NATIONAL QUALITY FORUM
+ + + + +
RESOURCE USE BONE/JOINT

TECHNICAL ADVISORY PANEL MEETING

+ + + + +

Thursday,
July 7, 2011

+ + + + +

The Technical Advisory Panel met at the National Quality Forum, Suite 600 North, 601 13th Street, N.W., Washington, D.C., at 8:30 a.m., James Weinstein, Chair, presiding.

PRESENT:

JAMES WEINSTEIN, DO, MS, Chair, The Dartmouth Institute for Health Policy
MARY KAY O'NEILL, MD, MBA, CIGNA HealthCare

ELIZABETH PAXTON, MA, Kaiser Permanente*
JOHN RATLIFF, MD, FACS, Thomas Jefferson University
CATHERINE ROBERTS, MD, Mayo Clinic
CRAIG RUBIN, MD, University of Texas Southwestern Medical School
PATRICIA SINNOTT, PT, PhD, MPH, VA Health

Economics Resource Center

NQF STAFF:
TAROON AMIN
HEIDI BOSSLEY, MSN, MBA
LAURALEI DORIAN
SARAH FANTA
ASHLIE WILBON

ALSO PRESENT:

DAN DUNN, PhD*

TODD LEE, PharmD, PhD*

LAWRENCE MANHEIM, PhD*

HOWARD TARKO, MD*

CHERI ZIELINSKI*

*Present via telephone

T-A-B-L-E O-F C-O-N-T-E-N-T-S

T-A-B-L-E O-F C-O-N-T-E-N-T-S

T-A-B-L-E O-F C-O-N-T-E-N-T-S

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

8:41 a.m.

MS. WILBON: So, Operator, we're going to go ahead and get started.

OPERATOR: Okay. You are connected. Go ahead.

MS. WILBON: Okay. So good morning, everyone. We are actually going to go ahead and get started now that we have everything. The technology is all set up.

So welcome, everyone. I know everyone came from near and far and we are excited to finally be able to discuss the bone/joint measures. We are about three-quarters of the way through our TAP meetings. We have got one more meeting in a couple of weeks with the Pulmonary TAP. So we are excited that things have been going well.

And hopefully along the way, we will be able to, based on some of the meetings that we have had already, offer some guidance on how to make things a little bit more

efficient as we go.

So my name is Ashlie Wilbon. I'm the project manager for this project. And I'll just I guess introduce my staff or let them -- my team, our team. And you guys can introduce yourself.

MR. AMIN: Hi, my name is Taroon Amin. I'm the Senior Director supporting this project also. I recently joined NQF from the Brandeis Team working on the episode group or software for the public sector program.

MS. DORIAN: Good morning, I'm Lauralei Dorian. I have also recently joined NQF. I've actually come from New Zealand and I'll be working as a project manager on this project.

MS. FANTA: Good morning, everyone. I'm Sarah Fanta. I'm project analyst on this project.

MS. TURBYVILLE: Good morning. I'm Sally Turbyville and I'm serving as a consultant in helping supporting this effort

with the staff.

MS. BOSSLEY: Hi, I'm Heidi Bossley. I'm the Vice President of Performance Measures at NQF. And we are thrilled to have you here. Truly appreciate all the work that you have done and what you are going to do today. We know it is not a small amount of work we have asked you to do.

So it's very much appreciated.

MS. WILBON: So actually, I'm going to throw it back in Heidi's corner. We are going to have you each go around and introduce yourselves to each other and at the same time, Heidi is going to give you instructions on how to -- about the disclosure of interests that we will do before we start evaluating measures. Thank you.

MS. BOSSLEY: Okay. So as you all may remember, it was a while ago, but we asked you to fill out Disclosure of Interest Forms. What we are asking you to do today is just orally provide information on anything that

may be directly related to the work here.

So you don't have to give a list of everything. You don't have to say what every membership that you have, but anything that may be funding that you received or any work related to this project, I would disclose.

The other thing I would remind you all as you are sitting as individuals, not representing the organization you work for or who nominated you. It's just a reminder we like to give everyone. You are here to give your expertise.

So we will just maybe start around the room and give some introductions as well as any disclosures you may have.

DR. RATLIFF: Hi, I'm John Ratliff. I'm a neurosurgical spine surgeon from Thomas Jefferson in Philadelphia. Can you guys hear me okay? I don't have any direct conflict of interest related to outcomes assessment and neither conflicts I

stated in instrumentation development and royalty payments for same. And we're glad to be here.

DR. O'NEILL: I'm Mary Kay O'Neill. I'm the Chief Medical Officer for CIGNA in the Pacific Northwest. And I'm board certified in PMnR. And I don't have any conflicts.

DR. SINNOTT: I'm Patsy Sinnott. I'm from San Francisco. I'm from the VA, the Health Economics Resource Center. I'm a physical therapist originally by training. And my only disclosure would be that when I was at PBGH for two years after my graduate work, I worked with the Ingenix Episode Grouper and the Cave Episode Grouper, so that, you know, I have my experiences with both of those.

DR. RUBIN: I'm Craig Rubin. I'm from the University of Texas Southwestern Medical School in Dallas. I'm the chief of the geriatric section and I have no conflicts

to report.

DR. ROBERTS: Good morning. I'm Catherine Roberts from Mayo Clinic in Arizona. I don't think I have any pertinent conflict of interest, but I do work on -- as a musculoskeletal radiologist. I do work on the appropriateness criteria for the American College of Radiology and also for national quality improvements, metrics for the American College of Radiology.

CHAIR WEINSTEIN: Jim Weinstein from Dartmouth. I think my conflicts are probably the Dartmouth Atlas, which works on lots of claims data for Medicare data mostly. I also have several NIH grants related to spine and will probably want to state some of the literature issues that are missing probably biasly because of my own work.

So, please, forgive me ahead of time. I also am the editor and chief of Spine, so I have some other literature knowledge. I currently serve as President of

the Dartmouth Hitchcock Clinic, a 1,000-physician group. I'm a spine surgeon and I'm the Director of the Dartmouth Institute, which does the Dartmouth Atlas.

MS. BOSSLEY: Okay. Anyone on the phone, any Committee Members? Liz Paxton?

MS. PAXTON: Hi, Liz Paxton. I'm the Director of our National Implant Registry for Kaiser Permanente and that does include both cardiac and orthopedic and I do not have any conflicts of interest.

MS. BOSSLEY: Okay. This is the usual question we ask. Does anyone have any questions for your colleagues or anything you would like to discuss that they have disclosed? That's the typical answer, too.

So we are going to -- thank you very much.

MS. WILBON: Okay. Great. Thanks. So what we have is just a brief kind of introductory slide presentation for you guys to kind of get everyone on the same foot

this morning, go over the criteria and kind of some operational things that we will encounter as we go through the day.

So I think everyone has the slide packet in their folders as well, if you want to follow along.

So today, essentially through this presentation, we are going to be briefing giving an overview of the consensus development process. You can get an idea of how this meeting and this project kind of fits into that overall process.

Obviously, we want to make sure that you have a good understanding of the evaluation criteria. You have already started evaluating the measures that you did before you got here, so I'm assuming you guys already have some understanding of it, but hopefully we can clarify any questions that you had in that process as well, obviously, to evaluate the sub-criteria of the four bone/joint measures.

And then throughout the day, at the end of the day, if you have any suggestions on process improvements or, you know, how we might be able to do things better in the future, we do have one more TAP meeting. So to the best of our ability, we are trying to carry forward any, you know, efficiencies and lessons learned along the way. So we are definitely open to that input.

So the consensus development process is, approximately, an eight-step process. The two steps that are grayed out have already been completed.

We are in the Consensus Standards Review step at this point. Once these TAPs have finished evaluating all the measures, and that input is forwarded to the Steering Committee, staff will put together a draft report that summarizes all the discussions of the TAP and the Steering Committee, all the recommendations that was forwarded to the public and Member comment period.

We send those comments back to the Steering Committee and to the TAP, if necessary, to see if there is anything that might impact the measure moving forward or any changes in recommendations.

We then put those back out for Member voting and then it goes to our Consensus Standards Approval Committee, which we call the CSAC, which is an oversight body that we have here at NQF that reviews the recommendations. It makes sure that the process that we use for project was followed and so forth.

And they will make recommendations or confirm the Committee's recommendations and then the Board will ratify that.

So this is just a pictorial of the process here. And, obviously, the technical advisory panels and work groups feed into that Steering Committee review process.

So just a brief kind of overview, we -- actually, this project has been going on

for two years now. So we started in 2009 working with the Steering Committee, of which both Dr. Weinstein and Dr. O'Neill were a part of, in really thinking through, you know, this was our first time evaluating resource use measures, how are we going to define them, how should we evaluate them, what are the important aspects of resource use measures that we should be aware of before we start evaluating them?

And this is a definition that we landed on for resource use measures, that they are broadly applicable measures that compare health services counts in terms of units or dollars. They can be applied to a population or event.

And those counts of frequency of defined health system resources, some may further apply a dollar amount, amount for charges, paid amounts and so forth for each unit of resource.

So keeping that in mind, I'll just

kind of go back a little bit about how this project is structured.

Again, because it was our first time doing -- reviewing resource use measures, we wanted to kind of break it up and not do it all at once. We ended up with about 36 measures to put through this process. And as you can see, they are huge measures.

And so we wanted to kind of focus on one condition area, if you will, which we selected the cardiovascular diabetes and non-condition-specific measures. So we have one TAP for cardiovascular and diabetes measures. And the Steering Committee reviewed the non-condition-specific measures.

So that Cycle 1 is still ongoing, but it's kind of a parallel process of this now. This bone/joint TAP is actually part of Cycle 2. And so we are expecting that the measures will go -- within the Cycle 2 will be through the process by the first quarter of 2012.

So this is just a brief time line of each of the steps for Cycle 1 and Cycle 2. And you guys can take a look at that. I'm not -- I won't spend any time on that this morning.

The review process that we set up for this project was essentially that, as the measures came in, staff would review them, make sure they were complete. We did a lot of work with the developers up front, although they are still not the perfect submissions, we did try to have conversations with them up front to make sure that they understood what we were asking for and that to the best of their ability, they were providing information that we were asking for before we pass it on to the TAP Members and the Steering Committee.

We simultaneously after that would send it to our statistical consultant, who prepared those summaries of the scientific acceptability for you. And he reviewed them and then, obviously, we passed that on to the

TAP for review.

And as mentioned before, your evaluations and issues that you identity with the measures will be passed onto the Steering Committee for their review and final recommendations for endorsement.

So in terms of the role of the TAP, we are looking for you to evaluate the measures against the sub-criteria, and we will talk a little bit more about what that is. But particularly to identify the strengths and weaknesses of the measures. And we are hoping, you know, that you guys are going to focus, obviously, on the scientific acceptability section where all of the clinical construction logic, the -- all that stuff where, obviously, your expertise is needed to kind of make sure that the episode construction, you know, not just the -- you know, your expectations of how clinical course should go.

And then that guidance, obviously,

is passed on to the Steering Committee. And the composition of the TAPs is very different than the composition of the Steering Committee. Obviously, we have seen the people on the TAPs with very specific clinical expertise that aligned with the type of measures that we received.

The Steering Committee is composed of a little bit broader expertise. Obviously, there are some physicians on the TAP or on the Steering Committee, as are seated here. But they tend to be more kind of maybe policy or higher -- a little bit further removed sometimes from the clinical level. So we wanted to make sure that we had the specific clinical expertise as well as the methodologists on the TAP to provide that specific expertise to the Committee.

So what we are going to do today is a very systematic review of the evaluation criteria. We will move through each of the criteria in order sequentially from importance

all the way down through feasibility.

And again, we will be looking at how well the information that the developer submitted meets the criteria that are outlined in the table that we will refer to. We will be asking you to rate the sub-criteria on a scale of high, medium or low or insufficient. And we will talk a little bit more about how the voting tool is used, but each of you should have a little black remote that we will be using to capture your votes and they will show up on that screen up there as we go through the day.

And we can decide along the way, but we can -- what we have been doing is kind of talking through all the sub-criteria for importance and then voting on each one and then go through scientific acceptability and then vote or sometimes if we vote on a couple from scientific acceptability, discuss, go back and vote, discuss.

So we can kind of see how that

goes, but, essentially, we will be voting along the way.

The ratings that you submitted on-line are really just preliminary ratings. We expect that when you get here and you hear your colleagues discuss some of the same things, that you may change -- want to change some of your ratings. So what we capture here are your final ratings that will be submitted to the Steering Committee.

So don't feel bad if you feel like you rated it one way and you want to change your rating; that's okay. We expected that will happen along the way.

So to just talk a little bit more about the sub-criteria. So again, we talked about how we will be rating those sequentially. And as you probably are already familiar with, we have four major criteria: importance to measure or report; scientific acceptability of measure properties; is the measure usable and is it feasible.

So for importance to measure or report, we are really talking about the focus area of the measure. So not whether or not the measure itself, the way it is constructed, is important, but is the topic area that they have chosen important? And is the information they submitted, does it support that it is important, that focus area is important?

What we found in the other committees and TAPs is that because this project is so focused and we chose the conditions, that everything is pretty much going to be important.

So what we are going to do is have Dr. Weinstein lead the group through that discussion to try to keep it as brief as possible. We expect that the discussion for scientific acceptability will be where the bulk of the, you know, discussion will be, so we don't -- we want to try to use our time wisely and not kind of belabor an issue that is going to end up being important anyway.

