NATIONAL QUALITY FORUM

Moderator: Measure Developer Maintenance
January 6, 2017
12:00 p.m. ET

Operator: This is Conference # 44802032

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

Katie Streeter: Hello. Good afternoon, everyone. This is Katie Streeter, Senior Project Manager here at NQF. We'd like to thank you for taking the time to be with us today for the second Musculoskeletal Standing Committee Off-Cycle Review Webinar.

So today I'll be taking a quick roll call to see who's with us on the line. The main reason for our webinar today is to review measure 0514, the CMS measure, MRI lumbar spine for low back pain.

Before we do that, NQF staff here would like to make a few opening remarks related to the review of that measure as well as the imaging measure for low back pain that was discussed last time that's an NCQA measure NQF 0052.

So I wanted to take a quick roll call. Do we have Roger Chou?

Roger Chou: Yes, I'm here.

Katie Streeter: Kim Templeton? Thiru Annaswamy?

Thiru Annaswamy: I'm here.

Katie Streeter: Carlos Bagley? Steve Brotman? Craig Butler? Sean Bryan?

Sean Bryan: Here.

Katie Streeter: Kelly Clayton? James Daniels? Christina Dodge? Katherine Gray?

Katherine Gray: Here.

Katie Streeter: Marci Harris Hayes? Mark Jarrett? Puja Khanna? Wendy Marinkovich?

Wendy Marinkovich: I'm here.

Katie Streeter: Jason Matuszak?

Jason Matuszak: I'm here.

Katie Streeter: Catherine Roberts?

Catherine Roberts: Here.

Katie Streeter: Art Schuna?

Art Schuna: Here.

Katie Streeter: John Ventura? Christopher Visco?

Steve Brotman: Hi, it's Steve Brotman. I got turned off from the phone originally. I just want to let you know I'm here.

Katie Streeter: Hi there. Thank you. Anyone else that I may have missed? OK. Great. OK. Thanks.

So with that, I'd like to turn it over to Dr. Helen Burstin who is our Chief Scientific Officer here at NQF to make a few opening remarks.

Helen Burstin: Great. Hi, everybody, and thanks and if you noticed that we only have -- we don't yet have quorum yet on this call. It's OK. We would -- if we don't reach quorum by that time we need to vote, it's fine, well just do that offline.

We'll send you SurveyMonkey to do your voting and we'll summarize the discussion for anybody who couldn't make it today. So I don't want you to be concerned about that.

I apologize I couldn't be with you for the first call. It was the same time with our measures application partnership. But I know there were some questions raised as I went through the transcript of, you know, sort of the understanding of what happened the last round, why you're being asked to look at this again.

So I thought I might just take a couple of moments to give you some context for how we got to we were doing these measures at this time. And I'm going to turn to Karen Johnson to give you a little bit more of a discussion about some of the testing issues that we've gone back and looked at for the NCQA measure you discussed a few weeks ago.

So first, I just want to have some context what happened last time when your endorsement decisions for the two low back pain imaging measures went to our CSAC, the Consensus Standards Approval Committee, the board level committee that approves measures for NQF for endorsement.

There were some concerns about the application of NQF's criteria and in particular, the guidance around exclusions and in particular NQF's criteria specifically state that exclusions should be supported by clinical evidence and obviously, that was a very important part of the discussion last time and this time.

But that in addition, there should also be evidence that this is sufficient frequency of occurrence such that if that exception wasn't -- if that exclusion wasn't part of the measure that the results would be distorted.

So there were some questions about how many of the exclusions that were listed would potentially change the measure results certainly beyond the question of phase solidity which is so important. And so the decision was made that they wanted to allow the developers to have more time to go back, take in to consideration all the good advice the standing committee provider

and very important points of view about potential changes that could get made to the measure.

So that's what we're back with at this time, the last time you discussed the NCQA measure and today, we'll talk a little bit more about the NCQA measure in terms of next steps but also wanted to just give you a bit of an overall context for the works that we're doing.

And so the logic again was specifically about the application of NQF's criteria and the frequency of exclusions and allowing the time for the developers to revise the measures which we have now given them.

And as you saw in your discussion last time about the NCQA measure, measure 0052, they did take much of your advice to heart and in fact, made substantial changes to that measure and that health plan level measure.

And I'm not going to turn to Karen to give you a brief update of some of our deeper dive into the testing questions that you raised on the call part of voting at the last call. Thanks.

Karen Johnson: Thanks, Helen. So you guys may have recalled that on the last call when you voted on the NCQA measure, it pretty much went down on these validities of criteria and the votes were five low and 12 insufficient.

So we are -- we believe that the majority of the insufficient note maybe…

Male: Maybe turn up your volume, it's a little low.

Karen Johnson: OK. It's OK.

Helen Burstin: We'll also push the back phone a little.

Karen Johnson: Yes. I want to closer to the phone here. We are wondering if the majority of the insufficient notes that came through the last time on validity for measure 0052 is because we had NQF staff had indicated to you that the results that were provided were insufficient.

And we've actually spent some time looking again at what we would like to see and what they actually did provide and we've changed our mind to be honest with you and we've decided that we no longer really consider what was they had insufficient for data element validity testing.

So what that means is that we are going to ask you to do some reloading on validity and reliability as well. But we're not going to do that today. We'll hold that until next -- until the post-comment call.

So in thinking about this and actually talking with NCQA and what happened, you know, NCQA acknowledges that the information that was provided was based on old testing data I believe from 2002 and they are actually going to try to see if they can find some newer data for you guys to look at in terms of data element validity.

So they're not sure at this point whether they are going to be able to do that or not but by pushing things off until the post-comment call that gives them time to do that and then you will be able to either look at their new data and vote based on that or if they don't have new data then we would just ask you to look at the data that was presented what you have in front of you right now and don't think of it necessarily insufficient because of our guidance that it was insufficient but instead look at the results and base your rating on the results that you see in front of you. So let me pause there and see if you have any questions.

Helen Burstin: And this is Helen. I'm just glad -- so there's no action needed today. We just want to give you some context for the discussion you would likely have on your post-comment call.

No revoting. No re-discussion today necessarily but just the one that at least point out what would like come your way potentially in terms of the ability to look at some new testing data either at the score or the data element level depending on what's available on the post-comment call.

Jason Matuszak: This is Jason Matuszak. So just to clarify, so we're looking at the specific form that you're talking about right now with the validity testing would be the 0052 testing form under 1.6 data element validity testing. Is that what you're talking about that they're going to give information on?

Karen Johnson: Yes. Either in that section or if you look a little bit further down in your form under Section 2b3, they had added in some stuff since the last time and I think that was in red font so you can see what was new.

They -- I'm not sure if they will update that table with brand new step or the step under the testing section. But it will probably be one or the other if they can obtain some more current data.

Jason Matuszak: I'm sorry, is that on the -- is that on the SharePoint site or is that just sent by e-mail?

Karen Johnson: It's on SharePoint. So it is part of the materials that you've presented. And the last time, we specifically talked about Table 7 in their submission form. So that's what you have in front of you.

Again, depending on what kind of data they will be able to provide, they may do something similar to what they did in that Table 7. They may do something a little bit more akin to what's in the 2b2 section or if they can't get newer data, they may not be able to provide anything next. So not quite sure yet what's coming.

Roger Chou: Yes. So this is Roger. Just I maybe help re-orient people. I mean, I think what happened last time was there was quite a bit of discussion about the exclusions and whether we thought that had been addressed adequately.

But I think in terms of the process, what happened is that we went through the algorithm, which as you look reliability and validity, there wasn't reliability stuff because this was (substantive) change so they haven't been able to test it yet but supposedly you can just look at validity and do the algorithm.

And then we got there and we kind of got stuck because, you know, NQF had given us a preliminary rating of insufficient because of the way the data were presented. They don't present kind of sensitivity, specificity, CAPA stuff. They presented information kind of standard deviations and T tests and stuff like that.

And I think as a group, we accepted -- we pretty much accepted the insufficient rating and based on the algorithm you stop there. And I think what NQF is saying now is that there are different ways to look at validity and they felt that maybe they were applying their criteria too stringently and that they're -- I think they're basically asking us to relook at the validity testing to see if we think it's acceptable or not.

And if they can pass there then we can go to the rest of the algorithm. But -- that's kind of my understanding. I think the other piece of this is that we're getting -- we may be getting additional data.

Helen Burstin: Right.

Roger Chou: The data that was presented before was on validity testing based on 2002. So that's why we're going to give NCQA a little bit of time and then -- yes. So I think that's where we're at. So, hopefully, that helps clarify things a little bit.

Helen Burstin: That's perfect. Thanks, Roger.

Katie Streeter: OK. Thank you. So now we'll turn our attention to the review of Measure 514, MRI Lumbar Spine for Low Back Pain. Similar to last time, we're working in and out to have developer present on the phone to introduce the measure and answer any questions that the committee members might have.

