NATIONAL QUALITY FORUM

+ + + + +

MUSCULOSKELETAL MEASURES
STANDING COMMITTEE

+ + + + +

THURSDAY
MAY 8, 2014

+ + + + +

The Standing Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 9:00 a.m., Roger Chou and Kim Templeton, Co-Chairs, presiding.

PRESENT:
ROGER CHOU, MD, FACP, Co-Chair
KIM TEMPLETON, MD, Co-Chair*
THIRU ANNASWAMY, MD, Dallas VA Medical Center
CARLOS A. BAGLEY, MD, FAANS, Duke University School of Medicine
STEVEN BROTMAN, MD, JD, AdvaMed
SEAN BRYAN, MD, Greenville Health System
CRAIG BUTLER, MD, MBA, CPE, American Academy of Orthopaedic Surgeons
KELLY CLAYTON, BS, Arthritis Foundation
LINDA DAVIS, BSN, Minnesota Health Action Group
JAMES DANIELS, MD, MPH, FAAFP, FACOEM, FACPM, Southern Illinois University
CHRISTIAN DODGE, MD, Bastyr University
ZOHER GHOGAWALA, MD, FACS, Tufts University School of Medicine
V. KATHERINE GRAY, PhD, SAGE Health Management Solutions, Inc.
MARCIE HARRIS HAYES, PT, DPT, MSCI, OCS, Washington University in St. Louis School of Medicine
MARK JARRETT, MD, MBA, North Shore - LIJ

Health System
PUJA KHANNA, MD, MPH, University of Michigan
WENDY MARKINOVICH, BSN, MPH, RN, Blue Cross and Blue Shield Association

PRESENT:
JASON MATUSZAK, MD, FAAFP, CAQSM, RMSK, Excelsior Orthopaedics
CATHERINE ROBERTS, MD, American College of Radiology
ARTHUR SCHUNA, MS, RPh, BCACP, American Society of Health-System Pharmacists
JOHN VENTURA, DC, American Chiropractic Association
CHRISTOPHER VISCO, MD, Columbia University College of Physicians

NQF STAFF:
HELEN BURSTIN, MD, MPH, Senior Vice President for Performance Measures
ANGELA FRANKLIN, JD, Senior Director, Performance Measurement
KAREN PACE, PhD, MSN, Senior Director, Performance Measurement
ANN PHILLIPS, Project Analyst
KATHRYN STREETER, CHES, Project Manager

ALSO PRESENT:
KELLY ANDERSON, Lewin Group
MARY BARTON, MD, MPP, National Committee for Quality Assurance
DALE BRATZLER, DO, MPH, Oklahoma Foundation for Medical Quality*
CHARLIE BRUETMAN, MD, Lewin Group
JOHN FITZGERALD, MD, American College of Rheumatology
MELISSA FRANCISCO, American College of Rheumatology
COLLEEN McKIERNAN, MSPH, Lewin Group
RACHEL MYSLINSKI, American College of Rheumatology
JENNA WILLIAMS-BADER, MPH, National Committee for Quality Assurance

JINOOS YAZDANY, MD, MPH, American College of Rheumatology

*** Present by teleconference**

TABLE OF CONTENTS

P-R-O-C-E-E-D-I-N-G-S

9:03 a.m.

MS. FRANKLIN: Hello and welcome to Day 2 of the Musculoskeletal Standing Committee Consideration of Candidate Measures and I just want to do a quick check and see if we have one of our co-chairs on the line. Kim Templeton. Okay.

And, operator, can you check to see if we have her dialed in by any chance?

OPERATOR: She's not dialed in at this time.

MS. FRANKLIN: Okay. Thank you. So, with that, I will turn it over to our Co-Chair Dr. Chou.

CHAIR CHOU: Welcome to Day 2. Thanks, everyone, for coming back after a long day yesterday.

So, I think just wanted to do a -- are there any kind of logistic things we need to talk about today? Cabs or anything like that? Okay. Not yet.

Just wanted to do a quick recap of yesterday and then talk about what we're going to go over today.

So, yesterday, we reviewed I think a total of nine measures or seven measures. Four gout measures and three RA measures.

One of the gout measures was passed as a trial measure and a second one also passed. Kind of that gray 40 to 60 percent. Not quite consensus yet and two of them didn't get passed our initial three screening criteria.

All of the RA measures passed and we have a couple more to consider today.

I understand we have five more measures today. We'll be starting off with the Disease Modifying Anti-Rheumatic Drug Therapy Measures. There's one on pain management for long-bone fracture and then a couple on imaging for low-back pain.

So, I think unless there's other issues that we needed to discuss, we can

probably move on.

Great. So, do we have the developers for the Anti-Rheumatic Drug Therapy Measure?

MS. FRANKLIN: Please feel free to come to the main table.

CHAIR CHOU: Those two spots are reserved and if you can give us a brief overview of the measure, that would be fantastic.

MS. WILLIAMS-BADER: Hi, everyone. My name is Jenna Williams-Bader. I am Assistant Director of Performance Measurement at NCQA.

Mary, would you like to --

DR. BARTON: And I'm Mary Barton, Vice President for Performance Measurement.

MS. WILLIAMS-BADER: That was brief. Thanks.

All right. This morning, we are going to start talking about the Disease Modifying Anti-Rheumatic Drug Therapy for

rheumatoid arthritis. This measure is specified at the health plan level and I think we covered the importance of this measure quite well yesterday.

Just so you know, in this measure, we do look for at least two claims with a rheumatoid arthritis diagnosis to try to verify that the patient does indeed have rheumatoid arthritis. We're also looking for the DMARD to be dispensed rather than ordered or prescribed and we do have exclusions for HIV and pregnancy.

Oh, I think -- as I said that we did a good job of covering yesterday the importance of rheumatoid arthritis, it is a rare condition, but it does have quite a significant impact on patients who have the condition.

The guidelines very strongly indicate that patients should be prescribed and put on DMARD as early as possible.

However, the guidelines also

continue to recommend DMARD therapy even for patients that have had the condition for a while. There is actually no recommendation to completely stop DMARD therapy for those patients and at the moment, there isn't much evidence to show that non-medication remission is really possible in these patients. So, while the guidelines do recommend scaling back the DMARD therapy, they do say that patients should continue to stay on DMARD therapy.

This measure was first developed in 2002/2003. The concept was actually supported by and recommended by two different panels that NCQA had brought together. One being to define measures for rheumatoid arthritis and another to define measures for pain.

When that measure was eventually developed, we pulled together a musculoskeletal expert panel to help inform the development of the measure.

It was field tested in 2003 and

you'll see the results of those -- of that field test in the information that we gave to you. The measure did perform well during field testing. We saw that there was variation across the plans and the rate was not as high as we would hope. You will have noticed that during the field test we found that the diagnosis of RA was -- there was high agreement between medical record review and claims.

There was unfortunately a bit of a discrepancy in one of the plans for the actual dispensed DMARD. However, I would remind you this is from 2001/2002 data and these plans have had much more experience to date with dispensing data. So, we do expect that the plans are performing better now and actually as you can see, the rates are relatively good.

However, we do still see variation between plans and we do see variation between the different lines of business as well. The commercial plans are performing better.

However, we do still see that there's variation between the plans at the 10th and 90th percentile and the importance of this measure I think is emphasized by the fact that the Medicare Stars Program has picked it up.

The measure is also included at the physician level and the physician quality reporting system and there is strong interest in the physician level measure as well as we'll hear about during the next presentation.

I think that covers it. Thank you very much.

CHAIR CHOU: Thank you. Puja and Art, do you want to provide a brief overview from the Committee.

DR. KHANNA: Sure. Thank you. That was actually a great start to it.

So, to briefly describe what we said is as far as the evidence goes, we thought it was high and it was based on a systematic review of empirical evidence

presented to us. So, we did not have any, you know, concerns about it.

One of the things that was brought up was that the work group members questioned whether there is, you know, a continued opportunity for improvement in this measure. Because like you pointed out, 90 percent of the commercial plans are already meeting the measure. Whereas, you know, although the Medicare and Medicaid plans were not quite there yet.

So, the thought was that how much do you really think we can get to 100 percent because that would be the ideal. So, if you could, you know, --

MS. WILLIAMS-BADER: Yes, I can certainly speak to that.

So, as you point out, the average is high. However, even in the commercial -- across the commercial plans, we do see that there is variation and actually the variation -- there is variation by region as well.

So, while there are some regions where the average is around 90 percent, there are others where it's as low as 82 percent and then between the -- as I said, between the 10th and 90th percentile, there is also still variation.

So, in different regions, the 10th percentile is as low as 75 percent and the 90th may be as high as 95 or 90 to 95 percent.

So, while plans are performing well even across all three lines of business, there is variation indicating that there's room for improvement still.

CHAIR CHOU: Thanks. Art, do you have anything to -- any additional comments?

MR. SCHUNA: No, it's just that I think this is a very important measure and clearly, there's data supporting the importance of this. So, I think the evidence is pretty strong on this one compared to some of the things we looked at yesterday.

CHAIR CHOU: Great. Any questions

for the stewards at this point? Jason.

DR. MATUSZAK: The way I understood the literature and the evidence that's out there, there actually is a significant body of evidence that shows that people can take at least long holidays off their DMARDs or even up to a year without having a significant increase in some of their flares.

How do you guys address that evidence?

DR. BARTON: I think the issue is that the best way that we can be sure that patients are receiving attention for their rheumatoid arthritis is to use a measure that looks over a period of time and I think, you know, your point raises the good question of whether one year is the ideal period of time to look for evidence of a prescription to address disease modification.

Unfortunately, we're -- in a programmatic way, we're kind of tied to that

kind of interval. But, it may be that it's a -- as close an approximation as we can get to the perfect interval which might be longer. I don't think that the evidence would suggest that people should be taking several years long holidays.

DR. MATUSZAK: Actually, the study I'm looking at right here shows that they can take up to five years in some cases without taking their DMARDs and stuff.

So, I think that the evidence does exist out there that certain people and you can predict in some cases who those people might be that might be able to do better.

But, if we're doing something like a functional status assessment on their people on a routine basis, is that enough for outcomes on these patients as opposed to a prescription for a DMARD every year?

CHAIR CHOU: We have a couple of rheumatologists on the panel. Maybe they can help address this. Puja or Mark? Yes.

DR. JARRETT: It is true, you know, in practical terms, there are patients who go into a natural remission and go years without taking a DMARD and what I would look to say -- ask is I don't know if ICD-10, if it ever comes to fruition, would help us because that might allow us to differentiate if it goes down another level in terms of active versus inactive disease and if you had that, then you'd be able to do that. Differentiate those patients you've seen twice a year who are doing great basically on nothing or a non-steroidal and coasting and then wouldn't get counted then in the denominator.

So, I don't know if you can explore that in terms of seeing if ICD-10 will help you with that.

But, I still think overall this measure is very critical. Because even with those exceptions, you still need to get the bulk of the patients, 95 percent, on a DMARD. So, you know, that -- the small group, we got

to figure out how to get them out of the denominator.

CHAIR CHOU: Anyone have any other questions or comments? Oh, Thiru.

DR. ANNASWAMY: Just a quick question about what changes to the measures has been made, if any, since 2009 because this is a maintenance measure.

MS. WILLIAMS-BADER: This measure was reevaluated in 2012 and when we do every evaluation, we look at the performance data, we look at the guidelines to see what updates are there and we also look to see if we've received a number of -- if we've received questions from the plans that are reporting the measure to make sure that we've covered all of our bases.

When we did the reevaluation of this measure in 2012, there were no significant changes. I think we might have made a few tweaks around the edges for coding, but nothing significant to the measure

specification and that was supported by a bone and joint measure advisory panel that we pulled together as well.

DR. BARTON: I think we would also say every year we review the pharmacy data that's in the measure.

CHAIR CHOU: Linda.

MS. DAVIS: Have you looked at what health plans do to actually try and improve their score on this measure and what impact that has and what there is variation from plan to plan?

MS. WILLIAMS-BADER: That is an excellent question and there's definitely strong interest from our Committee on Performance Measurement which is a large and multi-stakeholder group that reviews all of our measures that are reported by plans and they've asked the same question.

We do try to do that for certain measures. We have over 70 in our set and that kind of review does take quite a lot of time.

So, I think as I said there's interest in that, but we haven't done it for this particular measure.

DR. BARTON: I think I would say we have anecdotal evidence from a couple of experts who have provided us with some insight about how their plans use internal monitoring of this to actually provide feedback to the clinicians who are caring for patients with rheumatoid arthritis. Which in some managed care plans is, you know, entirely rheumatologists, but there are places where there's a mixture of internists or geriatricians who are also taking care of these patients and so, by providing feedback to them, they can at least ask the question for someone who might not have had the thought raised, you know, to encourage them to consider.

I think to Mark's point every clinician is in the position of having to sort between the 95 percent and the 5 percent and

make that individual decision. But, if they haven't even thought to consider the question, you know, they're not going to -- there's no opportunity for the patient to get on a therapy.

CHAIR CHOU: I just had a question about the exclusions. Just kind of harmonizing with some of the stuff we talked about yesterday about the PPD kind of exclusion. People who haven't been tested or for whatever reason or who have a positive PPD and haven't been treated yet. Seems like that should be an exclusion. It's going to be a small number of people, but yes. Yes.

MS. WILLIAMS-BADER: Mary's indicating that's a very small number. Also, we are including -- this measure does not only look at biological DMARDs. So, there are still options for those patients as well. Making it an even smaller number.

CHAIR CHOU: Why don't we move into the evidence? We're already, I think,

talked about most of the evidence. Again, this is for re-endorsement. Right?

If there are additional questions or comments about the evidence, this is one of the required sub-criteria.

If there isn't more discussion, I think we can go ahead and take a vote on this one.

MS. PHILLIPS: Okay. We're voting on measure 0054. You have five options. One for high. Two for moderate. Three for low. Four for insufficient evidence with exception. Five for insufficient evidence and the voting begins now.

Okay. We have all 22. We've got 13 high, 8 moderate, 1 low, 0 for insufficient with exception and 0 for insufficient.

CHAIR CHOU: So, this one passes. Let's move to the opportunity for improvement. Again, we've already discussed this. That the average rates are high, but there is variation both among the different health plans and then

among the different types of health plans.

Other additional questions or comments regarding the opportunity for improvement piece.

All right. I think we're ready to do a vote on this as well.

MS. PHILLIPS: Okay, 0054, for the performance gap, you've got four options. One for high. Two moderate. Three low and four for insufficient. Voting begins now.

Okay. We're at 22. And we've got six high, 13 moderate, 3 low and 0 for insufficient.

CHAIR CHOU: So, this passes here as well.

Let's move on to health care priority. Again, we've discussing RA the last afternoon and this morning.

Are there other comments or additional things to say about the health care priority? Puja.

DR. KHANNA: So, this is actually

data that has come from CDC as well. That it's considered one of the leading causes of morbidity/mortality in the country. So, it is one of the top 20 priorities. So, and definitely a cause to treat -- start DMARDs early and treat the disease aggressively.

CHAIR CHOU: Additional comments or questions? Yes, Linda.

MS. DAVIS: So, this measure has been in place for several years and have you seen improvement over those several years and how much improvement has there been?

MS. WILLIAMS-BADER: Again, this is one where the performance has -- there's been shifting among the plans that are reporting. We don't see a large amount of improvement from when it was first reported to now, but we do see incremental increases and especially for the plans at, you know, at the edges.

So, we have seen some improvement. Because it started at a somewhat higher rate

than some of our other measures, there hasn't been as large of an improvement as there might be.

MS. DAVIS: And that's just improvement in the actual measure and actually, the number of prescriptions that are written for people with DMARDs? None of the outcomes --

MS. WILLIAMS-BADER: The number of patients that have been -- that have been dispensed at DMARDs.

MS. DAVIS: But, not in morality or outcomes or any other?

MS. WILLIAMS-BADER: Oh, I see. No, we don't do that type of analysis. The plans aren't reporting that data to us.

MS. DAVIS: So, I'm wondering if the continuation of this measure will continue to have the kind of mixed improvement in the actual utilization measure, the process measure over time and how impactful that actually is.

MS. WILLIAMS-BADER: And I think that's actually a very good idea to think about comparing that to rates of morbidity/mortality to see if there is some connection for those plans where they are improving their rates.

CHAIR CHOU: Great. Thanks. So, some of that discussion I think we'll consider later when we get to the usability and use criterion.

Are there other comments about the health care priority?

Okay. I think we're ready to take a vote on this as well.

MS. PHILLIPS: Okay. Measure 0054 for high priority are your four options. One for high. Two moderate. Three low and four insufficient and voting begins now.

Okay. We're at 21 and we've been holding for a couple of seconds. We have 22. Thank you.

Sixteen for high, 3 for moderate,

2 for low and one for insufficient.

CHAIR CHOU: Okay. So, we passed the three initial criteria. So, we're going to move on here.

The next area is the reliability. So, I think we've kind of touched on this already.

Puja and Art, did you have other questions or comments about either the specifications or the liability testing?

MR. SCHUNA: I guess one question I have regarding this is the patient factors which might influence whether patients will get a prescription or not. There can be a big variance in copay fees from place to place. Some patients may be paying $200 to $300 or so for biologics a month and they may choose not to be treated for that reason and the question comes then should that be the plan's responsibility or the provider's responsibility in that setting. That's a concern I have about this.

CHAIR CHOU: Did you guys want to comment on that?

DR. BARTON: Well, I think demonstrating the fact that many plans are able to get rates up as high as the 90s would say to me that there are ways to -- either those are, you know, completely made up of populations that are -- have benefits that are, you know, well designed or they found ways to address and equalize the opportunity.

And I think that if we are -- as we see this being used in the Stars Rating Program, there will be more of an appetite for plans that are doing really well to talk about how they're getting high rates and for, you know, sharing of best practices. Because I think you're pointing out something that's really currently, you know, a tremendous issue in this country, but not necessarily going to be fixed by a measure.

MR. SCHUNA: So, you say that plans can go up to 90 percent or better, but

you also mention that certain plans do not come close to meeting those goals and certain regions of the country didn't do -- perform as well and I have to wonder whether patient factors may be responsible for that and not the institution.

MS. WILLIAMS-BADER: I guess I would also say that at the health plan level we definitely think that the plan should be responsible for making sure that patients are getting the medications that they need to get and as has been verified and discussed by this Committee today, DMARDs are very important for patients with rheumatoid arthritis. So, I do think that we hold the plans accountable for making sure that the patients are getting medications that they need.

If the patients are refusing for reasons besides cost, maybe they just -- they choose not to, then we do think that that might be some that would apply across the plans and those might be the ones that -- I

see you shaking your head. I'll let you --

MR. SCHUNA: Well, I mean there's differences between plans. I mean some plans, for example, in the VA system, $8 is all a patient pays for any drug per month. Patients going to some outside insurance company pays hundreds of dollars a month.

You know, so, those kind of variations exist and I guess what I'm concerned about is that this doesn't necessarily deal with those questions and it becomes the plan's fault or the provider's fault that the patient doesn't -- chooses that they can't afford $300 a month.

DR. BARTON: If it's not the plan that creates the benefit design, who is it?

I would say that the plan is the place where you should hold the accountability for insuring the fair opportunity to access therapies.

CHAIR CHOU: Mark had a comment.

DR. JARRETT: One last comment

following up on what Art's saying and we also have to look at it. This is saying that they're going on a DMARD. It doesn't say which DMARD and many patients because of their plans and they can't afford the copays wind up being stuck on just methotrexate which is relatively inexpensive rather than getting a DMARD which might cost them because of their copays $600/$700 a month and they can't afford it.

So, at some point not in the purview of here, but we need to think about asking the clients if they're collecting this data to slice and dice it and then compare it to what their copay structure is. Because I would bet that we'll find lots of patients are on a DMARD, but probably not the appropriate DMARD. Something is better than nothing.

So, we should measure it and get the docs to do it, but then I think that information is a treasure-trove that we should look into to really see to really truly

improve the care.

CHAIR CHOU: Linda.

MS. DAVIS: I heard verbally that this does not include biologics. Is that accurate or does it also include biologics? I'm on biologics.

MS. WILLIAMS-BADER: Oh, I just meant to say that it includes all DMARDs both biologic and nonbiologic.

MS. DAVIS: Okay. All right. So, that changes. So, the price variation is substantial and can be up to, you know, thousands a month. Right?

And I work for a employer coalition and they're all self-funded and they all make their own decisions about their benefit designs. They're all very concerned about specialty drugs and biologics and scared to death of what's going to happen with that in the future. So watching it very carefully.

But, almost all of them to a T accept those with unions have high-deductible

health plans and I see those deductibles increasing and increasing nationally and locally and I don't see any trend changing that anywhere. So, I'm wondering about the implications for high-deductibles in this cost that's also been factored into some of the variation that you're seeing.

DR. BARTON: May we answer? So, this is a problem that's larger than this measure clearly and so, the NCQA holds the health plans responsible and the HEDIS reporting to include their ASO patients within there. Which is the self-funded employers set of patients that they're administering. So that the health plan would be actually held accountable for, you're right, a range of benefit plans and designs. Some of which it had actually designed and others which it had not.

Right now, that's the model for, you know 25 years that NCQA has used. It's the -- health plans are the locus that were

willing to take accountability and where we have found it feasible to hang accountability.

But, I would bet, you know, a quarter that you're right. That over the next ten years, we're going to have to find new models to address accountability so that the -- you know, whatever the organizations within the delivery system that are actually responsible for caring for patients are the ones who are being held accountable. But, I would say it's premature to know yet what that's going to look like.

DR. MATUSZAK: I just wanted to ask about some of your reliability studies that you had done again and I can't find the exact passage right now where I read it, but I jotted down a note here that shows that some of your field tests showed as little as only 57 percent concordance with the medical record and the data that you guys were pulling.

Can you maybe speak a little bit more about your field tests and what the

accuracy has been?

MS. WILLIAMS-BADER: Absolutely. So, the reason you might have had a difficult time finding it is that the field test data is actually under the validity section.

So, back in 2002/2003, we did a field test with three health plans. We do try to get a range of health plans and a range of lines of business if we have -- if we are specifying a measure for multiple lines of business.

As you do point out, particularly for the DMARD therapy itself, we did see that there were -- that the numbers on the rate of agreement between the medical record and the claims was lower than we would like. There was one plan where the rate of agreement between the administrative claims and the medical record was around 58 percent.

However, as I mentioned, this was quite a few years ago. Plans have had now -- particularly the ones participating in HEDIS

have had more than a decade to improve the ability to get claims -- the prescription data and I'd also say it's in their best interest to make sure they can because that's for the numerator event and we do know that plans are able to get that data since we have some plans that are performing so well.

The actual RA diagnosis, the validity and rate of agreement was much higher for the diagnosis. So, we see that the denominator is valid. Whereas, we did a little bit of discrepancy there on the numerator event.

DR. MATUSZAK: How do you know those high-performing plans are not part of the 57 percent that are having reporting errors and maybe are over reporting the -- or under recognizing perhaps the number of patients with rheumatoid arthritis, you know?

MS. WILLIAMS-BADER: So, like I said, the rate of agreement for the RA diagnosis was quite high. There was one plan

in the field test out of three that had a large number of missing medical records. So, their n was quite small. It was only 30.

However, in the other two plans where they had over a hundred patients included in the sample where they were able to identify the medical record, the rate of agreement between the medical record and the claims for the diagnosis was 94 percent and 99 percent.

So, again, we think that the denominator's being identified correctly.

On the numerator side with the DMARD therapy, yes, there are some plans -- there was one plan where they only had 58 percent agreement, but as I said -- and actually, the place where we're seeing the discrepancy is that there were about 35 percent the DMARD therapy was identified in the medical record only.

I would point to -- I would just reiterate that it's in the best interest for

the plans to get better access to the pharmacy data if they're not able to, if that's the problem and we do see that plans are able to since their rates are high.