So, again, the scientific acceptability is where we address the reliability and validity of the measure. And the usability criteria looks at whether or not the measure and the results of the measure are usable for the intended audiences. We will talk a little bit more about that.

And then we will also -- the last criteria is feasibility. And that looks at whether or not there is any sufficient burden on implementing the measure for any measure users.

So importance to measure report, I'm not going to spend a lot of time on these. We will actually go through them as we are evaluating the measure. We can address any questions that you have there. But it looks at whether or not it's a high-impact area that they have selected; whether or not the purpose and objective of the measure is clear; and whether or not the resource units and service categories that they have selected to measure

make sense based on the focus area that they have chosen to measure.

Scientific acceptability, again, looks at reliability, whether or not the testing -- the information they submitted on testing the measure demonstrates that the measure can be implemented consistently across different systems or users; that it is valid and credible that you are actually measuring what you say you are measuring.

And then the last kind of dangling sub-criteria for scientific acceptability is about disparities and that has come up across all the TAPs and with the Steering Committee as well. And I think there will probably be a separate discussion here as well about that. And I think what we found so far is that they are important.

Disparities are important, but that there are some limitations with administrative data in capturing that a lot of times. And so I think the TAP -- for each TAP

and committee, they have just been weighing the importance of that based on the type of measure and the condition that it is and how well the developer has demonstrated the ability to do that with the measure as it is constructed.

Okay. So particularly with the reliability and validity, we had a task force that was done, I think, last year that looked at evaluating reliability and validity. And they came up with some guidance, particularly for TAPs and Steering Committee, so that the evaluation of the reliability and validity across these groups is consistent.

And so as you are rating these sub-criteria, we just kind of want to give you an idea of what a high would sound like, what a medium would sound like and what a low would sound like.

So for a high rating for reliability and validity, you would tend to think in your review that all the measure

specifications are unambiguous and likely to consistently identify who is included and excluded from the target population; that the resources -- and the resources and costs being measured and how to complete the score is clear and unambiguous; that the empirical evidence that they have submitted about the reliability and validity of data elements and with the measure score is consistent; and that they have -- the appropriate method and scope of reliability and the statistics are within acceptable norms.

For validity, much the same thing, that you will be -- that the measure specifications are consistent with the intent described and importance to measure. Again, that the evidence of the validity for data elements in the score are unambiguous.

So they are very much the same for the reliability and validity for the high score.

For a moderate score, for

reliability, you would think that all the measure specifications are unambiguous as noted and that the empirical evidence is within acceptable norms. So not quite perfect. Maybe some improvements, but it could be workable.

With the validity, a moderate rating, again, the measure specifications reflect the intent cited in importance to measure; that the empirical evidence of validity is within acceptable norms and that there has been a systematic assessment of face validity, which is the minimum threshold that we have four demonstrating validity for a measure score of the measure.

And that the scores obtained from the measure, as specified, will provide an accurate reflection of cost and resource use being used to distinguish high and low resource use.

For a low score, there is one or more specifications that are ambiguous with

the potential for confusion on identifying who is included and excluded from the target population or that the empirical evidence that they have submitted on reliability is not -- is unreliable or the data elements or measure score are outside the acceptable norms.

For validity, again, the measure specifications do not reflect the evidence or do not support the intent of the measure; that the empirical evidence is not -- did not use the appropriate method or scope.

So again, with the low rating, you are not so -- you are not comfortable that, as constructed, the measure would be able to be repeatable or valid.

And insufficient evidence, the way that other TAPs and Steering Committees have been rating it is that based on the information they have submitted, you don't feel like that you could come to a conclusion on any one of those. So maybe there is a, I don't know, statistical score or something

missing that you feel like you would need that in order to determine whether or not the measure was reliable or valid.

So briefly, this is, again, kind of going back to some of what was discussed with the Steering Committee last year. And we broke up the construction of the resource use measures into five modules to accommodate not only our submission form, which you kind of got an export of, which is what we sent you, an evaluation form, but so that we could kind of better breakup the evaluation of the resource use measure to ensure that everything that we needed to evaluate was there.

And what we ended up with was five modules:

One for data protocol, which can be submitted as guidelines or specifications, which looks at the beginning stuff like data cleaning or aggregating the data necessary to implement the measures.

The clinical logic, which is

obviously what we are going to be looking for you guys to focus on.

The construction logic, which looks at, you know, triggering mechanisms, how they eliminate redundancy and overlap.

Adjustments for comparability, which is where the risk adjustment and any stratification methods will be addressed.

And then the reporting guidelines, which is where the reporting module, which can be submitted as guidelines or specifications, which would be where they would address attribution rules, benchmarking, how peer groups are defined and so forth.

So I'm going to kind of skip through this a little bit. And we have been through these. I'm going to just kind of --

CHAIR WEINSTEIN: Great.

MS. WILBON: Okay. Sorry. We had a brief discussion yesterday about this particular criteria and what public reporting really means. And I think it was more so

around the public reporting, right?

So we did want to provide a little bit of additional guidance on that, because I think that's something that a lot of our TAPs and Steering Committee have been struggling with. And it's also something that NQF, as an organization, has been discussing internally and how best to define this and make it clearer.

So, Heidi, I'm going to be looking to you periodically for your clarification here.

These -- the ability criteria has three sub-criteria. The first one focuses on whether or not the results are reported to the public at-large and particularly for the ABMS measures, because they are new measures and they haven't been in use, this becomes a little bit more of an issue.

We don't require that measure developers that submit measures to the project that they have been in use, but we do ask them

to demonstrate or describe how it would be in use or what their plans are for getting it out there or how they intend it to be used.

In terms of identifying the public, we do define that as the public at-large. Correct, Heidi?

MS. BOSSLEY: Yes.

MS. WILBON: And particularly for, I think, other measure developers like Ingenix, for instance, where they have other entities using their measures, it's not always clear exactly how other people are using their measures or how it is being reported. So I think it comes down to, you know, weighing how the information that has been submitted by the developer and whether or not you feel that that is sufficient, based on your scores.

MR. AMIN: Actually, I would just add something to that.

MS. WILBON: Sure.

MR. AMIN: Considering that resource use measures are sort of new, keeping

this criteria in mind that we really want to have the measure be -- whether it is meaningful and understandable to the -- to an observer evaluating this, you know, the outcome of the score of the measure.

And the process, the NQF process will be, after three years when it goes under maintenance, this specific criteria, expected -- there will be an expectation that there would be more provided on how the actual measure has been used over the three years.

So keep it in mind that although it is clearly a very important criteria, this is the first time that we are going through resource use measures. So, you know, as a building block to measuring efficiency, we may -- you know, they may not have had the opportunity to have it be published at the -- for the public at-large.

MS. WILBON: Sure.

CHAIR WEINSTEIN: Can I comment just for a second?

MS. WILBON: Sure, yes.

CHAIR WEINSTEIN: I'm sorry, did you want to say something?

DR. SINNOTT: No, go ahead and I'll go next.

CHAIR WEINSTEIN: It's just I think these measures are fairly complicated, even for me, let alone the average provider, let alone the public, so, I mean, there is a huge bunch of steps that have to occur here to make these decision tools, which I consider these potentially for patients at some point, to understand cross-benefit resource utilization around their treatment options.

So I would hope that NQF will have a process by which this gets decoded into something that is understandable.

I go through this every day with-- not every day, every week with our physicians on new episodes of groupers and trying for them to understand most have no idea what it costs to deliver the care they are delivering

today, let alone what an episode is.

MS. WILBON: Yes. Thank you.

DR. SINNOTT: So all that just brings up the kind of basic question for me and that is are we talking about the grouping function or the physician's scoring function when we are evaluating this?

Because we seem to be moving back and forth between that terminology. And when you talk about score, I'm not sure what the score is from or for. Is it a score of physician performance, resource utilization or is it a score on something else about the episode grouping function?

MS. WILBON: So, Taroon, you can clarify, if you need to. The grouping function is a part of the construction of the measure and what would actually be reported out of the measure as -- for the public would be that score. So whether or not it is an O to E ratio for physician, you know, costs, so if it is that's like a one --

DR. SINNOTT: It's a score of a physician activity.

MS. WILBON: It depends on the measure. There -- depending on how the measure is, I was using that as an example. There are some measures that are specified for a level of analysis of physician. There are some that are at health level. There are some at the director regional level.

DR. SINNOTT: Right. Okay.

MS. WILBON: So whatever that level of analysis is, you -- for most of these measures, you will end up with a score. Maybe it's a ratio. In some of them, it may be a dollar amount. But whatever that end result is is what would be reported.

DR. SINNOTT: So when we are looking at validity and reliability, we are looking at the validity of the episode grouping function as well as the validity of the scoring function or not?

MS. WILBON: I would say both.

DR. SINNOTT: Okay.

MS. WILBON: Because the logic behind the grouping function is about how the measure is constructed and whether or not that is a valid approach is what we are asking you to evaluate.

DR. SINNOTT: But you are not asking us then to evaluate how the episode -- Dr. X has 45 episodes. And there are mechanisms in the resource use compilation into a doctor's bundle of activities that then that score to give a physician a score --

MS. WILBON: Yes.

DR. SINNOTT: -- so they are two different things. Number one, are the episodes valid in their construction?

MS. WILBON: Yes.

DR. SINNOTT: And number two, is the analysis that goes into the OE or whatever it is --

MS. WILBON: Yes.

DR. SINNOTT: -- appropriate?

MS. WILBON: Right. You will be looking at both of those.

DR. SINNOTT: Okay.

MS. WILBON: Yes.

DR. SINNOTT: I just wanted to clarify that.

MS. WILBON: Yes, thank you very much. Those are on the table now.

CHAIR WEINSTEIN: But you are going to get -- this methodology is going to get us into trouble when we get down to a doctor who has a small end from any kind of --

DR. SINNOTT: Oh, I understand. Believe me, I understand.

CHAIR WEINSTEIN: Yes.

DR. SINNOTT: I spent two years at PBGH trying to instruct, help, provide the information to doctors.

CHAIR WEINSTEIN: Yes, yes.

DR. RATLIFF: I'll bring up one point that I wanted to bring up with each of these measures. The costs that we are talking

about are not really costs. They are like healthcare costs. They are like how much money the hospital is spending. How much healthcare resources are being expended in this treatment.

We don't talk about loss of work or time out of work. We don't talk about other societal expenditures in these measures. So when you are talking about physician costs, it's direct healthcare expenditures or related to what resources acquisition is expending, it seems.

MS. O'NEILL: Well --

DR. RATLIFF: Because I have looked at --

MS. WILBON: The charges.

MS. O'NEILL: Not the charges, no. Unless it is -- unless I missed it and it was completely different from what we have already discussed. There -- every measure except one was based really on account of services that was translated into a standard price.

And so it isn't true costs. It's just --

CHAIR WEINSTEIN: Resources.

MS. O'NEILL: -- it's a resource use. So, I mean, the health partners try to put that forward for one of their measures to actually allow people to understand that if you went here, it would cost you this much, actual dollars. So we do have -- I personally have a concern that we are putting a standard price out and -- by usability criteria, that people will not be able to interpret what that means outside of people that do this kind of work.

But there is no, you know, time loss productivity. I mean, there -- none that --

DR. RATLIFF: Not the real capture of societal costs.

MS. O'NEILL: No.

DR. RATLIFF: Or societal expenditure in each of these measures. And

going to your point, we are using a cost basis. Like, essentially, standardizing costs. So you come up with a number that we can work with.

MS. O'NEILL: Well --

DR. RATLIFF: But then we are going to have based like a physician score on that that is going to be reported to the public, that everybody is going to see, so they can see how efficient their physician is.

I mean, I think going back to the point that you are moving around, like that to me is dangerous. And then it becomes very pejorative in terms of how these outcome measures may be used five years of now to grade "physician efficiency."

CHAIR WEINSTEIN: Let me just help out. I think we are all having trouble grappling with some of the methodology by which -- and then advancing ourselves to the point of somebody using this in some way to determine the efficacy or efficiency of a

system, at some point.

And I think what you know from reading this stuff and from the Ingenix work, they have actually used this in some organizations to try to help physicians understand their resource use compared to their colleagues for certain diagnostic categories.

And it seems to have been helpful in those cases. For example, I know the Sutter system has done some of that in California. The point -- and I think most of the people who do this work, this methodology is not that uncommon using the BETOS system from CMS and other methodologies for looking at this resource use.

The problem is most people are worried that just like outcome measures, you know, or any of the standard measures that CMS is putting in place, process measures, you know, which are probably the most accepted, did you get a hemoglobin A1c? Okay. I did

that. I got it.

But when you start to then look at my outcomes compared to my other colleagues' outcomes, well, my patients are always different. And these things try to adjust, as you know, for the various difference in patients. But getting the sort of clear populations that physicians and/or the public will understand is very complicated.

And we were talking before we started, you know, I have been working on this for a long time, as all of you have, this does not get simple that people are willing to accept. I think we have to accept that for today. Try to do the process of grading these measures.

As good or bad as the grades come up, we just say what we think. But I think this is a long way from acceptability at the physician/patient level, because the people who work on these things in finance and working on the groupers.

What Ingenix, you know, owned by United Healthcare, is doing now is a business strategy. They want to figure out how to manage cost as the Federal Government does around efficiency.

So this is an exercise to sort of move towards that. Let's not hide that. On the other hand, you know, let's point out some of the issues that we have, that's our job.

But let's try to get through the process today and point out the shortcomings that we have and then we will have done our job.