We also have two of our lead discussants on the phone, Katherine Gray and Thiru Annaswamy to help us walk us through the measures. So to begin, I would like to turn it over to the developers to give a brief introduction of the measure.

Colleen McKiernan: Thank you. This is Colleen McKiernan. So thanks for the opportunity to speak today about NQF 0514, MRI Lumbar Spine for Low Back Pain.

As I said, my name is Colleen McKiernan. I'm a senior consultant at The Lewin Group and I'm joined by my colleagues Charlie Bruetman and Kelly Anderson also from Lewin as well as several members of the Yale Center for Outcomes Research & Evaluation or CORE team.

On behalf of CMS, Yale CORE and its partner, The Lewin Group, were to maintain NQF 0514, a measure originally endorsed by NQF in 2008 and implemented in the hospital outpatient quality reporting program beginning with calendar year 2010 payment determination.

The initial population for NQF 0514 includes lumbar spine magnetic resonance scan with the diagnosis of low back pain on the imaging claims. We then removed MRI studies from the denominator for patients with a history of a red flag condition. Examples of these red flags include cancer, HIV, trauma and other diagnoses, the full list which is provided in our measure submission package.

The imaging studies included in the denominator towards which antecedent conservative therapy was not performed that is a patient who underwent an MRI of the lumbar spine study for low back pain did not have claims fit evidence of chiropractor, physical therapy or evaluation and management proceeding the imaging study are then captured in the new (order).

NQF 0514 was most recently brought before this standing committee for endorsement and maintenance in the summer of 2014 during which time the committee made several recommendations for ways to test and improve the measure specifications.

CMS, Yale CORE and Lewin all appreciate the opportunity to speak with you today about our recent reevaluation efforts and want to kick off today's

discussion by highlighting some of those updates we've made over the last two years.

Since august 2014, CMS and its contractors have performed a thorough review of the clinical evidence to identify guidelines and articles most relevant to the measure concept and the red flag conditions captured in the NQF 0514 spec. Performed quantitative and qualitative evaluation of the look-back periods for several of the measure exclusions such as lumbar spine surgery, tested different ways to capture the evaluation and management numerator exception to ensure that the type of the antecedent therapy is appropriately specified.

Performed comprehensive measure testing of the NQF 0514 specs, including evaluation of its scientific acceptability, feasibility and usability and convened periodic harmonization discussions with the team from NCQA to align NQF 0514 with NQF 0052 to the extent possible.

Most recently, we provided you with updated testing results for the reliability of NQF 0514 in response to preliminary feedback provided by NQF staff to us last month. Again, we look forward to this discussion and I'm happy to answer any questions you may have about our submission. Thanks.

Katie Streeter: Thank you. I'd like to turn it over to our lead discussants to walk us through the measure and we'll start with Criteria 1, importance to measure and report and specifically we'll focus on the sub criterion evidence.

Thiru Annaswamy: Hi, this is Thiru Annaswamy. The three of us discussed this measure and I'll be happy to lead the discussion. The other members on the lead discussant panel are welcome to chime in at any time. And subsequent to my summary, obviously, we'll be open to discussion with the rest of the committee.

So this script that I was given to lead this discussion essentially follows the questions provided by NQF after reviewing the data provided to them by the measure developer. So going through it step by step, the first step is this an outcome measure or a process measure.

This is a process measure so that hasn't changed. In terms of evidence provided, the evidence essentially is unchanged except for an update. The measure developer provided an update using the 2015 American College of Radiology Appropriateness Criteria on low back pain that provided a systematic review of 12 studies, which included three well-designed studies, two good quality and seven average quality that may have design limitations.

Essentially, uncomplicated low back pain and/or radiculopathy with no red flags were rated as one or two, which means MRI was not appropriate and unlikely to be indicated, you know, in this specified clinical scenario or that the risk-benefit ratio for patients is likely to be unfavorable.

So that is the level of updated evidence provided by the developer. Based on the questions for the committee, the question is, is the evidence provided by the developers updated but directionally the same as presented in the previous review? Does the committee agree and if so, is there a need for any repeat discussion or re-vote on the evidence?

So upon review, my comment is yes, the committee agrees or at least the lead discussants agree and there does not appear to be any need for a significant discussion and/or any need for revote. I'll open it up to other comments from my fellow lead discussants in the committee.

Katherine Gray: This is Katherine Gray. Actually, I had questions. When we discussed this previously, I basically said that I would, you know, sort of be the questioner person on the team as Thiru goes through and gives us the a summary.

I -- can I sort of just backup and also make somewhat of an opening statement to the developers that I believe that the overuse of MRI for low back is an important issue in healthcare. So I want you to know that I'd like to see this measure continue.

However, there are some troubling issues by the fact that there's been no impact over the last six years. So that kind of the background of which and I

think it's kind of goes with the discussants' previous discussions two years ago and more recently.

First of all, do the developers have any just overarching comments to make about why they think the implementation hasn't helped improve things in terms of the data changing?

Colleen McKiernan: This is Colleen McKiernan. We're happy to talk about that now but I know that there are similar comments in the usability in new section. So I defer to NQF if you want to hold that discussion now or we should save it for usability.

Katherine Gray: You can save it for usability. That's OK. I just thought if I got that out, you know, we can deal with this now. But then let's go to the evidence.

I have a question and I just want to clarify that these six variants that you are saying that you are using are -- they're on page 26 on our -- whatever the big packet is called, I don't know what that this is called but anyway it's a very like (echelon) packet. But it's the 2016 submission for what it is that you're using. It's also where the footnote is about the elderly.

Male: Yes.

Katherine Gray: OK. Everybody with me? OK. Variant two in this is part of what causes the problems before and if you read what it says, it says any one or more of the following and one of the ones was elderly and that's what we got into the last time where it might have even said 70 and older or something, I don't remember what it said previously.

But nonetheless, it, you know, it's seems to me that it's a bit confusing at the very least. I mean, if people –if ordering providers are trying to follow this, if you look at this, this is a (seven) which says it is appropriate to do imaging.

And so I think that there -- that this type of guidance may be part of the reason that you're having lack of impact showing up because if they're not clear to the

providers what they should be doing for patients maybe that they're also inconsistent across patients.

Roger Chou: Hi, This is Roger and thanks to Katherine and Thiru for the overview. Before moving on, I just wanted to again just kind of orient people.

So my understanding is that that the measure itself hasn't really changed. So we're being asked to reconsider things essentially I think because the CSAC was asking us to think about the exclusion things and how important we thought they were.

So just to clarify that so people may be wondering what's different between their current measure and what was presented in 2014 and I think the measure itself has not really been altered as far as I can tell. Some of the supporting stuff has but the measure has not changed.

Katherine Gray: Right. I think that's…

Kim Templeton: I hope -- this is Kim. Hopefully, everyone had time to review to the e-mail that was sent yesterday with the synopsis of the discussion from the last time which may help guide what we're discussing today.

Katherine Gray: That's true but I just assumed if it said there was 2016 submission that that was slightly different in terms of evidence then what has…

Roger Chou: Yes. And like I said I think that's supporting information and that's correct. I'm just saying that the actual measure hasn't changed in terms of the specifications and all that. So just so everyone is clear about that.

Katie Streeter: Hi, this is Katie with NQF. We'd like to see if the developer would like you provide clarification on if any and what changes were made to the measures since the last review.

Colleen McKiernan: So this is Colleen McKiernan. I can kick this off and then others from our team or Yale CORE can squeeze in.

So really our primary focus since the last review was to more precisely specify some of the things about what they concerned. So looking at the evaluation and management, numerator exception to kind of see if we needed to tighten up, looking at the look-back conditions for things like surgery since I know there were some comments about that during the previous discussion.

And then one of the challenges that we had last time was that our submission did not include some of the updates that have been like because it takes the timeline from when we submit to when we discussed. It didn't include some of the exclusions that have been added to the measure by CMS but that were like officially in the measure yet based on the timeline for making updates.

And so really it's the -- there's a couple of exclusions that have -- that were added at that time that we talked about verbally that weren't in the packet. For 2014, that are now officially recorded in the documentation as well as our significant testing efforts and then really reviewing that evidence since I know there were some discussion last time about the, I mean, inconsistencies across guidelines for identification of the red flags since there's no kind of gold standard for a document that this like one list of red flags.

So that's kind of a high-level overview of our -- of the works that we've done in the -- where we are since we've last discussed in 2014.

Arjun Venkatesh: Hi, everyone. This Arjun Venkatesh here from Yale CORE. We function as the prime contractor for CMS on this measure and I think Colleen's description of some of the, you know, specification and alterations is valuable.

I think the only other thing I wanted to make sure that standing committee was aware of is since the CSAC decision in 2014, we have also -- it's not just been within this -- the group of asset -- the developers that has reviewed this measure but as a team, we also have a technical expert panel that includes many national leaders and radiology variety, orthopedic surgery, you name it.