DR. BARTON: I would add that HEDIS has a pretty rigorous audit that would be -- that would make me say that it's unlikely that health plans are committing fraud on a measure like this because of the way that we use both statistical methods and actually chart review methods to audit performance as it's reported.

CHAIR CHOU: Linda.

MS. DAVIS: So, PBM data is what they're dealing with. Right? For the numerator when they're looking for the prescriptions from the pharmacy benefit managers. That's where pharmacy claims come from.

You know, Kaiser probably has their own data. So, it's not uniform. It's not everybody doing it that way.

But, most have PBMs and the PBMs, they can have carved out PBMs which are -- they may have members on their medical plan that have up to eight to ten different PBMs across their population and so, they need to get data from all of these other PBMs as well.

Do you know the completion rate? Like how many of those other outside PBMs they're able to get the data from in order to have accurate and complete information about the data.

DR. BARTON: I guess the two ways that we have access to this information are one, when plans don't have a sufficient sample size of patients, we expect them to report that fact rather than reporting a rate. So, that's one kind of non-report that they're permitted to categorize and then another would be if they had enough patients.

So, if the denominator, you know, ascertainment went well, but they were somehow not reporting because they didn't have access

to data. Plans are rarely using that option.

So, the only information that I have would be to say that somehow by whatever method is appropriate to their setting, they're getting the data.

CHAIR CHOU: Karen.

DR. PACE: Yes, I just want to do a little clarification here so that we can kind of move through our criteria.

The reliability testing that was done was done at the plan level and those data are provided and indicate pretty high reliability and what that is telling us is that the scores can distinguish among the health plans and that is -- the reliability question is about systematic error and the ability to identify differences between the measured entities.

A lot of what you're talking about gets into the validity question. At the data element level, are the data valid that are going into the score? Are the appropriate

exclusions, et cetera?

So, you may want to kind of deal with the reliability question first and then move on to validity and, you know, that's possible. I mean you can have a very reliable measure and still have concerns about validity or may not be valid, but we try to kind of split those out because they are -- they're related, but different properties.

CHAIR CHOU: Yes, Puja or Art might want to comment on this, but the reliability testing they had mean scores of like .87 to .93. Which are supposedly quite high. I guess it's on a -- up to a 1.0 scale.

And then I mean the other piece of the reliability is just the specifications. If we have -- you know, if we think that the numerator, denominator and exclusions are specified clearly in what we want them to say.

Yes. Go ahead, Puja.

DR. KHANNA: So, the numerator and denominator are definitely specified. I think

we had some questions about the exclusions.

The work group members thought that malignancies could be potentially put in as an exclusion because there are -- you are including biologics and in that case, you will have, you know, patients who have had biologics.

DR. PACE: And again, I know this may sound like splitting hairs, but the reliability question on the specifications is are they precise? Are they defined? Are they identified?

The question about whether something should be in as an exclusion and isn't really does again come to validity. You know, what does the evidence say about, you know, is this the right denominator population? Are the right people excluded and included? So, I know it kind of splits it a little bit, but, you know, the specifications related to both reliability and validity.

DR. KHANNA: So, I guess no

questions about the numerator or denominator.

CHAIR CHOU: Okay. Thanks. Other comments or questions about reliability aspects?

All right. Why don't we go ahead and take a vote on this?

MS. PHILLIPS: Okay. Measure 0054, we're voting on reliability. You have four options. One high. Two moderate. Three low. Four insufficient. Voting begins now.

We are up to 22. We've got 11 for high and 11 for moderate.

CHAIR CHOU: Okay. Let's move on to the validity piece. So, that passed there.

We already started the discussion on validity. There was some discussion about the variations. It sounded like there was some -- that some prescriptions were missed. So, you had to capture that in the medical records. They weren't in the electronic record or whatever. But, you guys were pretty confident about the denominator.

Yesterday, actually one of the ACR folks presented some data that showed that ICD-9 codes were very inaccurate for various, you know, rheumatological conditions. They said the positive predictive value was like 60 percent. Which just seemed very shockingly low to me.

But, it sounds like that hasn't been an issue when you guys have checked the -- when you guys have done your checks. Is that correct?

MS. WILLIAMS-BADER: Right. I -- yes, I would say that. Yes, we haven't found that and again, we are using two claims. So, they -- and they need to be on different dates of service.

And I think we also think that this is some -- an issue in which it is beneficial for the plan to make sure that the patients are really being diagnosed accurately so that they're not having false diagnoses in their denominator. Because those would be

patients you'd want to -- wouldn't want to necessary give this prescription to.

CHAIR CHOU: So, let's say that a PCP calls something RA which isn't really RA. That gets picked up somehow and thrown out? There's a way to do that?

DR. BARTON: No, there would -- if it appeared in two different visits on two different days, they would be included in the denominator.

But, I would imagine that we would be hearing from plans if they were -- because of our -- the feedback loop that we utilize and the opportunity for plans to ask questions during the measurement year. That we would be hearing about that.

We certainly hear about that kind of issue on other measures. So, it may be that that's a generic ICD-9 problem, but perhaps less so in plans that have chosen to address this as a quality issue.

CHAIR CHOU: Karen.

DR. PACE: Yes. So, maybe you want to just -- because I know and I think this was your question that the validity testing was done on three plans and you have the sample sizes for those plans. But, you also -- so, I think that's where the question has come in with the one out of the three.

But, you also mentioned some ongoing auditing. Do you have data related to this measure from your ongoing auditing or how --

MS. WILLIAMS-BADER: I don't believe that we provide that information. That's part of our accreditation process and it's to make sure that we have confidence in the results that we're getting from the plans. But, I don't believe that information is publicly available because it's -- it goes back to how we're accrediting the plans.

DR. BARTON: But, the published rates are the published rates.

MS. WILLIAMS-BADER: Exactly. The

published rates have been audited. So, we have confidence that those are the right rates.

DR. PACE: Okay. Right. So, then what they have is these kind of -- the small actual empirical validity testing and then also their face validity.

So, according to our criteria, this would be eligible for a moderate rating on validity or less if there were issues. But, that would be where it would fall in our algorithm.

CHAIR CHOU: Are there --

MS. WILLIAMS-BADER: Sorry. Can I just say? When the measure was reevaluated in 2012, like I said, we pulled together another measure advisory panel and then we also discussed this with the multi-stakeholder group, the Committee on Performance Measure, that I referenced earlier and there are plans on that panel as well as a number of different stakeholders including you. We have

Government representatives. We have consumer representatives. We have physician representatives.

And the misdiagnosis of RA was not one I believe that came up during that reevaluation. So, about -- that would be seven or eight years after the measure was first introduced into HEDIS. That wasn't something that came up during that reeval.

CHAIR CHOU: I just have another denominator question. So, are people -- and this is for the rheumatologists in the group. Are people with seronegative RA managed differently?

I've had at least one patient where we've had trouble getting biologics approved. It's been a -- kind of a battle, you know, just to get prior authorization and stuff.

DR. KHANNA: Right. So, usually the seronegatives get treated with a nonbiologic DMARD to begin with and it is hard

to get prior auth for the biologics and that's across the board, across all health plans and the rationale obviously is that you're not seeing disease progression and we don't have markers.

But, there are patients who they still get a DMARD. Absolutely. It depends on the clinical presentation of the patient.

CHAIR CHOU: So, earlier, just to remind everyone of a few, I think, issues that were brought up with the validity. Whether patients with malignancies should be excluded. This issue about people who are in remission with. Whether there should be somehow something built in for them and then this issue about patient preferences particularly related to out of pocket cost and things like that.

So, just to remind everyone that these issues had come up and then this -- kind of this measurement thing particularly with getting the claims data for the drugs.

Are there other issues or things that people wanted to bring up here? Art.

MR. SCHUNA: Yes. I'm not too sure about the malignancy as an exclusion because you can use rituximab. You can use nonbiologic therapy in those patients.

I guess the people I'd be concerned about would be patients who might be under active treatment for their cancer in which they -- but, I guess I don't see that necessarily as a denominator issue and even if it were, it would probably be a fairly small percentage of patients.

DR. JARRETT: No, I agree with Art. It's a small number. The ones who don't want to take the nonbiologicals, it becomes more patient preference or sometimes their oncologist will say I don't want you taking anything. But, most of them are -- you know, if they have active disease, they'll usually take a nonbiologic. It's usually not a problem.

Methotrexate's about the only one that they're not going to rush to take because it might interfere obviously with perhaps previous chemo that they've gotten and such.

But, there are enough other choices that most will take something.

CHAIR CHOU: Great. Other questions or comments from the panel? Go ahead, Linda.

MS. DAVIS I have seen lots of advertisements for biologics and get support from the drug companies for coverage for these. I don't know how many people get that actual coverage.

Have you done any analysis to see how that might impact the numbers of people you're missing in your claims data because they're getting money from some other manufacturer?

MS. CLAYTON: Throwing the patient side out there, a lot of the nonbiologic DMARDs, you're looking at the copays of

anywhere from 100 to 300. The biologics, however, you're looking at anywhere from $500 to $2500 a month.

So, as patients, we rely heavily on those copay assistance cards and I typically I mean you're on a biologic for maybe two/three years and you're -- essentially, it fails treatment. So, then you end up switching.

But, I've had cases, you know, that I've encountered and other patients have encountered where, you know, it may take up to three months to get a copay assistance card from a pharmaceutical company. Therefore, you're not being treated those three months or four months while you're waiting and then most of them, it'll cover nine to ten months of coverage only.

So, for many of us, it's like paying an additional mortgage payment. Which we can't afford. So, then you're on the -- not on the treatment.

There's cases -- United Healthcare, they will no longer allow their patients to use a copay assistance card up front. So, they're forced to pay their copayments up front and be reimbursed directly by the pharmaceutical companies rather than using that at the time the prescription's filled. So, once again, that leaves patients who are -- unable to fill their prescriptions as well. So.

I would say most people -- patients I've encountered through the ACR and Arthritis Foundation, you're probably looking at what 60, 70, 80. Most patients I would say rely heavily on these copay assistance cards. So.

DR. KHANNA: Majority of patients, it's not 10 or 20 percent.

MS. DAVIS: Those data aren't in the claims date.

CHAIR CHOU: Other questions or comments? Jason.

DR. MATUSZAK: I just wanted to ask how the plans have responded to provider panels not having enough rheumatologists in a given coverage area and stuff and how valid do you think this measure is if the majority of rheumatoid arthritis care is being provided by primarily care physicians in a given area as opposed to rheumatologists because of access issues?

MS. WILLIAMS-BADER: I know that we have had some discussion about this with experts. The measure certainly doesn't specify what type of provider needs to be seeing the patient and we believe that, as the evidence shows, these patients should be receiving the DMARD therapy.

So, if there is a lack of rheumatologists in a particular area, then I think we hold the plans accountable for insuring that these patients still are able to access care and receive the medications that evidence tells us they should be receiving.

CHAIR CHOU: Karen, did you have a comment or is that old? Okay.

All right. Are there additional comments or questions or I think we may be ready to put the validity criterion to a vote. All right. Let's do the vote.

MS. PHILLIPS: Okay. We're voting on validity for 0054 with four options. One for high. Two for moderate. Three for low and four for insufficient. The voting begins now.

We're at 22. We have 0 for high. We have 16 for moderate, 4 for low and 2 for insufficient.

CHAIR CHOU: All right. So, we've passed the validity criterion. So, we're going to move on to feasibility.

Puja or Art, do you have comments for the feasibility criterion?

DR. KHANNA: So, the work group thought that feasibility would not be an issue. Obviously, the data elements are being

already captured and generated in the EHR.

One issue that I thought, you know, thus far since we are still dealing with ICD-9 codes is that we should be able to capture that under ICD-10 as well and obviously, we wouldn't be going into ICD-10 for a little while. So, that's something to keep in mind.

But, otherwise, no issues.

CHAIR CHOU: Art, anything to add?

MR. SCHUNA: No, it's just that they've already been doing this for some years and it appears like it is feasible.

CHAIR CHOU: Questions or comments from the rest of the panel?

Okay. I think we're ready to do a vote on the feasibility issue then.

MS. PHILLIPS: Feasibility for 0054, you have four options. One for high. Two for moderate. Three for low. Four for insufficient. Voting begins now. Yes, we do.

We have one Committee Member who

stepped out of the room. So, we have 21 votes. We have 13 for high, 8 moderate, 0 for low and 4 insufficient.

CHAIR CHOU: All right. We're going to move on. That passes. Then let's move on to usability and use.

I think some of this information has already been provided, but again, I'll turn the mike to Puja for additional comments.

DR. KHANNA: No, we thought that it was definitely usable. You know, commercial plans are using it and the benefits definitely outweigh the negative consequences. So, no issues regarding that.

CHAIR CHOU: Art, anything to add?

MR. SCHUNA: No, nothing further.

CHAIR CHOU: I just had one question. Have you seen improvement in the places that had low rates? I mean over time has that -- I think you alluded to that earlier, but the numbers have come up a little bit.

MS. WILLIAMS-BADER: They have and if I -- just a second here to see how they are coming up. Let's see.

Yes, we have seen that the lower plans are increasing over time. I'm --

DR. BARTON: I would just say that, you know, statistically, of course, that's going to happen. More likely that the 70 percent plan will show the ability to improve more than a 90 percent plan.

So, I'm not sure how much to make of it, but we have seen improvement in the 10 percentile over time.

CHAIR CHOU: Yes. I mean I think it's just you'd like to see something. Whether it's -- you know, whether it's just a statistical aberration or what. But, it's nice to see a move in the right direction at least.

Okay. Other questions or comments regarding the usability and use criterion?

All right. I think we can do a

vote then.

MS. PHILLIPS: Okay. For 0054 usability and use, you have four options. One for high. Two moderate. Three low. Four for insufficient. Voting begins now.

Okay. We're at 22. We have 12 for high, 9 for moderate, 1 for low and 0 for insufficient.

CHAIR CHOU: All right. We've passed all of the criteria. So, now, we're going to the final vote.

So, this is a recommendation for reendorsement. Yes or no.

MS. PHILLIPS: Okay. We're voting for overall suitability. One for yes. Two for no. Voting begins now.

Okay. We're at 22 and we have 21 yes and 1 no.

CHAIR CHOU: I think Thiru was trying to get in a final comment. We'll see if he sways the -- turns the whole group around here.

DR. ANNASWAMY: No, I was wondering if there was any discussion to be had similar measures.

MS. FRANKLIN: We've identified this measure as related to our next measure and we will be discussing those differences at the end of the day.

DR. KHANNA: You keep stealing my questions, Thiru.

CHAIR CHOU: Sorry. We can move to the next measure which is 2525. Which is also on DMARDs and this is a new measure. Okay. And it's for full endorsement. Okay. Great.

So, I would invite the developers to present the measure.

DR. YAZDANY: Right. Good morning, everybody. All right.

So, this is the last measure submitted by the American College of Rheumatology for this session.

And this measure, unfortunately,

there was actually a typo in the submission materials on the very first page in which we made it sound like this was for newly diagnosed patients. Actually, the wording was for newly prescribed DMARD therapy and I'm sorry, but we didn't catch that earlier.

But, the measure should actually read percentage of patients greater than 18 years with rheumatoid arthritis which we define as two face-to-face encounters with a diagnosis of RA. I'm sorry. Say that again. Who are prescribed, administered or ordered a DMARD in the measurement year.

So, this is prevalent RA and we're looking at DMARD use and this is actually the way that we tested this measure.

We've already talked about the rationale for this measure and in terms of the why is this important criteria, I think the most important thing that we've learned over the last decade of using the DMARD measure in the health care system is that this is a

disparity sensitive measure and we now have, I think, very good data to show that racial ethnic minorities, low income individuals, older patients and also certain geographic regions are much less likely to be dispensed a DMARD. That's based on the NCQA measure.

I just wanted to take a moment and point out that we are interested in harmonizing this measure. It's -- we have no intention of creating duplication.

The difference with the NCQA measure is that this is an electronic measure. It's an eMeasure and that we are intending this for physician level accountability.

But, what's the same is that we've aligned the DMARD list. We've aligned our definitions of rheumatoid arthritis. I think the measure concepts are the same.

What's different is that the specifications reflect the data source. So, for example, whereas the NCQA measure is looking at pharmacy claims and whether or not

a medication was dispense to the patient, not all EHRs are connected to pharmacy data and so, therefore, we find DMARDs by looking at whether they were prescribed, administered or ordered.

The other thing is that for physician level accountability, we've included a code for inactive rheumatoid arthritis. This is not a code that's commonly used we admit. It means that the diagnosis that was on the problem list is no longer active. It has an end date and it uses the diagnosis inactive quality data model element.

I think people need to learn how to use this code. I don't think a lot of people are using this code. I just wanted to point that out up front.

However, up to 10 percent of patients with RA over time may go into remission and so, I think that's something that we've heard loud and clear from clinicians. Is that we need to include that

and performance is not expected to be 100 percent at the clinician level.

Our exclusions are aligned with the NCQA measure HIV and pregnancy. The only difference is its inactive disease clause.

There are a lot of questions which I'd be happy to answer in more detail about including all possible contraindications to DMARDs and our expert panels had extensive discussions about these issues. But, we sort of went, you know, item by item and I think there were good reasons not to include a long list of exclusions.

For example, with TB testing, some DMARDs are safe for people that have latent TB. Even biologics can be started one month after initiation of TB therapy. So, that one was out.

Cancer, some DMARDs are safe in patients with cancer and the literature regarding the relationship between DMARDs and cancer even with biologics is evolving and we

couldn't reach agreement about that being a universal exclusion.

I will stop there.

CHAIR CHOU: Thank you. So, Art and Mark are the lead discussants on this. They're trying to defer to each other, but I think Mark has the hot potato.

DR. JARRETT: All right. I'll guess get started.

I mean a lot of the issues have already been discussed on the -- in the last measure. I'll say something politically incorrect, but I would actually like to see this done by ACR because perhaps there will be more slicing and dicing of the data and figuring out why some people aren't getting it.

So, for example, filling a prescription is great. But, a lot of physicians won't write the order even because they know it's not covered if it's a biological. So, we know that's the reality of

what goes on.

So, getting down to that level will be helpful and will also help us figure out why people are not prescribing certain things. So, I think it is useful in that regard.

One of the -- you know, the work group had looked at things and it's the same questions that came up before.

One of the issues had come up about when the diagnosis is made. If it's in that measurement period and they got seen, you know, almost at the end of the measurement period and then got seen once more, perhaps a DMARD wasn't -- the first visit was that was the diagnosis clinically. Second visit was let's talk about the drugs and, you know, there may be some slip up on that in that regard in why it wasn't started right away.

The TB testing and everything, although it was brought up, I agree with all the discussion and the same thing with the

cancer issues.

And the last thing was the DMARD list. Although it's very long and very inclusive and matches with the NCQA, it is a little bit of a problem. That some of the drugs many of us would look at and say is that really appropriate therapy for somebody who's got active rheumatoid arthritis. It does offer an option for that small select few perhaps that do have a malignancy and are afraid of taking certain drugs or where the oncologist may prevent it.

But, many of us would feel that for people with real active RA that's not really the appropriate decision. But, again, I think that's delving down deeper.

Art, anything else?

MR. SCHUNA: No, I don't think so.

CHAIR CHOU: John has a question.

DR. VENTURA: Yes, if I could just ask for clarification. How does capturing this data at an individual level allow you to

dig deeper versus capturing it at plan level?

DR. YAZDANY: I'll take that one. So, this is another measure that is programmed into our qualified clinical data registry and through that effort, we will be collecting much more data than just these quality measures and that's really going to be the primary reason and I actually think that the ACR's been very proactive about getting qualitative feedback from membership and, you know, sort of building that into our quality measurement loop and so, I think we can expect to see some granularity in the future.

CHAIR CHOU: So, just so I can be a little bit clearer on the difference between this and the prior measure, so neither of them are focused on how quickly you started. It's just whether you've been on it within the last 12 months. Right?

And so, the main difference is that one is an eMeasure focused at individual clinicians and the other one is really health

plan level. Is that correct? Okay.

Are there other examples of quality measure you guys do this kind of thing for?

MS. BURSTIN: There are many like that. Some of which are actually both at the health plan and at the clinician level NCQA, but there are some others where NCQA has the health plan level measure and the clinician level measure is done by the specialty society.

And the key thing there is, of course, the harmonization which is why that's what's been brought up. The last thing you want to do is have two measures out there at different levels that, in fact, are different enough or just different enough that what you're seeing is measured in noise as opposed to really getting at the details that you want.

So, appreciate the efforts already to date to try to get them permanent.

DR. YAZDANY: Just one question that I have and this is really for the NQF staff is I think it may lead to maybe a cleaner perception of measurements if there were sort of umbrella measure concepts, but then with different pathways to application or implementation and I think that would lead to the perception perhaps from the health care system and clinicians in particular that this is not just a proliferating of many, many different measures and it seems like this measure is a perfect candidate for something like that.

I understand there isn't precedent, but if there's a way that we can participate in this being the first example, I think we'd be very motivated to help with that.

MS. BURSTIN: That's exactly what we try to do on these harmonization examples. Perhaps not that explicit about seeing this as the umbrella concept, but that is very much

what we try to.

CHAIR CHOU: Other comments or questions from the rest of the panel? Jason.

DR. MATUSZAK: I just want to ask the same question I asked to the previous group and again, in reviewing the evidence that you guys submitted for this measure, I didn't see any mention of the studies that talk about long-term remission off of drug therapy and I'm just wondering, you know, why that piece is often times or not really accounted for well in this measure.

And completely off topic, how are you guys -- in your registry, how are you eliminating patients that have passed away? Are you continuing to keep their information in your registry in perpetuity? Are you resetting the data every year? Are you kind of pulling those patients back out? Are those going to continue to count in your denominator, but not in your numerator because you're not prescribing for them? Just --

DR. YAZDANY: Good questions. So, this issue of long-term remissions off of drugs, we do have some medical literature regarding that and in some studies, it's as low as 9 percent. In other studies, it's as high as 29 percent. I'm not sure I believe that because that doesn't actually coincide with any of our clinical experience.

But, in any case, it was very important to include that inactive disease data element because we need to capture the fact that some people go into remission and don't need drug therapy. That's our baby step in dealing with this issue.

But, I think that this Committee actually took an important step in endorsing a concept that will allow us to do that in a much more sophisticated way in the future. Because if we actually are able to capture disease activity scores and those fall into the remission range, that's actually the true denominator that you want, you know, and in

the future when everybody's doing disease activity and recording the score, then we can actually calculate something that's more meaningful.

But, I think this is the first step, inactive code and then eventually, the registry, you know, hopefully will be able to do that in a more sophisticated way.

In terms of patients who have passed away, we have not explicitly dealt with that issue. With our registry developers, obviously, at some point, we will have to actually test whether or not the electronic health record is reliably capturing that information and is transmitting that as a data element to the registry.

I don't think that testing has happened yet, but it's really important.

CHAIR CHOU: Linda.

MS. DAVIS: I just keep thinking about what we were talking about yesterday where the actual process of measurement

improves outcomes and the connection between the two and we haven't really talked that much about that with these two measures and I'm wondering if there's been any kind of analysis or thinking about how to learn more about that. Whether actually measuring this will make an impact on the outcomes of the patient.

DR. YAZDANY: So, I think that we have two types of evidence in terms of the relationship between DMARDs and outcomes.

The first is randomized controlled trials in which there was a placebo arm. Admittedly, many of those are old and the newer trials have combination therapy as the comparison.

But, from those older trials of which there are actually a substantial number, we know that for disease activity damage, often it's radiographic erosions and function DMARDs improve outcomes and I would argue that those are outcomes that matter. So, those are the randomized controlled trial data.

