreciate what you're saying about the -- our intents for the measures to be at the level of the health plan, and yet when people find measures that they like, they sometimes use some other places.

When we talk about a 28-day period, that's not from the initial presentation, that will be from the image. No, it would be -- we look 28 days after ...

Jenna Williams-Bader: We look for both the image and the exclusions, 28 days after the initial diagnosis.

Jason Matuszak: OK.

Mary Barton: OK. So, I guess what -- certainly, your point is that there could be a sequence where you thought somebody had mechanical low back pain, and it wasn't until they had failed conservative therapy that you then went on to image them and that you then found the serious pathology. I can...

Jenna Williams-Bader: I think the reason why we kept the 28 days is because we're saying, "You shouldn't be doing the imaging study within the 28 days after the diagnosis." If you do suspect some things, because again, it's a measure, it's not a guideline, we're -- so we understand that physicians are going to use their clinical judgment. If a provider does do -- does order an imaging study, and they do happen to find something, then again, as Mary was saying earlier, we would expect them to act on it pretty quickly because (either) very serious conditions. And that you would still see the claim.

But then if they did wait, then those patients aren't even going to be counted against the provider. If they were doing conservative therapy first, then the imaging study might be done six weeks later, well, that patient won't even be

counted as the numerator hits. So the provider in that case is doing exactly what they're supposed to do.

James Daniels: (Jady) Daniels here with a question. And I apologize (but it's) in there, I didn't see it. When you talk about trauma, you -- are you going to be specific, going to be any trauma? And also, was there a time limit, like, if it was a day versus three months on it?

Jenna Williams-Bader: Sure, so we have -- we did submit the value sets which has more specificity and has all of the codes we use for each of these different exclusions. For trauma, it is a specific list. It's including things like fractures and other indication -- other injuries that might indicate that there would be a fracture that you'd be looking for. We look for trauma within the three months prior to the diagnosis of low back pain.

Catherine Roberts: Good. Any other thoughts or can we continue on to 2A2 because we're not voting right now, is that correct, Katie?

Katie Streeter: That's correct. Yes, we can open it up to other comments before we do vote on reliability.

Karen Johnson: Right, we need to talk about testing.

Katie Streeter: Testing.

Catherine Roberts: Yes. So ...

Christopher Visco: So ...

Catherine Roberts: ... I was going to move into -- oh, sorry, another comment?

Christopher Visco: Yes, I'm sorry. So Chris Visco here. To further the comment on and expand on the comments regarding exclusion, one area that I would like to hear the developer comment on as well is regarding patients with known

anatomic spinal anomalies, (physio) scoliosis or (expandable feces). That certainly is a prior radiograph demonstrated and abnormality or an anomaly.

Again, this is something that would probably fall to area where we wouldn't want to get additional imaging, and this is not covered by ACR meaningfully as far as I can tell.

Mary Barton: So thank you for that comment, you know, we are not orthopedic specialists here at NCQA. So we've really relied on the guidelines that we reviewed. And there was -- there were a -- there was enough of a long list of potential red flags to keep us busy as we did this reevaluation.

And, there was also, you know, you could consider two (axes) on which the whole set of potential red flags can be rare or common, and they can have relatively weak evidence and relatively strong evidence linking them to an underlying pathology.

So, when you arrayed the potential red flags on a graph like that, you find, you know, unfortunately -- or there's just the degree of our confidence about the impact of a particular "red flag" is variable. And especially when you're talking about some, you know, we've already mentioned fever. That when you have potentially nonspecific marker that you want to put on the list of red flags and it has a weak link to an underlying pathology, then we're -- that doesn't sit well with the stakeholders that review, you know, that use our measures.

And so, we were opting to try and find the right balance point for red flags that had enough of an evidence link to be -- to really make people sit up and look. And, we tended to -- if they were operational -- if we were able to operationalize them in claims, we tended to do so even if we knew that they would be rare or super infrequent, because as ...

(Multiple Speakers)

Mary Barton: ... earlier we ...

Christopher Visco: So, understood on certainly that that would pertain to issues (there), you know, areas where there is, you know, a lack of evidence or things that were rare. But in this, in particular, especially with known spinal anomalies, it's not rare at all.

In fact, you know, scoliosis affects about 3 percent of the populations, (enough of) -- especially in the age group that we're looking at here as 18 to 50-year-old set, we're honing in on them. (Expandable feces) when you add that in, it's another chunk especially when you add in prior radiograph demonstrating lumbar anomalies, you're going to get up to about 20 percent of the population once you add all those things in. And then the incidence of degenerative lumbar disease, especially once you get into the 40-year-old set, starts to increase pretty dramatically. I think this might have been a big miss on your graph that you're describing.