A variety of specialties that touch this measure and have reviewed with them the specifications, the denominator exclusion, the potential red flags that

needed to be considered and reassess their phase validity around these elements.

And so hopefully, that provide some additional reassurance regarding the specifications of this measure and its consistency with guidelines including the ones that have come out.

Kim Templeton: And this is Kim Templeton. I guess a question that I have that if you're using guidelines as was mentioned in this, you're referring back to the ACR guidelines which note that the patients in this position with uncomplicated back pain, the first lines of treatment are self-management, NSAIDs, Tylenol, massage therapy, acupuncture, physical therapy which maybe a direct referral.

And this is part of our concern the last time, too, is that there's no way to capture that in claims data. So the vast majority of non-surgical measures and the vast majority of measures that the ACR is recommending can't be picked up in claims data so there's no way to actually follow their guidelines based on this -- based on what's in this measure that you've written.

Thiru Annaswamy: Yes. This is Thiru Annaswamy. I think that is absolutely the fundamental cracks of the issue. I completely agree.

The evidence is there and we would want to be supportive of any attempt at reducing overutilization. But the way the measure is designed is just -- it does not capture it adequately on both sides of the equation whether or not it captures antecedent conservative care adequately or does it exclude people adequately if they have other conditions that are not -- that do not come under the umbrella of uncomplicated low back pain.

Kim Templeton: And this is Kim again and in terms of exclusions, the last time, yes, there was a concern about prior spine surgery but it's currently written that it's spine surgery only within the past 90 days. It's not necessarily a cut-and-dries time period in which people are going to have issues after spine surgery that could be long term.

I would say the same thing with the cancer. I'm an orthopedic oncologist and we see patients who developed metastatic disease to their spine long after the 12-month timeframe that's noted in this measure.

Kelly Anderson: This is Kelly Anderson from The Lewin team. Do you mind if I respond to a couple of these comments and hope I can provide some insights into the way the measure is structured?

Kim Templeton: Sure.

Kelly Anderson: So to start with the discussion around the antecedent conservative therapy, we certainly agree there's a lot of different ways to appropriately manage uncomplicated low back pain whether that be physical therapy or rest and Advil.

Most of the (physicians), they start with an encounter with the physician who would then recommend the appropriate course of treatment. So some of the ways that we define whether an MRI lumbar spine study is considered appropriate or inappropriate for this measure is by looking back in the weeks leading up to this MRI study in order to see whether or not that patient has had an E&M claim with that provider.

You're absolutely right, using claims data, it's not possible to then track exactly what that follow-up care is following that E&M visit but we are using that as a proxy for appropriate care at attempts that conservative management before imaging would be considered appropriate.

Thiru Annaswamy: Thank you for the clarification. If I may have a counter to that, this is Thiru Annaswamy again. The issue is it does to some extent service a proxy and I think we do agree with that overall.

But issues that were raised and we can probably get into this in a little more detail when we talk about validity but the issue that was raised in the previous discussion was, you know, for example, self-referral or if it was a telephone encounter and a prescription called into a pharmacy or electronically

prescribed and continued referral to acupuncture and other treatments mentioned earlier by Dr. Templeton.

If those occurred and they were considered adequate according to the evidence supported here but did not get captured by an E&M visit, you would be missing all of that in the consideration of antecedent conservative care.

Charlie Bruetman: Can I clarify? This is Charlie Bruetman from Lewin. One of the -- while we understand, you know, (clubbing) and we can -- we're not going to have a long discussion about the charges with the claims as we know that is an ongoing issue and we -- if question with claims and challenges, we're trying to avoid and this is expert panel's perspective and it was the perspective also when it was approved and endorsed and re-endorsed by NQF was the need to avoid let's call -- what we would call or I think to physicians I think you have called it may be quick trigger doctors.

We are trying to avoid a person that goes to get an MRI to go to a doctor with a low back pain and immediately get an MRI and this is because the physician either lacks understanding or clarity to make that decision and therefore does immediately an MRI and it leads to a lot of consequences as described in just few medical associations in choosing wisely.

And that's why one of the proxies this is having was we understand that we can't capture what was done if they got Advil, if they rested, heating pads or nothing happened, there's a proxy of saying there's a -- we looked back to see what happened from the day of the MRI backwards to see if there was some intervention that would have led to antecedent therapy.

If the doctor ordered -- if you go today to the doctor to -- well, you go and you get MRI today and we see that they got a visit like two days ago but nothing had happened before, the intuition would be that that was immediately ordered at that time and nothing had taken place and that what was inappropriate.

And we're not here to determine which antecedent therapy is appropriate or not. We're not here to determine if Advil is better than Tylenol or is better

than heating pads. The question is we want to avoid the doctor that said, here's MRI because you have low back pain, get it tomorrow.

And that's what we're trying to determine by looking back at claims between the days I think it's 28 to the day 60. And then we do address some of the other concerns about low back pain and exclusions, you know, at the top and I think it was at the committee last time.

Yes, there's a discussion surgery, prior surgery if somebody had a surgery five years ago. We determined immediately that they're appropriate. The discussion was that it had to be related to a surgery and that's why the -- well, that's the surgery -- the 90-day period to see if it's related to the surgery that had taken place instead of a previous surgery that took 20 years ago.

Kelly Anderson: Yes. I'm certainly happy to elaborate on this during the discussion of validity as well but we did conduct some additional testing going back and comparing what a 90-day look back versus a five-year look back.

We've considered that five years is about the maximum you can get using that care claims data and we've only saw a few additional cases using a five-year look back versus a 90-day. So, really, it's (just a significant difference) in the performance where it's coming out of this alternate time periods.

Karen Johnson: Thanks, everyone. Any other comments before we move to agreement that there's no need to revote the evidence this time around?

And again, we most likely will not be voting live today like we did last time. We will be voting using an online survey tool due to the fact that we don't have quorum. And if no other comments then we can move on to gap in care opportunity for improvement.

Thiru Annaswamy: Yes. This is Thiru Annaswamy again. Like alluded to earlier, the performance gap actually demonstrated worsening of performance from an inappropriate use of MRI rate of 32.5 percent in the 2009 year to 39.5 percent in the 2014 to '15 year according to data presented here.

The question is is there a gap in care that warrants a national performance measure? Yes, there is a gap in care but the question that was raised earlier why is this worsening and if the theoretical claim that this limits or reduces MRI -- inappropriate MRI the trigger-happy doctors is true then there should have been a movement in the opposite direction. So that is a little troubling trend if you ask me.

The second question is does this measure provide information to understand disparities in this area of healthcare? So a comment that I had was are the disparities related to the availability and access to conservative care options as opposed to the actual trigger-happy doctors being more prevalent in areas of poor healthcare access.

If so, does this -- if the disparities are not related to availability or if they are related to availability and access to conservative care, does this measure lose a little relevance because the reason why the performance gap in areas of disparity maybe worse is because of access not because of increased presence of trigger-happy doctors in those areas.

So those were my thoughts on it and I'll open up to the rest of the lead discussants and committee.

Katherine Gray: This is Katherine. I guess along with Thiru's comments, I just wonder from developers had they thought about trying any smaller scale trial to find out more about what are the reasons that this seems to not be working the way you would expect and you're actually getting more inappropriate care.

Female: So I think there's a couple of different factors that fit into this and some of this is a little bit of the color that goes beyond just the standard questions on the forum.

Since the implementation of this measure, there's been a couple significant differences in the way that it's calculated that I think affect these numbers. The first of this is that responsive to the ACR guidelines, we've continued to add a red flag exclusion conditions.

So when the measure was initially implemented back in 2011, there was very limited list of exclusions and based on feedback from the test and feedback from organizations like NQF, we have continued to expand that list to make sure they're only looking at cases of uncomplicated low back pain. It was also…

Thiru Annaswamy: Sorry, can you pause there? Shouldn't adding more exclusion should make those numbers go down not go up, should it?

Female: It depends…

Thiru Annaswamy: If you include people who might be getting appropriate MRIs that were previously categorized as inappropriate that should make those rates go down.

Female: That will be the case if you're seeing a proportional shift but it doesn't -- when you remove cases from the denominator, not all of those were going to be inappropriate studies and so some of the hospitals are shifting disproportionately to others.

Katherine Gray: Can you explain that a little bit? This is Katherine. I don't quite understand that. What are the hospitals doing?

Female: Sure. So when you're removing cases that are red flag conditions, in a perfect world where providers were only imaging appropriately and you removed those red flags, you would expect performance scores to go down.

In a world where there's (a norm) of a mixed bag which is what we're seeing with the performance score -- median performance score around that 35 percent to 40 percent mark, providers are not only imaging in cases of red flag conditions and for some red flag conditions because we're being more broadly inclusive, they may actually not have been imaging in all these circumstances.

So for some hospitals, you're moving more cases from the denominator relative to the numerator than you are in other hospitals. And so it's not kind of our clear movement from (remove excluded cases) everyone's performance score goes down.