Newer data that's really based in claims at the population level using, for example, Medicare fee for service data show that patients who are not on DMARDs -- so, for example, if the comparison is people that are treated with steroids alone have higher number of -- greater health care utilization and more hospitalizations even when you do a case mix adjustment for all of their co-morbidities and where they live and socio-demographic factors and so, I think the observational data also support the fact that DMARDs are actually perhaps decreasing health care costs.

And we can argue about the biologics because those are very expensive. I'm not sure I can make that claim with the biologics, but I think in general DMARDs are cost effective and they improve outcomes in both observational and randomized control trials.

CHAIR CHOU: Yes, Linda, were you asking if there would be efforts to see how

patients do in these -- I mean if they're -- I mean did that answer your question?

MS. DAVIS: Well, it did kind of I mean, but we looked at the actual process of measurement and does that have an impact on outcomes? Not whether or not giving DMARDs and improving that has improvement.

So, let's assume that increasing DMARDs improves outcomes. Does measuring use of DMARDs increase use of DMARDs?

DR. YAZDANY: I'm not aware of any study that's looked at that.

CHAIR CHOU: Other questions or comments before we kind of move into the criteria?

All right. So, the first sub-criterion is evidence. I think we've covered this for the most part, but if there are other comments, Mark or Art. Others on the panel.

All right. Let's go ahead and do a vote on evidence.

MS. PHILLIPS: Okay. We're voting

on the evidence. You have five options. One for high. Two for moderate. Three for low. Four for insufficient with exception and five for insufficient. Voting begins now.

Okay. We've got 22. We've 9 for high. We've got 10 for moderate. Zero for low. Two for insufficient with exception and 1 for insufficient.

CHAIR CHOU: So, this passes. The evidence criterion. Let's move on to the performance gap.

Again, I think we've covered this evidence. Do the lead discussants have anything additional to add? Other questions or comments from the rest of the panel?

I think we can move to a vote.

MS. PHILLIPS: Okay. We're voting on performance gap for measure 2525. One for high. Two for moderate. Three for low. Four for insufficient. Voting begins now.

Okay. We're at 22. We've got 4 for high, 16 moderate, 2 for low and 0 for

insufficient.

CHAIR CHOU: That passes also. So, we move on to priority. So, again, anything from the lead discussants to add to what's already been discussed? Anything from -- any other comments or questions from the rest of the panel?

Let's do a vote.

MS. PHILLIPS: Voting on priority for measure 2525. One for high. Two for moderate. Three for low and four for insufficient. Voting begins now.

Okay. We're at 22. We've got 12 for high, 10 for moderate, 0 for low and 0 for insufficient.

CHAIR CHOU: That's passes as well. So, we move on. So reliability is the next and let me see if I describe this correctly.

So, we want to make sure that we -- we're looking at the specifications, whether they're clearly specified and then if

there's been reliability testing. Which I'm not sure if we talked for this. Is it -- it's not the same as for the last measure. Right? Because it's a -- okay and is this -- because this is new there hasn't been reliability testing?

DR. PACE: So, this measure has had limited testing in only two sites in two EHRs and this is the eMeasure. So, they're using data element validity like what we saw yesterday. Which will suffice also for data element reliability. So, we can look at those results and again, it would only be at the data element level and in this case, it was only two sites instead of the three that we saw yesterday.

CHAIR CHOU: Okay. And the results? Can you summarize what the results were?

DR. PACE: Yes. So, let me pull up the -- yes, our -- right. Why don't you summarize your results?

MS. PHILLIPS: All right. Let me get my -- let me just pull this up. All right. So, part of the reason that we only testified this in two sites was that there was still a discussion going on at the NQF and at the national level regarding how many EHRs were required and so, therefore, we weren't from the start planning to do additional and given that this was done on a shoestring budget, it just gets very expensive. So, that's why there's only two sites.

We did data element validity testing and so, we were looking for agreement about the diagnosis of rheumatoid arthritis as well as the performance score overall. So, whether or not a DMARD was found in these two face-to-face encounters.

There was actually some disagreement or instances in which the patient was in a DMARD as recorded in the note, but it was not found in the three places in terms of administered, ordered or prescribed fields in

which we were looking at this and I think this does reflect that medication reconciliation is still not perfect, far from perfect in electronic health records. But, I think that there are lots of efforts to try to improve that and I suspect that agreement will increase just as that increases in electronic health records. It was still within the substantial range. Which I think is adequate for the purposes of testing.

The other piece of validity testing was just the expert panel ratings and those had a median validity score of 9 out of 9. Which is the highest rating.

CHAIR CHOU: Thank you. So, for yesterday for some of these new measures, we were kind of looking at reliability and validity together. Is that what we're doing here? Okay.

MS. PHILLIPS: Yes.

CHAIR CHOU: Because we weren't doing that correctly for the last one.

DR. PACE: No, this one's fine.

CHAIR CHOU: Okay.

DR. PACE: And again, it's because of -- we say if you're doing data element, you could do the validity, too.

CHAIR CHOU: Okay. Thank you. And then, Mark had brought up some issues with the exclusions and things. Did you want to comment on that some more? Yes.

DR. JARRETT: No, I think that at this point I think I would -- I think it's fine and I think we can't let perfection get in the way of success. I think we're all struggling with eMeasures all over and we're not going to find the correlation, the R value we would like, for probably another two or three years, but that doesn't mean we shouldn't start the process and gather whatever information we can where we can and I think I could still support doing this.

CHAIR CHOU: Great. Art, anything else you wanted to comment on with the

reliability/validity?

MR. SCHUNA: No, not really.

CHAIR CHOU: Okay. Anyone from the rest of the panel with questions or comments here?

All right. Let's go ahead and vote. So, this I believe is going to be both the reliability and validity.

MS. PHILLIPS: So, we were just having a sidebar. We're talking about looking at this potentially as a measure that's for trial use --

CHAIR CHOU: Okay.

MS. PHILLIPS: -- potentially.

CHAIR CHOU: Okay.

MS. PHILLIPS: Because there's only two sides tested which may not meet the sufficiency of testing.

CHAIR CHOU: Okay. How does that impact what we're doing right now?

DR. PACE: So, we should move on to just the question about the specifications.

CHAIR CHOU: Okay. So, we're just talking about the specifications.

DR. PACE: Sorry.

CHAIR CHOU: That's all right. So, I think we're ready to do the vote.

MS. PHILLIPS: All right. Voting on trial measures specifications for 2525. You have four options. One high. Two moderate. Three low and four insufficient. Voting begins now.

Okay. We are at 20. Been holding here so we can vote again. We're at 21. There we go. Twenty-two. Thank you.

We've got 3 for high, 17 for moderate, 0 for low and 2 for insufficient.

CHAIR CHOU: All right. We pass. So, we move to the last couple of criteria. So, feasibility. Again, lead discussants, Mark or Art, do you have comments to make about the feasibility issue?

DR. JARRETT: It was the same issues about defining the onset of the disease

and the process over the year and that may throw things off. That was brought up by the work group. But, everybody kind of felt -- and the TB question was brought up again, but again, that -- I think it's been resolved in multiple discussions and those are the only issues from the work group.

Unless I left something out, Art.

MR. SCHUNA: Nothing further from me.

CHAIR CHOU: Comments from the rest of the panel?

All right. I think we can vote on the feasibility.

MS. PHILLIPS: Okay. Voting on feasibility for 2525 four options. One for high. Two for moderate. Three for low. Four for insufficient. Voting begins now.

Okay. We've got 22 responses. We've got 6 for high, 15 for moderate, 1 for low and 0 for insufficient.

CHAIR CHOU: That's passes and

then we move to the usability and use area. Again, Mark and Art, any comments -- additional comments here?

DR. JARRETT: No additional comments that haven't been covered that I know of.

CHAIR CHOU: Thanks. Art, nothing? Any additional discussion or comments from the rest of the panel?

Okay. I think we're ready to vote on usability.

MS. PHILLIPS: Usability and use for measure 2525, you have four options. One high. Two moderate. Three low and four insufficient. Voting begins now.

We are at 21. So, if you can all vote again. Okay. We're at 22. We've got 4 for high, 17 for moderate, 0 for low and 1 for insufficient.

CHAIR CHOU: So, again, we pass and we move to the overall vote. So, this is a vote for testing for trial measure. Yes.

Okay. So, this is a vote for trial measure. Yes or no.

MS. PHILLIPS: Measure 2525 for overall suitability trial measure. You've got two options. One for yes and two no. Voting begins now.

Okay. We are 19. So, if you would all vote again. Okay. Now, we've hit 22. We've got 21 yes and 1 no.

CHAIR CHOU: All right. So, that passes. Thiru really wants to talk about harmonization, but we're waiting for that for later.

MS. FRANKLIN: I think we're saving that for later.

CHAIR CHOU: Okay. All right. So, you got to hold your horses. We're going to talk about it later.

MS. FRANKLIN: Actually, since we're --

CHAIR CHOU: You want to hear about it.

MS. FRANKLIN: It's fresh in our minds. We can do it right now.

CHAIR CHOU: Okay. Did you want to make some comments?

DR. ANNASWAMY: I want to see what the others thought about it.

CHAIR CHOU: What was that?

DR. ANNASWAMY: I would like to listen on the discussion.

CHAIR CHOU: Okay.

MS. FRANKLIN: Sure and there's just a -- just technically speaking, we have handouts and discussions from the developers that they sent to us regarding harmonization that we're going to pull up on the screen and we'll pull those up for you to take a look at and we also have a side-by-side identifying the differences between the measures and yes.

And we'd also like to have the two developers, if they could, come sit at the table for this discussion.

DR. PACE: And just to note to the

Committee, the reason we wait until after you've evaluated each measure individually because if for some reason or another one of the measures would not actually be recommended for endorsement, it's kind of a moot point. So, we first think that each measure needs to stand on its own of meeting the criteria and then if we have two measures that are related or competing, then to move to the next phase. If they're competing, which one's better or if they need to be harmonized in some way.

MS. FRANKLIN: So, we're pulling up the side-by-sides of the two DMARD measures. One from NCQA and the new measure from American College of Rheumatology and the developers have already provided us their discussions, information about their preliminary discussions.

If you could review those for us and then we'll have a discussion about the differences between the measures.

DR. YAZDANY: I'm be happy to do

it. All right. So, like I mentioned before, we've already actually tried to harmonized a vast majority of the data elements.

So, for example, the measurement period in both measures is 12 months. Both require two encounters. We've gone through the medication list line by line and reconciled everyone so that they match.

We have aligned the exclusions which are HIV and pregnancy with the one exception which is the inactive disease NR measure and at this time, we are not able to capture that in claims data. Otherwise, we would put it into the NCQA measure as well. But, it's just not possible.

And then the other difference is how we actually find DMARDs and we talked about that. So, one is dispensed. The other is ordered, prescribed or administered NR. So, I think that the concepts and the measurement period and the numerators and denominators actually align as best they can

given the data -- the different data sources. I don't think that we had any piece of this where we couldn't come to agreement.

Am I remembering that correctly, Jenna?

MS. WILLIAMS-BADER: That does sound right. Yes. Good summary. Thank you.

CHAIR CHOU: Questions or comments about the harmonization of these two measures? Mark.

DR. JARRETT: Do you see two years from now retiring one of them and just being an eMeasure and that's it?

MS. WILLIAMS-BADER: Again, ours is specified at the health plan level. We are developing a long-term plan for eMeasures. It would actually take quite a bit of work and as someone who also exclusively works on eMeasures, I understand the difficulty in that work.

So, it's not something that we can say in the next couple of years we'd be able

to do and we do think that the plan level measure is important because of the reasons we stated earlier. Which is that we do see the plan as an accountable entity that would really be able to influence this particularly given the issues around cost of the medications and access to care.

So, probably not within the next couple of years. Let's think maybe five or ten and see how we're able to do with the plan level eMeasures when we're able to start rolling that out.

CHAIR CHOU: Jason.

DR. MATUSZAK: I do think that it's valuable to hold the plans accountable and the physicians accountable and however we need to do that.

I guess my biggest struggle with all of this is still the idea that we're not holding people accountable to getting patients into remission or low-activity status or anything else.

We're holding people and plans accountable for the actual prescribing of a drug and to me, that kind of -- it's -- yes, they're effective, but we also know that there's ways to treat many people without using drugs and you know and I'd always just encourage us, you know, is it enough to treat our hypertensive patients by giving them hydrochlorothiazide or do you treat to some target and actually work towards, you know, holding people accountable for getting these people these functional outcomes that you really want to have and so, I struggle with that part of it.

DR. YAZDANY: I think those are important points and I think it's important to articulate what the vision for the future is and what we're doing to get there and I think, you know, the first step is endorsement of the measures which require measurements of the outcome, functional status and disease activity. That's the walking before you can

run.

If we can reliably actually get our health care system to do that, then that future that you're envisioning where there's an outcome measure that's risk adjusted and that actually, you know, really drives quality improvement quickly becomes possible.

So, I think it's a great vision. We just can't there before we have the measurement in place.

CHAIR CHOU: Can I ask a clarification? So, I thought you said that inactive RA was an ICD-9 code. It's just not used very much. Is that right?

DR. YAZDANY: It's not an ICD-9 code.

CHAIR CHOU: It's not an ICD-9. So, it's a code that you guys made up that nobody --

DR. YAZDANY: It's not an ICD-9 code. This is a problem that begins and then is --

CHAIR CHOU: It's an ICD-10 code.

DR. YAZDANY: Yes.

CHAIR CHOU: Okay.

DR. YAZDANY: So, it's an ICD-10 code and --

CHAIR CHOU: So, it will be an ICD-10. But, if it will be an ICD-10, there's no reason that it couldn't be on the NCQA measure. Right? Even if -- just people won't necessarily know to use it right away.

MS. WILLIAMS-BADER: Yes, over the past two years, we've been preparing for the ICD-10 and transforming our ICD-9 list into ICD-10 list. So, this is definitely something that we can think about and consider as we're developing those.

CHAIR CHOU: I think that would help address Jason's concerns in some ways. If there is -- you know, if that's -- if they're not included in the denominator, somehow people who have inactive disease however that's defined.

Just a couple of other clarifications for me is you had already talked that the title is going to be -- the description is different. Right? This is the old description. We are no longer talking newly prescribed or diagnosed or whatever RA. It's really the same patient population.

And then this numerator is really not people who received a DMARD as it says up here, but it's people who were prescribed whatever, whatever DMARD.

MS. WILLIAMS-BADER: I apologize for those errors leading to mass confusion.

CHAIR CHOU: Okay.

MS. WILLIAMS-BADER: Sorry about that.

CHAIR CHOU: Okay. All right.

DR. PACE: So, that is a difference. Right? Your NCQA is dispensed. Yours is just that there was a written prescription or the --

MS. WILLIAMS-BADER: That's right.

That's right. So, we talked about that.

And the error is in that -- it's newly prescribed. Which just we need to take out the word newly because I think it's confusing everybody.

DR. PACE: So, do you have any -- did you discuss that difference and what it relates to? I mean, you know, dispensed is actually more specific than there was an order in a chart.

MS. WILLIAMS-BADER: Yes, we did discuss that. I think part of this is definitely coming down to the data source.

I will say that NCQA has transformed some of our HEDIS measures into eMeasures and we stayed true to the intent of the HEDIS measure. Which is that medications should be dispensed and I will say it's been very unpopular among the users of those eMeasures. It's an eMeasure in stage -- there are a few in stage 2 of meaningful use.

So, we are hearing quite a lot of

push back from EHR vendors and physicians saying they just don't have access to that dispensing data yet. But, it's certainly somewhere I think we'd like to go. It's just that field is rapidly evolving and there are a lot of connections that need to be made and sharing of information that's just not done yet.

DR. YAZDANY: To give people a sense, you know, from the clinical front line, current ePrescribing is a one-way street. So, you write a prescription and it goes to the pharmacy, but you don't actually get any information back to your EHR and people are trying to close that loop. But, until they can, it's not going to be possible. But, maybe in the next iteration, we'll be there.

MS. BURSTIN: Just one suggestion perhaps for ACR in terms of the ability to harmonize particularly since this is going to be out there. You're testing. You're working on it. Actually to stratify your rate by

dispensed and ordered. So, you can actually match the level of dispensed to the NCQA measure, but then we can address some of those issues we just heard earlier about how often it's ordered and perhaps the patients don't have the financial means to actually get it.

So, you'd learn a lot to Mark's point about being able to dive deeper. If you actually had those two rates, perhaps something you could test separately.

And then just one more point. Not so much about this measure, but as long as you're on the topic of harmonization, the two you, I know NCQA's also working on a -- because I chair the Quality Measures Work Group, the Policy Committee functional status has been a goal achievement for patients with RA and as you know, ACR just put forward a measure yesterday on functional status assessment for RA. So, even though they're not directly -- this measure is not done yet, I would just again encourage you guys to as

you're working through your measure be sure that the registry-based measure and the eMeasure are coming together.

DR. YAZDANY: So, just to give people an update on that effort which we've been involved in, even though the title of that is functional status, the measures listed are disease activity measures. It's confusing, but so, that measure will actually map. At least the disease activity measurement tools map to our disease activity measurement.

I just want to make that clear because I think it's confusing in the title and we've responded it's not functional status and we're working on, you know, the ACR saying get the disease activity measures into the EHR and record a score and what we're working to is setting a shared goal about what the target should be. Sort of an untested and interesting idea. So, hopefully, we'll hear more about that in the coming years.

MS. BURSTIN: I just would encourage that continued collaboration because I think that's your real opportunity there.

CHAIR CHOU: Yes.

DR. GRAY: Yes. Can you explain what harmonization means? I mean what's the point of the discussion exactly? Have the two entities come together and harmonized this and we're just looking at the end product or are we supposed to suggest something or what are we -- what's the goal?

MS. FRANKLIN: So, the Committee can review the differences in the measures and make suggestions as to how the two developers should work together to make sure the measures are harmonized so that the reporting burden in the field is reduced. But, I think that's the ultimate end goal here.

CHAIR CHOU: Linda.

MS. DAVIS: I keep thinking about public reporting and patients like Kelly and others who might want some of this

information. Is that on the road map for either of -- I mean I know NCQA is kind of out -- I don't think it's available to consumers who are the good arthritis health plans, but it would be -- I always think about the ability for people to actually see the results of all this work as consumers as well.

MS. WILLIAMS-BADER: So, I will say our measures are actually -- you can access data. I unfortunately am hazy on the details and my NCQA staff over there are going to be really disappointed in me.

But, you can actually access information on the rates. I do believe you need to license that data, but it is available and we do rank the -- we rank the health plans. I don't know if that goes into the level of detail of each measure and their scores. But, that's one way in which the data is publicly available.

I think we both agree, both our organizations agree, that we'd like to get to

a point where the physician level measures are publicly available as well.

I know that in PQRS average score or summary scores across all the physicians that are participating in the PQRS program are available, but we've not gotten to the point yet where we are or CMS has not gotten to the point yet where they're reporting those PQRS measures at the physician level.

But, I hope that's the direction we're all going.

MS. DAVIS: But, it doesn't sound like it's in the road map. It's not a specific plan at some point in the future yet.

DR. YAZDANY: I think there is a road map and I may be getting the details wrong, but even in terms of the qualified clinical data registry, you know, the more advanced functions are fully transparent data eventually. So, I think we're working towards that, but we're not there yet.

CHAIR CHOU: Thanks. Are there --

Thiru.

DR. ANNASWAMY: I guess we still don't have the answer on what exactly we are doing with the harmonizing task and then the other one is going to Jason's point earlier. I think in the long-term vision of this whole idea of tying performance to outcomes perhaps, you know, I heard of the ICD-11 being planned which will directly be interfaced with the ICF, the International Classification of Function, to where coding will be matched with the functional stratification of patients.

So, perhaps that vision will come together in the long run.

MS. BURSTIN: In terms of your first question, in terms of what we're expecting today, we want to have this opportunity for the Committee to look at the side-by-side, see if there are any obvious differences there you think beyond what they haven't harmonized that they should work on. What we usually expect is that within one year

by the annual update for the measures that any suggestions made by the Committee will be re-reviewed to be sure that they've met those harmonization goals.

So, at this point, we've just -- this is really a preliminary conversation to see if there's anything obvious here that you think they need to specifically continue to focus on and apply.

For example, I raised this issue of dispensing versus prescribed.

If there are any other issues you think they should work on, otherwise, you know, we'll just let them continue to work on this and we'll kind of assess again as the measure moves forward and again, since the ACR measure is still being tested, there's a nice opportunity there to see if anything else emerges that may require further harmonization.

CHAIR CHOU: Jason.

DR. MATUSZAK: Just really quickly

and you guys probably already touched on this, but how come the differences between the drug lists and what are the individual differences? Why does one plan or why does one measure count certain drugs and others doesn't or are they included in there? Just summarize in a different place. Because I don't see some of these on both sides of the --

DR. YAZDANY: So, this is the first pass of the medication list side-by-side and some differences. For example, if I'm remembering correctly, the ACR list didn't include some Gold formulations or mycophenolate. For example, this is two differences and we just -- we reconciled those differences.

Like for example, the clinical discussion regarding the Gold formulations was that it would be preferable to have a harmonized measure rather than argue about the efficacy of, you know, one Gold formulation versus the other. Especially since claims

data tell us that, you know, there are two patients in the U.S. taking gold anyway. So, it's just -- you know, it doesn't make sense to argue about Ns that small.

And that's actually -- you know, that was true for anything that we disagreed on. So, no one's using it anyway. It's almost like, you know, what's the point. But --

DR. PACE: So, what does that mean then? Are you going to have lists that are the same? I guess that's the question.

DR. YAZDANY: We actually -- somebody made a spreadsheet. Right. And I don't know where that is. With the harmonized sort of side-by-side. I think we should, you know, email that out at some point then. Right.

CHAIR CHOU: It would be fairly straightforward just to make sure that the drug list is the same. So, that, I think, is an easy fix.

Other questions or comments? I mean I guess my other thing would just be that it would be nice for the exclusions to match up precisely especially with this, you know. However, you guys end up handling the inactive disease piece. So, that would be the other thing.

Jason.

DR. MATUSZAK: And one more. I see that on the ACR measure that you guys highlight that it's got to be two visits with the same clinician. Whereas, the plan measure is two visits with a diagnosis of RA regardless of the provider. Is that correct?

MS. WILLIAMS-BADER: That is correct. I guess I should speak that. Again, I will look for signs over there from NCQA staff to -- if I state anything incorrectly.

We don't have any measures that indicate it needs to be with the same clinician.

So, I know Mary's not at a

microphone. So, I'll repeat what she says and just add to that.

What Mary was saying is that at the health plan level, we're not using it for accountability of the physician. A physician level measure you do want to try to insure that the patient is really in that physician patient panel. So, you would want those visits to be with the same clinician.

On the plan side, we are using those multiple claims to insure the diagnosis correctly represents the diagnosis the patient has. So, it doesn't -- because we're holding the health plan accountable for the measure, it doesn't matter if that was with the same clinician. We're just saying that if the patient truly has RA as represented by two claims for the RA diagnosis, they should be receiving this medication.

Because as Mary pointed out, they might be treated by a team. So, it doesn't mean that the patient has to see the same

clinician.

DR. YAZDANY: So, let me just make one more comment. I think that's great. I agree with that.

And in terms of accountability, one thing that came out during testing is that it's not uncommon for a rheumatologist on the first visit evaluating somebody for inflammatory arthritis to have a code of rheumatoid arthritis and what they mean out -- what they actually mean is rule out rheumatoid arthritis and so -- and there were instances where we found that in testing and so, in order to increase specificity which is important if it's going to be a provider level accountability measure, we were able to do that with the two codes. So, it sort of reflects sort of the clinical practice style and coding I guess that exists right now.

DR. MATUSZAK: One more piece on the two clinicians thing. If it's you and your nurse practitioner or PA, do you count as

the same clinician or you count as different clinicians?

DR. YAZDANY: So, the measure authoring tool as far as I understand it which granted I may reveal my ignorance here, but we actually can't assign attribution in the measure authoring tool and so, that is done in implementation of the measure at the practice site.