MS. WILBON: Thank you. So I think Taroon alluded to this, but I just want to kind of piggyback on what he said, again, the resource use we recognize that this resource -- you know, evaluating resource use measures is not the whole picture. That we are kind of framing this in the context that one day they will be, hopefully soon, working towards bundling them with efficiency measures

or trying to figure out -- I'm sorry, with quality measures to try to figure out how to get a better picture of value and efficiencies.

So we are looking at this as a step in a multi-step process, but in order to bundle them, we have to kind of make sure that this building block of that bundle is valid and reliable. So that's kind of -- that's our approach to this point. Realizing that it's not there yet, but it's a first step in a process.

And this is just a pictorial of the spectrum of accountability and transparency and kind of how public reporting fits along that spectrum.

And this is just a brief slide. NQF has done some work in the past around efficiency measurement. And they established some definitions of quality of care, cost of care and efficiency and they defined efficiency of care as a measure of cost of c

are associated with the specified level of quality of care.

And that the value of care, as a measure of specified stakeholders preference, we did assessments of a particular combination of quality and cost of care performance.

So that said, this is kind of in the realm of where we are going. We recognize that again, this will be in the context of quality at some point in the future.

Feasibility is one of those criteria. Hopefully that will go relatively quickly. For 4A and 4B, we have --

CHAIR WEINSTEIN: Could you go back to the slide you skipped?

MS. WILBON: Sure. Sure.

CHAIR WEINSTEIN: I mean, really what this -- this is other work done by NQF earlier on as well. There is kind of phases of care and I'm not sure all of our group for the TAP, this bone and joint one fit into everything so neatly.

But the fact of the matter is there is a population of patients that have some series of diagnostics or diseases. There is a process by which, you know, you want to understand that the patient actually knows what is wrong with them, so that valid.

That there is going to be an intervention where the patient has a choice, another methodology that would get into preferences here, which isn't included in much of this work, but is another effort that people are trying to get into, preference-based decisions around elective kind of things, not emergency things.

That there is some measure of value with quality or cost in some way that people find acceptable. And so part of the denominator issue that we are working on now is this real cost issue and how you measure that in a way that would be acceptable as part of the value equation. And really, that sort of gets to what we need to get to in

healthcare and we could argue that that's right or wrong.

But I think what we all would agree to is if we could understand what the value is, quality or cost, in using some specified measures, then we could understand how we are going to pay for things, based on that method.

And when you get into some of the conditions we are talking about today, they are pretty much preference kind of decisions. They are elective for the most part for a back operation. A hip fracture is not.

And there is not a problem with hip fractures because everybody has them fixed. There is not a lot of choice. You know, 96 percent of people have them fixed. And the 4 percent who don't is because they are too sick to go to the operating room.

But the 30 -- and there is a one year 30 percent mortality. So it's a problem. When you get into hips and back, those become

a lot different. So how do you actually understand the value of that?

If you understand the numerator/ denominator and if a patient was given good information and had a preference, you would probably get to this episode kind of thought process. And that's where we are sort of thinking big picture. We are just taking a piece of this in the denominator and trying to get to the cost piece now.

MS. WILBON: Thank you. So the feasibility criteria is the last criteria. There is four sub-criteria. The first two we found tend to go pretty quickly. Most of these measures, I think, from both the developers today, are focused on admin claims data, so both of which you could say admin claims data is routinely generated during care and that they are available electronically. So for the most part, those tend to go pretty quickly.

4C and 4D, obviously, will render

a little bit more discussion, but just a brief overview of those. So transitioning now into a little bit more operational things, I'm going to hand it over very shortly to Dr. Weinstein, so you guys can get started.

We will open it up for public comment briefly, before we get started. And then hopefully the measure developers will be on the phone. We will ask them to briefly introduce each measure before you start discussing them, to kind of get you in the mindset and kind of explain what the intent of the measure is and so forth.

They will also be available to respond to any questions that you have during your discussion of the measure.

And then once you, obviously, have heard what you need to from the developer, then the TAP will go into their evaluation of the measures.

So each of the TAP Members are assigned, I think, one or two measures for in

depth review and then we have actually broken up the criteria even more. So when we get to those criteria, we will just ask that you, you know, identify any issues that you did, maybe refer to some of the other evaluations that were submitted before here and kind of summarize and recap and identify any issues that you think should be addressed by the entire TAP for discussion.

Again, we will have Dr. Weinstein kind of lead us through a brief discussion of the importance and 4A and 4B, which should go relatively quickly. And those are the measure assignments.

So for the electronic voting, so, again, everyone will -- has a little remote. We will decide -- we will prompt you at which point we should be using them. But for most of the measures, you will be -- all of them you will be rating on a scale of high, medium or low or insufficient. High is 1, low is -- these are all yes/no, but high is 1, low --

moderate is 2 and low is 3 and insufficient is 4.

And we also have a handout in your folder that gives you a little bit more instructions on what to do if you want to rescind, if you mess up, and you want to send a different score.

And as you vote, they will show up in real-time on the screen, so you can kind of see the distribution of how people rated it.

And what we have done for the past meetings, and Dr. Weinstein can talk a little bit about how he would like to do this, if you get like all highs and one low, particularly if it's not quite in alignment with how the discussion has gone, that we will kind of ask, hopefully not to call anybody out, but just ask you to kind of explain for our notes and for the developer, so they can kind of have an idea of how the ratings have been -- are justified essentially.

After the meeting -- well, first,

let me just say we are expecting that we are going to get through all the measures today. So we are hoping and crossing our fingers anticipating that there won't need to be any follow-up necessarily with this particular group. But there may be, you know, an email or two with follow-up from the developers, if you need additional information.

But we are very hopeful that we will get done today. So other groups have gotten through, I think, up to like six measures in a day, so --

CHAIR WEINSTEIN: This isn't a challenge, is it?

MS. WILBON: No. Not a challenge. It is somewhat of a challenge. I'll say that.

So we are -- we have a lot of confidence in you that you will be able to get through all of these in one day. So --

CHAIR WEINSTEIN: We have got a couple surgeons here, we're going to get it done.

MS. WILBON: Yes, all right. All right. If needed, we will schedule any -- we will schedule a follow-up call or two, but we are hoping not to have to do that. There will probably be some emails after, but so you've got your work cut out for you, we realize, but I think you guys can do it.

So that's the end of the presentation. Do you guys have any questions?

CHAIR WEINSTEIN: Just make sure you bring us through everything in the right order.

MS. WILBON: Oh, absolutely. I do want to refer everyone to this table that is in your folder that I think everyone has that, at this point. We will, essentially, be following this for each of the measures, kind of sequentially in order for the sub-criteria on the left side of that column.

So that will be pretty much your primary guide for the day. So that said, I'm going to go ahead and hand it over to Dr.

Weinstein to get started on the first measure.

And can I just --

CHAIR WEINSTEIN: What's the first measure?

MS. WILBON: The first measure is 1586.

MS. BOSSLEY: Radiculopathy.

CHAIR WEINSTEIN: Oh, good. Okay.

MS. WILBON: And that's an ABMS-REF measure. Robin, are you there on the phone?

DR. MANHEIM: Larry Manheim. I'm here. Robin might be on the phone, but --

MS. WILBON: Oh, okay. Great. Great. Can we just -- I'm going to hand it over to Dr. Weinstein, but can we just have you start off with a brief introduction to the TAP for this measure?

DR. MANHEIM: Right. I'll be very brief.

MS. WILBON: Okay.

DR. MANHEIM: This measure

measures resource use and cost associated with the management of an episode of care for acute, subacute lumbar radiculopathy with or without lower back pain.

I would note there is another measure being considered, which is unspecified lower back pain measure. And, basically, this distinguishes from that in terms of severity, because the work groups we used thought there was a difference in severity that was important and required separate measures.

The episode for this is triggered by an initial ambulatory care visit for radiculopathy, which is defined by ICD-9 Codes and it lasts -- in other words, the episode lasts for three months following the initial visit, plus we pull in non-E&M costs related with diagnoses related to radiculopathy for 14 days prior to the trigger visit, because it was felt that there may be orders done before that were done over the phone before the patient came in.

So it's a three month period, plus the 14 days prior to the initial visit. And people who had a radiculopathy diagnosis in six months prior to the initial visit are excluded from the diagnosis. There is a bunch of other exclusion criteria.

The age groups are 18 to 64, that's the age group considered in the measure. And people -- I'll just note that people are assigned to a physician, based on them having -- they are assigned to only one physician if 70 percent of their E&M visits were -- at least 70 percent were to one physician.

Otherwise, all physicians with more than 30 percent of the E&M visits during the episode receive assignment, so you could have multiple assignments if two physicians had more than 30 visits -- 30 percent of the E&M visits.

If no physician had at least 30 percent of the E&M visits to them, then it's

not a sign to any physician.

The only other thing I would mention is we include chiropractic and physical therapy care in those providers and we adjust BETOS Codes accordingly to make sure they are included.

I'll stop there.

CHAIR WEINSTEIN: Yes, thanks. This is Jim Weinstein. I'm going to just get us started on this. And I'm going to take the prerogative of questioning the inclusion criteria right away.

When you use ICD-9 Codes and when you get in to your chiropractic and other things and you look at the actual use in some of your tables of the most commonly used treatments, you know, you wonder if the diagnoses are actually correct.

And I guess I have raised this a little before the meeting, but I think this undermines the whole process. And I only want to have it clarified for the group, because I

think it's an issue in this particular diagnosis where, from my own work and again, I mentioned my conflict, the Sport Trial.

We know that the surgery actually works for the right patients, better than non-surgery, although not all patients certainly need to have surgery. 30 percent of our patients didn't and are quite happy even though they didn't do as well.

But I feel like when you include chiropractic, physical therapy and all these large numbers using that ICD-9 Code without any physical findings, confirmatory MRR, et cetera, you are including way too many people in this diagnostic code, which then starts to undermine the validity of the model.

And so that's a very core issue for me before we even move forward into the model. And I applaud you on the excellent work. I know how hard this is, so I'm not criticizing you or anybody else personally.

But I am criticizing the inability

of a data system to actually group patients in large cohorts in this particular diagnosis when, in fact, you show that the payment is better for this diagnosis for others, who tend to use this, which is a problem with the system.

And I would be curious what my colleagues on the panel think before we go forward with answering that question.

Who wants to start? John? Then we'll go over to May Kay.

MS. O'NEILL: One of my concerns with this topic area for a venture as opposed to concerning with the measure itself may be, as you are saying, really the problem is that in my experience, my clinical experience in rehab, I take care of a lot of people that had, you know, complex scenarios two and a half years out from their presenting back pain.

So I saw all kinds of story lines, if you will. And the concept of what

radiculopathy is and is not is not at all clear. I mean, people, I mean even within my specialty, certainly within primary care, certainly within some of the other types of healthcare professionals, will call anything that has leg pain radiculopathy.

And those certainly aren't documented nerve root, mechanical nerve root impingement, which would lend itself to a mechanical decompression. And a lot of people got surgery that should never have gotten surgery, for example.

So it's just starting at the very first criteria of can you look at a group of ICD-9 codes from this cohort of providers and think that you are seeing the same diagnosis in the patients is hugely problematic, which is very different than whether you have bumped your enzymes when you have had an MI or you broke your hip.

So, yes, just from the get-go, we are challenged.

DR. RATLIFF: I think that's an extremely well-put point. How an orthopedic spine surgeon or a neurosurgical spine surgeon may apply a group of ICD-9 Codes to a patient with radiculopathy is probably pretty similar. But at issue to widen that, as you have a more heterogeneous group of practitioners diagnosing patients, you are probably going to have heterogeneous use of the codes. And then extrapolating to form this from say the market scan database as done here. I mean, that's introducing one potential source of bias right at the outset with how you are defining your patient population.

In working with insurers through our national organizations, we find that Aetna has one definition of radiculopathy. United may have a different definition of radiculopathy. Some want straight leg raise, some want sensory changes, some want motor deficits. It is, as you point out, a free-floating term.

But still, I think it's something that you've got to work with. And all that we really have to work with are these ICD-9 Codes and I think the way the measure developers have put these together is not unreasonable with the caveats that we have offered.

I think they have done about the best that they can with kind of an imperfect definition.

CHAIR WEINSTEIN: Please, yes.

DR. SINNOTT: I have just a question. Are you concerned that the two measures are separated and not a single measure?

CHAIR WEINSTEIN: In what?

DR. SINNOTT: The two ABMS episode definitions are separated as -- rather than a single measure?

CHAIR WEINSTEIN: You mean the back pain versus radiculopathy?

DR. SINNOTT: Right. Because of the in --

CHAIR WEINSTEIN: No. I'm --

DR. SINNOTT: -- what is it the garbage can we throw our papers into?

CHAIR WEINSTEIN: Well, back pain is more of the garbage can. I mean, I think that's the problem --

DR. SINNOTT: Yes.

CHAIR WEINSTEIN: -- with that one. The radiculopathy to an orthopedic and a neurosurgeon that is a surgical indication is very different than all the --

DR. SINNOTT: Of course.

CHAIR WEINSTEIN: -- people I think included in this claims-based look.

DR. SINNOTT: Right.

CHAIR WEINSTEIN: Because they are just taking these codes that are written down by people who make that diagnosis for whatever reason and I'm sorry to say that you do get a better payment if you use that diagnostic code versus another.

And so I just -- I'm not -- I

don't want to undermine the process.

DR. SINNOTT: Right.

CHAIR WEINSTEIN: But we have to recognize the limitations. Because when you get into the episode, and what the cost -- these numbers are fairly low for what you would reimburse for an episode if somebody actually went to a surgical case versus somebody who had, you know, radiculopathy that was not.