Kim Templeton: And I think one the things to -- this is Kim again, one of the things to potentially going to look at is the medical legal climate because I would say the vast majority of these are, you know, would -- anecdotally would say that they're not "trigger-happy documents." These are people that, in our current medical legal climate, are trying to protect themselves or trying to take good care of patients and make sure that nothing is missed and that they're not sued.

Colleen McKiernan: Absolutely. I would say this is -- we know this is not -- having a higher score is not necessarily finger pointing. It's more saying that you're imaging more than other facilities.

So there's a relative comparison and because that -- as you mentioned because of medical legal concerns, because of a variety of great area of things that can't be quantified using claims. There's a climate out there that is going to driving everything up or down that some of that can be measured in claims and some that cannot.

So I think the real intent of the measure is to encourage providers and the hospitals in which they practice to really think about whether or not the imaging is appropriate at that time and whether or not other therapies have been tried first.

I do also want to talk through a few other changes that we've seen since the implementation of the measure that made drive this change in performance score. We did change or CMS changed calculation sources for this measure from one claim database to another and the largest spike that we see in performance score is in between those two years with the two different data sources.

And so even though hospitals have equal comparisons within a year moving from one data source to the next from one year to the next did lead to spike in the scores. And since we've moved to that new data source, we actually do see a small reduction year over year in the rate of overuse.

Jason Matuszak: This is Jason Matuszak. So let me give you a real pause as to the validity of your claims data overall to have that big of a change between two different data sources?

Colleen McKiernan: So I would that's a little bit outside of the bar where we are not able to fit correct to that. I think that one of the things that -- one of the good things I take from that is that all of the measures removed over that data source so there's equivalent comparison across measures within a year.

It's just when you look longitudinally especially given the changes in the exclusions and then the changes in the data source that say that, you know, there's been some -- there's -- the discontinuity is explained by a couple of those factors.

Jason Matuszak: No. I get that. I guess when we're trying to evaluate something that's, you know, that that's a process measure, that's not (Malcolm's) measure that, you know, we're trying to look for these sides of things and to have that big of a question of the validity I think is where a number of committee members or at least myself, you know, have worries.

Colleen McKiernan: Absolutely. It's understandable. That's why I'm glad we're discussing it.

Karen Johnson: Any other comments on opportunity for improvement before we move on to reliability?

Katherine Gray: This is Katherine again. Sort of have you guys considered any other smaller scale studies to figure out what's going on inside of this like doing something for those 50 percent that are, you know, supposedly inappropriate and seeing more about -- you know, last time we talked about for the other measure, we talked about trauma and trauma is in that variant, too, also low velocity trauma.

So it just occurs to me that there could be some things that are unclear for the ordering doctors to know what they should be doing and maybe we -- that's part of the problem.

Colleen McKiernan: I think that's a great point. Over the course of the last year, we have spent a lot of time trying to dive into what's driving the differences in performance.

We spent a while looking at differences and performance score for hospitals to grossly reduce their total volume of imaging versus those that stayed relatively steadier and even increased that imaging.

So we have spent some time looking how volume drives score. We do each year also look at factors like urbanicity of the hospitals, what their teaching status is to understand who they vary across some of those parameters.

We haven't taken a look specifically at how particular exclusion conditions might be adjusting the scores. We looked at the volume of exclusions but not how that adjust the performance score again but I think that would be an interesting question to look into the future.

Katherine Gray: Thank you.

Karen Johnson: Thanks. Thiru, do you want to move on to reliability?

Thiru Annaswamy: I'd be glad to.

Karen Johnson: Thanks.

Thiru Annaswamy: So going into reliability, the numerator statement for this measure is that the number of spine MRIs in patients with a diagnosis of low back pain without a patient having claims-based evidence of prior antecedent conservative therapy which was clarified earlier has to include physical therapy, chiropractic care and RA evaluation and management code.

The denominator includes MRIs in patients with a diagnosis of low back pain on imaging claim in a hospital outpatient department on -- in a 12-month time window.

The denominator exclusions include, as we discussed earlier, spine surgery within 90 days prior to the MRI; cancer within 12 months prior to the MRI; congenital spine or spinal cord malformations within five years; inflammatory autoimmune disorders within five years; infections within five years; spinal vascular malformations or subarachnoid hemorrhage within five years; spinal cord infarction, neoplastic abnormalities, radiation therapy, spinal abnormalities associated with scoliosis within five years, all within five years; syringohydromyelia and postoperative fluid collection soft tissue changes, syringohydromyelia within five years and postoperative fluid collection within one year; trauma within 45 days; IV drug abuse within one year; neurological impairment within one year; HIV within one year; and unspecified immune deficiencies within one year and to spinal abscess as an exclusion diagnosis in one of the diagnostic fields on the claim.

It also specifies here that no updates have been made to these specifications since the last review in 2014. I attempted to look through -- glance through the attached spreadsheet for the ICD-9 and 10 codes since we last had this review, the 10 has come in and there's been spreadsheet attached.

The calculation algorithm is pretty straightforward and then the questions asked are -- to us are, are the data elements clearly defined, are the appropriate codes included and my review revealed that neurological impairment, which is an exclusion, is not adequately captured in the ICD-10 and ICD-9 codes in the measure, which includes things like radiculopathy, spondylolisthesis, spondylosis codes which may have radiculopathy symptoms.

If they do not have a specific code for neurological impairment, those conditions for which the evidence supports appropriate use of MRI are not adequately captured in this measure. I can go through the rest of the questions or we can stop here for further discussion.

Katherine Gray: This is Katherine again. I mentioned my concerns earlier with the surgery and cancer. The other ones that concerned me on our list of exclusions, congenital spine and spinal cord malformations.

Those are typically diagnosed at birth so kids may not have issues with it. So if they sent to me within five years, are you talking about a five-year-old presenting with back pain?

So I don't understand the five-year limit. I would say the same thing with the spinal abnormalities associated with scoliosis. That can be diagnosed anywhere from infantile scoliosis to most commonly adolescent scoliosis.

Kids that have those issues typically don't have back pain. So again, I don't understand the five-year exclusion. You may have scoliosis when you're 15 but you may not have back pain until you're 60. That would then take them outside of this exclusion once.

So again, those are two another couple of things I just -- I don't see why they are set sort of limit put on the exclusion because clinically, it doesn't make any sense.

Colleen McKiernan: So this is Colleen McKiernan and I can speak to the -- your most recent comments about congenital spine and spinal cord malformations as well as scoliosis.

So the five-year look back is a limitation of our access to data. Unfortunately, we don't have data for a lifetime as a person. And so what we're doing is when we say five-year look back that's the maximum number of claims you have. We're just looking for any claims-based evidence of that diagnosis.

So someone were diagnosed with scoliosis when they're 12 and then in their 70s, they're having an issue with their spine. If the provider document that they have scoliosis and there's a Medicare claim within the five years of the scan, then it's covered.

So even if the diagnosis is decades old for a Medicare beneficiary, so long as there's a -- so long as there's a claim within five years, for Medicare, just marking that diagnosis, we'll go ahead and exclude the patient. To the counter of that, though, indicates where the patient had scoliosis when they were five and it's not relevant to their current encounter and we're not seeing them on

the claims data which you wouldn't expect to see unless it's related to the reason the patient is seeing that provider then that's something that we're not -- I mean, you actually wouldn't want to see that full history of the patient.

Male: Any comments from the developer regarding the point I raised about radiculopathy, spondylolisthesis, spondylosis, those that are not considered in the clinical realm as uncomplicated low back pain but would not have a diagnostic code of neurological impairment attached to it, how would those conditions, you know, they would be obviously not excluded but they would -- they would be in those conditions MRI would be indicated clinically speaking?

Kelly Anderson: So I think that's helpful feedback on potentially codes to consider adding to the list of exclusion. I do want to clarify how patients end up in this measure. The denominator doesn't only look for cases where the patient has an MRI lumbar spine study. The patient must both have an MRI lumbar spine study and have an opt-in code so they must have one of those ICD-9 codes that are categorized as low back pain in order that they can -- they're considered for inclusion in the measure. And then from that set of patients, we would otherwise consider to have uncomplicated low back pain, we then pull out red flag condition.

Colleen McKiernan: And I would -- I would add that the list of ICD-9 and ICD-10 codes that you reviewed was vetted by our internal commissions. We brought kind of a higher level list to our expert panel but your feedback is well taken and I think that, you know, we're going to take it back and take a look and make sure that if there is an error in the crosswalk, that we correct it.

Arjun Venkatesh: Also, this Arjun Venkatesh here from Yale. I think one thing that may be reassuring to the steering committee on these types of issues whether there are, obviously, numerous clinical scenarios we can reflect on and think about at the patient level that may be challenging to perfectly measure in claims, but that we can do a quite a good job mentioning claims, this neurological deficit example being one, is that I would think that -- two things, I think, are important.