MS. WILLIAMS-BADER: I might actually be able to help out here.

If the measure is being considered for the Meaningful Use Program, otherwise known as the CMS EHR Inventive Program, then the eligible professional is the one who is reporting the measure and except for some differences in Medicaid, I believe that is going to be physicians.

But, that doesn't mean that if other members of the care team provide the care that the eligible professional won't get credit for it. As long as the information is

available in the EHR, then the eligible professional gets credit.

So, you're sort of holding the eligible professional accountable for the care team and making sure the care team is delivering the care to the patient. But, I don't believe that it -- that they would be dinging in a sense the provider if it is a nurse practitioner who provides one of those visits.

CHAIR CHOU: I think Thiru had a comment and then Kelly and then Art.

DR. ANNASWAMY: So, at the end of this discussion, is there an action item that you expect this Committee to do like we did for the approval or the endorsement of the measure? Do we just discuss and say guys, work it out or do we say work it out with option A, option B, option C? That's kind of what I was asking earlier.

DR. PACE: No, I think it's primarily about making recommendations. So,

where -- that you still expect some harmonization and probably identifying if it's short term versus long. Like the medication list. Maybe that's a short term thing versus something that might be longer term.

But, ultimately, I will say if there's something that is really just a huge problem that you see, it could result in you going back and saying well, we don't think we should recommend the measure for endorsement. That would be kind of an extreme situation and probably relates to one not really closely following the evidence that exists versus something else.

But, I mean generally it's really about improving it and recommendations for them to accomplish that by the time it comes back.

DR. ANNASWAMY: -- the recommendation then?

MS. FRANKLIN: Not at this time. No. We're not voting today. We would capture

the suggestions for the developers so far that we've had around harmonization in terms of the denominator. I heard something about the exclusions as well as the descriptions of the measures in the medication list and we would capture that in that report, put that out for public comment and come back with crisp recommendations for the measure developers mostly likely after the comment period has ended.

CHAIR CHOU: Yes, so, I think we're into our break. So, I want to try to wrap this up. A couple of other people wanted to say something and then I think we can summarize recommendations from the panel and then hopefully close this out.

But, Kelly, I think you had your hand up.

MS. CLAYTON: I just had a quick question. For the ACR measure, is the only reporting party rheumatologists? Because, you know, sometimes patients are diagnosed by the

family practitioner, get referred to a rheumatologist, wait three to six months for a visit, don't receive a DMARD on that first visit and I didn't know if there was a percentage of the population that could potentially be lost in reporting.

DR. YAZDANY: So, our measure for performance at the rheumatologist level, but I think your point is a really important one. There are many areas of the United States where there are not rheumatologists and access has been a problem with long wait times.

Now, one thing that's interesting about going forward with implementation both at the health plan level and at the rheumatologist level is that our more groundbreaking health systems have figured out that those two entities actually need to work together and use tools like telemedicine, electronic referral systems and other innovative things to do, population health management and actually in some ways having

both of these measures go forward allows for that because the rheumatologists learn population health management through the registry and the health plans have an incentive to increase DMARD use.

So, we need to work towards an access measure, but in the meantime, I actually think that there's some interesting things happening on the ground to move this field forward.

CHAIR CHOU: Art wanted to make a comment.

MR. SCHUNA: Yes, I think it will be interesting to see how these two compare when you begin getting results because there are differences in data collection with one being electronic, the other including paper records.

The differences in these drug lists don't seem that great to me. I think in some cases they have listed drugs by category rather than exact name and when you consider

that, they tend to be pretty similar to me.

I see down the road eventually though that this could just be one measure because if the ACR criteria evaluation tool included plans, you could group the plan data and you'd have both pieces of information in one criteria and I think that was -- there was one other point, but I can't remember it right now.

CHAIR CHOU: All right. Thanks. So, let me just try to summarize and let's wrap this up here.

I think, you know, the description -- this isn't a harmonization issue, but just that they're going to, you know, fix it to what the current description is and then those will be fairly well harmonized.

We had talked about, you know, even though the lists are very similar, it would just be nice if they just looked the same so that it's easy for people to see that and I think that's a very easy fix there and

then I think that the other main issue was just how to deal with this inactive RA issue. If that could be something for the two groups to work on moving forward.

Were there other suggestions from the group for harmonization?

DR. PACE: I think the other thing that came up was the dispense and order.

CHAIR CHOU: Yes.

DR. PACE: Just to at least note it or --

CHAIR CHOU: Yes, that's one of the other things to look at.

All right. So, I think we are past break time. So, we'll take what? Ten minutes or something and then reconvene here at 12 after or so. Thanks.

(Whereupon, the above-entitled matter went off the record at 11:01 a.m. and resumed at 11:11 a.m.)

CHAIR CHAO: All right, we are going to move on to our next measure. It's

number is 0662. It's median time to pain management for long bone fracture. The measure steward is CMS, and this is a measure for reendorsement.

We have someone from CMS on the line.

Can you give us a brief overview of the measure?

MR. BRATZLER: This is Dale Bratzler. I'm not from CMS, but I work with the contractor, Acoma Foundation Medical Quality, for the measure. This measure was developed as a part of a group of performance measures for emergency departments that focuses on timely access to healthcare.

There were a number of performance measures for the emergency department that focused on throughput times, time to see a provider, provider decision to admission, and this particular measure focused on timeliness of pain management for patients because, as we know, many patients end up sitting in a

waiting room for prolonged periods of time. So the measure focuses on the median time from arrival at the emergency department until the patient receives pain medications.

We delimited the denominator of the performance measure to patients with long bone fractures for two reasons. The first was that there was little controversy that the vast majority of patients who have a long bone fracture need pain medications.

Secondly, we didn't want to have some vague description of pain or some pain syndrome that might have unintended consequences of getting people to pursue, you know, going to the emergency room because they knew there was a performance measure requiring timely administration of pain medications. We wanted to avoid that.

So, working with the Emergency Department Technical Expert Panel, we focused on long bone fractures because, as one of our ER docs said, if somebody consciously breaks

a long bone just to get pain medicines, so be it; we're going to give them pain meds.

So it is a median time procedure. Patients are excluded from the denominator if they've already received pain medications, if they refused to get pain medications, or if there is some explicit contraindication to getting the pain medication.

That's just a brief overview.

CHAIR TEMPLETON: Thank you.

Sean, Wendy, and Chris are the lead discussants on this.

I don't think we need to repeat the overview, so I think we can start talking about the evidence. Do one of you want to take the first --

DR. VISCO: Sure, I'll take the jump right in there.

Just to go back to the type of measure here, I note it's being listed as an efficiency measure. We were of the feeling was this was a process measure. And

certainly, we all felt that way on this unless we were, you know, missing something overall.

In terms of the evidence for 1A, we felt that there was moderate evidence supporting a delay in treatment. There is very good moderate evidence for ethnic disparities, especially in analgesia delivery. However, we felt that there was very weak evidence regarding a particular time frame for treating the pain, and there was really no clinical guidelines to support a particular time frame for treatment. So, in this particular measure, that's a 35-minute benchmark, and we really couldn't find the evidence to support that.

Shall I go into performance gap?

We're going to focus on evidence first.

And then Wendy or Sean, did you have additional comments to make your?

DR. BRYAN: No. I think Chris summed it up nicely. You know, the feeling of

the work group was that there probably is, taken in total, probably moderate evidence.

CHAIR TEMPLETON: Is there anything from the rest of the panel?

DR. BUTLER: Can I just have some clarification on this, from my understanding as to what an efficiency measure is. I understand it has two components, the resource use and the quality component, and there's no evidence required for the resource use component.

Beyond that, I'm having, I guess, trouble distinguishing really from the process of measure.

Sorry.

CHAIR TEMPLETON: Sorry, if you could be just a little bit closer to the mic, I think. And there was a helicopter or something.

So, if you could, repeat your comment.

DR. BUTLER: Okay. Let me see if

I can remember about I just said.

(Laughter.)

DR. BUTLER: So I guess I'm having a little trouble with the -- you know, they said they thought this is a process measure. And I tend to agree, especially when those different components are broken out to identify it an efficiency measure. So that's what I'm trying to distinguish.

Is there a higher guide rule?

MS. PACE: This is Karen Pace.

I would say that I guess we would tend to agree. I mean this is kind of about getting the right treatment at the right time for the patient, and if someone's in a pain, sooner is better than later.

I think -- Dale can speak to this -- they may have been thinking of efficiency in terms of the efficiency of their systems to make this happen. But, you know, it is about patients getting the right treatment.

Dale, do you want to comment on

why you are thinking efficiency?

DR. BRATZLER: Well, again, this is just one measure in a whole set that had been discussed by other committees looking at timely management of patients who present to the emergency department. This is just one piece of a group of measures that focus on efficient movement of patients from arrival at the emergency department through the emergency department.

There was a perceived need to focus on management of pain, and we spent long periods of time talking about this, about how we develop a measure to promote unintended consequences of inappropriate use of pain medications for patients seeking medications, and yet, looked at a group of patients where there was general consensus that timely treatment of pain is a reasonable exclave expectation.

DR. PACE: Right; I don't think anyone is disagreeing with the timely -- and

the classification of it, we can handle.

We probably need to look across some NQF measures about the timeliness. By I don't think it will change your evaluation of the measure at this point.

DR. BRATZLER: I don't think we ever tried to argue that there was an outcome other than patient satisfaction, perhaps, related to the median time to pain management and patient outcomes. I don't know that it changes the outcome for a patient and so I don't we ever tried to make that argument.

DR. PACE: Right, and you haven't really baked in a certain benchmark. It strictly is the median time for each institution; correct?

DR. BRATZLER: That is correct.

CHAIR CHOU: I guess a question I have for NQF is, is there a different algorithm for efficiency measures? I mean I don't really see that on here.

DR. PACE: No, there isn't.

I think you're well point is well taken. For strict resource measures, if it's just cost or number of visits et cetera, we don't have the evidence component. But if it were an efficiency measure that included cost and resource use plus quality, the quality component would still need evidence. So I think it's okay to just proceed.

CHAIR CHOU: So I think we treat it as a process measure, I think, is just basically my take from that. I think this was mentioned before, that most of this is indirect in the sense that it's based on evidence that, you know, some people don't get pain medications immediately and are therefore having some suffering. I know there's also some evidence about disparities.

Do other people want to make comments about -- oh, yes, Kim.

CHAIR TEMPLETON: This is Kim -- yes; I'm sorry. This is Kim.

I just have three questions. One,

in the definition of pain medications, does this also include regional blocks? There's increasing information in the elderly with hip fractures about the efficacy of using ileal fascial another regional blocks. Would that be included?

DR. BRATZLER: Yes, it is, including intrathecal, regional, nerve blocks, Biers procedure. All of those are included.

CHAIR TEMPLETON: Okay, and two other questions. One, what if you have someone who comes in with pain who does have an injury, but yet, when you're trying to make a diagnosis, you get a plain x-ray and your picking up an old fracture. This is not an acute fracture; their pain is due to something else.

Would those be included within this?

DR. BRATZLER: Here, it would depend on the billing for the diagnosis, so I can't tell you for sure whether it would or

not. But these cases are selected based on the diagnosis codes.

CHAIR TEMPLETON: And then I guess the last question would be, you'll have some young kids who will come in, as well as the elderly who are osteoporotic, who have a mechanism of injury, but because of their metallization of their bone, you can't detect a fracture on plain film. And so you're down to MRIs, CT, some other imaging modality to confirm the diagnosis, which is going to take time. And would it be beyond your time frame that's delineated for this.

How would that be accounted for?

DR. BRATZLER: Well, of course it is a median time, so that addresses some of the outlier issues, that if the case came into this denominator, perhaps it should be, you know, could extend to --

(Simultaneous speaking.)

CHAIR CHOU: Okay, thanks -- I wanted to try to stick with the evidence right

now we can talk about the speculation --

(Simultaneous speaking.)

CHAIR CHOU: Thanks for those comments.

Mark, would you go ahead?

DR. JARRETT: The rest of the specifications that --

(Simultaneous speaking.)

DR. JARRETT: I'm just getting a little -- especially because --

CHAIR CHOU: I have Dale on the phone.

Do you want to respond to that?

DR. BRATZLER: I can't say much to --

(Simultaneous speaking.)

DR. BRATZLER: -- by the American College of Emergency Physicians --

The PPE approach --

(Simultaneous speaking.)

CHAIR CHOU: Do you -- patient satisfaction is a reality. Do you have data

showing that -- whatever, whatever measurement -- some of that kind of kind of thing?

DR. BRATZLER: I don't know. I don't have any particular patient satisfaction data.

Again, this is part of an efficiency group looking at timeliness of care for a whole group of emergency department measures.

CHAIR CHOU: Okay.

Are there other comments from the group?

John, and then Jason.

DR. FITZGERALD: Yes, I guess it's to kind of tie in to what Merck said. And that is, it was originally endorsed in 2011, and now, three years later, we're going to keep this same work for the emergency department staff. I don't see that it's moving towards an outcome improvement.

CHAIR CHOU: Jason?

DR. MATUSZAK: I guess I just had

more of a question about for the developer about the research that they cited.

Which particular long bone fractures did you guys find that pain management was not being adequately addressed? And how do you address those long bone fractures?

I mean which long bones are we talking about? Everything the long bone? Hands? Feet? Distal fibula? Give me an idea here.

DR. BRATZLER: So it's primarily humorous, radius, ulna, femur, tib, fib.

I'm not sure if Wanda or somebody from NQF has an open mind. They know the denominator codes. It's all based on codes. So, hands, foot, would not be in the long bone group.

CHAIR CHOU: Are there other comments or questions?

(No response.)

CHAIR CHOU: All right.

So I'm just going to sum up where we're at. Although there is a reendorsement, there are some questions about the evidence supporting the measure. I'll just stop there. There are some questions about the evidence supporting it, and it sounds like it's quite indirect.

At this point, I see that a couple of the workgroup members did try to rate it, and some people -- I mean there's at least one moderate or high in the workgroup comments -- but again, there are quite a few questions that have been already reflected, I think, in the discussion.

So I'm going to pause there. If there are no further questions or comments, we need to put this to a vote. Remember, it needs to pass the evidence criteria to be able to move on.

So let's pause here and let's do the vote.

MS. PHILLIPS: All right. We're

voting on 0662. We're voting on the evidence.

You have five options: "1" for high, "2" for moderate, "3" for low, "4" for insufficient evidence with exception, and "5" for insufficient evidence.

The voting begins now.

Okay, we're at 21, and we have 21 in the room.

So we have zero for high, three for moderate, seven for low, two for insufficient evidence with exception, and nine for insufficient evidence.

CHAIR CHOU: This is below the 40-percent cut off -- yes -- we have three moderate and two insufficient with the exception, so that still comes out to 24-percent.

So I think we stop here, unless there's other --

MS. MARINKOVICH: We can stop here.

DR. MATUSZAK: Yes, just one more comment.

It was mentioned earlier, care coordination --

(Simultaneous speaking.)

DR. MATUSZAK: -- simple item direct from the patient.

John Watson once again -- in Kaiser Health -- has fielded a single item of patient report of harm. And once again, the rates are fairly high, and they do vary quite widely. So it's another potential way of looking at harm, using patient reports through some database on that from John Lawson.

CHAIR CHOU: So I'm going to open up open it up, maybe first to the lead discussants, and then to the rest of the panel, for further discussion here.

DR. MATUSZAK: Sure. I think part of the challenge was the complexities and subtleties in the evidence with regard to level of care and type of ER that care is

involved in.

For instance, in tertiary care centers in the Friedland study, head injury was associated with lower analgesic use, but that wasn't necessarily a bad thing. In multi-trauma cases, which -- I think it was, and I can't remember of the top of my head; I think the Brown study in 2003 -- at least 10-percent or maybe more of patients had complex cases, and these are cases where the potential administration of analgesia would either complicate the diagnosis or would also involve missing something with visceral or other traumatic causes. We can't really address underestimate those 10-percent because those are real dangerous types of cases that need, you know, a lot of thoughtful care.

So there was a potential for that, and again, that gets into a lot of subtlety, which mostly related to some thought that it was difficult to look at this without thinking about risk stratification for -- and that came

up with in a workgroup as well, which gets outside of the evidence -- but it's still related in terms of how the data kind of pans out.

I had a couple other thoughts, but I'll turn it over to Wendy and Sean.

MS. MARINKOVICH: So a couple of things I'll add is that we didn't see a lot of movement in the measure. There was a slight movement, but we didn't see a lot of movement, and that was one of the discussion groups that we had.

DR. PACE: That's right. Yes, we would just say that, what were the evidence issues?

MS. FRANKLIN: The evidence issues -- okay -- no, go ahead.

DR. PACE: So let's break it down.

This measure is about timely administration of analgesia for patients with long bone fractures. So the evidence they provide -- and we should talk about the

evidence they provided. And if we look at that of the process, getting the right treatment at the right time, the desired outcome for this would be, first, symptom control, pain control, and patient satisfaction, are probably the two desired outcomes.

So, you know, we can think about evidence of pain treatment, and does that influence pain control and patient satisfaction. We can look at if they provided any evidence about, you know, the time that that should occur.

But I think some of the other issues you're bringing up at the do with the performance gap, which, if there is none, that's where you should vote that down, or how the measure is actually is constructed, whether it needs solutions and risk stratification, et cetera.

But I think we need to really focus on the evidence first and then talk

about these other things under the appropriate criteria.

DR. BRYAN: And that's why I brought that up the way that I did, because we as a workgroup did kind of feel like, in general, this sort of hit at a moderate, and a couple of us felt that it was on the low side. But if this got to 1B and 1C, I think that we would see that's where we had more of this kind of discussion, which is why I wanted to bring in some of performance issues that we had, that maybe people are looking at that and already bringing this down to a low rating.

But that's, again, where we were going to come in and say, this is probably low in that realm.

DR. PACE: Right, and I think that's fair. But, you know, we really do want you to evaluate the criteria as they're stated because it's helpful for other reviewers to understand where potential differences in the measure are for the measure developer.

You know, the fact that this really isn't based on a systematic review of the evidence, it probably would fall into the 'moderate' category to begin with, or it could be, you know, perhaps meet the exception. But I think we just need to be clear on what criteria you're basing your vote on.

CHAIR CHOU: Yes, I mean I'll summarize some of the stuff that I've heard and my take on it.

You know, really, no direct evidence has been presented. So, at best, this can be moderate to begin with. But even the indirect evidence, I think, is problematic, to me at least, in that we don't really don't know what the optimal time for administration is.

I mean there's this kind of assumption that a sinister patient hits the door, they should be given analgesics, and I'm not sure that's necessarily the case. And then this in issue about heterogeneity, it's

kind of tied in with the specifications, but I also think it's also tied with the evidence because there is heterogeneity in the patient presentations, and not everybody should be getting analgesics within 10 minutes of walking into the ER, or whatever.

So I think that the indirectness, at least to me, is an issue and then, related to the indirectness, some of these assumptions that are being made. I think it's difficult. I actually think it's difficult to interpret what it means that the median time went down by three minutes. I don't know if that's bad. If it's good, it's only a slight improvement, but I don't even really know, really, how to interpret that.

So, I'll open it up to see what other people, with the other takes have been from either the leads or other folks on the panel.

DR. JARRETT: No, I agree with Roger. I think that if we just follow the

algorithm, you know, it's not a patient-oriented health outcome, and that takes us down that path. You know, I mean the evidence that they present is not a systematic review per se. So we're looking at these individual studies. Individual studies don't actually look at what the measure is testing.

And then the only thing we're really left with is whether or not that actually should be considered for an exception --

(Simultaneous speaking.)

CHAIR CHOU: Marcie?

DR. HAYES: So my comments would be similar to Jason's and may not --

(Simultaneous speaking.)

DR. HAYES: And what we concluded and recommended was that was a good measure for use within hospital over time for not sufficiently accurate for a cross hospital comparisons, given the results I just mentioned.

(Simultaneous speaking.)

CHAIR CHOU: I can't remember what the vote was, but were there people who did feel that the evidence was better, moderate at least, and wanted to comment on how kind of where they came to that?

MS. MARINKOVICH: I was one of them with the moderate.

CHAIR CHOU: Yes.

MS. MARINKOVICH: In going down the algorithm, I agreed that there was no systematic review, where they had a body of literature that they turned in.

And so, with the information presented, there did appear to be an issue, and this literature supported that there was disparities between them, that there was a need for pain medication in these large bone fractures.

I don't disagree, as well, with the, if it's trauma, you know -- but that was for specifications where I felt that there

would be some additional thoughts, you know, with exclusions of patients that had issues with other areas.

But with the evidence itself, I did read it as moderate following the algorithm that was in the indicia.

DR. BRUETMAN: Yes, I looked at it in much the same that Wendy did, and it was kind of between low and moderate. But I felt that overall, the evidence, although indirect, probably fell at a moderate level. However, you know, when you go down the priority, reliability, validity, usability, and just the fact that, you know, there is really is no direct evidence that giving the pain medication sooner really is a proving the quality of the care or the outcomes.

I felt, you know, there were enough issues with this measure that I was going to vote low on a lot of other things, but this was the one thing where I felt that it did need to meet that kind of minimum bar.

CHAIR CHOU: Thanks.

Are there any other comments or questions?

Mark.

DR. JARRETT: Since didn't get to specifications, looking at median time really doesn't address the issue of disparate care because, although the median may drop, there may be a large segment on one side of the curve that is still getting poor care. And looking at the median time, you may come out looking really good but still be giving disparate care to certain populations. So I think it's not achieving the purpose of what the measure was supposed to do.

CHAIR CHOU: That's a great point, thanks.

DR. JARRETT: And during our discussions, the developer agreed with us that this was really a reflection on time to diagnosis as well.

And you know, furthermore, one of

the discussions that came out -- just to give information back at this point in the developer -- one of the discussions that came out was that long bone fractures are such a small portion of pain in the emergency room that they really should think about the problem when thinking about this. It didn't seem to stand alone as a measure, to us.

Again, it's further down the algorithm. It's not speaking to the evidence, but it's speaking how the developer could consider this particular measure.

CHAIR CHOU: Marcy?

DR. HAYES: So I'm just trying to kind of -- I thought I heard the developer say that there was some consensus with a particular committee, but I'm not seeing that in the documentation.

Did I hear that incorrectly?

CHAIR CHOU: I think they worked with some of emergency room group.

Is that correct?

I think that's what he was talking about. I don't think there was necessarily ACER or whatever, that it was a professional society necessarily.

DR. BRATZLER: Yes. So most of the members of the tech technical expert panel -- not all, but most -- are members of the American College of Emergency Physicians.

CHAIR CHOU: Did you want more from us, or what is the process here now?

MS. PHILLIPS: We were just discussing if the committee, after this discussion, if the committee's concerns were with evidence or with some other part of the criteria.

That was our only discussion. I don't whether we wanted to tease that out more, or.

CHAIR CHOU: I think I heard expressed concerns about the performance gap as well as the priority things. We could do votes on those if you wanted feedback there.

DR. PACE: I think the only question is, after this discussion, whether anyone would change the vote that they gave on evidence. If not --

CHAIR CHOU: We can revote.

DR. PACE: That's up to --

CHAIR CHOU: If that's the appropriate procedure, we can revote and see what's happened.

DR. PACE: Either way -- I mean if people want to indicate whether they would change. If everyone feels like it's going to be the same -- you know that, I mean --

CHAIR CHOU: Well, that kind of puts people on the spot.

Why don't we just your revote and just vote how you feel.

MS. PHILLIPS: All right, we're voting on 0662 on the evidence.

You have five options -- starting now.

Okay, we have all 22 responses.

Zero for high, two for moderate, six low, one for insufficient evidence with exception, and 12 for insufficient evidence.

CHAIR CHOU: Okay.

(Laughter.)

Thank you, Dale.