Mary Barton: I'm sorry, where does this show up in a guideline? That -- I guess I didn't speak -- I was not clear enough. We are not orthopedic surgeons here, nor physiatrist. And so, when we look to the pool of potential red flags to consider, we were relying on the guidelines that have been published, where I think you started off by saying this was missed. Is that right? Did I hear you say that?

Christopher Visco: Exactly. It's not in the guidelines and it's not -- but it is in the literature.

Mary Barton: Well, thank you. I certainly would commend -- I know there are several representatives of guideline developers on the (expert) panel. So I hope that we would all be able to consider this in updates going forward.

So, we can't really put something in a measure if it's not in the guideline first. It's not NCQA's habit to get in ahead of the guidelines.

Christopher Visco: Right, but we can consider that that's majority of literature as we evaluate this measure.

Thiru Annaswamy: This is Thiru Annaswamy. I just wanted to interject real quick about mechanical low back pain guidelines and the conditions that Chris Visco may have alluded to may not necessarily fall under mechanical low bck pain, because they may have an associated radiological abnormality with them. But if they're coded as mechanical low back pain, and they are not truly mechanical low back pain, uncomplicated low back pain, then this measure may miss those patients, if I'm stating it correctly.

Karen Johnson: So this is Karen. I'm looking at the clock and we still have quite a few things to get through. So, let's go ahead and talk a little bit about testing, since our lead discussant will discuss testing for us.

Catherine Roberts: Sure. This is Catherine Roberts. So, reliability testing, you know, I think the bottom line here is that in the updates, the change in the populations (made) by changing -- adding physical therapy and telehealth visits. Changing some criteria that it would meet the NQF definition of a material change, and that when there's a material change, the NQF tends to require that you have some updated reliability testing using, you know, the specifications as rewritten.

And so, if that's true, preliminarily, we've got this down as insufficient waiting for some repeat measurements. Would that be accurately summarizing NQF requirements?

Karen Johnson: Yes, that would. So we were specifically noting that there were some additional exclusions added in. There's other things also, as you said, were added in.

So, maybe we can ask NCQA. Did you have -- if I'm remembering correctly, Jenna and Mary, you guys updated with some of your original testing, but maybe not for your new exclusions, do you have that data?

Jenna Williams-Bader: Well, I think that we did end up -- when we originally tested the measure, we did test several exclusions, not just the ones that had made into the final measures, but we did actually test. We looked at recent infection,

fever, unexplained weight loss, prolonged steroid use and immunosuppressant. So, we included that in our update here just to show you what the original testing showed.

As I mentioned in the intro, we do reliability testing with actually -- with submitted data from plans that Mary pointed out, we won't have that data until next year to do the reliability testing. So, we -- basically, we submitted what we had, which was the old testing data, and then reliability testing will be available, we'll be able to do that once we have the new data from plans next year.

Karen Johnson: So, in looking -- kind of jumping ahead and apologies to the committee, one of the things that we say is that if testing of data element validity has been done, then we don't require that you do additional data element reliability.

So, your score level, (your waiting) on your claims, so you had a bunch of data. And I think what you provided was percent agreement. Is there any way you can turn those into kappas or sensitivity specificity statistics, those kinds of things? And that would hit our requirements for data element validity and therefore get your data element reliability as well.

Jenna Williams-Bader: I don't know if we can. I think we'd have to take to look at what data we have exactly. I know this was basically the data we had submitted last time. So, we were, I guess, working under the assumption that that was going to work this time. It would -- we would need to go back to our analysis (staff) to find out if we have enough data that actually do kappa coefficients or sensitivity specificity analysis.

Karen Johnson: Yes, I mean, since you've added in the exclusions specifically, we would want to see that those can be consistently or accurately pulled. I -- my guess, if you have that data, but I can assume.

Jenna Williams-Bader: So -- well, we don't have the kappa scores. I mean, what we were able to show is a comparison of the claims data to medical records to show how much -- what's missing from either of those data sources. And so, that's

what we presented in table seven. That was what we had presented for the other exclusions as well, the other data elements. So, we did at least include what we have provided already for the other data elements.

Karen Johnson: Does the committee have any other questions about reliability, the testing or, any other questions for the developer?

Roger Chou: So, I had a question for the leads on the committee. So, it seems to me like, you know, we can't test the reliability, which is fine. I'm more interested in their assessment of the specifications because that seems to me to be kind of the crux of what we need to sort. And, I don't know if those are separated out in the voting, but it seems to me that they probably should be, if they aren't. But I wanted to see what the – what (Catherine) and the other leads thought about that.