And in our study, they had to have, you know, all the definitions that you would expect from a surgeon. They had to be a surgical candidate. They had to have an MRI. They have to have radiculopathy of leg pain below the knee. It had to be present for more than six weeks, those symptoms.

And it's not possible in the database to do that. I mean, not easily possible. But so I'm not saying don't use this or let's throw it out. I'm saying that this is a big disclaimer that we need to

recognize.

DR. SINNOTT: Yes.

CHAIR WEINSTEIN: Please.

DR. RUBIN: Just for more clarification. So this measure, the intent is only to be used for people less than 65? I mean --

CHAIR WEINSTEIN: That's now included in criteria.

DR. RUBIN: That's included in the criteria?

CHAIR WEINSTEIN: Yes.

DR. RUBIN: So if this moved forward, it would not be, I'm not used to my word marketed, but applied to other populations, because --

CHAIR WEINSTEIN: Well, their database was for people less than 65.

DR. RUBIN: Right.

CHAIR WEINSTEIN: We have done the same thing with Medicare over 65. And it is the same trouble. People over-utilizing that

diagnostic code, because they don't know what else to write down.

DR. RUBIN: No, but if you are evaluating this measure and endorsing it, then taking it from this point on would only be specifically for those who would be not valid or we're not talking about proving this for any other population than what is being recommended.

CHAIR WEINSTEIN: Correct.

DR. RUBIN: Or what was -- well, my concern is that you have --

CHAIR WEINSTEIN: A MarketScan is only --

DR. RUBIN: Well, right, but my concern is that you have a tool now, a measure that is "approved" and what is done with this after this point in time and if it's applied to patient population --

CHAIR WEINSTEIN: Well, this measure, if it's approved, will be for the specific purposes by which it was developed

and for the specific criteria.

DR. RUBIN: Okay.

CHAIR WEINSTEIN: And they showed their table.

DR. RUBIN: No, no, I --

CHAIR WEINSTEIN: Yes, yes.

DR. RUBIN: -- respond.

MS. O'NEILL: Jim, I just wanted to make one more sort of statement about a categorical concern that I have in this diagnostic group compared to the other ones we have looked at. And that is if we were looking at commercial databases, commercial administrative databases, for example, within CIGNA, you know, we have 13 million lives. We could look at who had the ICD-9 Codes and we could look at what the utilization patterns are.

But since we are dealing with working age adults, one big cohort of people with this group of diagnoses are injured workers. And they would not have data in the

commercial database.

And even some of the exclusion criteria, the look-back on the exclusion criteria, if those diagnoses and service delivery were under a Workers Comp payment methodology, they would be invisible to the analysis. And this is the leading diagnostic category in Workers Comp.

So just in terms of an understanding of what slice of the population we are able to look at by these criteria, I think people should just be mindful that we are missing that whole group of people.

CHAIR WEINSTEIN: Absolutely. It's a whole other issue that, you know, I'm sitting with the IOM looking at Social Security Disability, it's another issue as well.

Any other comments from the panel?

DR. SINNOTT: Just that they may not be completely missing.

MS. O'NEILL: I know.

DR. SINNOTT: And that they might be partially missing. And, therefore, the resource use is very limited for that diagnosis. It looks very efficient, but your -- because they have gone. And the reclaim process from the insurer, back to the Workers Comp to get repaid for the -- has not occurred.

CHAIR WEINSTEIN: I mean, it gets to John's point about a lot of these costs that are outside of the episode.

MS. O'NEILL: Sure.

CHAIR WEINSTEIN: The total costs.

DR. RATLIFF: I'll have to bring up one more point, since you bring up the sports study. I think -- and again, I don't think this is a bad measure just go -- I think they put a lot of work into this. It's pretty reasonable.

The key with a randomized control target, your standard RCT is at control. Like here, with using this measure, you don't have

that. You don't have control over who is coming in. The same way I can't control who is coming into my office.

The patient comes in bringing all their comorbidities. They have just put out a cigarette as they are walking in the front door. You can't control for all that.

And I think capturing this data for all of its inconsistencies and with the issues we have brought up, it's still a starting point. It's kind of a step one towards understanding better how we are expanding this portion of healthcare resources.

CHAIR WEINSTEIN: Yes. I think the way to make this better though is in your -- using a string of codes. So, you know, if they had an MRI, you know, they have that information. A lot of them probably did. Although, you will see, I mean, when you look at these databases, a lot of them don't and they go to surgery still, even without an MRI,

which is hard to believe.

So it is complicated. I just want to make sure that there is a disclaimer in our report that talks about these limitations, that we have recognized them, because our colleagues and the public would not want us not to.

It is not a bad place to start. This is an important measure, which is our first question --

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: -- of high importance. It's a diagnosis that is costing a lot of money that is not doing very well in its outcomes and costs, so it's very important. But I want to make sure we understand the limitations, but not saying that we throw it out.

Any other comments? Okay. So we should go on.

MS. WILBON: Yes. So it sounds like -- and actually, I think a lot of that

will probably come up again when we get to the scientific acceptability sub-criteria.

CHAIR WEINSTEIN: Right.

MS. WILBON: So with those caveats on the table, we could probably move pretty quickly through the sub-criteria for importance.

CHAIR WEINSTEIN: Yes.

MS. WILBON: Which asks you to determine whether or not the measure focus is a high impact area, whether or not they have demonstrated that it is a high -- that there is high resource use or cost problems or variation within this focus area, whether or not the intent of the measure is clear and whether or not the resource use service category selected makes sense for this particular condition.

DR. RATLIFF: Can anyone on the panel for that, are we for voting on the resource use, kind of Step 1? Can we pass that as a consent calendar if kind of

everybody agrees or do you want individual votes for each?

Because I think everyone agrees, at least for these first few measures, this is pretty important to investigate. Yes?

MS. WILBON: Oh, okay.

DR. RATLIFF: Do you want to actually have -- what is actually -- do you want us to actually have to push the buttons?

MS. WILBON: Yes. Would you rate them? Would anyone rate them as high --

DR. RATLIFF: We're moving quickly to the end of --

CHAIR WEINSTEIN: We think there is consensus --

MS. WILBON: Okay.

CHAIR WEINSTEIN: -- with what John says, that this is a highly important measure.

MS. WILBON: Okay.

DR. RATLIFF: I mean, correct me if I'm wrong.

MS. WILBON: I think we can actually move through pretty quickly, if you hit the button, rather than -- just I realize that --

DR. RATLIFF: You want to have it for the record?

DR. SINNOTT: Yes.

MS. WILBON: Yes. It goes pretty quickly. So what we --

MS. FANTA: You have to point and it's fun.

CHAIR WEINSTEIN: Uniform data collection.

MS. WILBON: Yes. So Sarah has a computer with a sensor on it, so if you could just kind of point your remotes to her.

CHAIR WEINSTEIN: Where is Sarah? Sarah, Sarah?

MS. WILBON: So when that timer starts, you can go ahead and hit that.

CHAIR WEINSTEIN: Did you feel anything, Sarah? We are all pointing at you.

Okay.

MS. WILBON: So --

CHAIR WEINSTEIN: There is a next question.

MS. WILBON: Is whether or not the measure demonstrated considerable variation across providers of population.

CHAIR WEINSTEIN: Yes, this is an important question, because I'm not sure it does that effectively. Was the data submitted that demonstrated considerable variation?

MS. O'NEILL: So you are saying that --

CHAIR WEINSTEIN: That's a good question.

MS. O'NEILL: -- they showed less variation than some of us who do this work --

CHAIR WEINSTEIN: Yes, exactly.

MS. O'NEILL: On the front line?

CHAIR WEINSTEIN: Exactly.

DR. RATLIFF: Exactly.

CHAIR WEINSTEIN: Is this an

observed variation or is this what they are bringing to us? Do we see variation?

MS. O'NEILL: This is an area of huge variation, but they are called to find as much as I see.

CHAIR WEINSTEIN: Correct.

MS. O'NEILL: Correct?

CHAIR WEINSTEIN: Yes. So that's the issue. So I wouldn't want everybody to just say this is high again, just being cautionary, because I'm not sure that the measure did do that. So vote your conscience. Okay. You have an official vote.

MS. WILBON: Oh, yes.

CHAIR WEINSTEIN: Do you have another one for us?

MS. WILBON: So there was one high, four moderate and one low, for those on the phone.

And the next one that we will be evaluating is whether or not the intent of the measure was clearly described in the

submission. Right. So, Liz, sorry, can you-- we will send you an email to kind of delineate how you should submit your ratings throughout the process. Okay?

MS. PAXTON: Okay.

MS. WILBON: Sorry about that. Okay.

CHAIR WEINSTEIN: Did you -- are you -- can we vote now?

MS. WILBON: Yes, unless there was some -- any discussion about whether or not this --

CHAIR WEINSTEIN: I thought I did, but we didn't see the clock running.

MS. WILBON: Yes.

DR. RATLIFF: I can't vote more than one.

MS. WILBON: Right. So there is four high and two moderate for 1þ.

And 1(d) asks whether or not the resource use service categories that they identified for this particular measure,

basically, makes sense for this condition in the focus of the measure.

So there is three high and three moderate.

So that wraps up importance, which we thought would go relatively quickly. And we will move into scientific acceptability, which I think is going to be where the bulk of your discussion is.

And it looks like Dr. Ratliff was assigned 2(a)(1). So we will start with you, if you want to kind of summarize what you found.

DR. RATLIFF: So I didn't prepare any slides. I'm not sure how you want to go--

MS. WILBON: That's fine.

DR. RATLIFF: -- or move through this.

CHAIR WEINSTEIN: Thank God.

DR. RATLIFF: So we are going to go over leading off the 2(a)(1), which is scientific acceptability. The idea here

really is what I took from our definition or the definition offered by the NQF was whether or not this measure specified a patient population that you could generalize.

So not just the MarketScan data, but the measure is based upon whether somebody could be measured or spread out to all of the U.S. Healthcare System or any place where you have like an EHR.

And I guess we would open up the discussion with that. I mean, these are episode-based resource measures, but for this, again, I kind of looked first at the patient definition and maybe we should talk about that again, briefly, since we are already talking about radiculopathy and then talk about how they defined the episode itself.

If I look at the discussions that were emailed out on what other folks thought, I must have been in a good mood, because I thought they did a pretty good job of defining this measure and I seem to be the only one who

like forwarded responses back who thought so.

I should take that back. It looks like some of the like highs were recorded.

Craig, you had comments that you brought up in the email about the ages of the patients. And that was one of the few written comments that I saw that got emailed about. I don't know if you guys want to start with the discussion of the definition of radiculopathy or discussion of particular issues from the 2(a)(1).

DR. SINNOTT: I just had a question about, besides the definition, the truncation at 90 days. And why that was selected when these events, generally, are recurrent and prolonged. So I don't know if the developer is still on the call?

DR. MANHEIM: Yes.

DR. SINNOTT: You might talk about the truncation. I mean, I appreciate the subacute "ends" at 90 days, but --

DR. MANHEIM: Well, that was the

point was that it was subacute ends at 90 days. In fact, the distinction they made was whether to do six weeks or three months. And we presented them, the work group, some data for both. And it was decided three months.

But they were trying to not get into chronic and to have recurrent episodes. I'm sorry, that was separated more than six months as new episodes, so that was the rationale.

CHAIR WEINSTEIN: Could you ask that question again? Could you ask your question again? I'm sorry.

DR. SINNOTT: Sure. My question was why truncate the episode at 90 days. I'm just repeating the question.

CHAIR WEINSTEIN: Yes.

MS. SINNOTT: And I mean, I recognize that, in general, the literature says subacute ends at 90 days and chronic starts at 90 days. So I was concerned that we are losing some of the recurrence in that

particular.

So let's say somebody has treatment for six weeks and then stops treatment, which is all you can tell from the administrative data. And then at 89 days starts back again. It's, essentially, the same episode of six months is your definition of absence of care.

But then that it ends up being neither a second episode nor a prolonged first episode.

CHAIR WEINSTEIN: I mean, John, what do you think in the sense of, you know, a patient who you watch, you end up operating on at 12 to 25 weeks, they don't get back within, you know --

MS. SINNOTT: Right.

CHAIR WEINSTEIN: -- another 12 weeks, potentially, you know, so is that episode too short in the context of the ideal patient even? Is what I'm asking. And is the idea of the measure, and I'll direct this to

the developer, to try to capture the initial presentation of radiculopathy?

DR. RATLIFF: Through treatment.

DR. MANHEIM: Well, that's right. A part of it is to capture the variation, so is there unnecessarily high variation? So if -- I think part of the reasoning is there may be a surgery down the line, but that's not part of the initial presentation. In fact, if it's in the code within the first three months, and I shouldn't use the word, but, maybe that's more appropriate surgery than if you had a surgery within those three months before to watch the patient some more.

I think that's the rationale for eventually cutting it off, that you are looking at variation. You know, it's not a research study you are looking at for disease, but given the data limitations, and the, you know, amount of time you can actually observe them, basically, like use two years of administrative data, it's felt that that was

the best time limit.

DR. RATLIFF: I guess I would open up following up on the other points made. Is that really modeling our patient or is that the patient that is coming in -- is that the patient that I'm ending up doing surgery on? Is that representing and capturing the overall group of patients or are you already restricting it down to such a small subset that you may not be able to generalize your outcomes from that?

DR. MANHEIM: Can I just state one more thing and I'll stop, which relates to that of an earlier question is we might not have made it clear enough, but the intention is always that any physician will be compared to its peer group.