CHAIR CHOU: I think we're going to move on to the next measure now, which is 0052, use of imaging studies for low back pain. The developer is NCQA.

Let me pull this one up. We have -- oh, there we are -- Mary.

(Off microphone comment.)

CHAIR CHOU: So, actually, I guess I should mention that.

I wasn't involved with the development of this measure, but I have been involved with a lot of ACP stuff related to low back pain imaging. So I don't know if I need to recuse myself from voting or anything -- no? Okay -- so, just as a disclosure.

Mary, can you give us an overview of the measure?

DR. BARTON: I'll just start by saying that among the measures in the HEDIS set, a very small number are focused on overuse, and this is one that has been around for several years. We've seen some progress on this, but really, my hope is that over the future, as issues like choosing wisely gain steam, and as clinicians think about how they might apply care most judiciously to patients who need it, that these kind of measures will be more successful in changing practice.

But I'm going to let Jenna give an intro to the measure.

MS. WILLIAMS-BADER: Thanks, Mary. You did cover a little bit of what I was going to say.

This is, as Mary pointed out, an overuse measure. Here, we're looking for the percentage of patients with a primary diagnosis of low back pain who did not have an

imaging study within 28 days of diagnosis.

A couple of things to note about the description: We are looking for a patient who did not have an imaging study, and we do this by actually -- we create an inverse measure where we actually look for claims of the images, but then we subtract that from one to come up with the measure weight. So, although it is an overuse measure, a higher score is actually better because we are looking for those who did not, and that would be the appropriate treatment for those patients.

Again, it is a health plan-level measure, and we are relying on claims data for the measure reporting. This measure was actually developed that a similar time as the DMARD measure that we were discussing earlier. It was actually developed by a panel that had, or, it was informed by a panel that had been developed that had come up with pain measures before. I mentioned them during the DMARD

discussions.

So this was identified as an important topic by a group that was looking at possible pain measures, and that was the group of NCQA, the American Medical Association, and the joint Commission. It has been and continues to be an area of strong interest, and the guidelines continue to support that imaging within the first few weeks of a diagnosis is not indicated, and that it is better to engage in watchful waiting until those four to six weeks have elapsed. we do have exclusions for red flags, however, which would indicate that imaging is appropriate, and those red flags are that there are signs of neurologic impairment, trauma, IV drug abuse, and cancer.

Other points about this -- we do ask for plans to identify the earliest encounter for the episode, and we have a 180-day, what we call negative diagnosis history in which we look back at claims to ensure that

there was not a low back pain diagnosis in the 180 days before the index encounter, index episode. So, while we recognize that low back pain is a chronic condition and it may flare at certain times, we're trying to identify the patients at the beginning of their episodes so that we are addressing the acute back pain.

The measure was reevaluated in 2012, as was the DMARD measure, and was reviewed by the same measurement advisory panel. They strongly supported this measure moving forward. We again looked at the guidelines, and the guidelines had not changed. There does continue to be strong support. And as Mary noted, this measure is actually, or, this concept is actually in the choosing wisely recommendations; it was one of the first ones that was recommended. So it continues to be of strong interest.

There is some flatness in the scores. Although, we do see a variation when you start looking at the percentiles. And

again, we do hope to see some more renewed interest in improving the scores for this measure in that it is, because of the choosing wisely initiative and also because this measure has been included in CMS's HER incentive program.

So we do feel that it continues to be important. There's room for improvement.

I think that covers it.

CHAIR CHOU: Thanks, Jenna.

So we have three lead discussants. I think Zoher was going to start off, and then we'll take comments from the others.

DR. GHOGAWALA: Thanks, Roger.

So we have a number of comments about this measure. This measure is, as stated, a process measure that is used to look at the overutilization of lumbar spine imaging -- plain films, CT, and MRI -- for uncomplicated low back pain.

You know, the first thing we do is

we'll look at the evidence and what our study group said about the evidence. But the big picture here is that, you know, low back pain as an overall health expenditure in the United States is an enormous issue, and the developers shared some of that data.

I think one of the most widely quoted is the Brook Martin study in JAMA from 2008, which looked at the cost of all spinal disorders being in the range of $80- to $100 billion a year, compared to patients in the United States who don't have low back pain disorders. As you drill down on that data further, the outpatient costs associated with low back pain are $20- to $30 billion. Now, a lot of that is the cost of pain medications, but certainly, a lot of that is certainly the cost of diagnostic imaging; in particular, MRI. So the workgroup felt that the significance of this matter, as a matter of overutilization, I think, is very, very clear based on the data.

The opportunities are also there in terms of the variation in care. They show a difference -- and again, Jenna, you described the score being one minus the numerator over the denominator -- so a higher score is better. But the difference between a 10th percentile and the 90th percentile was 75 to 90, so a 15-percent differential, and for a healthcare matter as big as this, that's a significant matter.

One thing that is disappointing as we evaluate this measure is that over time, that practice or performance variation has not changed, and there are many issues behind why that might be.

But the first order of business is the evidence, and they present six randomized studies from which two meta-analyses were performed and guidelines have been generated. I think the important point about this is that high-quality studies have been performed that have asked the question, if you use early

imaging for low back pain versus not, you don't see a difference in outcome. Okay? So one can then infer that perhaps that early imaging is not required.

However, if you look at this measure, one of the things that is different -- I think it's always important when you're looking at randomized studies and then looking at process measures to ask the question -- is the process measure in fact directly related to the randomized study evidence? And here, it's close, but it's not directly related. And let me just explain why that is.

Usual care implies that there's still follow up with these patients so that, if somebody had a significant problem that was not imaged initially, they could be imaged because there would be further care.

If you look at this measure, this is looking at patients who are coming to either a clinic or to an emergency room with a diagnosis of low back pain absent these red

flags, and the measure is advocating that these patients not be imaged. Now, if these patients in a single event are not followed, it's possible that some of these patients in fact had a problem that would require imaging.

So the big picture -- when you look at it, the evidence base is there, the guidelines have been generated, and I think there's very little question that there's overutilization of imaging for low back pain. But you know, our committee felt that this is a moderate degree of evidence that directly relates to this measure, but a systematic review has been generated in terms of the evidence.

Overall -- we'll get the other points -- the other major issue, from our study group perspective, is that the study population that is defined by this measure, our study group felt, was inadequate. Again, the point here of usual, uncomplicated low

back pain and overutilization of imaging is all true. But how you define that study population is critical.

I think some of the red flags that are included in this study population -- history of cancer, history of trauma within a year, IV drug abuse, and neurological impairment in the last year -- are all reasonable.

But maybe, Cat, if I can turn this over to you, the American College of Radiology has published and studied this type of study population and has come up with some other very important criteria for identifying patients that may in fact not have uncomplicated low back pain, which would be excluded in this measure.

DR. ROBERTS: I think some of my comments refer to validity, so I might hold those and bring those up then.

But just speaking to the evidence, there is a very large body of very high-

quality literature on this topic, including systematic reviews. So this could be rated as highest high. Our group together had a consensus of moderate.

I will speak to the other topics on the validity later.

CHAIR CHOU: Thanks, Cat.

Again, just for full disclosure, we did one of the systematic reviews. It was published in Lancet a few years ago, so I'm one of the authors, or the lead author.

Carlos, did you have additional comments?

DR. BAGLEY: I would agree with that. I think, because they didn't get the imaging early doesn't mean they didn't get the imaging, and I think there's no direct study the shows that.

So, again, I think the data is there, but based on the specificity of this particular measure, it's kind of more of a moderate level.

CHAIR CHOU: JD?

DR. DANIELS: Are the other speakers done talking?

CHAIR CHOU: Those were the three, yes.

DR. DANIELS: Okay, sorry.

I may have just sort of missed it, but I just sort of went on my own -- you know more about this than I do, Roger -- Gilbert, from the UK, "Does early imaging in influence management and improve outcome in patients with low back pain," a programmatic randomized controlled study, which is actually a little bit larger than a Kendrick study. And what they ended up doing, you know, just to kind of mix things up again -- but they actually found that, although it did not appear to affect management, it did improve cost of clinical outcome. I don't know how much a pound is now, but it was 870 British pounds per quality.

I just wanted to mention it

because most of my job all day is to convince most people that back pain is part of the human condition and that you can prove that no one's perfect; just look at an MRI.

The main thing that I'm worried about with this measure is what you're looking at. If you go and say, you know, how you're going to pull this out if somebody comes into a primary care office, most docs, I don't think, want to do an x-ray. And so what happens is that there's a whole bunch of other issues, psychosocial issues, and sometimes, that's what really ends up adding the cost up.

And if you're going to measure, like, say, you've got a guy who -- just last week, this happened -- a firefighter comes in and he has night pain and he's a tough guy and you kind of know him, and he's got a tumor. Okay?

So when I get x-ray him -- so it was like, boom, I'm going to x-ray him that

day, and he didn't fall off a ladder or anything; so how do you justify that? What I'm probably not going to do is, Charlie, I'm going to x-ray your back because I'm worried you've got cancer, and you know, that's probably not going to be the diagnosis. It going to be, like, nonspecific kind of back pain. And you'll say, well, those are far and few in between.

But you know, you could get populations -- let's say you've got a geriatrician, and you've got nursing home people, and they fall, and they're on a lot of meds, you're probably going to want to do that, especially if they can't kind of express that they're having pain.

So, you know, I think that although I agree with this whole kind of issue here, maybe part of the reason that nothing's moved is what you're measuring -- I'm talking a little bit too much; sorry.

CHAIR CHOU: Yes, thanks, JD. No,

that was great.

Just a few comments about the evidence. One is both the Kendrick and Gilbert studies were done in the UK, where care is very different from here; for example, they do five times less spinal fusion surgeries than we do in the United States.

If you actually look at the costs across the studies that have been done in different places, it's actually all over the place. In some places that increases costs; in others, it decreases. So I think it's very hard to interpret the costs, especially for the non-US studies.

And then if you look at other outcomes, like anxiety -- right? There's this idea that you would reassure patients by showing them that they don't have cancer or whatever. You actually don't show that there's any positive effect, that in some cases, it actually gets worse because you see all these degenerative things, and so many

patients think there's something wrong with them.

And so, you know, really, it's essentially a wash with any outcome if you look across all these studies -- pain; function; quality of life; patient satisfaction -- and if anything, there's a trend towards worse outcome in people who get immediate imaging. So, hopefully, that clarifies things a little bit.

I think there are still issues in terms of how you manage patients day to day and all that kind of thing.

The other piece of this, I think, is that the RCTs don't capture some of these downstream harms, but we have other studies that people who get early imaging are much more likely to undergo surgery and things like that without any clear beneficial impact. Right? It would be fine if they were doing better, but they're also not doing better. So there are a lot of downstream costs that

aren't captured aren't really captioning many of these studies.

So I just wanted to make a couple comments.

There are a bunch of hands up. We'll get to you. I think John had his up first.

DR. VENTURA: I think Webster's study addressed that issue of cost well and showed a fivefold increase if you have early unindicated MR, as well as a delayed recovery --

DR. DANIELS: I'm sorry; I was misspeaking; x-ray, not MR.

DR. VENTURA: Oh, okay. Yes, I thought he was addressing the MR issue, also.

DR. MATUSZAK: I guess what I worry about with this measure is that you're not really looking at what they're looking at in most of these studies. In the studies, you're looking at the time from the onset, really, in saying that most back pain resolves

within four to six weeks, uncomplicated if you work it. But really, you're not; you looking at claims data.

And nowadays, especially with high deductible plans and things, I couldn't tell you the last time I saw an acute back pain that came on acutely. I mean most of the patients I've seen have already had it for three to six months, and there's nothing in here that really accounts for that.

You know, you talk about radicular findings and stuff, but I work with the young athlete population, and actually, in my 18- to 21-year-olds, my most common ultimate cause of back pain is a stress fracture in the spine, and that's not something that, you know, again, is caused by a trauma, and it has this long course before it comes up.

So, without having any information about being able to subtract out for the long lead time before that initial diagnosis is made, I worry about how accurately this

reflects the evidence that's being presented.

CHAIR CHOU: Again, I can help answer that little bit. So, in the six trials or so that are out there, the inclusion criteria are actually really varied. So a few of them were focused on people who truly had acute low back pain, but several of them actually included patients with subacute or even chronic back pain; so, longer than 12 weeks. And we tried to stratify by duration, and there's really no difference. I mean there are not a lot of studies, so it's hard to see anything. But it's as far as we can tell, there's no real difference.

Now, that being said, there is not a whole lot of data about what to do with somebody who's had back pain for two years. I mean that's considered a red flag in a lot of places. And so you're right; somebody could present with back pain, and even though it's their initial presentation, they could

have had it for two years or something, and they would be -- I don't think they're captured directly by the evidence. But we also have no evidence showing that, you know, they need to have an x-ray or whatever.

So anyways, there were a few comments here.

Zoher?

DR. GHOGAWALA: So, just one point of clarification because, JD, I see you're concerned with the geriatric population. But this measure -- correct me if I'm wrong -- is 18 to 50. So I think we should just understand what, the measured population group.

CHAIR CHOU: Are there other people -- yes, John?

DR. FITZGERALD: I just have a question about -- so we want to bring up an issue about the exclusions. Is that under specifications, or?

(Off-microphone comment.)

DR. FITZGERALD: I think Cat raised some issues there too.

Are there other comments or questions about the evidence?

Yes, Linda?

MS. DAVIS: Where is the data that 50 is the right upper limit?

CHAIR CHOU: So I can speak to that, I think. I mean it, it's simply based on the fact that the red flags -- being over 50 is a red flag for cancer. It's not a specific red flag. We've traditionally had a bunch of "red flags," and together, they have a very high sensitivity, and age has been one of them, traditionally. So that's the reason for it.

Are there other questions?

Yes, JD?

DR. DANIELS: I don't want to belabor the point, but I guess this is more of a question to the staff.

When you say you've screened for

cancer or like that, when you do the selection, so I assume that what happens is, the way you found out is that someone had an x-ray is that the diagnosis for back pain was, like, rule out cancer, and that didn't count.

But otherwise, if the clinician -- what kind of happens is that when you're worried about something -- it has to do with the type of population you take care of. And so, if you take care of a mostly healthy population, and you see a serious problem, most of the time, you don't just say, well, Ms. Jones, I think you have cancer; we're going to do these tests and find out.

Usually, what we do as we say, I'm concerned about this. And so, when we do this diagnosis -- it would be like a nonspecific -- we're going to kind of come back and kind of talk to you. Where, if you go with a different population, where most of the people who are walking in your door presumably have

the disease or have a higher prevalence of the disease just because they're there, it kind of almost works backwards.

So I guess my question really, or, my concern -- because I really agree with everything Roger has said -- is not so much that it's the wrong thing to do. It's more on how you're collecting your data. And that might have a reason why it hasn't moved, because, I think there are some cases where, if you came in with certain physical or history-type findings than the one that Dr. Dao had his big one a long time ago, where he kind of went through and said, here's the meta-analysis, and what really counts, and that's one of them. So that's my comment.

CHAIR CHOU: JD, I want you to hang onto those thoughts because I think we're going to get to that when we get to the feasibility and use kind of stuff. I think those are kind of important questions, and I'm going to have you guys respond to this when we

get there.

Are there other comments or questions about the evidence. I think we'd probably like to move it towards a vote here.

(No response.)

CHAIR CHOU: All right, let's vote on the evidence.

MS. PHILLIPS: We're voting on Measure 0052. We're voting on the evidence.

You have five options, "1" for high, "2" for moderate, "3" for low, "4" for insufficient evidence with exception, and "5" for insufficient evidence.

The voting begins now.

Okay, we've got 22. We've got five for high, 15 for moderate, one for low, one for insufficient with exception, and zero for insufficient.

CHAIR CHOU: All right, so this passes the evidence criteria, and let's talk

about performance gap now.

I'll hand it over to Zoher:

DR. GHOGAWALA: The workgroup felt there is a performance gap. Again, as I mentioned before, they've looked at the rates here. Whichever way you want to look at it, the way they've scored it, it's a 75 for the 10th percentile and 90 for the 90th percentile. So there is a range of practice and probably for proof.

CHAIR CHOU: Cat or Carlos, do you have additional comments to make here?

DR. ROBERTS: There were additional disparities noted for minorities and other ethnic populations.

CHAIR CHOU: Thanks.

In addition to the data that was presented, I'm aware, at least, of other published data showing that there are high rates of imaging in Medicare populations, for example, which I guess would mostly be excluded from this measure. But other studies

have shown that as well.

Are there other comments or questions about the performance gap issue from the rest of the panel?

(No response.)

CHAIR CHOU: All right, let's go ahead and vote.

MS. PHILLIPS: Okay, we're voting on 0052 for performance gap.

You have four options, "1" for high, "2" for moderate, "3" for low, and "4" for insufficient.

The voting starts now.

We're up to 22.

We've got 10 for high, 12 for moderate, zero for low, and zero for insufficient.

CHAIR CHOU: So this passes that criterion as well.

Let's talk about the high priority.

Again, Zoher.

DR. GHOGAWALA: So, again, the workgroup here felt like this is a very high-priority item. Overutilization of radiographic imaging for low back pain has been identified in the literature repeatedly.

The other thing that was pointed out by the developers is that the rate, in particular, of CT imaging is sufficiently high enough that if you look at a broad perspective across the American population, there is probably a meaningful oncologic risk associated with that.

So, again, we felt it was a high-priority item.

CHAIR CHOU: Are there additional comments?

Carlos or Ted?

(No response.)

CHAIR CHOU: How about the rest of the group?

(No response.)

CHAIR CHOU: All right, why don't we go ahead and vote on the priority criterion.

MS. PHILLIPS: Okay, we're voting on priority for 0052.

There are four options, "1" for high, "2" for moderate, "3" for low, and "4" for insufficient.

Voting begins now.

We are at 22. We've got 19 for high; we've got three for moderate.

CHAIR CHOU: All right, so we passed the three must-pass criteria, and now we're moving into the others.

First, we'll do reliability, and again, I think we want to try to separate out reliability and validity, which had been a challenge for me. But I think the specifications are mostly referring to whether they are clearly specified. Right? So, not whether we think this is necessarily the right specifications -- that will come later -- but

whether it's clearly specified and whether they're, and then the reliability is just how reliable the testing is. And then we'll talk about the validity stuff in the next section.

So, Zoher?

DR. GHOGAWALA: So, from a study group perspective, again, separating the two issues, from a reliability perspective, data was presented by the developers from, I believe, two sites. And the reliability, we judged to be moderate based on our assessment of the data.

CHAIR CHOU: Thanks.

Carlos or Cat, are there other comments here?

(No response.)

CHAIR CHOU: Are there comments from the rest of the panel regarding reliability?

(No response.)

CHAIR CHOU: Would you be able to briefly summarize what the reliability testing

showed, just so that everyone is kind of on the same page here?

MS. WILLIAMS-BADER: Sure, happy to.

We did the same type of analysis, beta-binomial, as we did in the DMARD measure, looking at and comparing the signal-to-noise, and here we see that in the commercial plans, there's actually a very high reliability score of 0.99. In Medicaid, the reliability score is, the average is the 94.

We also see that even in the ranges of between 10th and 90th percentile, that for commercial plans, it's .81 to .99, with .7 being the threshold as high. It's showing that the range for the 10th to the 90th as high reliability. For Medicaid, we do see a lower end for the 10th to 90th of .64 to .98. But again, .7 is the threshold here that we're looking at.

If you look at the histograms that we provided, you can see that there are still

the majority of the plans falling at the .7 and above.

CHAIR CHOU: Thanks.

Are there any questions or comments about reliability?

(No response.)

CHAIR CHOU: Let's go ahead and vote on this criterion.

MS. PHILLIPS: Okay, we're voting out reliability for Measure 0052.

You have four options, "1" for high, "2" for moderate, "3" for low, and "4" for insufficient.

Voting begins now.

Okay, we're holding at 21. So if everyone could vote again, that would be great.

We're still at 21 -- there we go. Thank you. Okay. We have eight at high, 14 at moderate, zero at low and zero at insufficient.

CHAIR CHOU: That passes.

Moving on to validity, this is where I think there are more concerns or issues to discuss. So again, I'll start with Zoher, and then we'll move to Cat and Carlos.

DR. GHOGAWALA: I think I'll turn this one directly to Cathy.

The workgroup, just as a summary, had significant concerns about low validity here.

DR. ROBERTS: Thank you.

So I'd like to refer to the American College of Radiology appropriateness criteria from 2011, and this is on appropriate imaging of low back pain.

Those appropriateness criteria list the exclusions that are listed in this measure. However, they include several more, and we feel that the absence of these exclusions make the data that's being collected less meaningful. Some of those exclusions, which we think are missing and should be there, are unexplained weight loss,

insidious onset; unexplained fever; history of urinary or other infection; immunosuppression; diabetes mellitus; prolonged use of corticosteroids; osteoporosis; and prior lumbar spine surgery.

Our group felt that it was important to exclude any clinical findings that suggest neoplasm or infection, and the way that this is currently written, it does not.

CHAIR CHOU: Are there other comments from the group before we let the developers respond?

DR. GHOGAWALA: You rock, Cat.

(Laughter.)

CHAIR CHOU: Jenna, would you like to respond?

MS. WILLIAMS-BADER: Absolutely.

I definitely welcome those comments, and I think they are worthwhile for us to think about in the future. To give you history, we did actually look at some of those

during our field testing, and we found that their rates for those were quite low.

Now, one reason why they might have been low is that claims might not be the appropriate place to capture some of those types of exclusions. Unexplained weight loss is jumping out at me as one, and we actually did look at that one during field testing and only found 28 cases out of 22,000. So I'm not saying that it's not more common than that. It's just that in claims, that might be difficult to capture.

We're actually really interested to see how the e-measure performs. Again, we do have an e-measure that's included in meaningful use. And the electronic health record does give us the opportunity to look at symptoms and other types of information that's not readily available in claims.

I will say, though, that as much as we want the EHR to be telling us about symptoms, it does not, as well as it should.

The availability of that information is actually quite low. But we do think that it's a promising data source for these types of overuse measures, where the claims may not help us to identify all of the key exclusions.

CHAIR CHOU: Go ahead, Cat.

DR. ROBERTS: Thank you.

I really appreciate those comments, and I'm glad you that you'll look into them. I do feel it's important for these measures to be as meaningful as possible, and I really feel bad for practices that might have a disproportionate number of post-op patients or immunosuppressed patients. And if we look at the metric, it will show that they are giving substandard care, when they're not; they're giving appropriate care.

So people spend an enormous amount of time writing these. They spend an enormous amount of time collecting data for these. So I still believe we should endeavor to be as meaningful as possible when these are written.

CHAIR CHOU: Go ahead, Mary.

DR. BARTON: One of the things that we have found in health plan measurements is that the threshold of potential exclusions that we try to incorporate in measures -- we usually use two percent as sort of a rough estimate.

So things that happen super-rarely, it doesn't, the health plans have told us it's actually not worth their while to track down, because it's not going to affect -- and also, when we imagine across health plan -- I'm not saying this is true for provider groups. I think you make an excellent point about provider groups -- but across health plans, the idea that immunocompromise in people under the age of 50 with sort differentially and one health plan could have more than another, that's hard to imagine.

So I think that, you know, while these are issues that absolutely play into

clinician decision-making on a one-on-one basis, they have not been at the level, at a health plan level, where we have heard feedback. And believe me, when health plans believe that we are measuring something that is unfair, they're so not shy about letting us know that.

(Laughter.)

DR. BARTON: So I would be curious, and I think I'd be interested to go back to some of our health plan partners and find out about their thoughts about these issues because, certainly, you make a compelling case. And I think, you know, in particular, the provider group assessment or the individual physician assessment, those exclusions would help immensely with the credibility of the measure.

CHAIR CHOU: So Helen and Jenna and Zoher all want to say something.