The first is that there are -- is when we have gotten these scenarios in the reevaluation of measurement and we've looked into various scenarios in the past, one of the things that we find in general is that there are so many denominator exclusions in this measure that it is often highly unlikely that there is another patient that has an exclusion that is not captured by claims that does not also have one of these other things. And so, you know, in this example, you'd have to have a patient with a neurological deficit that has no cancer, did not have surgery, a variety of other things that put them at higher risk that are also red flags because red flags often don't come as one, they often come as pairs, right? These are often highly associated and related to each other. And so they tend -- when you have these individual conditions, not to have much systematic effect on the measure and I think that's reassuring. I think that having the TAP review, it is reassuring.

And then the third thing I would think about it from the perspective of the validity of the measures, its ability to classify a hospital outpatient department's performance on MRI imaging. You really have to believe that that systematically now distributed between facilities. And if not, then the measure is still able to identify those high outlier facilities were imaging is more prevalent that where we'd want to see increased efficiency.

Roger Chou: So, this is Roger. Just a couple comments. You know, getting back to what Thiru said about -- I mean, sciatica radiculopathy is one of the most common reasons why we order MRIs for back pain and I don't know if that's counted. Oftentimes, there is no neurologic impairment, it's just pain.

And I do not know how that's being -- what kind of, you know, it just depends what your -- if you're -- if you're calling that neurologic impairment, then all those are excluded. It's just not clear to me what's being done with that but that -- that's relatively common and it is not -- it often doesn't come with other stuff. I mean, isolated sciatica, herniated disc is, you know, 5 percent to 10 percent of people with low back pain. So just, you know, some clarification there would be helpful.

I still have some questions about the age, you know, not excluding older patients. I actually happen to agree that that age is a relatively weak predictor

but it is a predictor and this makes the measure inconsistent with guidelines. And so I'm having some trouble kind of reconciling that.

My other comment is just about the ability of a measure to identify people who have had chronic low back pain because that is another indication for MRI in most guidelines and, you know, somebody can show up with back pain that's is been there for two years. I do not see that there is any way for this measure to pull out those people. And you know, this is, you know, clinical practice. This is relatively common. We see people who have had back pain for a long time and you make a decision to image or whatever. It's not, you know, it's not somebody that's been coming in regularly necessarily for that issue. These are kind of ongoing, you know, chronic things.

So just a few things that, again, I've been just having trouble trying to reconcile them in my head in terms of the guidelines and the clinical stuff and what's been the exclusions.

Jason Matuszak: This is Jason, Roger. I'm just -- because I'm looking at the Excel spreadsheet right now and just to clarify what I'm seeing and I don't know if this helps to explain what you're talking about but as of as what I'm looking at here in the ICD-10 codes is that if you have spondylosis with radiculopathy or sciatica, what I'm seeing is that that's counted both in the numerator and the denominator. So just sciatica, without evidence of neurologic impairment, would be considered inappropriate use. Am I interpreting that correctly, measure developers?

Colleen McKiernan: I don't have a value set in front of me at the moment. I know that the ICD-10 codes are still in draft and refinement so that's something I can certainly go back and check and get back to you on but I'm not sure off the top of my head.

Thiru Annaswamy: This is Thiru again. I don't think the crosswalk is an issue. This was there in the ICD-9 set as well. When we discussed the measure in 2014 and they just cross walked over to the ICD-10 set as well. So those diagnostic codes have not been removed as specified earlier. No updates have been made in the specification since the last time. So they remained a problem

then, they just remained a problem now with the -- with the new title called ICD-10.

Colleen McKiernan: This is Colleen McKiernan from Lewin to tackle a couple of the things. So, first the neurological impairment code. Our note taker is actually writing down the codes about what your concern because we're going to circle back to those as soon as this meeting gets over to make sure that if there is a problem with the code, as you mentioned in both ICD-9 that that's corrected.

For the exclusion, as you've said, no changes have been made to the measure. As I noted earlier in the call, the -- during the 2014 meeting, we had changes that were pending that we discussed verbally which we talked through with the committee and I can give you a list of the things, things like -- let's see -- inflammatory and autoimmune disorders and new plastic abnormalities and a series of other exclusions that were discussed verbally but not in the package that we submitted back in 2014 just because of the timing of ntep review versus updates made by CMS such as to clarify that in terms of changes that have been made.

And then finally, for the old -- for the older adults situation. So as you noted Variant in ACR appropriateness criteria for low back pain, it does indicate that one or more of the following, including elderly individuals, I wanted to clarify that when you go back to the literature review that accompanies the appropriateness criteria ratings, it does state that there's a recent study that found no statistically difference in primary outcomes after one year for older adult. So a spine imaging within six weeks after an initial visit for a care for low back pain versus similar patients who did not go undergo early imaging.

And so the panel that review -- that developed the appropriateness criteria states that, quote, this panel does not include age older than 50 as an independent red flag. So it's old -- I think the description of the variant is a little too succinct because it's older patients are a consideration that a provider should take but it should not be an independent red flag. So an older patient with symptoms of something else that is a red flag is cause for a greater concern than a younger patient with a symptom of a red flag but it's not -- you

can't independently -- you shouldn't independently just image adult over 50 if they have low back pain.

Male: So this -- that's not quite a correct interpretation of the evidence. I mean, older age is a risk factor for red flags by itself. What the study was looked at was whether imaging impacts outcomes in people with older age. That's a different issue. And as I said, older age is a relatively weak predictor but it's a predictor. It's always been recognized as a predictor.

If you look -- go back and look at (Rick Dio's) studies and everybody else's, I mean those are the people that end up having cancer that is undiagnosed. It's people who are older. So they don't have a previous cancer but that's how you pick it up, right?

So and the issue about the effects of imaging, you could say the same thing, probably about any of these conditions that if you try to do a study that looked at whether people, you know, who have had a, you know, history of cancer, whether doing imaging impacts outcomes versus not doing imaging, you probably couldn't find any clear difference. I mean, it's -- that, I think, is kind of a red herring. And, I said, it's -- I think -- I think that's over interpreting the evidence and as I -- and if the guideline people thought it was so important, why did you put it in the guideline? I mean, that doesn't make sense to me. So, like I said, it think there's a -- there's a problem that this is inconsistent with the published guidelines and really presents a mixed message to clinicians.

Charlie Bruetman: So my question would be -- and I have a few comments that were raised and I think it's somewhat different perspective on the measure. But, I still think -- this is Charlie Bruetman from Lewin. So, in the perspective, that would be that adult would -- with low back pain. The first episode, they say low back pain if they're over 650, they would be appropriately scanned.

Male: No, it's not saying that it's appropriate but it could be appropriate. You know, we're not trying to tell people what to do, we're trying to measure whether what's being done as appropriate or not. I think that's, you know, it's -- when you -- when you're not including something, it's not that you're saying that

that's necessarily an appropriate thing to do, it's just that we can't measure accurately when it's appropriate and when it's not and there are certainly times when somebody who comes in, who's 65 years old and is losing weight and is, you know, having night sweats which you're not going pick up in the NM codes, it would be very appropriate to do an MRI. I mean, who would argue with that, clinically?

(Crosstalk)

Male: …when you might do a trial therapy. But the point is that you need that, you know, if you're going to be consistent with the guidelines, you should be consistent with the guidelines.

Male: Can I -- I have two concerns there. One is, as you've said, you know, we respect and we think it's part of the measure that takes into account and, you know, we understand that there's a challenge with the client's base measure but in that case, as you've said, if somebody -- the doctor sees all the -- they had weight loss, you know, there is another compounding factor that makes the elderly with a weight loss that maybe (is not on the) younger, so there's another and we're not restricting that physician decision-making.

I also think one of the -- the intent here, as we mentioned before and as, I think, was repeated, the quick trigger, I think the goal here of this measure, we're not expecting it to be a zero, actually. We're not -- and that's why the values are -- and I think we need to look at the child with MRI lumbar spine. I think we all agree in the medical community and provider community has agreed that it's an overuse. We identify that there's an overuse because there's doctors that immediately order because I don't understand the rationale for this or when it's appropriate.

And then there's also the challenge we are facing with a measure that has a huge range of performance. And what we're were trying to do by having this recognizing that it's not a perfect measure and claims data will always have challenges is we're not expecting to be zero, we're trying to avoid quick triggers and avoid and try to reduce that large range of performance.

In addition, I mean, just for the measure, we have to understand it's a measure with low burden, so it's not that it's created a significant burden on providers and which also, just to clarify, it's a pay reporting, not a pay for performance. So we're not looking at them and finding you're above average, you're (penalized). This is how do we get practicing communities, in this case hospitals, to look at where there has been saying, why is -- let's say on average and we know it's high at 40 or 35, why am I at 72? That's what we want this measure to create. Just to have the people to look and say, it seems that we have a lot of doctors, yes, we might have been a couple patient that, you know, the cancer or they had scoliosis when they were 12 or some other factor, but there's a -- there's a reason that not everybody is at 40. People are at ranges and that is what we're trying to reduce saying, hey, you're at the very high end, what am I doing differently, what we should be looking at, what would guidelines help us to reduce that obvious range of performance because of everybody went at 40 exactly we'd say, hey, obviously, there are some concerns, we can't get it to zero but everybody's there.