Chiropractors would only be compared to chiropractic and so on. It might not really solve your problem, but there is no intention of chiropractors being compared to surgeons in terms of the patients they see.

DR. RATLIFF: Now, wait a minute. You give a table at the end of your little presentation where my specialty is the most expensive in the entire group. So that kind of invites comparison. You may say you are not comparing, but it's certainly there.

MS. O'NEILL: Well, we are comparing them.

DR. MANHEIM: I would call it validation, rather than comparison. It would be surprising that we are not --

CHAIR WEINSTEIN: You know, I think the problem is you are still comparing the apples and oranges, which gets to these ICD-9 Codes and that you can't get away from that. But the reality is if it's true radiculopathy, surgical or non-surgical, that a lot of these patients can get better on their own anyhow.

And it might take more than 90 days for them to go through a -- we know from the sport data again that over time, these

patients can get to a point where they can function. 30 percent of them never had surgery, even now, eight years, nine years later.

The episode can't go that long, but I'm not sure that 90 days is enough. When you are going to -- realize what we are saying here, at least in my opinion. And Heidi should correct me or somebody should, because what you are getting to is a public reporting, us supporting a policy that potentially will stop payment for this episode after 90 days.

If -- or at least it is going to be bundled potentially by somebody.

MS. O'NEILL: Well, I mean, I guess, part of the problem when I was wading into this, on a number of the different measures, is to think about what it is we are measuring. And so are we measuring resource utilization as it tracks the natural history of these back pain cases following the patient over time?

You know, what is the natural history of back pain? And one of the problems we are going to have with any measure in any time frame is that has less -- there is a less standard story --

CHAIR WEINSTEIN: Well, let's stick to radiculopathy though, first.

MS. O'NEILL: But if we are comparing with the measure, the performance of whoever is delivering care, the individual or the system, over the first 90 days of the onset, then that's different.

Now, I know that there is a -- you know, you could have some kind of thing develop in the future where people are being paid on this basis and everybody would get surgery on the 91st day and they would look real cheap in the first 90. I mean, any measure can be gamed, right?

But I think it is a little hard if we are looking at these measures as, essentially, payment policies, which is, I

think, where we go the push back on one of the health partners things earlier.

It was like oh, no, you are going to tell me I can't do this, but our -- I don't see that that is what we are doing with these measures.

CHAIR WEINSTEIN: Yes, but, you know, even in your own company, there are risk contracts now for, you know, managing certain populations for a specific bundle of payment. And all I'm suggesting is we have to be careful.

You know, if we think 90 days is right, then great. Let's agree that that's the group, the episode and say that's okay. And maybe that is okay for this sub-population that they have studied in this database.

The problem is with the -- which I go back to the original question, they are mixing a lot of patients in this that where 90 days, on average, looks okay.

MS. O'NEILL: No.

DR. RATLIFF: I would ask the developer that as well. Did you model this from like your data and look, are most of these patients finishing their treatment before 90 days? Is that why you chose the 90 day cutoff? Is the developer --

DR. MANHEIM: Ninety days was chosen based on the work group and some subacute -- there is never any notion that this would be used to come up with a payment scheme. The noting was going to be used -- it would be used for quality measures to come up with comparing physicians in terms of their cost.

CHAIR WEINSTEIN: No, but his question is did you model this with the data showing some set of patients were done with their episode within 90 days? Yes or no?

DR. MANHEIM: We took the 90 day and we didn't go -- we looked -- compared --

CHAIR WEINSTEIN: Was that arbitrary or did you actually base it on --

DR. MANHEIM: It was arbitrary based on --

CHAIR WEINSTEIN: I think it's arbitrary.

DR. MANHEIM: -- physicians that--

(Simultaneous speakers.)

DR. RATLIFF: It kind of arbitrarily pulls out 90 days, too. So there is foundation, but it's following our -- would we suggest a different time course? Would we suggest a longer period, a shorter period? What would be consensus of the panel?

CHAIR WEINSTEIN: My sense is you would want to validate that against sub-populations within this diagnostic group.

DR. RATLIFF: Other comments on that with regards to the 90 days?

DR. RUBIN: I just think it is a dirty, it's a messy clinical problem. And, you know, if it's reasonable, I mean, with all the limitations, I think somebody could come out and say 120 days and there will be

problems with that.

But to sort of try to grapple with this in a measured way, in a -- I think they have done reasonably well in, at least, trying to characterize the problem and trying to assess it. I think there will be criticisms, you know, whatever number you choose. Certainly, more information would be helpful, more evidence.

CHAIR WEINSTEIN: My only comments for you, John, are I thought there were a lot of pros to this. I like their use of the claims. I like their standard pricing list for costing. I thought they had good detail on how to standardize cost in patient, outpatient and pharmacy.

I thought that they used the categories of services, the BETOS thing, which is easy for people to get to from CMS. I thought the methodology was easy to follow, in some ways easier than the Ingenix were going to come up with.

Their trigger events were very clearly defined. I think they excluded patients without pharmacy benefits, which was good. Their winsorization methodology was good. They excluded patients, you know, with the kinds of diseases that would confound this significantly.

I thought their group cost is related to the diagnosis and unrelated to the diagnosis were good. And I think their risk adjustments were good. I thought some of the cons of what they did was their coding. It assumes coding is consistent across facilities, which generally it isn't.

Time limits on episodes, I said may be artificial. There is no mention of software automation of this process, so I'm not sure, but we'll get into usability. And it does not address specific resource utilization within a procedure or E&M visit, so the type of provider is not addressed to me in the model specifically.

And it's not to address non-billable activity. So those are the pros and cons from my perspective.

DR. RATLIFF: I'm picking up on one of those points and kind of one of the talking points that I wanted to bring up.

The ICD-9 and CPT Codes that are included on page 12 of the PDF that they sent out for 1586, I'm okay with those from the panel. I wanted to make sure the other panels thought that that was an inclusive list and that we think that we are capturing the data that we want to capture with regards to treatment of radiculopathy.

With the caveat that we discussed earlier, there is not going to be a standardized use of those codes in between different practitioners and possibly even between the same practitioners in different institutions.

But I think that if you get into that with any kind of population or database

approach to assessment, is everybody okay with the CPTs that they chose?

CHAIR WEINSTEIN: Those top 20?

DR. RATLIFF: Yes.

CHAIR WEINSTEIN: The thing that that pointed out to me, again, was that these subcategories of patients are probably pretty different, because the top two by far, you know, 14 and 7 percent of those, so 22 percent of those are therapeutic exercise and manual therapy.

That's almost routinely used. The efficacy of that is questionable in this kind of diagnosis, but that is a huge expense that, to me is questionable efficacy, but, obviously, you know, only 6 percent had surgery, 6.8 percent as a code.

And so, to me, again, these ICD-9s, if you do some sub-groupings or different method of breaking these down, you would probably get more specificity. But it just pointed to me again that these are different

populations across this database potentially.

MS. O'NEILL: On the list of the common non-related diagnoses and procedures, there are columns that are entitled "Related and Non-Related," so that in certain -- do those columns indicate that the related costs were grouped to the episode and the not-related ones were not grouped, so that these non-related E&M Codes, occasionally, are related? Do you know what I'm saying?

DR. RATLIFF: Are you directing that to the developer?

MS. O'NEILL: Yes.

DR. MANHEIM: If I understand right, I hope I'm stating this correctly, but, what we looked at was for cases where there were a related diagnosis in terms of having the correct diagnostic categories to include them versus those where those codes came out, they had a different diagnostic category.

DR. RATLIFF: Does that answer your question?

MS. O'NEILL: Well, I mean, there are some small numbers here that confuse me. I don't want to get off the main point of it, but, I mean, some of the non-related E&M Codes and procedures are things that, you know, I think people might use in this patient category.

So just for example on that ICD-9 list, there is a pain in the limb and there is 279 of these are under the related column and 4,000 are on the non-related column.

DR. MANHEIM: Yes.

MS. O'NEILL: And is that because that diagnosis occurred by somebody who provided care to this patient and it turned out that they did not have a diagnosis of radiculopathy? I'm just trying to --

DR. MANHEIM: Right. That's right.

MS. O'NEILL: Okay.

DR. MANHEIM: Within the episode, those were cases where that was the CPT Code,

but because they did not have related diagnosis, it was not included as part of the cost.

MS. O'NEILL: Okay. But that was --

DR. RATLIFF: You lost me there.

MS. O'NEILL: Those in the related column are included?

DR. MANHEIM: Yes.

CHAIR WEINSTEIN: So the way I'm understanding this is that when they did their -- your algorithm for inclusion of patients, you went through these different coding exercises. And when you found out that they had -- they didn't have a back pain code, but they had a leg pain, you know, it wasn't related, because of the coding, it wasn't.

MS. O'NEILL: Right.

CHAIR WEINSTEIN: Yes.

DR. MANHEIM: Right. So what we did is, you know, we had a number of meetings, mostly by telephone and we presented tables

saying well, here is the CPT Codes and they are or they aren't included, based on criteria. Does this look okay to you? Should we be including something else, via expanding diagnosis codes or including this regardless of the diagnosis code, et cetera?

So they would look at this and scratch their heads and talk about it and decide whether it needed to change, which we already had.

CHAIR WEINSTEIN: I think it was their grouping methodology that, you know, right or wrong, that's how they made their rules. Yes. Any other questions? John, do you --

DR. RATLIFF: Slowly advancing.

CHAIR WEINSTEIN: All right.

DR. RATLIFF: Can we discuss age and the fact that you said the cutoff was 64, because Craig did bring up a good point? And what we emailed around, I would like him to voice here for the minutes, just with regards

to the MarketScan data versus general population data.

DR. RUBIN: Yes. My concern, I think, are major limitation, even though they clearly state that it will include the age of 64, although there are some errors in some of the paperwork provided. And the reason for excluding people over 65, I don't think there was sufficient explanation.

DR. RATLIFF: It was your database, right? That's what you had access to?

DR. MANHEIM: Yes. In fact, through, I think it is probably my error, sometimes 84 mixed in -- which is in the original work group, 84 was mentioned, but there was a question about how people 65 to 84 differed. And given that the only data we had was through 64, we felt we could not go beyond that.

DR. RUBIN: Well, right. Well, so it seemed to be a convenience issue. And I

think that this is a non-reason to state that people over 65 would be treated differently.

The point of these measures, from my perspective as a clinician, is to try to identify variations, so we can identify better outcomes, identify poor outcomes to try to develop interventions to reduce poor outcomes.

And if your -- I realize that this is, again, you know, fine to be limited to less than 65, but from a national basis, we have this huge population of people. And we don't know -- we need to assume that just because they are going to be treated differently, I mean, you can say the same thing for any age group. It doesn't seem to be a scientifically valid or clinically valid approach.

And I just want to say it's a very shortcoming of the tool and would have been an opportunity, unfortunately I think, to look at this age group to measure important comorbidities and to identify either regions

or practitioners who performed better in terms of -- and this is a lot of the issues and we repeat this, but if the surgery is involved, you know, wound infections, this kind of surgery, pulmo-emboli, very valid comorbidities. That is applicable for all age groups, but particularly in this group over 65.

So I guess I would encourage that the developers would include this group and not, you know, sort of refrain from measuring and assessing this group.

CHAIR WEINSTEIN: Just one clarification. Epidemiologically, this is a diagnosis that mostly occurs between 33 and 55. It doesn't mean it doesn't occur in over 65. It does. And it is often diagnosed and it's another problem with ICD-9 coding or whatever, but you are exactly right.

But the reality is from an epidemiology standpoint, this is not a common diagnosis in people 65 and older.

MS. O'NEILL: But --

CHAIR WEINSTEIN: For which there is good studies that suggest it is treated well by surgical intervention or any other method. So that's all I'm saying.

MS. O'NEILL: But I would say that you could certainly call it out on your criteria from a scientific perspective that exclusion is not serving the greater good. However, on a feasibility criteria, when we get to that part of the measure, the fact that it is so expensive for most people to get access to the Medicare Database, it is untenable.

And so I think that when they limit their analysis to the data that they have available to analyze, then they have to give those metrics, because that is the limitation of their database. And Medicare has not made that easy for anybody who is trying to understand that.

DR. RATLIFF: We are getting ahead

of ourselves, but it's a good point. Have you validated this measure in something besides the MarketScan Database or have the developers looked at this outside of MarketScan?

DR. MANHEIM: No. There is some-- no, not to this point. So, you know, it's just no.

DR. RATLIFF: And moving ahead through my submission items that I was assigned to discuss, we are on page 14 of like 8,000 in your like PDF, so I'll try to move us forward.

Is everyone okay with the trigger visit or the idea of a trigger visit for the episode or what they choose as a trigger for bringing in their episode? Patricia?

DR. SINNOTT: Now, this is a two-part comment. Number one, am I right that you are attributing episodes to both physical therapists and chiropractors as well as physicians?

DR. MANHEIM: Yes. I mean, not to

be attributed to all three.

DR. SINNOTT: Correct. Okay. Just very much a side note, in the PT Codes that you include for identification of the provider visits, you don't include the PT Evaluation Codes 970001 and 2, even though they show up as high utilization codes in your report of utilization. So they just need -- if you are going to include them, they should be correct.

DR. MANHEIM: Okay.

DR. RATLIFF: Any other issues with the trigger? Hearing none, very good. Do we want to talk about relative risk and comorbidities modeling? Should that come up in this portion? I mean, obviously, we need to discuss it. I'm open to the panel's input.