Mary Barton: So, you're saying, separate the voting out like the 21 -- or is it QA1 specifications from voting on the QA2?

Roger Chou: Yes, I mean, at least from my perspective, that seems like those are very separate things and at least with the last -- on the -- you know, in the last committee meeting, I think the major discussions were around the specifications.

(Multiple Speakers)

Mary Barton: There's definitely a lot of discussion. You know, it's very challenging to -- as NCQA (said) have a meaningful balance versus a perfect metric. I don't know if NQF really allows us to separate those out as separate votes under reliability.

Karen Johnson: So this is Karen, unfortunately, we do not. Reliability takes into account the specifications themselves as well as the testing. The last time, your votes actually did go through on reliability based on the score level testing that they did last time around. And, the measure -- you actually had trouble with the measure under validity.

So while you were discussing all the exclusions under specifications this time around, it wasn't that the specifications weren't necessarily clear. I think the last time, it was probably more of a discomfort with the exclusions that weren't included, if that makes any sense. And that actually came under your discussion for validity and your voting for validity.

So, in other words, when we take a vote here in a few minutes, we're going to ask you if you feel like the specifications are precise and whether the appropriate testing has been done for reliability. And then we will go through and talk about validity after that.

Roger Chou: Yes. I mean, I guess, if that's the process, then we can stick with the process, but it seems a little odd to me. I mean, you know, it's basically going to be -- I mean, you know, we know there's no reliability testing so it has to be insufficient, I think. But, that doesn't get at all at this issue of whether the specifications are reliable.

So, I -- you know, I don't know if that can be addressed in the future but that just seems to -- those seemed to me to be very distinct issues and just to (mount it) all under insufficient seems a little unsatisfactory. But if that's the process, then, you know, so be it.

Katherine Gray: This is Katherine Gray. I was just wondering, so how does -- I mean, before we vote, what is the insufficient suggest to us, I mean, in terms of our voting? I mean, if -- OK, let's just say from the last time, you know, the committee agreed that it was -- it passed. Does this -- do insufficiency makes some difference in our voting or do we just rely on what we did last time or what?

Karen Johnson: Well, when measures come back to NQF for maintenance, they're expected to conform to our current criteria and our criteria say that if a measure has been changed in a material way, testing has to be updated to reflect that. So, we are expecting data to show that the measure is reliable and valid with the new specifications. And in this case, it's mostly about the exclusions that were added.

So, that's our criteria. And an insufficient would take down a measure if everybody kind of leaned in that direction. I think what's a little bit tricky about this, and I think there might still be time although not so much on today's call. But, we've already heard that in terms of replicating the testing that was done the last time around for reliability, the developers cannot do that. They don't have that data yet. So they can't do that.

What they potentially could do, and they've given you a little bit of a flavor of that, is they could do (and consider) of what we call score level testing, they could do data element level testing. And, again, they've shown in their validity section. They've shown some data element testing and they actually did add in three extra pieces for you to reflect the new exclusions that they provided, so they've given you what they have.

I think the NQF (squibble) with this, is that it's not exactly in the form that we expect. We generally would like to see things like sensitivity and specificity or at least something more than simple percent agreement.

So, they've given you something that probably can get you there, but it's not quite there yet. It might be enough to satisfy you if you're happy with the percent agreement. And again, if you're happy with that testing under data element validity, you can apply that to the reliability testing.

So, I realized that that's confusing for those of you who don't work at NQF on a day-to-day basis, and I apologize.

Female: Are we going to vote on them sequentially, first reliability, then validity?

Karen Johnson: Yes. I think I'll have you do that. Well, let's see. Why don't we go ahead and have a -- let me think about it for a minute.

Why don't we go ahead and vote on reliability as it stands, then we will discuss validity. And then we may decide -- depending on where you land on validity, we may decide to augment a report for reliability.

Roger Chou: I'm sorry, this is Roger again. So, for reliability, are we voting for the reliability of the current updated measure because as you just said, we don't have any data on it in terms of the reliability testing? Or are we voting -- are we using the data that was on the previous measure which we previously said was moderate? I'm very confused about that.

And then what does it mean if we say it's insufficient because we don't have data for the current measure? That seems like it's always going to be an issue with new measures or updated measures. So I'm not -- I'm kind of unclear what we're voting on here.

Karen Johnson: Right. You would be voting on the new measure.

Why don't we go ahead -- let me reverse my decision just now. Let's go ahead and talk about validity and let's talk about what they provided in terms of validity testing, and see if you're happy with that testing. If you are, then you can apply your feelings about that testing to reliability, so apologies. We will swap it out.