I just wanted to say -- this isn't going to help because it just makes more

things more complicated -- the risks factors are very complex. I mean some of these are very weak risk factors. So, unexplained weight loss, insidious onset -- the predictive values are very low. You know, you increase your likelihood of cancer from .8 percent to like 1.2 percent. It's trivial, almost. Whereas, having a history of cancer is actually a much stronger risk factor. It brings you to 1 percent from 1 percent to 10 percent.

And so it's very hard, actually, to do this in a measure to try to incorporate all of these different risk factors, some of which are very minor and uncommon -- you know, having vertebral infection is actually quite uncommon, but having a stress fracture is relatively common in people who have risk factors for it. And how do you kind of put all this stuff in? So this just speaks to some of the challenges that measure developers face in trying to deal with this, and kind of

some of the decisions that have to be made in terms of what they can do and what they can't. I mean it's a huge issue.

We struggle with this trying to even provide clinical guidance; what we tell somebody if you have a patient over 50? Do we say that they all need to be imaged? They used to say they should, and we've actually came back, the ACP at least, has said it's reasonable to manage them first if they don't have any other red flags and that kind of thing.

But, Helen?

MS. BURSTIN: Thanks very much, and actually, just to bring us back to the criterion, specifically that the exclusions are supported by clinical evidence and that there should be evidence of sufficient frequency of occurrence, such that the results are distorted without the exclusion. So what we don't want to do is overburden the measure with lots of things that may be clinically

logical, and at the doc sitting in front of a patient, you would consider some of those. But in terms of fairness and measurement, you don't want to overburden the measure.

We specifically want to make sure that it would only be those exclusions you would include that, otherwise, if you didn't have them, the results would be distorted. So I think that gets at some of this issue of the key issues around sensitivity analysis, essentially, and understanding issues of positive predictive value.

MS. WILLIAMS-BADER: Yes, I just have one more thing to say, and that's actually, going through this NQF process and initiating some discussions with the woman group about their MRI measure, which is coming next, has given us the chance to compare the exclusions that are in their measure to the ones in ours, and we do want to continue those discussions?

We think that making those changes

to our measure or us are significant enough where we would need to run them through our typical process, which is to pull together experts and discuss them, and potentially even put it to public comment. But we do see that there is an opportunity for us to think about some of these other exclusions.

DR. ROBERTS: I'll just say I'm surprised that lumbar spine surgery is a rare event.

CHAIR CHOU: JD?

DR. DANIELS: Oh, again, I'm just going from the stuff I saw. I think they are urinary thing did pan out, and I also thought it sort of is, you know, in how you word this. You might need a different word. But night pain -- those were two that I thought separated out.

Part of the reason that at least I'm bucking so hard on this is that I think what you've got here is you've got a problem. There's a hole in the boat, and instead of

trying to fix the hole, you're throwing a bail. And you're not really fixing the problem.

So what happens is, in primary care, you know, they talk about shooting zebras and all this stuff, so we don't really see -- you know, the way we decide if it's a zebra is -- we're not up on the high bluff with a high-powered rifle that can see the zebras like the specialists, if you want to say it that way. What we really know is what a horse really, really looks like. So what happens is, these people come in and they go, you're not a horse. I'm not really sure what you are today, and you may just be a weird looking horse.

But the big thing is that most people come in with back pain, they just need to be reassured. But what happens is the thing that kind of keeps me sane when I get up in the morning and look in the mirror -- it's a hard job -- is that, you know, I really try

to do with the best. And so, on those cases that kind of pop up, you almost have to do it.

And I think what this may be doing is kind of -- you know, doctors pay attention to this stuff. I know you guys are looking at the big healthcare level in stuff, but you know, it's sort of like part of the fiber of what makes you do what you do. That's my whole issue on this because we're spending a lot of time and effort measuring this, but I think you're looking in the wrong spot.

CHAIR CHOU: So JD was doing all the football allusions yesterday, and now we are to animals and boats.

(Laughter.)

DR. DANIELS: Yeah, the Navy, yeah.

CHAIR CHOU: So I think what you're saying is that it has to do with, there has to be some additional exclusions, that you're concerned that the denominator is

incorrect. Is that right, basically?

Did the developer -- Zoher, and then we'll -- go ahead.

DR. GHOGAWALA: You know, one of the things, just to follow on what JD is saying, is that this is a really hard problem because the patient population here is incredibly heterogeneous, and one of the issues, I think, is if you're trying to improve quality in this area -- and I would submit to you that so far, it looks like you're not -- is that you need to do something that actually is giving the type of feedback that would improve quality. And one of the ways to do that, I think, is rather than saying will come up with this complex list of inclusions and exclusions for a very heterogeneous population, is to recognize, I think, that there is a role for judgment. I think that's what JD was getting at. There's a role for physician seeing a patient with back pain, and exercising judgment. Okay?

And certainly, there are very well known risk factors, and there's good evidence here, but it seems to me, if I were just sort of stepping back from this after looking at the entire measure, what we want to do is we want to avoid imaging patients that don't have anything wrong with them that need further treatment.

And so -- again, looking at the RCT data, which is slightly different from this measure, to me it seems that what you want to do is you want to sign a measure that allows for the identification of patients, the number of patients of the percentage of patients, that were imaged that didn't have anything wrong, and that requires follow-up.

There's some more recent data from Canada, which I think is doing this very, very nicely. They're looking at nurse triage or nurse practitioner-type triage mechanisms where the patients are evaluated, no red

flags, no imaging, but they're followed up four to six weeks, mandated. Okay? And then they determine whether in fact there is something that's not the usual course, and then they make those decisions. That has reduced MR utilization dramatically. That was published a few months ago. So I think there's an opportunity here. But I worry that in its current iteration, it does just as invalid.

Furthermore, relying on claims data to say that certain things are very, very incompetent uncommon, I think, is dangerous because certain things may be in fact more common than we can cull from administrative data.

But my concerns are significant with validity.

CHAIR CHOU: I'll let the Jenna respond to that.

I just wanted to say I understand everyone and I agree with a lot of these

issues. But I also think that this is an issue with all measures. I mean there's always exclusions that aren't accounted for, and that's why you're going to get to 90 percent. There's going to be 10 percent of patients without an obvious red flag who are imaged, and we'll actually say that those patients are probably doing pretty good if 90 percent of the patients -- but there's a problem with 50 percent of the patients with no red flags are getting imaged.

My understanding is that that's how a lot of these measures are designed to work. You know, that it's not expected the places won't necessarily get to 100-percent. There is a way to compare one site versus another. There is this case mix issue, of course, but there's ways that that can be addressed and, you know, with the way that things are set up.

But anyway, I'll let Jenna respond if you have any other responses to Zoher's

comments.

MS. WILLIAMS-BADER: I think you actually covered what I was going to say, which is that we don't expect performance to be 100 percent, and so what we are looking for is differences between variation among plans. And there is that variation, indicating that that while we don't know what the upper rate should be, we know some plans are able to get to a certain rate, and if there are plans below that, then we see that there's room for them to improve.

CHAIR CHOU: I think Thiru and then Jason.

DR. ANNASWAMY: Just along the lines of what you just said, Roger, I think the way we can reconcile this and introduce the possibility of judgment and uniqueness of your health plan your practice would be the usability and use area. So we're not saying that we're going to come down with a big sword or axe on your health plan if you are at 75 or

85, but maybe there is opportunity for education, introduction of new practice patterns, like a nurse practitioner follow-up, et cetera, to know to that number up.

So the measure is a measure of the measures; right? I mean how you use it would be where some of these other factors would be introduced to make a difference to the outcome.

DR. MATUSZAK: Yeah, those interventions are every time I have to get on the phone to get a prior authorization and talk with some doctor in a distant place because I want to get an MRI scan for somebody less than 28 days after the ICD-9 code goes in for the first time. You know, those are the interventions that they put in place, which really just burdens us even more, which I think does factor into this.

But I'd submit that I think that maybe some of the movement that you're not seeing in this is because we've reached that

noise level here where the heterogeneity has reached a point where there's so much that comes down to the doctor's discretion, that these are patients that, I don't know how much better we're going to do with these. I think that most docs do a pretty good job of trying to do this, and maybe the randomized controlled trials that show that the outcomes don't matter is beside the point.

But I don't know that these few catchall ICD-9 codes that kind of get used for any imaging study that we want to do is concerning the low back, in most cases, is the most valid method of capturing the type of information that you're trying to capture, and that, I would submit, is almost fatal flaw with this.

I just don't feel that the validity is there for what you're trying to accomplish.

MS. WILLIAMS-BADER: Yes, I did want to speak to the face validity because I

haven't covered that well.

When the measure was first developed, as I said, there was one panel on pain that recommended it, and went through musculoskeletal workgroup. It went to public comment period. Then those results were discussed both with our musculoskeletal expert workgroup as well as with our committee on performance measurement.

When the measure was reevaluated in 2012, again, we pulled together bone-joint measurement advisory panel and also took the measure to CPM as well. I think there was an acknowledgment that we are approaching a time where we might want to think about whether this is as good as we're going to get. But because we are still seeing variation -- and I'll just point you to the fact that both the 10th and 20th percentile are 70 to 75-percent across the two lines of business -- that is showing the, I don't know that the CAC would be at 70-percent.

Lastly, as we mentioned before, this is a choosing-wisely topic, indicating that even though physicians think they might be doing this, they're still not quite encouraging the watchful waiting, as we would hope.

DR. BARTON: To say that the societies who put this forward under choosing wisely didn't think that it's been solved already because they're saying to patients that they should rethink and ask the question, do I really need to be imaged?

So I can't imagine who, other than the ordering clinicians, would know better whether they are using this at a frequency perhaps higher than is clinically indicated.

DR. MATUSZAK: But that's not really what you're measuring here. You're measuring from the time that ICD-9 code gets going, and whether or not there's an imaging study in the next 28 days. You're not measuring from the time of onset. You're not

measuring, you know, anything else other than that. That's why I'm saying I don't know if that's the most valid way of capturing the data, which I agree support, but I don't know that that's it.

CHAIR CHOU: Doesn't that 180-day exclusion prior to the first visit prevent that?

DR. MATUSZAK: Only on the ICD-9 codes. If it's not input in the 180 days prior to that, then -- you know, again, so your patients are coming to see you not anymore because they have an acute back strain that happened a few days ago. Because of the high deductible plans, now they're coming in six months afterwards, and they've already tried three different things to try and work through the problem themselves. But you're not capturing a lot of that. And I would submit that I think that's an increasing problem, especially with the way people have been delaying care more so in recent years,

with these high deductible plans.

CHAIR CHOU: Christian?

DR. BARTON: I guess my concern -- I agree with all the points, but what is the concern about the measure going into place without these caveats and these additional exclusions? I mean I think we would all acknowledge that the over-utilization is absolutely a problem, that the amount of yield we're getting from these images is relatively low. So, sure, we can all cite examples of times where we found that the zebra and it was really important to that individual patient. But if we include these exclusions, we're not fundamentally changing the yield on all those other exams that come out without clinical significance.

And so, while I agree with all of those comments -- I also agree with the limitations logistically and including low prevalence exclusions -- I was wondering if you could just articulate what is the end

deliverable here is that you're looking for.

DR. ROBERTS: Sure, and that's a great point because I get what you're trying to do. I do. And I completely agree with, you know, the standards that are in the literature I absolutely do.

My concern is eroding provider confidence in our metrics. We've all come together today, we've come long distances, we've spent a lot of time, to try to make these measures as good as they can possibly be. And I don't know if you've had the personal experience, but I certainly have, where my institution gets a metric back that says I'm giving substandard care. And it's the way they do the metric.

At that point, some people just turn off. You know? They say forget it. The way you do that metric, you're not even counting people who -- they have cancer? I'm not going to let my patient be hurt. Of course I'm going to image them. And when

these metrics are messy, that's what happens. We erode confidence of our providers, of our patients, and that's why it's important, I think, we bring this up and we talk about them and try to hold everyone to the highest level we possibly can.

People are getting more and more savvy about this every day, and it's not just the providers, not just the health plans, it's the consumers too. So, if they realize that these aren't meaningful measures, it's a slippery slope. So that's my concern.

CHAIR CHOU: Linda?

MS. DAVIS: I would echo that for employers as well.

I work with employers, and they look at, in our market in particular, and we have several health plans -- not several; a few -- that all overlap in terms of their provider networks, but there are different NCQA ratings, and they look at that and scratch their heads going, so we've got the

same providers in every network and they're all doing kind of the same thing, and we can't really tell the difference, but why are there differences in the scores of various health plans? The health plans get to dance and explain what that is.

But it's still kind of a scratch-your head kind of thing and lack of confidence in the measures.

CHAIR CHOU: I think we're going to try to close because we're overdue for lunch.

Just to summarize, I think, where we were at, there is some concerns about the, you know, how the denominator was defined in terms of, you know, presentation -- ICD-9 code, I should say, rather -- necessarily knowing exactly when the pain started. And then there were some concerns expressed about not having some exclusions on there. Those are the main issues that were brought up.

And then we also kind of talked

about how the measures used.

I actually, you know, I actually think 75 to 90 percent is pretty good for a measure in terms of people being able to comply with it, so it doesn't seem like most places have had a terrible time trying to -- I mean those numbers are actually better than I would expect looking at the literature in terms of inappropriate imaging rates. So, if anything, it seems like it's underestimating inappropriate in imaging, you know, however. But I think there were concerns that it would be overestimating inappropriate imaging by not having some of these exclusions and things like that.

But let me pause there. Are there final comments before we go to a vote?

I think we need to vote. Yes, solar?

DR. GHOGAWALA: Sure, Roger, just to follow up on that because there was a comment in the review from someone outside the

review group.

I think one of the problems here is that this an administrative coding kind of data, and so it probably does underestimate the problem. And I would submit, therefore, that it is not helping to address the problem because, when a clinician writes down, you know, spinal stenosis as the cause of the back, or herniation, it's not captured whether they know that that's the case or not. And a lot of people, on their MRI, will have these things.

So it is a tough one because this is relying on coding information to solve a problem that may not be solvable with a code.

CHAIR CHOU: Yes, I can't put suspected radiculopathy; it's not an ICD-9 code. You have to either put low back pain or some other, something else that you know won't cause a flag.

Cathy, do you want to say anything? And then I think there was a

comment everywhere.

(No response.)

CHAIR CHOU: No.

All right, so I think we're going to have to put this to a vote.

Remind me again -- they've done the must-pass things, but they still have the pass these, though, before everything is passed.

Okay, so we're voting on the validity now.

MS. PHILLIPS: We're voting on 0052, validity.

You have four options, "1" for high, "2" for moderate, "3" for low, and "4" for insufficient.

The voting begins now.

Okay, we are at 22. We got one for high, seven for moderate, 10 for low, and four for insufficient.

CHAIR CHOU: That seems to be our 30 percent threshold. We keep going, though;

right? Or do we stop here? All right.

MS. BURSTIN: We should probably -- this does stop it in terms of our evaluation today, though, but just in terms of the exclusions that are actually included in the measure, I just want to get a sense of the committee's appetite.

Certainly, some of the exclusions we've been talking about, like cancer, are actually exclusions in this measure. So it might be helpful to see if NCQA can go back and actually provide additional data back to the committee, if you'd like to see it, on the other rate of exclusions. What's included? What's not? How do the rates change? Because I think it's been, it's a great discussion, and your points are well taken, and we don't want to erode confidence in the measures. We don't want to erode confidence in the measures.

We also want to be sure there's a fair opportunity to look at what the data

actually shows in terms of the proportion of some of those exclusions that may already be included or may not be excluded and just get a sense of it.

CHAIR CHOU: Thanks, Helen.

It seems to me like you guys would have some of this data, so it would be interesting for me to see how many people who are, you know, being flagged as getting inappropriate imaging have had prior back surgery or have some of these other things. If we know what those percentages are, I think that helps in terms of assessing, you know, their importance.

And then the other piece of this -- I don't know if you're able to do this -- if you're actually able to see, you know, and patients are classified as not meeting this criteria -- so they get an image when they don't have any red flags -- how many of these actually end up having cancer, an infection, or something like that. I don't know if you

can do that, but that would be the other side of it.

Are there other comments?

I mean I'm not sure how to address Jason's concern about time from onset. That's not just an issue with measure development; that's an issue with just trying to do studies on low back pain. Like, if you have an administrative database or something, it's extremely difficult to sort out when the back pain started. So I'm not sure that these guys can solve that, but I think that was the other big concern.

Yes, Thiru?

DR. ANNASWAMY: perhaps if there are more data like that to where you can tell what is the difference, the true difference, between what you're trying to measure, which is uncomplicated low back pain, and the muddied water of complicated low back pain that presents as uncomplicated, potentially. If there is a percentage difference, then

perhaps the recommended benchmark would be a way to resolve this.

I can only give an analogy. For example, inpatient rehabilitation -- there is a 60-percent rule for admission. So 60-percent of the patients admitted into inpatient rehab units have to meet and inclusion diagnosis of five or six. There is a proposal to make it a 75-percent rule, which would be more restrictive. But that's being pushed back by inpatient rehab providers saying there's just not, we don't want to be too restrictive, or we're going to have to turn away patients that are not in these diagnoses, because we potentially see more of them in our facility than otherwise. So the 60 to 75-percent makes a huge difference in the amount of pushback you get.

Perhaps there is a benchmark that you could use to say, because we understand there's differences between different health plans and mixes, this is what we think we

should shoot for.

DR. PACE: But the measure as it's constructed doesn't set any benchmark.

DR. ANNASWAMY: Right.

DR. PACE: Okay -- so it's not saying what it has to be. It's a comparative performance -- okay. All right.

CHAIR CHOU: JD?

DR. DANIELS: I'd just like you to kind of go back because I may just be all wrong on this. But some of this stuff Cat mentioned, I thought, had shown -- it didn't use the exact words, but it may have shown that it made a difference. Like, urinary symptoms and night pain, those kinds of things, which, they're kind of indirectly -- but that's not an ICD-9 code.

So, you know, as far as you capture it, I'd feel a lot more comfortable with it if you kind of had that as an exclusion. I'd be fine.

CHAIR CHOU: So there are several

issues with the red flags. One is how well they actually predict the condition, so essentially, the likelihood ratios and positive predictive values with them. And with UTIs at least, there isn't a lot of data because we don't have a lot, there's not a lot of vertebral infections. That's what you'd be concerned about. But at least for constitutional symptoms, it's quite weak.

DR. DANIELS: Yes, and even just urinary symptoms, remember, they'll cover more than just a UTI. It covers other neurological things.

CHAIR CHOU: Yeah, with cauda equina or something like that. But my second point is it's not just the positive predictive value. It's how frequently frequent the condition is. So cauda equina is quite uncommon, and vertebral infections are quite uncommon.

So, even if you have something there, then I might not be particularly -- it

might be for a condition that just doesn't occur very much. And so both of those issues, I think, is what makes this complicated. There is a whole slew of things. I mean if you took all the red flags, you know, you would end up with, most people should be imaged, and we know that's clearly not the case. So that's the problem, is that there's so many that have been described and so many of these predictors that, how do you, you know, separate out the ones that are the most important?

I think that's what makes it --

DR. DANIELS: Well, there's a whole set of predictors we're not looking that wasn't even brought up.

The New Zealand folks are, I think, light years ahead of us. They've got the yellow flags, which are basically the psychosocial things that hit because I think, in real life, sometimes what happens is the average doc that sees them, they're not just

doing the x-ray to get the patient out of their hair. It's like the patient goes, I'm having trouble, and you've got to kind of convince them to go back and work or do whatever. And so that one might be a good one. I mean there's pretty good data about, that that's more like longitudinal, but kind of looking at that, you know, and that might be helpful so that there's, like adding that code.

CHAIR CHOU: Yes, I this is not addressing yellow flags. These aren't imaging things. But we certainly have yellow flags, too; we have purple flags and other things also. There's all sorts of colors.

But I would -- just one thing in terms of the evidence -- and again, a lot of my work has been in low back pain, and so the idea that imaging somebody, you know, reassures them just hasn't been proven, and it actually has been shown that it causes harm because people feel like they can't -- you

know, they start worrying about their back.

It's an attractive kind of idea, but it just -- if you look at the literature, if anything, it's a wash, or it makes people worse -- so I think I'd just be cautious about presenting that as a reason for imaging.

But I think all of these other -- you know, there are a lot of potential red flags. I think that beta would help to figure out which ones need to be in and which ones, you know, we don't have to worry about so much. And then, like I said, this issue about how do you define, you know, the population -- the 28 day thing -- or the onset of low back pain. I think that's a real challenging one. I'm not sure the best way to do that because I know that we deal with that just trying to do research and that in this area as well.

So are there other comments?

(No response.)

CHAIR CHOU: All right.

Are we supposed to take public

comments now, or do we break for lunch?

MS. STREETER: Operator, at this time, could you please open up the line and see if we have any members of the public that would like to make a comment?

OPERATOR: If you would like to make a comment, please press *1 on your telephone keypad.

(No response.)

OPERATOR: And there are no comments at this time.

MS. STREETER: Also, real quick, just one more process point.

As part of our new standing committee policy, we are appointing you two- or three-year terms. So this is a stagger on when people join them leave the committee.

You may have the chance to renew after your term, so don't get upset if you'd like to stay for three and we give you two.

Basically, during lunch, we will be walking around and asking you to draw a

number out of the cup, and that's how we will be appointing your term.

DR. ROBERTS: And two-year terms are renewable, of course, so don't worry you'll be with us forever.

CHAIR CHOU: So let's break for lunch, and then we have one more measure.

(Whereupon, a lunch recess was taken at 12:52 p.m.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

1:17 p.m.

CHAIR CHOU: The level of analysis and the care setting were reversed, so the level of analysis should be health plan, integrated delivery system, and the care setting or those other things, ambulatory care, clinician office, clinic, urgent care hospital, acute care facility, so I just wanted to make sure everybody knew that.

If there's a feeling that that would change your decision regarding the validity thing, please speak up. Otherwise we'll just move on here. Any concerns or issues here?

I think people are pretty clear on the conversation. We just want to make sure people - you know there wasn't confusion there.

Okay, so we're moving to our last measure which is Measure 0514, MRI lumbar spine for low back pain. This is CMS and we

have the developers here. Could you give us a brief overview.

DR. BRUETMAN: Okay, yes, I don't know if CMS is on the line, so good afternoon. I'm Dr. Bruetman, Charlie Bruetman, from Lewin, and we've been developing this measure with CMS.

With me I have my colleagues Kelly Anderson and Colleen McKiernan, and on the phone I believe is Dr. Nicholas Staedtler if we have any questions as well. Maybe Dr. - I think Dr. Nakano was going to try to join, but with the change in schedule, I'm not sure, and Fiona Larbi is on the phone.

So I will give an overview on the measure, and I know there is - hopefully it's not going to create more confusion because it's very similar to the prior measure that was discussed, so we'll clarify any issues as we go along.

Basically this measure is - calculates the percent of MRI lumbar spine for

low back pain without any prior evidence of conservative treatment. It is a claims-based measure and there's a very high prevalence of this low back pain, and it is based on, as was discussed previously and Choosing Wisely was selected as a top priority by the American Academy of Family Physicians, it is considered the fourth most common reason for office physician visits as they state, and as I said it was selected in Choosing Wisely initiative it is great - they are Number One and I'm not saying it's rated Number One, but it's top on the list of the 15 things they consider for - that physicians and patients should question.

They state that basically that you should not do imaging for low back pain within six weeks unless the presence of a red flag, for example a progressive neurological impairment.

American College of Radiology also states that uncomplicated acute low back pain is a benign self-limited condition that

warrants no imaging, and having the test performed soon after the start of a low back pain and this is evidence that was one of the reasons it was selected for Choosing Wisely is that if you do have that fares no better for the patient and they also state that an MRI in just a month increases by almost eight times the chance of having the surgery and five times the medical expenses incurred and with no faster recovery is seen.