CHAIR WEINSTEIN: I'm not sure how those were adjusted for in the model or whether they did or not. I can't remember. I'm trying to find my notes on that. Does anybody know?

To the developer, did you guys adjust --

DR. MANHEIM: Yes. The way they were adjusted was well, the final model chosen and provided Medicare instead of a Medicare-developed comorbidities were entered and those that were present more than 1 percent of the time and that were -- had a significance of P = .1, at least, were included in the model, controls the dose when comparing across physicians.

So it's a regression model that was used.

DR. RATLIFF: So I bring up as a point, and again, I like this measure, the risk adjustment model issue provided in your slides is various -- seems to go over pretty cleanly how you approach this data. But then it should go through your risk adjustment methodology in the PDF that you forwarded where you go through a lot more detail.

I mean, I get a little lost going

through this and I think even your statistician got a little bit lost when they reviewed this in terms of how you chose statistical significance for each model. We could bring up the point that you are using Medicare HCCs in a non-Medicare patient population, people that are under the age of 64.

I mean, this to me is certainly not intuitive. And even after reading it a few times and trying to study it, I'm not sure I fully comprehend how you are doing your relative risk modeling for this patient population, which, of course, is important from a surgeon's perspective, maybe not so much for chiropractic care, physical therapy or other aspects of this measure.

CHAIR WEINSTEIN: And the statistician had some comments about that as well, I was just trying to pull those up, who also felt that some of these things weren't managed well or, you know, I don't know if

that's the right word, managed, also forgive me.

DR. MANHEIM: Well, yes. Significance was used, I know he criticized that. And we did also look at predictability. The slides we used weren't included here. In terms of whether the predicted values varied, would be simply relative to the actual values.

The other thing we looked at, we used a large number of models to -- basically, it generally ended up in this measure using one where all the conditions were considered and then pared back based on what was significant or was not significant.

CHAIR WEINSTEIN: You used like 12 models or something, but it wasn't clear how you decided on which one, you know?

DR. MANHEIM: Right. It was stated --

CHAIR WEINSTEIN: It was a little bit of a fishing expedition.

DR. MANHEIM: That's right. I

would not -- and it wasn't made clear, but I think you're right about how it was dosed. Basically, it looked at how the -- they said the predicted value compared to the actual value in terms of maintaining the variability across physicians and not eliminating the variability across physicians.

CHAIR WEINSTEIN: Yes. I think for --

DR. MANHEIM: But I -- yes, and I actually did speak to the person who did it out here, so --

CHAIR WEINSTEIN: Yes.

DR. MANHEIM: -- I wouldn't want to say more.

CHAIR WEINSTEIN: I think just for the panel's sake though, John, it's important that we bring this out that there are these limitations and that's all.

DR. RATLIFF: I think it needs to be somewhere in the minutes with regards to the product of our panel that after they

caught their fish, I don't see where they compared it to other fish to make sure it was actually a fish.

Like whether or not this was actually validated through looking at different databases, validated through looking at it, I assume other approaches to modeling, which is essentially a medical condition, being low back pain with radiculopathy.

DR. LEE: So this is Todd Lee from ABMS. Actually, I'll jump in here. I did the risk adjustment modeling and I can speak to some of the questions that you all have raised.

We went through a process, and I apologize for sort of the lack of clarity in the submission, in which our work group identified conditions that they felt would be important in modifying costs for this patient population.

And then we also compared that to models where we included all other health care

conditions that were identified with the HCCs. Now, we don't use the HCCs that Medicare -- the coefficient ways that Medicare developed.

We use them only to identify the chronic conditions and then we estimate the relative cost of each of these chronic conditions through out modeling exercise.

We did this in a split sample approach. So we took 75 percent of the sample from the Med-Stat data and developed a model, tested the model fit in a 25 percent validation group. And what we ended up selecting was the model that fit the data the best out of all these 12 different specifications that we originally investigated.

Now, yes, it is, as you described it, a bit of a fishing expedition. We are trying to understand or sort of get rid of variability due to patient case mix, but we want to keep variability that is attributable to the episode and not completely wash away

all the variability that exists.

So we try and account for differences in case mix across these populations and we select the model that has the best performance. And we didn't provide all of the fixed statistics that the statistician would have liked to have seen and we have done -- we have subsequently done that for some of our other models or some of our other measures that have been evaluated. And we could certainly do it for this measure as well.

DR. RATLIFF: Thank you. Gently moving the discussion along, the costing method is something that is also assigned in this initial measure. Any comments from the Committee, comment from our group with regards to how they did their cost calculations? And I'm specifically looking at page 23 of the PDF that they have forwarded where they go through the standard cost calculation and then how they do standard units of service and standard

costs.

Would the developers like to comment on how they approached, just briefly, developing standardized units of cost for the therapeutic interventions we are discussing?

DR. MANHEIM: Well, I would just say that we -- the data we had from Med-Stat, we took the average cost for each code, for each outpatient code. And for inpatient codes, we took the average cost for each DRG and we -- but we did it on a per diem basis.

And then we discussed those few cases where there wasn't a DRG, what we did, which is somewhat complicated for a small portion of cases.

But basically, we took the average cost within a specific category, specific CPT or DRG level. And the average cost -- I should say average cost, obviously, we don't know the specific economic cost in abstract terms, so the average payment, the average amount, the payment that was designated to be

received by the provider, that includes the payment from the patient and the insurers.

DR. RATLIFF: Well, how do you do the observe versus expected ratioing for these costs as you go into your provider scoring?

DR. MANHEIM: Right. Do you want to address that, Todd?

DR. LEE: You bet. So each individual we look at the expected costs based on their case mix form our regression model, so we calculate an expected radiculopathy-associated cost for each person. We compare that to the observed cost and across each physician that it would attribute the care, we calculate from summary statistics of the observed to expected.

The average, the median for their entire population to which the care is attributed to that provider. And then we can compare observed to expected across peer groups.

DR. RATLIFF: Any comments on

that? Because I know the statistician brought up the point that this isn't an episode-based comparison, but something taking a step away from that that may kind of confound how you are going to compare between groups.

DR. MANHEIM: Yes, I think the comment was that -- that I saw was that it was not an average physician-base, but was for each episode. I didn't really understand that, but --

DR. RATLIFF: Oh.

DR. MANHEIM: So I can't respond to it.

DR. RATLIFF: I have a couple other points that I wanted to bring up, again, not validated. We haven't looked at your exclusions and validated them through using something besides the MarketScan Database. I'm afraid I'm bringing up stuff that we have already discussed earlier. And your risk adjustment methodology, you haven't explored outside of the MarketScan Database.

DR. MANHEIM: No. Most of the exclusions we have are standard exclusions are -- were based on NCQA. But we personally haven't used other data.

DR. RATLIFF: Well, that, for me, gets through 2(a)(1). I don't know if anyone else has other issues they want to bring up before we go to 2(a)(2) where we talk about reliability testing. We are kind of moving around a lot.

CHAIR WEINSTEIN: Yes. Are there any other comments about scientific acceptability? I think we have hit most of the points that I wanted to bring up and a lot of issues that I wanted to have kind of noted. I'm comfortable with moving ahead to other aspects of acceptability or even to usability.

DR. RATLIFF: We have sort of gone through all three at once.

MS. WILBON: So I think --

CHAIR WEINSTEIN: We don't rule out anything here.

MS. WILBON: -- we have actually covered a lot of it in kind of going through the specifications to see if they were clear or not. We have actually hit a lot of the other sub-criteria. So what I would propose is that we go through each and bring them up on the voting screen and read them aloud and just make sure if any -- yes, make sure everyone has covered everything.

And if there is anything else to discuss, when we get to it, we can just have that discussion, vote and then move on.

So we will start with 2(a)(1), which asks whether or not you feel that the specifications they provided were clear, such that, you know, any organization could pick it up and implement it consistently.

DR. RATLIFF: That would be also just for the methodology that you can generalize this.

MS. WILBON: Right.

DR. RATLIFF: It's not just good

for MarketScan, but you can take this to NIS. You can take this to the Medicare Database. This is going to be translatable to a larger patient population.

MS. WILBON: Well, this particular criteria is more so whether or not it can be implemented for comparability across organizations. So are the specifications clear enough, such that it would be consistent?

CHAIR WEINSTEIN: It doesn't get into the validation issue.

MS. WILBON: Right.

CHAIR WEINSTEIN: I don't think.

MS. WILBON: Validity comes up --

CHAIR WEINSTEIN: A later section.

MS. WILBON: Yes, later on. But I think we did talk a little bit about that, so we can --

CHAIR WEINSTEIN: But it's not this question.

MS. WILBON: Right, not this

specific question.

DR. RATLIFF: So it's not validation of the measure. I misspoke, but that you could use this measure in partners in like Medicare. It is generalizable. You can extract it.

CHAIR WEINSTEIN: To totally understand this and follow it, I think is -- yes.

MS. O'NEILL: But you -- we're just saying that this --

CHAIR WEINSTEIN: Microphone.

MS. WILBON: Mike. Use your microphone.

MS. O'NEILL: Oh, I'm sorry. I'm used to being loud. So but this is really saying that you could take -- based on a commercial administrative data set with these criteria, this rule could be applied at some delivery system in Seattle, in some delivery system in LA and that you would be, essentially, measuring the same things in the

different delivery systems.

That's what I understand this to be.

MS. TURBYVILLE: It gets to that this sub-criteria is specifically focusing on whether the specifications are written in a manner that someone could then take it and apply it consistently when we start talking about the data systems at a place would support it, that gets more into the validity.

This is really as written, was it clear, were the diagnostic codes provided? Could a programmer program this measure and implement it?

CHAIR WEINSTEIN: Yes, to me, this is easier to follow than some of the Ingenix stuff actually. So it gets to this easy -- could somebody follow this? Whether it is right or wrong, inclusive or not inclusive, valid or not valid, isn't the question. Is it laid out in a way that you can understand it and try to do it?

That's the way I'm answering this question.

MS. WILBON: And that's correct.

CHAIR WEINSTEIN: Okay.

MS. WILBON: And that's correct, yes. So does everyone feel ready to rate it based on Dr. Weinstein's -- okay. So let's --

DR. SINNOTT: So what happened to the previous counts?

MS. WILBON: Yes, we -- I think we started talking, so we will redo it.

DR. SINNOTT: That's fine.

MS. WILBON: Okay.

CHAIR WEINSTEIN: Should we go into -- there are some issues here, you know.

MR. AMIN: That was two high and four moderate.

MS. WILBON: Okay.

CHAIR WEINSTEIN: Yes. What's the next question? Could we just see the next question? Because I think somehow if we know the question, we can have a discussion that

may be very focused.

MS. WILBON: Right.

CHAIR WEINSTEIN: So this question is does the reliability testing -- and I'm not sure they did reliability testing.

Does the group want to -- does the creator want to say something about that? Did you guys do any reliability testing?

DR. MANHEIM: Not the extent of did not have an independent programmer try and program it. They got the same results as us.

CHAIR WEINSTEIN: Yes. So I don't -- so how do we -- they didn't do it.

MS. WILBON: Insufficient.

CHAIR WEINSTEIN: Insufficient. Okay. Can we vote now or do you want to have more discussion?

MS. TURBYVILLE: So, yes, just a couple of things to think about reliability before you vote. There is in some cases some of the TAPs have presumed, at minimum, a date element reliability, because it is a

commercial database.

CHAIR WEINSTEIN: Oh, but this isn't --

MS. TURBYVILLE: And that is --

CHAIR WEINSTEIN: -- though.

MS. TURBYVILLE: This is a -- it was tested on a commercial database and it's administrative data, which typically goes through when you are talking about the data element --

CHAIR WEINSTEIN: Yes.

MS. TURBYVILLE: -- certain checks prior to it being in the database, so they have considered that. And then also, I believe, and correct me if I'm wrong, with the measure developer with all the ABMS-REF measures, they -- in their reiterative process in reviewing it with the work groups, because of the complex programming, they were using that as a proxy to demonstrate reliability.

How you rate that, you know. So insufficient would indicate that we would, I

presume, and correct me if I'm wrong, Ashlie, because this gets into NQF process, would we ask them to submit something? How would we in this context handle an insufficient on this versus a low or moderate?

MS. WILBON: Well, at this point, the Committee does have to -- or the TAP does have to rate the measure as it is. So even if they were to submit additional information, if you wanted to see that and then we could go back and you guys could rate it later, based on what they submitted, that's an option.

But today, you have to evaluate what you see in front of you as is.

Taking what Sally said into consideration, beyond the data element, I'm just looking at Carlos' analysis. He didn't find any other reliability testing that had been done.

CHAIR WEINSTEIN: That's where I was going.

MS. WILBON: So I did want to get

some guidance from Heidi on whether or not -- how we distinguish -- how we would distinguish between a low and insufficient if nothing was submitted versus it not being sufficient.

MS. BOSSLEY: Right. I mean, they have submitted something. So I think I would probably not do insufficient or make it more-- or you would have to really provide that explanation.

CHAIR WEINSTEIN: Can I say that it's different based on what Sally said? Maybe this will help us. Maybe this will help based on what Sally said and what I heard you say and you guys, I get the sense, don't want us to say insufficient, right or wrong.

But the issue is they didn't do reliability testing. I just want to be clear. What Sally said was that given the database they used and the coding they used and the process they went through to do this, it was a reliable process is what I heard you say, Sally. Don't let me say what -- this is what

I heard.

And so that, you know, because they did some windsoring and they did some other things that, you know, this is reliable. To me, reliability is test/retest kind of work, which they didn't do, to my knowledge. And they can correct me if I'm wrong.