So did that make sense? We're kind of taking it a little bit out of order. But we will come back ...

Thiru Annaswamy: That's fine with me. Thank you.

Karen Johnson: We'll come back and get reliability in just a few minutes.

Anybody have any questions before we proceed?

OK, let's talk about validity. So lead discussants.

Catherine Roberts: Sure, it's Catherine.

Karen Johnson: Yes, thank you.

Catherine Roberts: It's -- yes, it's fairly long. We've got sections one through seven. The certain points have already been made which is that at the moment, it would be nice to have some kappa values or sensitivity specificity statistics. There might be some explanation on their validation process just to meet the NQF's full face validity requirements.

So, I guess I would summarize in that -- their data wasn't quite specific enough to meet the requirements of the NQF.

Karen Johnson: Did anybody else from the committee have anything they'd like to say about validity?

So, let's go -- Katie, if you would, let's go to the critical data element testing. And let's just look at the stuff in red that they were able to provide.

Looks like maybe page 50, if you can find that.

And Jenna -- so, it's not that they didn't provide anything, they gave you some stuff. And it looks like they were able to tell you for -- I believe for all of their new exclusions, they were able to tell you -- sorry, I'm trying to figure out.

Jenna, do you want to help me with this table?

Jenna Williams-Bader: Sure, just what it actually shows, I'm happy to. So, on the -- for the exclusions included in the measure, and I believe we had another table like this that was actually for the low back pain guidance as well.

So, here, we were comparing what was in claims to what was in the medical record. So, in the first column on the left hand side with the numbers in it, that ...

Thiru Annaswamy: What table are you talking about? I'm sorry.

Jenna Williams-Bader: Sorry, this is table seven.

Thiru Annaswamy: OK, thank you.

Karen Johnson: And Katie does have that up on the screen, if you can see that.

Jenna Williams-Bader: So the first column shows you the rate of how prevalent that particular exclusion was using administrative claims only.

The next column shows the rate using medical record only.

The third column shows you what the rate is when looking at administrative and medical record together.

And then, the last one shows the rates of patients who did not have that exclusion either in administrative claims or the medical record.

Karen Johnson: So in order validate the claims data against the medical record, should the committee be looking at one column versus the other, Jenna, or -- so the question is, is what's on the claim correct. That's ...

Jenna Williams-Bader: Right. So, yes, I guess if you were to -- since this is an administrative claims measure, then the rate using admin only would be what we're able to capture with the measure. And then the next rates with the medical record only would show you what we might be missing because we're not using medical records for the measure.

And then the rates in admin only plus the rate with the admin and medical record will basically give you the total rate that we'd expect to see because it's the rate coming from administrative only and then the administrative, but also available in the medical record.

Roger Chou: Yes, so this is Roger. So, the data for recent trauma is very surprising to me and it's concerning. If 20 -- if nearly 20 percent of people with trauma, you're not capturing it in a administrative data. I mean, first of all, that number seems really high to me. But if it's true, that's a huge number of potential

exclusions that aren't being excluded based on the administrative data. Is that I'm what seeing here? Is that what it's -- this is saying?

Jenna Williams-Bader: Yes. That is what it's saying. I mean, I will know -- I -- this will probably not help to address your concern too much. This was testing that we did back in 2003-2004. So it is possible that the claims is better now, but you're right. It does show that we would -- at least back then, we're missing quite a few of the trauma exclusions by only looking at the -- or only looking at claims.

Unfortunately, based on the test results we have, it's hard for me to speculate why that was, what exactly we were missing. And I agree with you, that does seem like a high rate, but the results that we had don't really allow us to dig much deeper into that.

Katherine Gray: This is Katherine Gray. I would also suggest that that neurological impairment is a pretty big number too for it's -- you know, for its position in the world, too.

Roger Chou: Yes, I mean, you know, as a primary care doc, I know that when we code low back pain, you know, which coded as radiculopathy. It's not -- it's usually not coded with a specific neurologic impairment code and, you know, I'm not sure what codes you guys were using but, you know, we usually use pretty non-specific codes even if we are, you know, doing that, a lot of times, it'll just be sciatica or something like that.

So, I'm not surprised that we're missing some, that percentage is not as worrisome to me as that -- I mean, that 19 percent is huge. It seems so high to me. I mean, I can't believe that, you know, there's so many people with trauma that were -- you know, and I guess it's how trauma is defined, but that just seems way, you know, I mean, if it's true, it's very concerning. And then I guess the other question is, how accurate that is in terms of real, you know, significant trauma at least.