The goal of the measure is to - I would say is twofold. On one hand it is - there is no benchmark on the measure so we do a calculation of the rate, and it is to decrease the overutilization of this MRI for low back pain without conservative treatment, and it also - we are also are looking at the outliers and try to reduce the outliers getting them closer to the mean as it is a publicly reported measure by CMS.

This measure was endorsed in 2008 by NQF, and was re-endorsed in 2011 and has

had very strong support from the medical community either expressed through comments or discussions we've had along the way.

All our measures when we discussed it it's not only it's based on the evidence of literature reviews that we do, but also we have a technical accurate panel that's been working with us of clinical experts, and they have been supporting - very supportive of this measure.

Along the way we've also had discussions with American College of Radiology and other entities to ensure that we are addressing the major needs of this - in addressing the evidence presented.

The measure just to be sure - this is - one of the - just to clarify how our measure is done, basically this measure has a denominator includes the MRI for lumbar spine studies with a diagnosis of low back pain on the initial claim and the numerator is from the denominator cases for patients that do not

have claims-based evidence of a prior conservative therapy, and we have a series of exclusions among them - I'll cite some - there's cancer, trauma, intravenous drug abuse, neurologic impairment, HIV, other immune deficiencies and abscess and there are a number of additional exclusions that we also have provided.

So basically that has been an overview of the measure. If there are any questions for the committee.

CHAIR CHOU: Thank you, so the lead discussants on this one are Craig, Thiru, and Katherine. Does one of you want to take the lead? Go ahead, Craig.

DR. BUTLER: I've been appointed to volunteer. So just without being repetitive just I want to emphasize the level analysis here is the facility, and I have some questions because I seen imaging facilities and care facilities and maybe we can talk about whether that in the context of where

health care is moving affects our judgment of this measure entirely.

Additionally, it is classified as an efficiency measure, and as we heard earlier, we're going to process this as a process measure, so my job is really to talk a little bit about the evidence and just a few facts and bullet points about evidence.

It is not a health outcome obviously. It is based upon a systematic review of about 15 guidelines and the girding underneath those guidelines thee is a total of 48 studies, and the summary is provided so that I'm not suggesting everybody walks through this diagram or algorithm as I did, but as I walked through it, it still leaves us with three options to rate as high, moderate, and low.

Obviously when we try and make those distinctions we're looking at three things - the quantity, the quality, and the consistency, and I think we can agree that

there's plenty of quantity here and the consistency in those studies is pretty high as well.

I do have some questions because they do break down the categories in those 48 studies which would be the foundation of the quality of the evidence and I see what, 40 of the 48 studies are rated in Category 3 and 4, so it suggests that the overall quality of the evidence is relatively low, and in fact they have zero in Category 1, and so that was my - and, again, you know, everybody makes your own judgment about where you - how you weight that factor, but we get pretty far down the road and we have some choices and I think that's the distinguishing feature.

CHAIR CHOU: Thanks, Craig. Thiru or Katherine.

DR. ANNASWAMY: I don't have anything to add to that.

CHAIR CHOU: Okay, Katherine?

DR. GRAY: Nothing to add either.

CHAIR CHOU: All right. I'm going to open it up to the rest of the panel. Do you have questions for either the lead discussants or for the developers? John?

DR. VENTURA: I guess this is for the lead discussants. One of the comments was there are no harms identified from overutilizing MR to counter or support the benefit of the measure. I would cite what I had referenced to JD. I didn't realize he was doing plain film.

Webster's study clearly showed higher costs, five times higher costs, delayed recovery, and higher disability associated with early unindicated MR, so I think it does address that issue.

CHAIR CHOU: So the study that John's referring to a study by Barbara Webster which was an observational cohort study. I think it was in a workman's comp kind of setting, and they followed patients out and, you know, they did find these associations

after trying to control for confounders, but it was an observational study just so people are aware.

Other questions or comments here? So, you know, the evidence for this one, it's not quite as direct actually I don't think as the one for the last measure where we have studies of routine imaging without routine imaging.

This is more about, you know, whether people got kind of what are considered to be appropriate therapies before jumping to MRI, and there's no studies that have really looked at that directly. I think that's a fair statement that you can make a lot of assumptions that based on the RCTs that doing a routine MRI doesn't really help patients, and the fact about these MRIs these days is that they're so - the detail is so fine that you pick up stuff in so many patients I mean Kat probably can speak to this some more, but the - you know, if you're over 50, the

likelihood that you're not going to have something on your back MRI is pretty low just because that's what happens as people's backs get older, and, you know, we've talked before about kind of all these kind of downstream harms which I think are some of the concerns that people have about MRI including, you know, additional surgeries and things like that which may be unnecessary, so, you know, again I think that some of the evidence to me at least seems a little bit indirect, but I think there certainly is some evidence to support what's in the measure here.

Do we have other comments or - yes, Thiru.

DR. ANNASWAMY: Yes, I'm glad you brought that up. I agree with the directness of evidence or the lack of it in terms of the conservative care of therapy recommended before getting an MRI.

The other issue was also - the measure developer clarified this during the

work group call where their intent is to try to reduce the number of MRIs that are ordered at the first visit and not necessarily trying to have them uncomplicated back pains always get therapy before ordering an MRI.

So perhaps the measure is stated in a way that doesn't completely capture the primary purpose of the measure so that was part of my thought that I wanted to share.

CHAIR CHOU: Okay, so bear that in mind because I think that will come up later again when we talk about - I'm not sure whether that fits into reliability or validity or whatever, but I think it comes up later again.

Okay, are there other comments about or questions about the evidence because otherwise I think that we're ready to take a vote on it.

All right, let's -

MS. PHILLIPS: All right we're voting on Measure 0514 evidence. You have

five options, one for high, two for moderate, three for low, four for insufficient evidence with exception, and five is insufficient evidence.

You may begin voting now. Okay. We have 21 present, and we are at two for high, 12 for moderate, four for low, two for insufficient evidence with exception, and one for insufficient evidence.

CHAIR CHOU: We meet our 60 percent threshold for high or moderate, so we're going to move on now, so, Craig, do you want to talk about research gaps and opportunities for improvement?

DR. BUTLER: Sure, so we have a performance gap between the tenth and 90 percentile, somewhere in the neighborhood of between 14 and 16, 15 percent.

There I think strikingly is a lack of significant percentage change between the 2007 and 2011 data that you presented, but those small percentages may in fact represent

big dollars, so there's two ways to look at that admittedly.

Nothing much in the way of patient level disparity data, but they did note what they termed a performance disparity with respect to the rate of the appropriate - according to their criteria MRI in the chiropractic hands versus surgeon hands and with the chiropractic being much lower rate of inappropriateness compared to surgeons, so that - I think again looking across the two years that they give you comparative data I think it's pretty striking at the inability to move the needle very much.

CHAIR CHOU: Do you respond to that?

DR. BRUETMAN: Yes, we understand and I just want to clarify how - maybe it's just an issue of how the data is collected.

We are looking - these are publicly reported measures of how hospitals compare for facilities in outpatient care for

hospitals.

The way the data is collected, and I know when you look at 2007 and 2011 it looks like a small - and it is a small number, the issue is that these measures are collected with data from two years behind so - because of the way - these are paid claims, so we can't collect in 2011 in 2011. We collect 2011 really 2009 data.

Public reporting started in 2010, so we're starting to see now the evidence of - or at least we're starting to see the ability of hospitals to react to the public reporting process which we did not have when we reported in 2010 because they could say any change I do will not be reflected until two years from now.

So that is why the variation has not moved enough. We hope that it will be moving as you get more information and more years of public reporting and hospitals look at this, and we're also looking at, as I said,

the higher end to decrease slightly - start decreasing towards the mean.

CHAIR CHOU: Yes, so some of this I think that Craig brought up will come back when we talk about usability and use. I think what we want to try to focus on a little bit is is there a performance gap. Are there people that aren't getting treated appropriately before they get their MRI, and at least according to this data that a substantial proportion of patients are - look like they're going pretty directly to MRI.

I have another comment that I think - well it's probably more related to usability and use also, but just I can't get patients PT or chiropractic care if their Medicaid or some of these other, you know, payers, so I think there are some challenges in terms of the usability thing, but again we'll talk about that when we get there.

Other comments about the performance?

DR. BUTLER: Just a question. I'm just wondering if the level of analysis has anything to do with the failure to see a change in performance. Any consideration in that this level of the facility is at the best level to measure and to really sense changes?

DR. BRUETMAN: Can you - what would be - maybe if you can clarify a little bit more to capture that concept.

DR. BUTLER: We've looked at other measures there at the level of an individual clinician or a group or plan, and this is a facility and especially how do you define a facility in a three hospital system and how that reports out or five hospital system. I don't - again, some explanation around what - how that's really done and called facility might help answer my question.

MS. ANDERSON: So facility has captured the individual outpatient facility level but also reason to capture at the facility level rather than the individual

clinician level is you increase the size of your denominator counts. You have a lot more power with the information you're saying.

If you're only capturing at the individual clinician level, most of your clinicians won't meet minimum case count requirements.

MS. MCKIERNAN: And other point, the reporting program I which this is included is the health hospital outpatient quality reporting program, so we are a bit constrained by the requirements of the project.

DR. BRUETMAN: So it's all hospital outpatient, I know hospital outpatient has been getting broader, but it's not at the individual physician, it's at the hospital level reporting, so we capture - really we're looking at the hospital's approach to developing to doing another study.

CHAIR CHOU: Any other comments about the performance gap?

DR. GRAY: I have one question and

that has to do with there was the notion that you would - and I think you just referenced it now that you would bring in the variants if you kept on going.

You didn't at least nothing that I could see describe what the outliers were and what - you know, were they things of disparities, you know. What characterizes the outliers and why would you think that they would change?

MS. ANDERSON: So our outliers tend to be more rural facilities. They're also ones with smaller bed counts and they tend to be non-teaching facility hospitals, so you do see some disparities based on hospital characteristics.

DR. GRAY: You're saying that they are overutilizing the MRs more?

DR. BRUETMAN: Yes, there's more use in those facilities, and we - I mean there are multiple hypotheses why this could happen. One of them is that - and this was discussed

by expert panel, you know, their concerns. Are we, you know, there are two ways of looking at this.

One is, well, are we putting rural facilities at a disadvantage, and on the other hand and that was the majority of the expert panel's thinking, well, it is also true that many times and probably with many of these measures you see much more on the primary care or family physician that might not be using completely appropriately the latest information on guidelines and that might be also part of why you can see the non-teaching hospitals or rural areas versus urban which might be using more frequently and more specialists on this, and I think that probably one of the reasons this measure was selected by the American Academy of Family Practitioners and not maybe other specialists who believe that they have much more an understanding of the details of why you needed or not an MRI.

CHAIR CHOU: I think that 31 percent is - it's not - I mean it's people that get an MRI without prior - you know, without these other care markers, so people who aren't getting PT and spinal manipulation and stuff, so it's - I'm not sure if it's really a measurable overutilization or it's more measured inappropriate that you're not doing the other stuff before ordering the MRI.

DR. BRUETMAN: I understand the question. I think you look at the idea is that you should not immediately do an MRI when a patient comes with low back pain and one of the approaches is really to say, well, there has to be a timeframe which people get some type of therapy or that do it before they get it and they were trying to avoid the doctor just going for the first time and the doctor saying, okay, you have low back pain, get an MRI because they say there's a significant chance that you're going to find something which - because that's - everybody's spine as

somebody said, they're all very different and even that was said that that's why some - there was a study and I don't have the exact who did the study, but the American Academy of Family Practitioners say - stated that they looked at older patients, and a study of older patients demonstrated - and these are people without back pain and that 80 percent I think were stating had some abnormality. It doesn't mean that they required something, so that was why they felt that even for older patients there's not - we can't say that you should do it immediately. There's always going to be findings on these patients and probably on other populations as well.

CHAIR CHOU: Linda, you had a comment?

MS. DAVIS: I have a question actually. The population doesn't have an age - or age range that I can see in the - so is this exclusively Medicare which means 65 plus or are there other populations that could use

this or do use this measure now?

DR. BRUETMAN: There's two parts to this. One is our population is 100 percent Medicare claims. We use Medicare, and it's for the Medicare program.

Now although we look at Medicare and it is - the majority of the people are over 65. It includes non - under 65 as well, and not in this case but in some that we use for emergency departments you see even more used under 65 than in the over 65 because the under 65 are either sicker or are dual eligibles or Medicare and Medicaid eligible, so they use more those departments, but in this case you have the study, the majority is over 65, but we do not restrict the measure to just apply only to the Medicare population.

We do not - although people live longer now, but, yes, it does include - fortunately they live longer, but it does not include - we do not restrict the age.

Also there's - well, we'll get to

that when we get to the numbers.

CHAIR CHOU: Thanks. Jason.

DR. MATUSZAK: I just wanted to clarify, the performance gap that you guys see and I took a look at some of your information in there between different large groups and small groups, rural areas and some of the other things, family practitioners and internists and things like that, but what I wasn't clear about I guess is is it the performance gap in between, you know, you're talking about the utilization of MRI at the first visit or is it underutilization of specific services prior to getting an MRI?

I mean are you measuring the conservative services ordered versus when they get an MRI or is it specifically based on that first?

DR. BRUETMAN: We are measuring the use of the MRI and when - and we use the use of conservative treatment as - in addition to exclusions that are included in the

measure, we also use that as an - if they had conservative treatment and evidence of this, then it's an appropriate study.

We're looking at the use basically. We're not measuring the use of conservative treatment. We're measuring the use of overutilization of MRI.

DR. MATUSZAK: So you're measuring the use of the MRI and then you're going to subtract out basically that they've had conservative management.

DR. BRUETMAN: I don't know if we have the - we don't have a presentation of the algorithm right? Okay, but basically we - let's see if we have the algorithm, we don't have it in front of us, but we can explain it if we have it in front of us, but basically we take in denominator everybody that gets an MRI and then we start taking the exclusions plus the people that we find evidence of prior conservative treatment.

DR. MATUSZAK: So I guess my

question is the gap that you see in the - the performance gap measurement that we're talking about is strictly just in the number or the percentage of patients with this diagnosis that get MRIs.

MS. ANDERSON: So there's two parts to the performance gap. The first part is, yes, what patients get MRIs and don't, and then the second part is actually with the data that you see in our measure which is you do see a wide range of rate of MRI imaging on our measure, and so you both have a performance gap in the order and then also within the facilities we actually do calculate the scores for.

CHAIR CHOU: Zoher.

DR. GHOGAWALA: I see how this is structured. Have you had some experience to see what the cost implication here? That is, you know, physical therapy is not inexpensive, and I just am curious. I know you have an option for evaluation and management, 28 days

to 60 days, but how is that breaking out in your measure?

DR. BRUETMAN: I'll be - you know, we'll clarify that and if CMS is here they can provide you further information. This program is strictly based - it's called pay for reporting program. There's no - CMS does not have the authority to use it as based on any performance at this point or pay for performance or looking at this outcomes of, well, but if you do physical therapy there might be an additional cost. We do not look at the cost implications of an alternative at this point.

We know that obviously the overuse of scans do lead to higher costs, but - and they do also lead potentially to overutilization of other services like surgery and other medical expenses as highlighted before, so there are costs but we do not calculate and CMS at this point was not looking to calculate costs on this. It's more

on the efficiency of the measure, and measure is defined and I think we used what NQF or ION had defined as do not provide - avoid use of resources that do not provide benefits to patients, and then the RAND which uses also the clinical ways of which the cost outweighs the benefit, and it was looking at really the - in this case the use of MRI not the subsequent cost.

CHAIR TEMPLETON: And this is Kim. I just have a question that refers to the exclusions especially since this measure as opposed to the last one does not have an age limit, should we also - or could there be inclusion of or why was there was exclusion of - exclusions such as signs and symptoms of infection such as fever or unexplained weight loss, osteoporosis, chronic use of corticosteroids, all those things especially if this is going to apply to an older population may need to be included because greater there - an exclusion for a history of

trauma, but if you have an elderly patient with osteoporosis, they may not have a defined episode of trauma that would lead to a compression fracture or something else that would need to be imaged.

DR. BRUETMAN: Can I - if I can - we understand that, and I just want to clarify that we have a number of exclusions followed most of the - I would say one of the major sources is the ACR guidance provided for this.

I'm not going to say it's 100 percent following, but probably you'll hear that, but we want to clarify that in the - and there's no reason for you to know this, but from the time we presented this we were still working as our angle maintenance process and we presented this in February.

We had our expert panel discussion in March, on 14 it was March 10th, so it was like a few days afterwards in which we presented additional exclusions with the - which the expert panel approved which included

many of the ones just mentioned - infections and osteoarthritis and we - many others that I'm sure are going to come up in the next few minutes, so those are going to be - CMS is in the process of internally approving this additional sources.

CHAIR CHOU: So, I just wanted to keep us focused on performance gap. We'll get back to the specifications and all that validity stuff when we get to those sections, but any other comments about performance gap?

So really what we're voting on is there a performance gap that this measure is addressing? Yes, Karen.

DR. PACE: Yes, and I just want to emphasize what you said. Obviously, you know, we're interested in improvement when we get to usability like you said, but one of the things to look at here is within a year the difference between like the tenth percentile and the 90th percentile, and you say see that there's differences in performance across

these facilities.

CHAIR CHOU: Okay. Additional comments about performance gap or shall we go ahead and do a vote? Let's do the vote.

MS. PHILLIPS: Okay. We're voting on performance for Measure 0514. Your options are one for high, two for moderate, three for low, four for insufficient. Voting begins now. We are at 22 - okay we have seven for high, 13 for moderate, zero for low, and two for insufficient.

CHAIR CHOU: So this passes. Let's move on to the priority. Craig, any opening comments?

DR. BUTLER: I think we can agree this is high resource use. Nothing else.

CHAIR CHOU: All right. Let's vote on priority.

MS. PHILLIPS: All right. Measure 0514 on priority. Your options are one for high, two for moderate, three for low, four for insufficient. You can begin voting as

soon as the mouse cursor comes back. Here you go. Begin voting now. Okay. We've got 19 for high, three for moderate, zero for low, and zero for insufficient.

CHAIR CHOU: Again, it passes, so we're going now to reliability, so again, focusing here on the specifications, whether things have been specified clearly and then anything about reliability testing. Did the developers have any comments to make about the reliability testing before we get into this discussion?

MS. ANDERSON: So I can provide an overview of the reliability testing we did if that would be helpful.

We - similar to the last measure you all heard about conducted a signal to noise analysis, and what we found is that we had a 53.1 percent median value for that analysis, and this is slightly lower than the target value you might want to see for a signal-to-noise analysis, but part of that is

reflective of both the way we structure the measure and the intent of what we're trying to achieve, and so when you're trying - the intention of the signal-to-noise analysis is to see how well you can distinguish one facility's performance from another.

We're never comparing head to head two facilities against each other. What we're instead trying to say is here's the median value of performance. Is a facility significantly different than this median value, and so when we ran a similar intent analysis but instead to look at how well we can distinguish facilities that are performing statistically significantly differently than the mean we were able to categorize over 50 percent of facilities having performance that was statistically significantly different from the median, and so I think that's a better measure of reliability given the intention of this measure.

CHAIR CHOU: Craig, any comments

here?

DR. BUTLER: This point, one of my other lead discussants now to take the floor.

CHAIR CHOU: Great, Thiru.

DR. ANNASWAMY: Yes, we traded, so with the reliability testing, there were not too many concerns. Most of the concerns under reliability actually applies to validity. The one concern expressed in the work group call was - well, concerns about addition of the new denominator exclusions, and there's no data but what the current data would look like if you use this as part of the analysis and that could impact changes in reliability over time because the criteria changed, so I guess it remains to be seen because you alluded to a lag period between when the claims are submitted to analysis to when you change some of the denominators so this is probably referring to that, but outside of that, there were no major concerns expressed.

So I think the consensus among the

work group and the lead discussants with reliability was at least moderate.

CHAIR CHOU: Katherine, anything or, Craig, go ahead.

DR. BUTLER: I did have a question about the denominator exclusion regarding prior surgery. Obviously it reached enough of a threshold that they made an exclusion. I was just wondering what was magical about the 90-day timeframe.

DR. BRUETMAN: Yes, I expected that and we heard this and this was a discussion with our expert panel and probably there's going to be potential differences in this and I'll leave to the expert clinicians to decide.

The way this was reviewed, the expert panel did not consider any prior surgery, and I know that's different than the ACR as an automatic exclusion. Their view was that what they were looking at was a prior surgery and within the 90 days as one could be

an arbitrary number is is there any issues related to that surgery that occurred, and they're putting the 90 days to see well maybe there was a problem and we want to go back to see there, and they want to put a timeframe so it's not, you know, we put a year would be completely maybe unrelated. That's why the prior surgery and the 90 days was based on follow up to a recent surgery that had occurred.

DR. BUTLER: That Number 5 guideline to get there it was just a - what you use for all of your validity and reliability as a condition of low back pain, and so when you exclude surgeries of 90 days, my point is I don't think someone who's had surgery in say 91 days after can be called just low back pain, and, you know, even a year or two after, probably is not uncomplicated low back pain, and I kind of speak to the issue as a multiple back surgery patient, so none of it's uncomplicated.

DR. BRUETMAN: I understand - and I know and I understand the concern and we're open to bringing this up again to the expert panel and to CMS.

We did look at the number of surgeries and it wasn't -- there is a number of surgeries, but it wasn't affected that much the measure.

MS. MCKIERNAN: So that's approximately 11 percent of the cases were associated with surgery within the 90 days prior to the - and the other point that I wanted bring up is I think that - so Charlie's explanation about the reason for having going back for surgery was to look for something related directly to the surgery.

I think that you're speaking to a different reason to exclude surgery. As Charlie mentioned, I think that it's valid and it's something that we can discuss.

DR. BUTLER: Just to understand you that of the claims that you looked at, 11

percent have had surgery within 90 days and they were excluded. Is that -

MS. MCKIERNAN: Apologies. I was thinking of a different exclusion. One percent.

CHAIR CHOU: Okay. Thanks, so keep that stuff in mind because I think that - again, a lot of that will get the validity stuff. I want to see if there's any final thoughts about reliability, so again, whether the specifications are clear and whether their measurements are reliable. I want to focus on here before getting into the validity stuff. Oh, go ahead, Kim.

CHAIR TEMPLETON: Sorry. I forgot to lower my hand. I'm okay.

CHAIR CHOU: Okay. I think it's time for a vote then, so we're voting on reliability.

MS. PHILLIPS: Okay. Reliability for Measure 0514. You have four options. One for high, two for moderate, three for low, and

four for insufficient. You may begin voting now.

We are at 21 holding there for a couple of seconds so if you can all vote again. There we go. Thank you.

We have one for high, 19 for moderate, one for low, and one for insufficient.

CHAIR CHOU: Okay, so that passes the reliability test. Excuse me. Now we're going to move to validity, and I think some issues have already been brought up with the exclusions, the surgery exclusion in particular. Do the lead discussants want to make some more comments here?

DR. ANNASWAMY: So the validity of the measure was the biggest point of concern. The purpose of the measure, the intent, is to look at uncomplicated low back pain and figure out how many of them got an MRI as kind of a trigger-happy clinician at first visit or as early as possible without trying conservative

care, so the issues with numerator and the denominator.

The denominator is you're calling uncomplicated back pain, but the diagnostic code is essentially claims based on ICD 9 coding. Perhaps things would be a little clearer for ICD 10 but with ICD 9 you didn't include just the diagnosis of uncomplicated low back pain or 724.2 or 724.5, but a slew of degenerative codes are also included, so all the degenerative disease, spondylosis, sprains, strains, radiculopathy, they're all included in the denominator which makes the measure in my view and the work group's view very low validity.

In the numerator, you're looking at people who had antecedent conservative therapy.