And I think by not having a measure that looks on MRI from claims would, you know, not help the medical community looking into their practices recognizing that there are going to be challenges, it's not going to be a perfect world and we're not going to identify everybody from the perfect claims but I think that's something that -- that's something we need to consider when we look at this measure.

Kim Templeton: And I think we agree that this is an issue of concern. However, based on the -- but I still think -- but you've heard all the concerns that have been raised by the committee members and looking at your -- to your differences in performance up to this point, how do we know the people on the high end who were ordering more MRIs, perhaps, are providing better quality of care just because you order fewer MRIs does not necessarily equate with -- with better quality of care.

I mean, it was mentioning that what about other conditions like, you know, fevers, chills, weight loss, so you know, you've got an intraspinal abscess listed as one of the exclusions and it needs to be elicited on a lumber spine claim but that may not come up until later and so if you have somebody that

comes in with an acute intradural abscess or an acute discitis, you're not going to know that until you get the MRI and, again, even with that, it may take a while to make the diagnosis.

John Ventura: This is John. I have to disagree with that in the context that I don't think there's any question that over utilization of MRI has been well-documented in the literature and that using that, in many cases, does not change or improve the outcomes and I think -- you have to agree with the previous doc's comments about the way to -- the intent of the measure and that we may be risking throwing the baby out with the bathwater if we get too hung up on trying to make this measured, you know, be able to score a perfect score, right? You know, there are going to be exceptions always sin healthcare and as long as the exception rate is low enough, it's not going to impact scores and people's ability to come out in a good way on this measure. So I think we're running the risk of, again, throwing the baby out with the bathwater.

Katherine Gray: This is Katherine Gray again. I honestly don't -- I mean, you know, to set our standard at perfect or near perfection or whatever, I don't think we're trying to do that. As I look at all of these things, what we're troubled by is we seem to be missing the target. There's something wrong with what we're doing and that's, I think, what the panel is struggling with. How do we improve this? How do we make it? We know there's a problem. We know there's a gap. How can we actually do something better?

And what we're -- at least, for me, maybe I can only speak for myself, I'm concerned that if we simply move forward, we're just continuing for, you know, doing something that useless, you know, that it isn't really dealing with the issues that we need to be facing. I realize all the constraints of claims data and so on but, you know, even trying to measure better, you know, sort of the homogeneity within facilities or, you know, doing something, you know, are we reducing the variant of what people are doing or can we find a way to be more clear within the ordering doctors, you know, what the rules of the road are. I mean, what can we do that will actually help make a difference in the real world with patients?

(Crosstalk)

John Ventura: I'm sorry. I'll just make one other quick comment. This is John again.

There was a wonderful study by John Mafi in the JAMA Internal Medicine in 2013 that looked at -- looked at guideline concordants for back pain. And the information was disseminated and yet in surveying physicians -- gosh, I can't remember, 500? Half PCPs, the other half specialist who deal with spine, it was horribly guideline discordant. So, you know, it goes way beyond the issue of just this -- is this measure going in the result we hope that would get. It would have to be how well are we disseminating the information in implanting the information of the guideline. And again, I have to go back to that statement to blame the measure for the fact that providers are not compliant. It may not be an issue of the measures qualifications but more just the nature of healthcare today.

Roger Chou: Yes. This is Roger. I mean, I completely agree with that. I mean, I think that whether people are doing more inappropriate MRIs now than before, we may have very little to do with the measures and I don't think that, you know, there's, you know, there's all sorts of things that impact clinical practice. So that's not where my concerns or issues have been.

And I agree with everybody. I mean, I've written many of the articles that talked about overuse of imaging as well as many of the guidelines that talked about it. I think the problem is that when you have a measure that isn't, you know, that includes people where, say, older patients, where many of the guideline say, you know, MRI may be appropriate, how do you interpret the results? I mean, I have a really hard time understanding what 30 percent or 40 percent means when it's -- it includes, you know, a number of conditions where many of the guideline say MRI can be appropriate.

So all I'm saying is that, you know, the cleaner you can make it, the more interpretable it's going to be and no, we're not asking it to be zero or even 10 percent or whatever but, you know, if you're getting numbers of 40 percent and, I mean, how can anyone interpret that or use that? I don't really understand that perspective to be honest.

Karen Johnson: OK. Let's move on to reliability testing if there aren't any other comments and we can turn this over to our lead discussant, please.

Thiru Annaswamy: I'll be happy to continue to Reliability Testing. The Reliability Testing was done to measure score and it was assessed during -- using data obtained from the 2013 Medicare data. The range of testing, reliability scores range from 22.4 percent to 86.6 percent with a median score of 44.9 percent.

The question for the committee is the test sample adequate to generalize for widespread implementation and the second question is do the results demonstrate sufficient reliability so that differences in performance can be identified. There was a new updated measure testing packet that was sent to us by e-mail. Upon looking at that, essentially, the measure testing intraclass coefficient reliability was came up with 0.59, I think I'm talking about the same thing here. But it is below the target value, similar to what was reported in 2015 and the -- the award description of the value was moderate. And the explanation submitted by the developer says that the intent of the measure is not to identify differences in performance between individual facilities, but rather to identify differences from the mean or the threshold performance value.

Thus, we believe that testing reporting in section 2B5 to determine statistically significant, meaningful differences in performance, is a more appropriate test for this measure. So those were my thoughts on it. My initial answer to the question, does it demonstrate sufficient reliability is no which is no different than the -- what the committee arrived at in 2014, but I will leave it there and open it up to comments.

Kelly Anderson: This is Kelly Anderson from the developer's team. Do you mind if I provide a little bit of clarification on the two different tests we ran for reliability?

Thiru Annaswamy: Please go ahead.

Kelly Anderson: Sure. So NQF typically requests the developers conduct signal to noise testing for dichotomous measures similar to the structure of this measure. And what signal to noise testing really looks at is whether or not you're able to

tell differences between two individual facilities. So, can you tell the difference between Johns Hopkins Hospital and George Washington Hospital or something like that. So looking at difference in between those two individual facilities.

And as you have seen, the score using that particular test is a little bit lower than what we would like to see. However, what we're really looking at with this measure is what is the rate of overuse with an intention of driving all hospitals closer to a rate of zero. And so what you really want to understand at that point is can you accurately calculate a hospital's rate of overuse consistently if you were to do that once, you do it again, do it again a few months later.

And so the updated result we shared with the committee use a test-retest approach which have tiers of data and randomize that into different samples to get approximately years' worth of data in each and look at the agreement in performance scores for these two different samples of data and that's where we get the reliability of 4.59 and using that intraclass correlation coefficient approach that is considered to be a moderate reliability score.

Karen Johnson: Additional comments about reliability testing?

Katherine Gray: This is Katherine. Just a question, a follow-up on the test-retest. Were you able to find, you know, in the much information about the ones who had higher test-retest numbers than the average?

(Crosstalk)

Female: …something we looked at but I think that's a great suggestion to try to understand what's driving particular high reliability versus not. We have certainly something we can take back.

Katherine Gray: Yes. I think that's where, you know, because it's clear with what you've got, the reliability, there is a lot of non-uniform processes going on. That's why it's not…

Female: Right. I think that's a great suggestion.

Katie Streeter: Thanks. This is Katie. We wanted to bring to your attention that we do now have 13 committee members on the phone and on the webinar. So that is quorum. What we would like to ask, is there any of the 13 members that now they need to sign off before 2 o'clock, and if so, we won't bother with voting today. If everybody think they can stay on until 2 o'clock and the co-chairs agree, that 13 -- makes them feel comfortable enough to go ahead and vote, again, that is our quorum then we can go ahead and vote today during the call.

Roger and Kim, do you have any comments about that?

Roger Chou: No, I think if we can do some voting on the call, that's ideal.

Katie Streeter: OK. So what we will do is go back to get and vote on gaps because we did accept the evidence so we won't be voting on evidence. So we'll go ahead and vote on gap and then we'll move to reliability.

Male: I'm sorry, one question. I don't know -- I don't know when the additional committee members join. Does anybody need a recap or any brief? Because they do it without a conversation. I don't know if anybody needs any update or we should summarize before.

Carlos Bagley: That would be great because -- this is Carlos. I came in about -- about 30 minutes in. So, that would be helpful.

Male: OK. So if we -- it might be a biased update but I'll try to make it as objective as possible. There is, when you presented a new package with some additional information, there are concerns with the committee that or the working group that provided regarding -- mostly, I would say on some of the exclusion issue, there's been no long discussion and if it is appropriate or not for elderly patients to be directly included or not and that can lead to some concerns on guidance or guidelines and how physicians address this issue.