Jenna Williams-Bader: Right, that's an absolutely fair point. And I do wish I had more details to give you about it. I will say as far as issues that come up because we do have a system that allows us to get feedback from health plans that's called our policy clarification support.

This is not an issue that's coming up, so I don't know if it was an issue that was released specific to the testing we did, the codes we included at the time, but we're not hearing now from plans that we're missing a ton of trauma exclusions. So I don't know if somehow it's self corrected and it's not actually missing as many in the real-world implementation of the measure as it did during testing, but it's a little -- it's hard for me to speak to because I don't have that level of detail from the testing results.

Puja Khanna: Could I ask a quick question? This is Puja Khanna from University of Michigan.

You know, I second what Roger is saying. My question is maybe I am, you know, you can correct if I'm wrong. The ICD coding, I mean, are these diagnoses based on ICD-9 versus 10? Could you, you know, provide some clarification to that?

Jenna Williams-Bader: Certainly, the testing was done with ICD-9 because it was from 2003-2004. All of our measures we have transitioned over to ICD-10, so there are now ICD-10 codes in the measure that we actually use. But yes, since this older data, it's using ICD-9.

Puja Khanna: OK. OK, thank you.

Karen Johnson: So did anybody else have any questions for Jenna about the testing that was done or any other pieces of the validity? As our speaker noted earlier, there is lots of things under validity, but I think the developer did address most of those things.

Very quickly, Jenna, one question that we had, if you had your empirical testing, we just looked at that in that table. But you also talked about your

face validity and it was a little unclear to us even though you guys have a very sophisticated commenting and feedback process.

Can you just explain how that would actually meet our criteria for testing of the measure score in terms of face validity, specifically that the score itself can differentiate quality, that didn't come through in your submissions but maybe verbally, you can make a (compare).

Jenna Williams-Bader: OK. Thank you very much for giving me the chance to do that.

So, we have several different committees and panels that I mentioned that we used in opportunities for plans and stakeholders to provide feedback, and some of them are more structured than others.

So I as mentioned, we do have our policy clarification support that's a year-round real-time mechanism for us to receive questions from -- and feedback from plans and to provide them with responses. When we're doing a reevaluation of the measure like we did for this one in 2014, we did (pull) all the comments that we'd received and -- recently and looked at them to see what issues there might be with the measure.

Then we go to measure specific expert panel. So this is our bone and joint measurement advisory panel. They provide feedback on the measure, suggest changes, discuss the evidence with us. We then post the measure up for a 30-day public comment. And for this particular measure, I believe, during the most recent public comment, we got 55 clients. And then we review that and present a final measure to our committee on performance measurement, which is a large multi-stakeholder panel who actually votes on all of the measures that are included in NCQA programs.

So they actually not only provide feedback the way our advisory panels do, but they do actually vote on whether or not the measures should be included in our program.

So this was one we -- after we took it to them for public -- after public comment, they reviewed it and did vote and approved on continuing to include the measure in our program, which would indicate that they believe the measure is able to distinguish quality, that there's still a gap in care and that the measure meets our measure characteristics including that the measure is reliable and valid, and things like that.

So, I don't know if that helps exactly, Karen. I hope it does. So that's our process.

Karen Johnson: Thanks, Jenna. That helps.

Committee, anything else on validity that you'd like to discuss?

Is everybody comfortable with going ahead and voting on validity and then we'll go back and get reliability?

Katherine Gray: This is Katherine again. The point about the 2003-2004 data was true in the second chart we were just looking at as well, is that correct? So it's all of that admin data, plus medical record (that's all old) data?

Jenna Williams-Bader: Yes, that's right.

Katherine Gray: Thanks.

Roger Chou: So, again, this is Roger. And, you know, the preliminary rating was insufficient for validity also, and I think the reason was because it's an updated measure.

I'm still confused about, you know, if the rest of the committee agrees with that or if that ends up being a vote, what happens to the measure? Does it just stop here, I mean, that wouldn't make any sense to me. So I don't -- I'm just trying to figure out what -- how we're supposed to vote on this.

Karen Johnson: Right. Well, remember, the preliminary rating is simply preliminary. It's done by NQF staff looking probably -- sometimes a little harder and sometimes a little easier than committees may, but we're looking mostly at our criteria.

We have an algorithm for validity and part of that algorithm says that testing needs to be done for the measure as specified, and that we are particularly interested in sensitivity specificity kinds of statistics.

In -- when data element validity is done and that percent agreement isn't quite good enough, but for validity, we also allow face validity and -- but it needs to be done with particular -- particularly looking at whether the score itself is a measure of quality or not.