DR. MANHEIM: We have another program -- a look over the program, but we did not have someone do specification and run it. You know, we do rerun, reprogram everything without having the program in front of them and see if they get the same answer.

DR. RATLIFF: And what you offer as reliability testing again goes straight to like MarketScan and just to like MarketScan and saying MarketScan is reliable, therefore, our approach is reliable.

And I think considering the impact and the power of what the NQF product is, we have got to be cautious with appropriately scoring like this measure. And if it is

insufficient, it's insufficient.

And then your argument could be offered that, okay, well, that installation doesn't really mean anything, because MarketScan is reliable. That's okay.

But in terms of assessing this measure, I think we have to assess this measure.

MS. BOSSLEY: So I would say if you all are feeling that it is insufficient, you should say it's insufficient and staff will just need to ask you, if they don't feel that they have enough information, to write the rationale of why you scored it that way. They may ask you that.

But I think it is perfectly appropriate for you to feel this is a tough one. Insufficient, typically, is when we say they haven't given anything. But it sounds like they haven't given the right thing or enough information.

So or if they haven't given

anything, then you just say that it's insufficient. So you just need to provide a good rationale to the staff, so that they can provide it to the Steering Committee.

So it's truly your call on whether you want to say low or insufficient.

MS. O'NEILL: That's -- the reliability definition up there is pretty narrow. So it pretty much is saying if you ran the same tests on the same population at the same time, you would get the same result. So it's not like some capricious process.

And so I think we -- it meets this, but that the point that you are making is if we go out into the general public and use the term reliable, is this what they are going to think we mean or are they going to think we mean something else?

CHAIR WEINSTEIN: But you would imagine that if somebody brought a program on running some data with these elements, they get the same result. But you yourself said,

Mary Kay, early on about health partners and comparing. You can run into problems. And so I think without being capricious, I think we can say that they didn't run reliability data.

So it isn't that it wouldn't be.

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: It's just not there.

MS. O'NEILL: Right.

DR. RATLIFF: Shall we vote?

MR. AMIN: That was three low and three insufficient.

CHAIR WEINSTEIN: But I think the question here precise specifications, I think they did a great job. But then when you take reliability testing, you run into the -- so -- I mean, in this one, I would give a little more levity, because I think that the measures they used were reliable.

MS. BOSSLEY: No, I understand. Right, no. And here again is where I think you need to use your judgment --

CHAIR WEINSTEIN: Yes.

MS. BOSSLEY: -- as to how you will rate this.

CHAIR WEINSTEIN: Yes.

MS. BOSSLEY: And then --

CHAIR WEINSTEIN: But I feel this is a little easier to --

MS. BOSSLEY: Yes.

CHAIR WEINSTEIN: -- rate, because they did have precise specifications. And they probably figured them out with some algorithmic testing that was reliable. Benefit of the doubt here. So, okay.

DR. RATLIFF: The only thing they offer for the reliability testing is that they ran the same assessment again using the same database where they measure the same thing with the same ruler and they came out with the same number, so it's entirely reliable, but then they didn't go measure something else with the same ruler to see if it was reliable or not.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: Or if it was generalizable to --

CHAIR WEINSTEIN: But I think we can answer this one, as a group. So can we score it?

MS. WILBON: So we had one high, two moderate, two low and one insufficient.

CHAIR WEINSTEIN: Are the measure specifications consistent with the evidence?

MS. WILBON: That actually should be intent. Like is the intent of the measure -- I'm sorry. Are the specifications consistent with the intent of the measure? What they are saying that they are measuring.

CHAIR WEINSTEIN: Let's have a little discussion, so we are all feeling like we are answering this based on our group discussion. Do you want to say something, Mary Kay? Use your microphone.

MS. O'NEILL: Yes. Well, I think this is the one that should reflect our

feelings like is this the right time interval? Are we counting things the same way? Are we comparing different provider types?

And, you know, I guess part of the conceptual framework that seems -- that we seem to be moving back and forth between is this intent of this measure to measure the resource utilization as driven by a particular physician or other healthcare professional. And is the unit of organization around that, are we really like people are concerned about some may use this measure to figure out if somebody is going to get paid for what they do or are we trying to look at what is the most efficient or, you know, what are the resources used to provide care organized by the individual patient through an episode?

And so when we have these thing saying the comparison stuff is between peers, surgeons-to-surgeons, chiropractors-to-chiropractors, PTs-to-PTs, that's one purpose. But if we are going to say if somebody, you

know, walks into your hospital or your healthcare delivery system in Dallas, are they cared for well, then it really doesn't make sense to then just compare the surgeon-to-surgeon.

What makes sense is to compare episode-to-episode and whether that is four PTs, an average .5 surgical, you know, X number. You know what I'm saying? So I have a hard time trying to figure out if we are talking about the performance of an individual physician or the care of an individual through an episode. And those are really different kinds of things.

And O & E, expected and observed--

CHAIR WEINSTEIN: But these are things, you know, in models you could adjust for, if you characterized that. And you could understand the variance based on that specific variable. So it could be done. It wasn't done, but that's okay. And they are saying they should correct this if we are

misinterpreting.

They are saying that they are doing this by comparing apples-to-apples. I'm not sure that's so easy with the coding issues, but I think your point is well-taken.

MS. O'NEILL: Well, the intent is to look at the episode of care. So then some of the issues around comparing physician type-to-physician type moves me away from thinking that supports the resource use of the episode with the organizing principle being the patient as opposed to the provider.

DR. RATLIFF: Yes, I would like to touch on that. Again, their end result seems to be more physician or provider centric. A little less a group of patients say in Dallas versus a group of patients in Philadelphia, does Philadelphia do a better job than Dallas? Not so much. Nor does a physician at HUB do a better job than a physician at Jefferson in terms of resource utilization for a given set of patients' episodes of care. Is that

getting to what you are asking?

MS. O'NEILL: Yes, yes.

DR. MANHEIM: And that was our intent.

CHAIR WEINSTEIN: If you did turn this to the patient, independent of the provider, which ideally would be the case, because a patient should be treated, you know, fairly uniformly in a system, given a diagnosis. You know, if they have hypertension, they are going to get X. If they have an MI, they are going to get Y, independent of who the treating person is.

In this case, the multidiscipline confounding that occurs makes this very hard to discern. And that is where I think you have to do these sub-categorization analyses, because what you would probably find is that the outcomes could be the same, if you had some systematic approach, which we are not seeing here and it's not really addressed.

But I think for this particular

question, as we have been instructed, are the measure specifications consistent with the method or consistent with -- what term did you use, other than evidence?

MS. WILBON: The intent.

CHAIR WEINSTEIN: The intent.

MS. WILBON: Or the focus of that.

CHAIR WEINSTEIN: So I think they laid out what the intent was. I assume they were consistent with their intent. Is that intent going to help the measure be more valid or not? I don't know. We have some questions about that as a group.

Any more discussion? Patricia or anybody else about this?

MS. TURBYVILLE: Just for -- to capture, so it was two moderate and four low. So was the voting -- the rating of this based on some concern of the administrative data as well as some of the -- so that the diagnostic codes perhaps aren't -- so if you could rephrase what -- for this particular validity

issues are, so we can --

CHAIR WEINSTEIN: Somebody who had low -- well, this isn't a judgment. Can somebody who picked low speak to why they said low?

MS. WILBON: It would be helpful.

MS. O'NEILL: So if the measure intent is to measure the resource uses in the episode of care, and you -- and as we have established with our earlier discussions, that there is a lot of variability in what kind of resources can be put forward to a given episode, if we start sorting things then by physician type and comparing people to peers, you will end up with an analysis that says that whatever provider type is driving the episode is the appropriate one and that will not come into question.

So surgeons will be related to surgeons, whereas, I think as Jim points out, there is a subgroup within this population that are surgical cases and a subgroup that

are not. And there would be no way to differentiate whether the surgical services or the extensive or minimal PT services or whatever is the right application of resource to the particular episode.

So I think we lose the ability to critically look at the resource uses on an episode from an appropriateness perspective by the way it is constructed. And that's my concern.

CHAIR WEINSTEIN: Anybody else want to comment for Sally's question?

MS. TURBYVILLE: So that would be then shared across the others who rated low. And the other reasons that we should be sure to capture to understand that rating.

DR. RATLIFF: I voted moderate, but I don't disagree with that at all. I think that's a pretty succinct explication of one of the major weaknesses of this approach.

CHAIR WEINSTEIN: Is that helpful, Sally?

MS. TURBYVILLE: Yes. Thank you.

DR. SINNOTT: It doesn't mean that any of us have a better idea of how to do it.

CHAIR WEINSTEIN: Well, I disagree. I disagree, because I think you could validate this. In validation, you could look at subgroups treated by different specialties and actually do some, you know, chart reviews. There is ways to validate this.

And people have done those kinds of things. So we shouldn't suggest it is impossible.

MS. SINNOTT: I'm not suggesting it isn't possible, but I'm assuming that we haven't -- if we are limited to administrative data --

CHAIR WEINSTEIN: Yes.

MS. SINNOTT: -- as currently known, then we haven't figure it -- we may not have figured it out yet.

CHAIR WEINSTEIN: But I think

there is an algorithm you could apply to this that might be more acceptable. And what I alluded to before is, you know, beginning of episode with symptoms, MRI, time to surgery, length of stay, did they go -- you know, did they have other visits?

You could look at their -- a cohort of patients with an administrative database and get a sense of are they different than those treated by chiropractors or physical therapists or even surgical differences.

So I agree with the limitations of the database for sure, but there are some other kinds of codes and other codes where we could actually probably get more specificity around a cohort of patients.

MS. SINNOTT: My only concern about that is what we refer to in California as the Redding effect, which is that people get heart surgery when they don't need it and, therefore, the outcomes look great.

CHAIR WEINSTEIN: Yes, that's the Dartmouth data. We reported that.

MS. SINNOTT: Right.

CHAIR WEINSTEIN: Yes, so I'm very familiar. We see that all --

MS. SINNOTT: We'll call it the Redding effect.

CHAIR WEINSTEIN: Yes. But that's pervasive. The issue really is, and that's why I brought that up in the very beginning, the indications and the way we use these codes. And NQF is very interested in patient preferences. We just talked about it. None of these things are captured giving good information when patients have chosen those kind of treatment algorithms.

And we know from our studies the answer is no. 30 percent wouldn't have. So we are taking the best we have to look at something in a phase and we are going to continue to make it better.

But I think our job is to try to

congratulate the people who are doing this work, because it's really hard, to try to help us get to a new level of understanding. And then improve the database, so that we can get more specificity and more validity of sub-populations.

Until we include patient preferences, so informed choice I would say, until we include outcomes and the diagnostic testing that validates, including the physical exam, we are not going to have the physician groups, anyhow, agreeing that this is a valid sub-population that is like my patients, you know.

So we all understand that.

MS. WILBON: So this question is about validity testing and whether or not what they submitted reflects that they have demonstrated that the measure score or the data elements are valid.

CHAIR WEINSTEIN: Any comments on this before we vote from the group?

MS. O'NEILL: I just have to make my standard comment on costs. So if you want to know what that is, I mean?

CHAIR WEINSTEIN: We do.

MS. O'NEILL: I think actual money spent is a resource used and so standardized pricing while understanding that they even out market differences and contractual differences and look at utilization decisions, I do think that it needs to be really clearly put forward, first of all, that if something looks like a dollar figure on the results, that they aren't real dollars, that they are standard dollars. And it think that is hard for the public to interpret.

And that there is value to be able to crosswalk these things in different situations to actual dollars, because those are the resources people are using for care.

CHAIR WEINSTEIN: Just to be clear, are you suggesting that resource utilization is not a surrogate for cost?

MS. O'NEILL: It is not a completely accurate surrogate for cost, no.

CHAIR WEINSTEIN: But a lot of people use that methodology?

MS. O'NEILL: Oh, I know that.

CHAIR WEINSTEIN: Yes. Yes. No, but I want to understand why it is -- I mean, it does -- again, I go back to the notion it gets us started on a path. You know, Kaplan uses TD ABC, you know, activity-based cost accounting, where you actually have to measure every time that a nurse is there for 30 seconds or a radiologist spends two minutes on an x-ray film.

MS. O'NEILL: Yes, but he is talking about his business costs under his own roof. It has nothing to do, I'll tell you, I contact with them, with what he is charging me.

CHAIR WEINSTEIN: Yes.

MS. O'NEILL: Or the employers that we represent or the out-of-pocket of our

membership.

CHAIR WEINSTEIN: No. I understand the different --

MS. O'NEILL: Okay.

CHAIR WEINSTEIN: -- methodologies to costing, but I think what they are trying to simply do is say that resource utilization, which is being measured here, is a surrogate for cost in some way.

MS. O'NEILL: Well, just as it is -- other things that we are measuring are approximations and not completely accurate an we feel like to be fully transparent, you need to call that out.

CHAIR WEINSTEIN: Yes, yes.

MS. O'NEILL: You know, because I can tell you I did a little work on some spine fusion practices in the State of Wyoming and not only was the frequency considerably different, the cost per case was considerably different.

So if we did standardized costing

between a fusion and -- you know, I mean, we are losing 50 percent of the financial information, if you will.

MS. SINNOTT: But --

MS. O'NEILL: So but I understand why we are doing it. I just want it -- I want -- people look at a dollar figure and that is something that most people think they understand what it means.

If we are doing standardized costing, and we are reporting it out, it just needs to be clear that this has taken away the -- it has nothing to do with the -- what has--

CHAIR WEINSTEIN: It's the average versus the variability. Is that what you are worried about?