All the literature that talks about sufficient trial of conservative care but here you're looking at physical therapy, chiropractic adjustments, and E&M code. I've

never quite - I can't quite say it unless I read it. Greater than 28 days but less than 60 days before the MRI. I think it means to say between 28 and 60 days before the MRI. I understand that better than greater than 28 and less than 60, but be that as it may, I don't think it quite captures the intent because you may have a variety of different conservative options used by clinician that are very appropriate and stated in the literature as being very appropriate which does not include these three things, and the measure does not capture those.

For example, education, avoidance of activities that might have strained it in the acute back pain, modifications of the work place for example, things of that nature are not captured here.

If an evaluation was done three months before but there was a delay in getting an MRI or perhaps a follow up was scheduled and at that follow up visit an MRI was ordered

because sufficient time has elapsed, treatment has been instituted, patient is not getting better, so in the Medicare population an MRI was ordered at that point, and that will be counted against it because this is a trigger-happy clinician ordering an MRI at the visit. Ninety days ago the first visit was not included in the claim because of the time window, and there are scenario after scenario that I could think of that was brought up in the world group call that just makes this measure extremely problematic in terms of its specifications, its inclusion, exclusions, and all of the above.

CHAIR CHOU: Thanks. Other comments? Katherine? Craig?

DR. GRAY: Yes, I have one. I think the one that we got into the most lively comments had to do with - well first of all, if you look at all the information that you provide for the evidence, you have all these overlapping guidelines, and I think in itself

is very confusing for anybody to know what do they really follow, but what really took the cake was when we got into Guideline Number 5 which you actually pull out and list in a separate thing as being important, one of the variables in it for Variant 2 is age greater than 70, and the group would just - freaked out at that and because it's a Medicare population and so we went back and re-read it, and it says one or more of the following, so just being over 70 means that the ordering doctor would be legitimate in ordering an MRI.

Now what's that do with the whole purpose of the measure, you know. That was what really caused us the greatest concern.

CHAIR CHOU: Do the developers want to respond?

DR. BRUETMAN: Yes, as we said, we had mentioned, we - the way we look at that there's overlapping measures and not all address the same issues and I wouldn't say the conflict some, but there are minor

differences.

We follow significantly the ACR guideline and when we looked at the age, and I think it was brought up even the discussions as we saw on the transcript and we heard last time, the expert panel thought that it was not inappropriate - it was not an appropriate exclusion that a person over 70 is automatically should be excluded.

In their perspective it's not a sufficient reason to say that you're 70 to not have it and this is might be a disagreement with ACR's guideline.

I think it was presented even by - I think, yes, because I'll say Zoher, I think he mentioned that he wouldn't - as an orthopedic surgeon he would not - he does not believe that's an appropriate exclusion. I think, and I don't want to mistake, but I think it was part of the saying that we would never have a surgery - we would not immediately scan somebody just because they're

over 70. The expert panel felt the same way, and I think as we said our development is based both on evidence from guidelines as well as experts that come to the discussion and that's why that specific area was not included as a direct exclusion in our measure.

MS. ANDERSON: And just to add to that, we also did review the evidence table behind that ACR guideline, and only two of the studies included in the evidence table discussed age at all as a factor when considering the appropriateness of an imaging study for MRI of the low back, and both of those studies only noted that you were likely to have - or a higher likelihood of abnormal findings, but they didn't correlate that with a need to do this, so even though as you get older you expect to find more abnormalities in the back. It's not actually the MRI might be necessary based on the evidence we saw in the evidence table.

DR. GRAY: And I've looked at that

evidence table, and it's over 50, so I spent some time actually finding the over 70, and I think that what I found when I looked it up is that it's not in isolation but that guideline, the information that is being laid out as the validation for the measure does not have that subtle - all the information that the literature actually shows, so there's a real conflict in terms of what - you know, what you're actually trying to - I mean we - you and your expert panel may know, but how are people out there using this measure and being judged by it, what does it all mean to them if they actually are mixed up enough to have picked Guideline Number 5, Variant 2, to be their thinking?

CHAIR CHOU: Other comments from the panel? JD.

DR. DANIELS: Yes, I guess I'm stuck on a horse fetish today. This is like the Wizard of Oz. It's kind of like a horse with a different color. You know, different

people come into the doctor for different reasons, and dependent on their age and what happened to them, we look at it differently, so if a real young person comes in the office and they just didn't fall off a swing set or something or having back pain, most of the day you kind of go along with it, but every once in a while your antennas go up and you go, wow, I got to kind of pay attention, and then you get to kind of the middle aged people who kind of like don't need to see you unless they've got one of these big burning red flags, but when you kind of look at the elderly group when you get over like 50 it doesn't count any more, you know, like 50 is the new 30, so, you know, when they're up around the elderly, you know, there's some simple things like they fell when they were coming out of church, and there you can do an x-ray and see if they've got a compression fracture or something, but more sinister stuff comes on to bring those, so most likely what's

going to happen, and I'm just talking practical, is an older person comes into the office or having pain and you talk to them. This seems okay, or, you know, they actually examine them and take their chart. Well they've got shingles - oh, boy. Sometimes that happens. People don't look, but the bottom line on it is that if a period of time goes by and they've got those, they're a lot higher risk now. That, I don't know if you can find that in - you know, teasing it out, but in general, the MRI would be one of the few times because the way I sort of look at an MRI as a primary care doc is not to make a diagnosis because they're going to give me 50 diagnoses to choose from when I order the MRI. You go A, B, C, D, E and kind of tell the patient a nice story. What I really look at is for two reasons.

One is to make sure that I'm not telling the person, hey, you're going to be fine almost no matter what you do, okay, and

the other one is you need surgery or you need - and, you know, very few times I have to call someone and say, hey, can you get him worked in, so that's kind of how it's clinically.

CHAIR CHOU: Thiru.

DR. ANNASWAMY: I just - I had more things to say, but I didn't want to hog the mic, so Round Two.

So in the guidelines that you have so elaborately summarized contradicts your measure directly. I'll just read out some underlined points here.

A specific clinical diagnosis of suspected lumbar disc herniation or sciatica, radiographs are not indicated unless patient is aged greater than 50 or has progressive neurological deficit, so that itself, the guideline states that in several of the instances that you're trying to capture, it might be appropriate to get an MRI.

Another one, Guideline 2, for patients with back pain, clinicians should

consider the use of medications with proven benefits in conjunction with back care information and self-care, so that counts as appropriate conservative care before getting an MRI and that's not captured in your numerator.

Guideline 7, MRI is not recommended for acute radicular pain syndromes in the first six weeks unless they are severe and not trending towards improvement and both the patient and surgeon are willing to consider prompt surgical treatment assuming the MRI confirms ongoing root compression.

So in this case, six weeks, if they're severe and there is a plan for surgery, it's appropriate according to this guideline to get an MRI.

MRI at three to four weeks may be reasonable if you're planning an epidural steroid injection with is also accepted as appropriate conservative therapy. It's evidence Level B, a moderate recommendation.

Guideline Number 10, spinal cord infarction or degenerative conditions, degenerative disc disease or its sequlae in the lumbar, thoracic, and cervical spine are indications for MRI.

For radicular pain without weakness by greater than - at greater than or three weeks, if there is no improvement obtained in MRI Level 2B evidence or diagnostic obtained in EMG.

If pathology is proven on an MRI, consider evaluation by specialist or surgical eval, so it's just on and on and on about how it may be so many instances that are not captured in this measure where obtaining an MRI is appropriate and also it's instances that are not captured here where obtaining an MRI is not appropriate that are not excluded - I mean that are excluded.

CHAIR CHOU: I want to go the developer the chance to respond and then come back to Jason.

DR. BRUETMAN: Obviously there is difference evidence and different guidelines that state, and we try to make the best case of understanding. That's why we bring some experts.

One of the things and I understand some of the denominator codes, there is a number of exclusions that eventually these patients are excluded from the analysis, and as I said, there's some additional exclusions that were not initially involved and are now currently in the measure or will be in the measure, but we had a lot of the number of these would-be - although they're included in the initial code, but then when that diagnosis is there, they will not be included so it will not be telling any provider that you're doing the wrong thing because they would eventually be excluded, so that addresses may of the issues.

The second is, yes, it is a claims-based measure, and, you know, one of

the challenges we have with claims-based measures you're not going to get all the information, but that would probably then leave matters to be almost, as I said, the trigger-happy doctor to justify and could almost justify any study because there is a way to almost justify any study, so that's why also the measure does not have a benchmark and the goal is not to have zero.

If you look at the rate, it's not a zero, so we understand that many cases are going to be, you know, well, but you don't have in the claims all the exact symptoms that you can capture and that's why, you know, we are not telling people, you know, oh, you're at 30 percent, it's really bad performance. You're at 22, you're bad performance.

There are - that is that I think was discussed previously, there's no expectation to get a zero on these measures, and as we call it, we try to work in giving I'll call the benefit of the doubt of why and

that - there's long discussion on how E&M - I know it's not very clearly captured. At least I know there have been a number of questions regarding our E&M approach and ability to capture that as - the claim for that as a conservative treatment, but obviously if a patient - we try to use proxies as, you know, they did not because our goal is to avoid a person going the first time to the doctor and getting the MRI immediately, and if they took 60 days - and one could say, well, what happened if they waited 60 days to get the MRI, yes, we will say they were included but if you have 100 percent of your patients that have that reason, I think there's another problem. That's why I think, you know, we're trying to play with getting people to acknowledge that there's an overuse as was agreed by the whole committee. There is a significant utilization capture as much as possible of the claims data, give I'll call leeway to the doctors to understand that

there's not going to be perfectly black and white cases because medicine is not a black or white science, and also to understand that, you know, your facility is not going to be if you want to call it punished or seen as a bad provider because you have a number of these cases.

Now when you have that 60 percent of your patients that come through your facility get an MRI for low back pain, you have to question yourself and that's what we're trying and CMS is trying to work on is trying to say how can we do the best way possible within the limitations of a claims-based measure and the hospital community and the medical community in general has reacted very positive to the measure. There's not been in many other measures I can tell you there's not been positive reactions as you are probably all well aware, and in this case we have not had any comments saying, well, we're not being captured. There's a mistake.

Everybody has their unique cases.

You know, I have a lot of this type of population or that type, and that happens all the time and we understand that. That's why you explain to them we're not paying for performance. We don't expect it to be zero, and there's no established benchmark.

We are working on people being at a rate which is called the average rate and try to get people when they're very high to see - recognize that their facility is very high on that to question themselves, are we doing something that is not appropriate that we need to work with our physicians to ensure that, you know, if I'm in a rural facility and maybe don't have access because we don't want to make a measure and nothing against - I mean - John Hopkins and we have people from John Hopkins in our panel.

I don't want to make a measure that address Johns Hopkins because it's like that's not what we're looking for. We're

looking to see if there is a facility that is saying, wow, why am at 60 percent and others are at 30? What is going on? I'll have to talk to my physicians to see are we just trigger - trigger to get or other cases, and every hospital we provide every single data point to every hospital so they can see because it's 100 percent of paid claims so I mean this is 100 percent of reality assuming that what was claimed - we'll assume that case and everything that was paid and claimed is reality, but, you know, we're not going to get into that discussion, but we are assuming that the facilities did perform the study. We provide every single patient and every facility has a time to come back to us and say, look, we reviewed this, and there's something wrong, and we go to that case and clarify that.

I have to say for this measure I don't want to say we've never had a concern, but it's not been a problem. Recently there

was a change in the way for multiple reasons, different way of calculating. There was a problem with the measures, and hospitals immediately called us saying we didn't meet the minimum case count. Our cases are not right. What is going on, so we had to explain there was a change in not our system, another systems, and that was clarified and fixed, but - so hospitals are aware of - they look at this and again, I think the goal is to not have perfection but to get in this case facilities to react to say what am I doing differently, what can I learn, what should I be doing to be more aware that why am I having, and I have to say, not this case where we facilities with 100 percent cases.

CHAIR CHOU: I want to have some of the panel members give their input, so, Jason, and then John.

DR. MATUSZAK: Well, I think your point is well taken that we want to raise awareness around the issue. I think that we

all accept that - that is part of the problem is making sure that people are aware, but I do think that Thiru did a wonderful job of pointing out just how conflicting your guidelines are and how physicians might be dinged for performance or facilities might be dinged for the physicians' performance even when they're following Medicare's own guidelines about it.

I also would just, you know, just in case by some chance this goes through I would also ask that you guys if you're looking at conservative manage tests, this stuff really be very inclusive. You don't have osteopathic manipulation, but you have chiropractic manipulation. Why?

You don't have athletic training services even though them doing a home exercise program could be just as effective as physical therapy in some of these things. You don't include that. You don't include anything about nutritional services, and you

know that dietary interventions can make a big difference in low back pain, so, you know, I would just encourage you to take a look at being much more inclusive as to the conservative measures that you would deem as being part of your qualifying things, but certainly I think you did a fantastic job of pointing out just some of the flaws and the validity here.

CHAIR CHOU: John.

DR. VENTURA: My question is actually along the same lines, but it's actually more for Thiru.

I could be reading this wrong, so please help me if I'm just off base here. The intent is that in essence you're not ordering an MR in the first 28 days in the absence of an exclusion, so if there are no exclusions, they don't want you ordering within the first 28 days. Am I reading that correctly?

DR. ANNASWAMY: The measure developer probably can clarify, but, yes, the

idea is that if you had an E&M claim which means a visit, a clinic visit, less than a month before you got the MRI that most likely means that the MRI was ordered at that first visit for the back pain which you don't want to have, but if you waited - if the MRI was 60 days after the clinic visit, it's likely that there was a good set of conservative care that happened in between. That's the assumption.

DR. VENTURA: I also get the impression though that if an E&M code is submitted after 28 days -

DR. ANNASWAMY: So it's E&M should be looking back from the MRI date, so that's how you do it. Looking back from the MRI date, you go back 60 days - no, wait. No, looking forward from the clinic visit.

DR. BRUETMAN: It's backwards. You're right.

DR. ANNASWAMY: So from the clinic visit, your MRI should be not before the first month, and it shouldn't be - it should be

between that first month and second month. That's basically it.

DR. BRUETMAN: And just to clarify. I know it can look confusing why we didn't say between this and that. It's just the way the program has to be done. You have to look at the way you put it's more than this, less than that, and that's why it was not in the perfect English wording, but that's how it's programmed.

What you said, it's very clearly not easily interpretable.

CHAIR CHOU: I want to ask one clarifying question and then I'm going to kind of summarize where we're at because we need to come to some resolution here, but I'm actually - now that I look at this Item Number 3, it's actually confusing to me, so in terms of the description, so we're saying if they got PT in the 60 days prior, it's okay. If they got chiropractic evaluation and manipulation, it's okay. I don't understand what this third one

is saying that if they were evaluated sometime between 28 and 60 days whether they got PT or anything else, it's okay also.

DR. VENTURA: Yes, that was my question actually that I thought that encompassed a lot of the concerns that Thiru had that, you know, that they could have had education during that period. They could have gotten nutrition. They could have had anything else.

CHAIR CHOU: Well they could have had just had come in to see the doctor and that would be considered that they got conservative care.

DR. BRUETMAN: The proxy is that the first part is pretty clear. You know, you had physical therapy in the 60 days. Obviously we can determine there's been some conservative treatment. The same with the second one.

The third one is so you - it goes to this issue, well, we don't want you on the

first visit to do it, so there's a time lapse that is between the last E&M and an assumption that it wasn't done immediately because there was some type of conservative treatment or some treatment, education, done by a physician that is not a chiropractor or a physical therapist, and unfortunately there is - for those who work with claims, there is never - you know, the claims, the way they're put in place, they will not put certain codes where they'll say we are doing or giving them this type of therapy or doing that. It's not capturing the claims, so we - to avoid - and this is a very broad approach, to avoid giving doctors that might be doing this or saying if you identify and E&M claim that has been more than 20 days were related and that's when it was ordered, were related that something occurred that was conservative before just providing.

It is true that if somebody gets in Day 60 their order and goes 60 days later,

yes, that would be we're excluding a physician, let's say in this case, although we're not measuring physicians that shouldn't have been excluded, but that's why it's not going to be zero.

CHAIR CHOU: All right, so if you're seen in the doctor's office one to two months before the MRI, that's your proxy for getting conservative care. Okay. All right, so I just want to summarize where we're at, and then we'll take a couple of final remarks and then I think we need to come to some resolution here, so I'm hearing some concerns about how the denominator is being defined, in particular about the exclusions and that there are some discrepancies between guidelines and I think it's not clear to everyone how the developers kind of reconcile those discrepancies or decided how to apply them.

Then I'm also hearing concerns about the numerators, both in terms of I guess the specific, you know, interventions that are

discussed as well as whether this other one to two month thing is an appropriate proxy.

Other final - a couple of final comments if there are some, and then we can maybe come to some resolution here. Craig.

DR. BUTLER: Maybe just one final comment. I see this is as a measure that become increasingly less valid especially when you consider the evolving models of care and you've kind of constrained it to three different ways to basically document conservative care, and I submit that you might have an app, you might have a video you refer a patient to that gives them a new care model to initiate conservative therapy which means that this becomes again less and less valid over time.

CHAIR CHOU: Yes, I was actually thinking about telemedicine for like people who are in rural communities who may not have access to PT or something. Telemedicine is probably something that's going to be widely

used in the future whether you can - how people are going to bill for that and all that other stuff is a question, but any other final comments? Yes, Thiru.

DR. ANNASWAMY: I'm not sure how many clinicians use the neurological impairment exclusion code in their coding. I've practiced spine care for 15 years. I've never once used it, and I've seen a lot of neurological impairment.

I would use radiculopathy or sciatica which are inclusions not exclusions here, so I just wanted a clarification on that neurological impairment exclusion.

CHAIR CHOU: Okay. I think we need to come to a vote here, so let's put it up for a vote for the validity criterion.

MS. PHILLIPS: Okay. We're voting on validity for Measure 0514. You have four options - one, high; two, moderate; three, low; four insufficient. Voting begins now.

We are at 22, and we have zero

high, four moderate, 15 low, and three insufficient.

CHAIR CHOU: All right, so this doesn't pass the validity test. I think we want to maybe provide some feedback and then -

MS. PHILLIPS: Do we have additional feedback.

CHAIR CHOU: Yes, so I mean, Thiru, did you want to say something here?

DR. ANNASWAMY: Yes, I think the intent that this is a very important priority and that you want to prevent the trigger-happy clinician from getting too many MRIs too quickly for uncomplicated back pain is extremely valid, and that bore out in the way we rated the evidence and the priority and the performance gap.

The issue is how the measure is structured, so I believe there is room for refining the measure so that you adequately capture that intent.

CHAIR CHOU: Craig.

DR. BUTLER: I would also point out when we are trying to do this thing using the guidelines for evaluating validity, the first question is are the measure specification consistent with the evidence provided in support of the measure, and I would submit that some of your evidence was inconsistent and some of the evidence wasn't presented witness the technical expert panel input on some of those ages and things like that that we didn't always have access to in the materials that we got a chance to review.

DR. GRAY: And I would suggest that the validity issues that we discussed and were raised may very well be the reason you're seeing no change over time.

CHAIR CHOU: Yes, so, again, I think everyone - the panel generally felt this is important and that there is evidence to support the overall aim.

I would bring up another issue which is that it seems to me that there are

two - I mean one problem we have is that not enough people are doing appropriate therapies for low back pain, that people are jumping to opioids, unnecessary surgery, etcetera, etcetera, and it's kind of addressed here but not really.

I mean I think that's almost a separate issue. I mean I think kind of confuses the picture here, and I wonder if there's some way to try to separate out the appropriateness of MRI separately from whether they're receiving appropriate conservative care. It may be the only way you can do it, but it's just something worth thinking about.

The other suggestion I would have is just to stick to one or two guidelines. I mean you're not going to be able to reconcile 12 guidelines. It's just - you've got to - because the methods that people use and the standards that they rate the evidence with and all this other stuff are completely different, and I would say that the ACR appropriateness

criteria are completely different from like a clinical guideline in terms of how they're meant to be used and interpreted and so to try to develop a measure that's based on all these disparate things, I think it's going to be tough, and so I think if there's a way that you're able to come to some - evaluate the guidelines and come to some, you know, agreement about which ones you think, you know, are the best developed or whatever, I think that would help kind of clarify some things, so, Thiru.

DR. ANNASWAMY: I agree with you about the ACR. I think ACR, a lot of the sentiments expressed here are try to get the patient in the middle of this.

ACR - not ACR, AUC, Appropriate Use Criteria, sorry, is focused on the modality as a focus, so you're trying to figure out for the modality in my hand, for the hammer in my hand, what should be the nail? How should it look like?

On the other hand you want to focus on the nail and you try to figure out what's the best hammer to use, so that would be - so you shouldn't - I would actually extend what Roger said and not use AUCs at all in determining this kind of measure. Patient focus should be the priority.

CHAIR CHOU: Any additional comments? All right, I think we're done with the measures. Thank you.

MS. STREETER: Just take a quick break to see if we have any members from the public that would like to make a comment on the phone.

OPERATOR: At this time if you have a comment, please press Star One, and there are no comments.

MS. STREETER: Thank you. So thank you all for joining us for our in-person meeting.

Just so you know, next steps as I mentioned yesterday, staff will be preparing

a draft report that summarizes your recommendations from today.

We'll be posting that report for public comment in a 30-day public comment period in June, then we'll be reconvening in July, I believe it's July 31st, to discuss the comments received.

MS. FRANKLIN: And I just wanted to ask if folks are interested in going around the table quickly to talk about gaps in the portfolio.

CHAIR CHOU: I'll just mention I think management of chronic pain should fit under the musculoskeletal thing, and I think opioids are one of the major issues, public health issues, right now. Gaps that people think that there should be measures to address?

CHAIR TEMPLETON: This is Kim. Getting back to some of the imaging modalities, perhaps use of MRI in management of chronic knee pain.

CHAIR CHOU: Thanks, Kim. Yes, Sean.

DR. BRYAN: One of the things we see very commonly in primary care sports medicine is tendinopathy, and it might be nice to have a measure on that.

CHAIR CHOU: Management and treatment, so evaluation management, okay. Yes, Linda.

MS. DAVIS: Common procedures outcomes from knee replacements, hip replacements, and spinal fusions.

CHAIR CHOU: Okay. Thanks. Other thoughts? Yes, Zoher.

DR. GHOGAWALA: I think the overutilization efforts or the efforts to create measures to address overutilization didn't do as well as we might have liked, but I think that there's a real role for more measures that look at some of the very costly overutilization procedures including in our field lumbar spinal fusion surgery with the

effort of not necessarily limiting but clarifying for patients and for providers when these kinds of expensive treatments are appropriate.

CHAIR CHOU: Yes, I think for the developers we wanted to be clear that we are interested in measures that address overutilization, that's is just challenging to develop them, but we'd like to see them again or see more of them. JD.

DR. DANIELS: Yes, integration of different disciplines in the management of musculoskeletal problems, I mean it's exactly what we're talking about today with the back and that.

I mean I don't think that's - everything is siloed, you know.

CHAIR CHOU: Thanks. All right

CHAIR TEMPLETON: This is Kim. I guess maybe one more is evaluation or secondary fracture prevention, the evaluation of bill of health after the initial fracture.

CHAIR CHOU: Thanks, Kim. All right I sense people are flagging and ready to go out jump out the door, so are we ready to.-

MS. FRANKLIN: Yes, you can also feel free to email staff with your additional suggestions, and we'll be in contact with you. We still have to - we're not - I don't think we have any - yes. At this time we don't anticipate that, but let's leave the dates open in case we have some issues to reconsider. We will as soon as possible. As soon as possible.

CHAIR CHOU: All right. Thanks, everybody.

(Whereupon, the above-entitled matter was concluded at 2:34 p.m.)

F

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Musculoskeletal Measures Standing Committee

Before: NQF

Date: 05-08-2014

Place: Washington, D.C.

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

Neal R Gross

Court Reporter