So, we felt in our preliminary analysis that the data there, even though it's old, that's OK as long as you guys think that, you know, applying 2004 data to 2016 is OK. We just were more concerned with maybe not having kappas or sensitivity specificity. So, that is specific to our current guidance for evaluating testing.

If you guys feel that what is there, even though it might not be this particular statistics that we would like to see, if you feel like that it's answering your question about validity and you feel like that, you know, the data elements are valid on claims, then that's how you should vote.

If majority of you do not lean towards high or moderate, or in this case, I think moderate is what we offer, then that would be, at least for now, the end of discussion. That isn't necessarily the end of the whole conversation because we have more of our process to go.

So it could very well be that if you don't land where NCQA hopes, they might be able to come back and bring you some statistics that would further convince you perhaps, but we'll see what happens when you vote.

Katie, let's go ahead. Any other final comments on validity?

Let's go ahead and look at -- that's reliability. Do you want to go to the next one?

And remember, we skipped reliability because you may find that the data that was provided for data element validity testing is reasonable, and if so, you can use that when you vote for reliability.

Katie Streeter: OK. Voting is open for validity on measure 0052. One is moderate, two is low and three is insufficient.

Karen Johnson: And Kimberly, I was going to try to vote for you. And cast your vote, if you would tell me.

Kimberly Templeton: Insufficient.

Karen Johnson: OK. Katie, I'm having a hard time getting...

Katie Streeter: I can add her in ...

Karen Johnson: Yes, if you would add her…

(Off mic)

…it's not working for me. I'm not seeing the voting.

Katie Streeter: OK. So, with Kim's vote, we have 17 votes casted. Zero percent moderate, 31 percent low, oh, just can't do the math, and 12 people voted insufficient.

Karen Johnson: OK. So right now, it looks like it's landed on insufficient. The measure does not pass on validity.

To wrap up everything, we need to go back and do voting on reliability. And Katie, a couple seconds to find that testing or that slide. For reliability, you

would apply the -- what you felt about the validity testing that we just discussed as well as the discussion about specification. And...

Katie Streeter: Voting is open for reliability on measure 0052.

Karen Johnson: And Kimberly, if you want to give us your vote verbally.

Kimberly Templeton: Insufficient.

Katie Streeter: OK. We have 17 votes. One vote for moderate, zero for low and 16 -- 17 votes for insufficient.

Karen Johnson: OK. Because...

Roger Chou: So, I'm sorry again, this is Roger. I keep coming back to this. But I'm very -- you know, I -- my understanding is that if it can't pass on the validity stuff, then basically, we can't move forward and that is troublesome to me because, you know, the group went back to try to address the comments last time, and a lot of the stuff, it doesn't seem like is really testable or for some of the validity things, I think that, you know, there's ways they could have tested it but maybe, you know, there wasn't awareness, that's exactly what NQF was looking for or what. So this just kind of concerns me about the process.

I don't know if there's -- anyway, so is my understanding correct, if it doesn't pass on the validity, then it can't move forward, but it seems to me that this is a process issue that, you know, in terms of the new measures, number one, it seems to me that this is always going to be hard for new measures to pass this stuff. Because if they are going to have a lack of reliability testing, et cetera, and maybe validity testing as well and then -- but then the other piece of this is that if there were specific things that NQF required for the validity testing, it seems like, you know, that should have been conveyed to the developers.

Karen Johnson: So this is Karen again. It's not really the end, although it is the end of today's discussion. What will happen next is we will go out for public comment and

the measure will go out for public comment. And, we will bring you guys back for what we call a post-comment call.

Now, during that time, NCQA may be able to take a peek at those numbers that they provided and potentially (we jigger) them a little bit to use a very scientific term to see about, you know, what kappas or sensitivity specificity. They don't know if they can do that or not, but they can check and see, and they can certainly bring that back to you.

In terms of, you know, can any new measure or any updated measure get through our process? Sure. We don't have to have, you know, claims data necessarily. You can do testing as a small subgroup which is what they (intended), you know, the last time. So, it definitely is possible. It's -- there isn't an inherent bias against new or updated measures getting through a process.

So, hopefully, NCQA will be able to maybe do a couple things. Number one, maybe get something beyond this percent agreement. We'll see if they can do that. They -- while they're at it, they may also go back and see if there may be some literature or something that might give some input into that trauma thing that seemed to be concerning to the committee. And, they can certainly bring that back at post-comment call.

And, you guys, if you feel like -- that you would like to see some of that data again assuming that they could provide it, then you are certainly welcome to look at this again at post comment.

What our process usually is, and we -- our discussion went way longer than we anticipated it going with the exclusions. And what that means is that we really don't have time, I don't think, to adequately let you discuss the, you know, CMS measure. And I apologize about that.