There's been also -- I know I might not be going into perfect order. There's been some concerns on some of the codes and personally while we believe that the code, yes, can be improved, we think that a code is or some exclusions

or inclusions can be changed and we're looking for feedback to change that, I don't think those speak to the specific concerns of the measure, but rather the coding or some inclusion or exclusion that can be taken care of.

The developers, we did add some testing and since the last time, we looked into more detail in some exclusions. We also, well, there was some additional exclusions added and there's been no performance change or significant performance change and that's been a concern and rightly so, by the committee. One of the issue that we expressed that there's been a change in data sources and that led to some changes in -- and there's been changes along the way and the longitudinal approach and looking at the changes in values, that we see the that both changes in specifications or exclusions has led to some changes or lack of change, and also some sources from CMS that changed from one system to another and a contractor also increased a little bit the numbers and we might not have -- that might be hiding some potential improvement.

One of the things, also the reliability which was low based on the previous signal-to-noise analysis, we provided additional analysis which -- I'm not an expert so I don't want to inappropriately explain but it's the intereliability…

Female: Yes. So we didn't use the test-retest approach to look at intraclass correlation coefficient per hospital, (using both two randomized) in my samples there.

Male: And that led to a moderate instead of low and it was on the top end over the moderate level and which was 0.59, I think, or 0.58 and it's moderate till 0.6 and substantial after that. So that showed with this new testing, it went from, at least in that case from low to moderate.

We also -- and, you know, I'll do my opinion that one of the things that we wanted to make a case is that intent of the measure is to avoid two issues. One is what we'll call quick trigger doctor, doctors that see a patient and immediately or an MRI, and also, this measure has a significant large range of performance between facilities and the goal is not to get to zero, but to decrease and for the higher -- called the (high-use) facilities to look at what are they doing differently that might be the reason and to at least have a look

and that we think that everything it plays an important role in describing potential problems in the hospital and not having a measure in our view of the -- for the hospital's will, let's say, in some ways provide for no accountability and hospital performance. This is also a pay-for-reporting, pay-for-performance measure and it's claims based and we acknowledged it for some (time) with claims that can be fully captured.

So this is, I mean, I'm trying to put them through that there are a lot of things that I missed and I appreciate if some other committee members wants to provide some of the other challenges that you -- you addressed probably much better than I would.

Kim Templeton: That's a great summary. I guess -- this is Kim. The only thing I would add in is expanding on your last point such that this is claims based and the -- the protocols that were -- that are being utilized or referenced as far as standard of care and provision of conservative interventions before an MRI as -- as is recommended of using self-care or over the counter medications or massage or acupuncture or whatever, there's no way to capture those in claims data. The developers mentioned that a surrogate measure would be to have a prior E&M code. But again, there's no way to confirm that the patients have done any of this and maybe there's a concern of the committee that we're perhaps missing a significant number of conservative measures that have been tried but just can't be documented through claims.

Thiru Annaswamy: Yes. And that's an important issue and we also clarify that we're not looking at, you know, conservative treatment. What we're trying to do when we look back in the claims by looking from the MRI backwards, if we see that the request for the MRI has been, let's say, in the 28 to 60-day period, we would determine that there's some action or assume some action has occurred and what we're trying to avoid is somebody going through a doctor today for the first time with low back pain and getting an MRI tomorrow and we don't know what happened and we -- it is almost like giving a prescription, we don't know if a patient takes it or not. You know, doctors could provide antecedent therapy or suggest and the patient may not do it but we know that the intent of the doctor was, hey, I know that I should not be doing an MRI immediately. I have to provide some other approaches except for some conditions like if

patient had cancer or HIV or other conditions which are excluded immediately from the denominator and, of course, from the numerator in which case those red flags or exclusions would eliminate those patients from consideration. But if not, the patient -- we are assuming that something happen in that term and, of course, we can't get the prescription or we can't get the Advil note or if they took it. But that's a little bit how we address the issue of this -- the antecedent therapy.

Karen Johnson: Great. Thank you for that wonderful summary.

I think we can go ahead and vote on performance gap and this is for measure 0514

Thiru Annaswamy: My slide didn't advance. This is Thiru Annaswamy.

Karen Johnson: Hi, do I have the right slide up for the voting?

Operator: You'll actually move to the next slide. The voting options will be available for everyone there and our voting committee members simply click in the box.

Karen Johnson: A is high, B is moderate, C is low and D is insufficient. And again, we need 13 votes.

Female: We do still have 13 folks on line so it looks like we're just missing one.

Carlos Bagley: It may be me. This is Carlos. I can't figure out how to vote on my computer screen.

Female: So, Carlos, all you need to do is click in the box, do you see a little box next to the alphabetic…

(Crosstalk)

Carlos Bagley: No. I don't know if it's not advancing on my screen or something, but no, I'm not seeing anything.

Female: OK. Could you try just one time to refresh your session for us by pressing F5 on your keyboard or Command R if it's a Mac?

Carlos Bagley: OK. It's working now.

Female: Excellent. Thank you.

Karen Johnson: OK. We have 13. 23 percent high, 77 percent moderate, 0 percent low and 0 percent insufficient. And now we move on to voting on reliability.

OK. So we're voting on for reliability for Measure 0514. Reliability includes precise specifications and testing, appropriate method and scope with adequate results. Voting is open. Thanks. Zero percent high, 62 percent moderate, 38 percent low and zero percent insufficient for reliability on measure 0514.

So now we can move on to discussing validity.

Thiru Annaswamy: This is Thiru Annaswamy again. I'll be happy to lead the discussion on this.

The validity and specifications, it says that this section should determine if the measure specifications are consistent with the evidence. Some of this discussion, we have already had. The questions for the committee are are these specifications consistent with the evidence. Again, we have had these previously.

Regarding validity testing, the developer has submitted new data on the testing where essentially they reviewed -they interviewed a technical expert panel and provided information from a survey of the technical expert panel to see if -- and I'm trying to flip my pages to see where -- where they are. So the two questions, as to this panel, panel were does this measure capture the most appropriate and prevalent types of antecedent conservative therapy available through claims data and the response was 72.7 percent yes, 9 percent not sure and 18 percent no.

And the second question was the measure helps assess inappropriate use of MRIs, lumbar spine tests. Do you agree and 9 out of 11 said strongly agree or

agree. So the questions posed to us were that the developer address concerns regarding exceptions that were noted during the prior review, so to recap the discussions we had in the 2014 review for this measure included some of the things we already talked about whether or not it excluded patients over 70 years of age, it doesn't and we talked about some of the issues related to that other conflicts noted included a disc herniation, sciatica, reticular pain and degenerative conditions with which may not include euro logical deficit and the point is that we discussed earlier, it continues to not exclude those.

The second point was concerns about surgery exclusion look back period. Presentation of conflicting guidelines used to identify red flag conditions and use of claims data to identify antecedent therapy. The issue of conflicting guidelines have somewhat improved because there is not that many conflicting guidelines cited in the update 2016 ACR but the other issues, in my opinion, remain and history of back surgery and trauma about the 90-day exclusion as opposed to an absolute exclusion which what the committee recommended in the 2014 review those -- those exclusions and issues remain.

And if I may add, the comment about if the trigger-happy doctor one more time -- if they do an evaluation and management and then order an MRI as it -- as a reflex, in this measure, it would capture that as antecedent conservative therapy because in E&M code would be submitted with evaluation and that would count as conservative therapy.

And so you would actually not capture a predominant majority in a speculative sense as the predominant majority of the cases in which a doctor or a trigger-happy doctor might actually see a patient and order an MRI when this measure wouldn't capture it. On the other hand, if an MRI is ordered in an appropriate manner in the elderly patient as alluded to earlier or in a patient with sciatica as alluded to earlier, but which who do not have neurological deficit, then this would show up as an inappropriate use. So I think this measure, in my opinion, continues to have validity problems which haven't been adequately addressed in the testing done by the measure developer.

I will stop there and I will -- you can comment.

Kelly Anderson: So, this is Kelly Anderson from the development team. I did want to provide one clarification on the comment you made about E&Ms and what we're calling quick trigger docs. So there is a time window between when the provider has that E&M claim built and when an MRI occurs in order for it to be counted as allowing for enough time for antecedent conservative therapy, so several weeks do have to past between that E&M encounter and the MRI study for us do counted as an appropriate MRI study. So an E&M claim just two days before the MRI or a week before the MRI is not considered a reason to call that appropriate within the measure specifications.

Thiru Annaswamy: Thank you.

Male: But there's, again, there's no way to know what they actually did, right? I mean, all -- it's just if somebody shows up in the doctor's office and they get an E&M code, they can do -- the doctor can do nothing and six weeks later, they order an MRI that's considered appropriate?