I think what we're going to have to do is have a second call with you guys so that we can look at the CMS measure. So before we go down that road, is --

would that be acceptable to the CMS folks, or would you rather we try to get through your measure this afternoon?

(Colleen McKernan): So good afternoon. This is (Colleen McKernan) from the (Loewen) group. I think I can speak on behalf of CMS, Yale and (Loewen). I think it makes the most sense for us to try to schedule a separate call, hopefully not before Christmas, but a separate call for us to discuss it because I know there'll be a lot of feedback and a robust discussion can't really happen in the next 12 minutes.

Karen Johnson: OK, great. What we could do with the NCQA measure is we could go ahead and talk about feasibility and usability and use looking forward.

Typically, we would stop discussion and not even talk about those two criteria because the measure didn't make it through reliability and validity. However, since there is a possibly a good chance that they can find some additional data that might be satisfactory to you, we would have to go through those criteria anyway, so we could go ahead and do that.

I think we have time to do feasibility and usability and use. And that way, if they can bring something back, then we don't have to do that on the post-comment call. Is that acceptable to the committee and to NCQA.

Female: Yes.

Thiru Annaswamy: This is Thiru Annaswamy. I would like to ask if we can also take a couple of minutes to talk about, you know, guidelines, the literature talking about uncomplicated low back pain and some of the exclusions that a couple of us were talking about which may not have been included such as prior radiological evidence of deformity.

If those conditions were necessarily included in the guidelines that talk about uncomplicated low back pain, because I'm not sure if I'm voicing the concerns that other committee members may have, but I'm certainly concerned that while it is very important and the guideline is very clear about not overdoing

imaging in uncomplicated low back pain, the way these measures, including the next one that we are not discussing today, have been specified.

I don't think quite gets that, captures the uncomplicated low back pain population adequately, which maybe the crux of the matter. But I'm wondering if we can have a couple of minutes to discuss whether the guidelines adequately capture those conditions.

Karen Johnson: Let's see how our time goes and we might be able to do that. Of course, we don't really have any control over what's in the guidelines or not. So -- and as NCQA mentioned, they do develop their measures based on what isn't (side). So, let's go ahead and talk about feasibility.

And (Kathleen), is there anything about feasibility specifically that you want to bring out?

Catherine Roberts: It's Catherine, right? Catherine Roberts?

Karen Johnson: Catherine, I'm so sorry.

Catherine Roberts: No, that's fine. That's totally good. I'm just checking. I just (threw it there). No, I think feasibility is pretty straightforward, you know, is the data relative -- readily available, can it be captured without undue burden. So, I mean, this is -- is a data that's routinely collected and all of this is, you know, yes in my mind. They've already putted in operational use, so. Comments?

Karen Johnson: OK, let's go ahead and vote on feasibility. Again, we're going on (phase) a little bit, but at some point in the near future, you'll revote on reliability and validity with happier results perhaps.

Katie Streeter: OK, voting is open for feasibility on measure 0052.

Karen Johnson: And Kimberly, if you would give us your verbal vote.

Thiru Annaswamy: I don't think...

Kimberly Templeton: Moderate.

Thiru Annaswamy: ...my voting.

Karen Johnson: Moderate. And we have somebody who's having trouble with voting.

Thiru Annaswamy: Yes, I don't -- Thiru Annaswamy, my box didn't shift, I'm still stuck on the reliability screen.

Female: Oh.

Karen Johnson: (Shan), I think there's a reset, is it F5? (Shan), are you on the line?

(Shan): You are correct. So just refresh your session by pressing F5 or refreshing your browser line. And then you should be able to, if you need to change your vote, just click in the appropriate box.

Karen Johnson: Did that work for you?

Thiru Annaswamy: It says connecting.

Catherine Roberts: While you're working on that, Chris or Carlos, would one of you be willing to take on the usability and use?

Katie Streeter: OK, it looks like we have 19 votes and then -- did we get Kim?

Karen Johnson: Yes, Kim said moderate.

Katie Streeter: OK. We have 12 voted high, and eight voted moderate. Zero low and zero insufficient.

OK, so moving onto usability and use.

Catherine Roberts: Chris or Carlos, one of you would want to take that?

OK, hearing none, I guess it's me again. That's fine, I'm happy to do this, it's just (doesn't need to all of) me. OK, so usability and use. You know, basically, this is just asking, could -- can you use performance results for both accountability and performance improvement activities. And I think as we've already discussed, the rate really hasn't changed in utilization from 2012 to 2014, that we haven't seen any big changes after using this. They didn't find anything unexpected. They didn't think they found anything that was harmful.

And, you know, part of our discussion should be centered around, you know, how can these results be used to actually further the goal of high-quality efficient health care, because that's what we're all trying to get to. You know, will this data help us provide better care for our patients?

And then, do the benefits of the measure kind of outweigh any potential unintended consequences from, you know, perhaps it not being perfect. And then how is that then -- how is the measure been vetted in real-world settings. And, how is it been used.

So, let's open the discussion on usability of the data from this as written.

Katie Streeter: Any additional comments on usability and use?

Female: This is...

Christopher Visco: I'm wondering what the developer has planned for in terms of usability and use on some of the disparities, that was (brought up earlier).

Female: Oh, you still...

Christopher Visco: And some of that (part of) the data.

Jenna Williams-Bader: Hi, this is Jenna. So just to clarify, you're asking what analysis we might be planning to do to identify disparities, is that the question?

Christopher Visco: Yes, yes, because that was brought up as a -- as one of the issues earlier.

Mary Barton: We note that, you know, Medicare Advantage is trying to shine a light on this. They've released data by race on the stars measures. And, we would be very glad if commercial plans, because of course, this measure -- our measure excludes the Medicare population that's younger than 65. You know, if we could engage the commercial health plans who report this measure to us to do so in a way that is connected with a patient level information on race or ethnicity.

And that is something I could assure you that we're working on currently. We think that there is, you know, where CMS has led, perhaps the rest of the country is ready to move as well. So we think that this is something, you know, doable within the next few years, absolutely.

Catherine Roberts: And then Jenna, maybe you comment on like unintended consequences. So, say, you know, maybe the measure is not perfect and it's not capturing everything, but what if one of my colleagues has an unusual patient population that's -- that actually needs imaging but isn't quite falling into the exclusion categories. Do you think there might be unintended consequences of their care being rated as substandard?

Jenna Williams-Bader: I guess (it's) all of our measures, that is -- there is the potential for that. We -- again, that's why we do have our stakeholders engaged in the measure development process and gives them opportunities to provide us with feedback.

It's one of the reasons why we actually ended up including telehealth visit, is because we were getting quite a lot of feedback from our stakeholder, from the plans about how the measure -- about how that might be an issue if we're not including telehealth. And, after talking about it with our experts and thinking about it carefully, we decided it was appropriate to include telehealth visit.

So, that really is one of the main mechanisms through which we tried to get information from the plan, from the stakeholder develop the unintended consequences.

Catherine Roberts: Great, thank you.

Jenna Williams-Bader: Sure.

Katherine Gray: This is Katherine Gray. I just -- I'd like the developers to talk about why they think that there hasn't been a change or what they believe might be a change in the future?

Jenna Williams-Bader: Well, I think that we actually, you know, (seek) for measures that aren't included in certain kinds of programs that plans have quite a lot that they're trying to impact and that measures included in certain programs are going to be the ones getting the attention from plan.

So, you know, Mary just brought up the stars measures and those are certainly measures where we see more investment by plans and more attention being given to them to try and improve rates. Although there are the Choosing Wisely recommendations around this, it's not a measure that's been included in one of those high-profile programs to this point. But I think we are hoping that with these many specialty societies being interested in this through Choosing Wisely, and greater attention being given to overuse now that we would hope that maybe that will drive plans to start paying more attention to this measure and trying to impact the rate.

Catherine Roberts: Any other discussion before we go to voting?

Katie Streeter: OK, hearing none, let's go ahead and vote on usability and use for measure 0052.

And Kim, did you want to give us your vote?

Kimberly Templeton: Yes, moderate.

Katie Streeter: Thanks. OK. So, we have 18 votes submitted, zero high, 18 moderate, zero low and zero insufficient.

I'll move on, OK. So, what I think we'll do over the next couple of days is, we'll setup a doodle poll to figure out what date we will schedule the next Webinar for and work with our developers to make sure they're available as well. And then we will proceed with discussing measure 0514.

So please be on the lookout from an e-mail from me to provide your availability for that Webinar, and we probably won't squeeze it in before Christmas. But we'll do our best to make sure that we find the time that will work for everyone.

And before we sign off for today, I did want to pause and see if we have any public comments if we could open up the lines if there are any public comments.

Operator...

Operator: At this time, if you'd like to make a public comment, please press star one.

And we have no public comment at this time.

Katie Streeter: Thank you. So, thank you all again for taking the time to join us today and please be on the lookout for my e-mail regarding the next Webinar. And, we will keep you all posted on the next steps.

Female: Thank you all.

Female: Great. Thank you.

(Multiple Speakers)

Operator: You may now disconnect.

END