Thiru Annaswamy: Yes. Well, yes. That's -- it's true and two issues here. One is but the intent were not this -- let's call it this way. There is a patient in appropriate management and there's a doctor in appropriate management and that case, as the doctor orders physical therapy and the patient never goes, we can't put it on the burden of the doc. He did the right thing and then the patient, you know, got an MRI that one issue. The question is we're trying to avoid the decision maker, in this case, the doctor to say, you came today with a back pain, here, get an MRI. That's the intent of the measure. It's almost like a prescription.

You know, we -- there's been a whole discussion, how do you determine a doctor prescribed but we don't know, even if it's filled the prescription, we do not know, first, if it's filled. If it's filled, they took the prescription. So, this -- we know that it's not the perfection on the claims that we can determine unless we get that chiropractor claim and in some cases, we do look at physical therapy and chiropractor so we can determine that as well.

Colleen McKiernan: The other comment, I would make, this is Colleen McKiernan again. The other comment I would make is that you are correct that if a patient goes

in, sees the doctor and then they say, we're going to just, like, try some stuff, like, try, like a heating pad or try over-the-counter pain management and we can't track that. That's information that's not available in claims. If this were an EHR measure, it would not be available in a structured field in EHR. There's no place to document, like a heating pad.

So I think that there is myriad ways that antecedent treatment can occur. We try as hard as possible to capture the most straightforward ones like chiropractor and physical therapies and claims and then for the remainder, we use that in the evaluation in management doctor's appointment as a -- as a proxy to identify -- to identify alternative options that are captured in claim -- in claims. And again, we do have that window between when they see the physician and when the imaging study should occur.

(Crosstalk)

Male: Right. My point is just that the doctor can tell them anything. We have no idea. They can tell them lie in bed for six weeks which we are telling people not to do anymore but they can steady-lie in bed for six weeks and the patient comes back in six weeks and gets an MRI and we say that according to the mensurate when you say it would be kind -- that's appropriate.

Female: And I think we all -- we're trying to cut down overuse in MRI and so it's great in theory but it's extremely difficult to put in practice and if the goal of this is to improve quality unless there's a way to measure all of these or to capture all of these non-code related interventions, then whether in order -- then there's no way that this actually reflects anything in terms of quality. It's just a number that really doesn't have any meaning to it.

Male: I'm somewhat -- I have a, maybe, different perspective but I just want to say, yes, it's true that if the patient is asked, OK, stay -- lie in bed for six weeks, you could say that's appropriate or not therapy, probably it's not. But what -- I think the challenge to that, as medical societies have put on MRI lumbar spine is that people don't recognize that a low back pain is not an immediate need for an MRI. And, obviously, I would say that a doctor that practices with and

appropriately saying yes, go and lie in bed for six weeks, there's a bigger chance that you're just going to do and appropriately order an MRI.

So and we -- and we understand there is going to be some of these cases. So we're trying to avoid the misuse. But in addition to the misuse because people tend to do it, the physicians have concerns that a lot of other things show up and then it leads to a lot of inappropriate surgeries, a lot of inappropriate challenges with the MRI taken that should not take place and that's what I think in choosing wisely that -- one of the challenges is there are other consequences than with doing an MRI.

So I think that's what we're trying to avoid. Yes, it does not make it -- again, as we said, it's not the perfect solution but I think it addresses some of the challenges of overuse that we identify. I also want to address the issue that we're saying that regarding the surgery, well, obviously, there are different opinions if it should be, you know, once we had the surgery to life event that for that rest of your life it's appropriate or not, there are different perspectives and our (temp) had a, you know, a discussion that no (internal) is going to be related, that's probably 90 days related to the surgery, that's why you want to look at it, that surgery that happened 20 years ago should not be sufficient to get an MRI immediately. So we're not saying you shouldn't but you shouldn't get it. If you have 20 years, maybe you should look at something -- it could be something different.

So that's a different of opinion and I understand that, you know, we're not going to get, say, one is right or wrong. Now, that is a coding issue that one that say we could go one way or the other, it's not saying that the measure is wrong, it's just saying that the seclusion should be changed to do it forever, do it for five years, do it for 90 days. And we explained doing -- we did five years just to do an analysis and it didn't change much in that case. It was very few cases. So it didn't change.

Female: That's very helpful. Let's try to have a community members have the last word because we're kind of running out of time here unless there's specific questions for the developer. Any other committee member comment?

Katherine Gray: This is Katherine. I just wanted to ask a question. You know, they -- the information they provided us was for phase validity which is a very subjective measure. So I was just wondering, since they ended up -- I mean, it was a small sample slide but they did end up with, you know, somewhere between, you know, 18 percent and 27 percent of the people who did not agree either with the antecedent therapy being identified correctly or that it measures the inappropriate MR test.

I was just curious and it says only two or three people, did you guys talked to them at all to find out why they said that? Do you have an understanding of that?

Colleen McKiernan: So, this is Colleen McKiernan from The Lewin Group. So, that's a great question. And so to give a 30-second summary of our process, so we spoke with our expert panel which was reseated, so these weren't -- there are a couple of older people who had been here forever, you know, worked on the measure from the beginning and until now and then there were a number of new people that we brought on to our expert panel.

So that mix of individuals has discussed this measure a number of times including many of the updates we've already talked about today. We sent them a survey to collect this information in a structured approach and then we walked through each of the or most of the questions with the expert panel to talk about, kind of the type of feedback they're -- that they had and if they agreed or disagreed. And so a lot of the comments that have been mentioned today, kind of were reflected in the one or two individuals who felt that the E&M wasn't what they wanted or, like, the various disagreements that you're -- that we note here in the couple of tables in our submission.

And so there's -- as we've all discussed, there's various perspectives from the way you should identify evaluation and management and look back for exclusions and I think the phase validity results reflect a variety of opinions in the clinical community on how to capture evaluation and management and whether we should be more precise and try to find a way to capture the NSAIDs and the other things that are not in claims versus people that are satisfied with the way the measure is.

So I think that, you know, when you look at the nine individuals who agree or strongly agree, that the measure helps assess inappropriate use of tests, really, we saw that the overwhelming majority, including qualitatively people that didn't vote in the survey but that were in the meeting, they feel did feel that overall, the measure is kind of hitting -- hitting the market and helping us reduce inappropriate imaging.

Karen Johnson: Thanks, Colleen. Is the committee ready to vote on validity?

Katherine Gray: So this is Katherine. I was just wondering, there were two questions that the NQF staff has talked to -- or, well, we're on A. Are we on A or B? Traceability or -- you guys had asked us two questions about the facilities, performance, that was significantly different from the mean and also if -- it was only specified for facilities to that have a minimum case count. I was going to ask if we can get those answered?

Female: So I want talk about the case count question first that you (referred back) to the other one. For a minimum case count, yes, we do only report the scores for hospitals that meet a minimum case count. That minimum case count varies depending on the performance score but it's around just the cases in order for hospital or have a value on hospital compare. All hospitals do receive for internal QI purposes though, their score, whether or not they meet that minimum case count, so the hospital has 30 or 35 cases. They're still receiving information about their rate of appropriate or inappropriate imaging.

Katherine Gray: But those would not be included in your data?

Female: There were not included in the data that we used in this package but we did look at those in our early phases of testing.

Katherine Gray: OK. Thanks. And how -- how about the other one about the significantly different from the mean?

Female: Could you remind me what that question was and I'm happy to answer it?

Katherine Gray: My note say can you tell us how many of the facilities had a performance value that was significantly different from the mean? That was the one where you did the top 10 percent and the bottom 10 percent.

Female: Give us just a moment and we'll see if we have that. I don't think we have that in front of us right now. I'm sorry about that but I'm happy to present -- provide that information after today's call.

Katherine Gray: OK. Thanks.

Karen Johnson: OK. Let's go ahead and vote on validity for Measure 0514. Voting is now open. A is high, B is moderate, C is low and D is insufficient.

Thirteen votes are in. Zero percent high, 23 percent moderate, 59 percent low and 8 percent insufficient. Measure 0514 did not pass the validity subcriterions. So with that, we will end discussion here -- about this measure. As far as next steps, all of your recommendations for this measure and for the NCQA measure will be written up in a draft report that will be posted for comments in mid-February, February 15th. We will be calling you next week for your availability to attend the post comment call. That will most likely take place the first week of April and that is when we will review all of the public comments that came in as well as revisit any of the additional information that may be submitted from NCQA. So please be on the lookout for the SurveyMonkey poll regarding your availability for that call next week.

We would like to pause and see if there are any public comments.

(Crosstalk)

Karen Johnson: Go ahead.

Operator: Thank you. At this time, if you'd like to make a public comment, please press star then the number one on your telephone keypad. And there are no public comments at this time.

Karen Johnson: Thank you. Do our co-chairs have any additional remarks before we end the call today?

Kim Templeton: I would just like to thank everyone for their time and their input. Thank you.

Karen Johnson: Thank you to all committee members and to our developers for attending today's call. If there are no additional comments, we will go ahead and end the webinar. Thank you.

Operator: Ladies and gentlemen, this does conclude today's conference call. You may now disconnect.

END