MS. O'NEILL: Yes. And then in given markets it may be way -- nowhere near average. So I mean --

CHAIR WEINSTEIN: Because I know your point.

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: I mean, spine fusion is a good example.

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: Where, you know, there is -- but the rates of those procedures in various areas are so different and the utilization or resources to get a fusion is very different, depending on where you live and who you see.

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: Is that your point?

MS. O'NEILL: Well, there is that. And there is -- some of this, I don't know if they were -- the NCQA methodology -- there was some discussion in an early measure looking at charge data.

CHAIR WEINSTEIN: Yes.

MS. O'NEILL: And we have, for the region around Seattle, a 20-hospital grid based on public available data on the differences between charges and payments and,

you know, there is completely different games that are played with charge-master and discount levels.

So there was one hospital that had huge discounts, but they still were more expensive than everybody else, because they started with such a high charge-master and the other hospitals said well, we don't charge very much, had a low charge-master, but almost no discount.

You know, I mean, there is lots of number games out there that are -- that end up being significant to --

CHAIR WEINSTEIN: I want to understand the variables of those number games, so that we can be clear for our reporting.

MS. O'NEILL: Right.

CHAIR WEINSTEIN: Because you mentioned contracting and everybody has got sort of a secret contract. And what they pay for things is different with CIGNA than it is

with United.

MS. O'NEILL: Correct.

CHAIR WEINSTEIN: It's different than Medicare. Is that your point in some ways?

MS. O'NEILL: That is.

CHAIR WEINSTEIN: Okay. I wanted to try to be clear.

MS. O'NEILL: So how much it costs to care for these folks, really costs to people who are really paying the bills, that actual piece of information is only vaguely approximated by standardized pricing and that's a --

CHAIR WEINSTEIN: Agree, agree. Thank you.

MS. SINNOTT: But I think we are also interested in the variation in utilization. So there is really a standardized cost that gets applied to the utilization variation, which is different from the variation in the contract charge or

contracted payment.

MS. O'NEILL: Correct. I mean, and when I first started this, I was trying to make the position that we should count things, instead of put a dollar figure on it that was an average, because it started leading us down to a path of having an apparently interpretable piece of information that was really inaccurate on the local level.

However, I understand standardized pricing also functions to relatively weight different types of utilization. So, you know, if we do standardized pricing, you can relatively weight over-utilization of labs versus over-utilization of surgery, which would have very different impacts.

So I understand the purposes of it, but I just think it needs to be called out that there would need to be a translation, if financial decisions or economic decisions are being made, there needs to be a translation, to the real number.

Not to stop it, but those are two different columns on the sheet.

CHAIR WEINSTEIN: And there are regional differences in those payments.

MS. SINNOTT: Yes.

CHAIR WEINSTEIN: At least threefold, at least.

MS. SINNOTT: Oh, yes.

CHAIR WEINSTEIN: So --

MS. O'NEILL: And within regions.

CHAIR WEINSTEIN: Right.

MS. O'NEILL: By the way.

CHAIR WEINSTEIN: Right, right.

DR. RATLIFF: So from a patient's perspective with this episode, it may be easier to use those kind of calls to base those or use like one unit cost, because then you may be able to see the patient's utilization of a given resource. So that's the kind of patient.

From a physician's perspective, that may be completely meaningless, because

what you charge for like a surgery, what a given physical therapist may charge for an intervention may be different than a physical therapist down the street, which is also irrelevant to what the person, the payer is experiencing, since they are seeing all these charges.

So again, I think it's a choice of like how you are approaching. Going back to an earlier point that we brought up, how you approach utilizing this evidence-based measure. Whose perspective are you looking from with regards to utilizing this?

MS. O'NEILL: But, I mean, from a choice perspective, increasingly all the national carriers right now have on their membership website the actual relative -- the actual different costs of getting different procedures at different facilities based on their benefit design and the contracted rates.

CHAIR WEINSTEIN: This is, you know, the whole tiering that is occurring,

which gets into that, you know, which then gets in to patient's copays, which gets complicated.

MS. SINNOTT: Yes, it does.

CHAIR WEINSTEIN: But let's just take the question now with those caveats. No, no, it's very helpful. Thank you. Thank you. It's important. It's important.

So does the validity testing demonstrate that the measure data elements are correct and/or the measure's score correctly reflects the cost of care or resources provided adequately distinguishing high or low cost or resource use?

Which I think is some of your point. You are not sure that it does.

MS. SINNOTT: Not the cost, the resource.

CHAIR WEINSTEIN: Yes. Any other questions before we answer this one? Okay. So are exclusions supported by the clinical evidence for analysis of frequency and

distribution? Is information about impact of exclusions for patient preference transparent?

Now, this is impossible. I'm sorry, because patient preference isn't really measured or captured, yes. Thank you. So it's another one where we have insufficient information.

Are you okay, Heidi, with this?

MS. BOSSLEY: Yes.

CHAIR WEINSTEIN: Sorry to distract you.

DR. RATLIFF: I don't think we measured all of them --

MS. BOSSLEY: Sorry, I'm multi-tasking.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: -- or discusses this even.

CHAIR WEINSTEIN: Right. Okay. Can we go onto the next one?

MS. WILBON: It was three low, five insufficient. I'm sorry, one low, five

insufficient.

CHAIR WEINSTEIN: This gets into risk adjustment, 2(b), for outcome measures. Is there evidence-based risk adjusted strategy or rationale data support -- no risk adjustment. So we think that there needs to be risk adjustment, so the second part of this isn't necessary, because if we didn't, then it wouldn't need to be there.

So the question is is there evidence that risk adjustment strategy was used? Any discussion about this before we vote?

DR. RATLIFF: We discussed this earlier in terms of a risk adjustment methodology and the complexities entailed there. Obviously, they have a methodology, I'm just not sure that it has been validated or that it is generalizable.

I mean, it seems reasonable from my interpretation of it, but, again, it's a relatively dense approach to risk adjustment.

CHAIR WEINSTEIN: Any other comments?

MS. WILBON: So I just wanted to point out, so the -- what we have on this slide is an abbreviated version of the criteria, so I just wanted to read the full 2(b)(4).

So it says that "For outcome measures and other measures, which includes resource use, when indicated, and evidence-based risk adjustment strategy is specified and based on patient clinical factors that influence the measured outcome and that they are not risk adjusting away disparities, that they are measuring patient clinical factors that are present at the start of care and they have demonstrated adequate discrimination and calibration."

So that's the whole criteria that we are evaluating, at this point.

CHAIR WEINSTEIN: But not including disparities?

MS. WILBON: Right. So NQF, basically, has done work and wants to ensure that people are not including disparity type factors, race, ethnicity, into risk models, which those things should actually be stratified for, so they can be addressed rather than adjusted away.

So that's just something we had in there for clarification.

CHAIR WEINSTEIN: I'm not sure they did that though. And I'm not sure their population addressed that. Could I have clarification on that?

MS. WILBON: Sure.

CHAIR WEINSTEIN: Did you stratify, based on race, in your mind?

DR. MANHEIM: No, we did not, because we cannot measure it in mixed up data.

CHAIR WEINSTEIN: Yes, that's what I thought.

MS. WILBON: So this question is just asking about their risk adjustment model

and what they actually -- there is actually a separate criteria for disparities that we will get to in just a second. But this one is asking specifically about their risk adjustment model.

So it was two moderate and four low.

CHAIR WEINSTEIN: Next question. This is about the scoring analysis. Are performance results reported? Do they identify differences in performance or overall less than optimal performance?

And, to me, they didn't actually compare performance. Unless this means -- they didn't do it across systems, because they only had one, but they did it across providers. Is that where we are at here?

MS. WILBON: Observe versus expected?

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: That would appear to be it, just observe versus expected per

provider.

CHAIR WEINSTEIN: Right.

DR. RATLIFF: As opposed to --

CHAIR WEINSTEIN: So is that okay?

DR. RATLIFF: -- really scoring the performance.

MS. WILBON: Yes. I just want to again read the full criteria. So again, these are just kind of abbreviated versions and it's not as robust as what we have on the slide.

So 2(b)(5), actually, asks "Whether or not the data analysis demonstrates that the methods for scoring an analysis of the specified measure allow for identification of statistically significant and practically or clinically meaningful differences in performance."

CHAIR WEINSTEIN: Yes, I just --

MS. WILBON: Or that there is --

CHAIR WEINSTEIN: Just go back to the statistician's problems, which we discussed, that they weren't adequate, but it

doesn't mean they didn't try. So that's all.

Any other comments by the group before we vote?

DR. RATLIFF: I guess the statistician's concern was that they were extracting from like the raw numbers to these ratios based on their distributions and that like extraction was an issue for the statistician.

CHAIR WEINSTEIN: Right. And they tried to address it.

MS. WILBON: So the score was six low.

MR. AMIN: Can I ask the Committee to give a little bit more clarification on this one, just for our rationale?

CHAIR WEINSTEIN: The last one?

MR. AMIN: This last one, the one with six low. Is the concern around the distribution of the ratio or how the ratios are actually developed for the scoring?

CHAIR WEINSTEIN: I think it's how

they were developed.

DR. RATLIFF: I would almost defer to your statistician's comments with regards to how they are extracting.

CHAIR WEINSTEIN: Which we are weighing some of our thoughts based on that as well.

DR. RATLIFF: Right.

CHAIR WEINSTEIN: We are weighing some of our thoughts based on Carlos' interpretation.

MS. O'NEILL: Yes. So it was hard to tell if the numbers were different, that's one piece. And, you know, I guess back to my more philosophic thing, some of the practicality of what is measured in terms of what intervention you might take within a system to improve things, you know, seems to me a little limited, because we go from -- we go directly to peer physician resource utilization and not episode of care of the patient in terms of efficient -- you know, the

utilization.

MR. AMIN: Thank you.

DR. RATLIFF: I guess for me it would work a little bit better if it was just kind of clean and here is your expenditures per episode, per physician as opposed to extrapolating out or kind of normalizing between different episodes and then giving that normalized data as an observer versus expected for a given physician.

I voted a little bit crisper and like here is what your payment was per episode. Okay.

CHAIR WEINSTEIN: And it gets into these, you know, sub-populations that may be different, too. So it's not bad, it's just the best you can do with this.

Oh, is that it?

MS. WILBON: There is one more. And this one tends to be not applicable, only because there are only -- yes, they are only using one data source which is the admin data,

so --

CHAIR WEINSTEIN: So do we have to vote?

MS. WILBON: No. We will just make this one not applicable.

CHAIR WEINSTEIN: Okay. So is it break-time?

MS. WILBON: Not quite. We have got a couple more.

CHAIR WEINSTEIN: Oh, right.

MS. WILBON: So we do need you to kind of give a roll-up score of the overall validity based on those five -- well, minus the multiple data sources, but those four bullets about the specifications being consistent, the validity testing, the risk adjustment and the identification of statistically meaningful differences.

So kind of a summary judgment on how they scored on validity.

CHAIR WEINSTEIN: Just for comment, because I think for the people from

NQF, I mean, I think this is complicated. And we will find this, you know, at least for me, that Ingenix did a lot more work with a lot more population, so you have more testing of it, which allows you to make some different interpretations maybe.

This measure has not -- this ABMS effort has not going through that sort of process. And I think they are early in their work. Maybe I'm wrong, but it's my interpretation.

But I want you to understand it's not we are trying to make this harder or easier, we are just trying to base it based on what we have seen.

DR. RATLIFF: And I would echo that comment. I don't think this is at all saying that this is not a reliable measure. It's simply that the testing hasn't been done.

I think the measure itself is like very promising. It just hasn't been exported.

CHAIR WEINSTEIN: Their stuff is

very, you know, I think, clearer than I think Ingenix in many ways.

MS. O'NEILL: And I think if some of the issues that we have raised here and the testing were available in many regards, I think the sort of philosophic structure of these measures is actually in a practical sense somewhat more actionable than Ingenix.

You know, because as a clinician, I look at Ingenix and I'm like what would I do next? I don't know. So anyway, I guess I also would like to put it -- if there is an encouragement -- is there an encouragement vote? Keep going, keep going.

MS. WILBON: So the overall, for those on the phone, the overall validity rating was six low. We are just going to vote on the last sub-criterion which is 2þ for disparities and then we will take a break.

DR. RATLIFF: And if their database didn't give them data to assess disparities between different ethnic groups,

then we ought to opt out of this one, also.

MS. WILBON: Right. I mean, it could be insufficient and this, again, is something that other committees and TAPs have weighed and whether or not it is a limitation of the measure or a limitation of the data of the admin data itself and just kind of where we are with collecting disparities data, in general.

So I think, you know, --

DR. RATLIFF: This is a limitation of -

MS. WILBON: -- weigh that -

DR. RATLIFF: -- the database they used.

MS. WILBON: Right. And so, you know, weigh that in your consideration and then we will just make sure, depending on the rating that we get rationale for why that particular rating was as such.

MR. AMIN: That was one low and five insufficient.

MS. SINNOTT: I just wanted to say something about the validity scoring just to reinforce that it is not a belief that it couldn't be good, but it is a criteria for making it better. You know, and that the group has strong feelings that it is very interpretable and would be very well-received by physicians or other providers.

DR. RATLIFF: And I would echo that as well. I think we are more or less answering the questions you are posing. So we are not at all saying that this is not a valid measure or that we would all imply that there is low validity applied to this measure.

I think it's a very good measure. It's simply that it was explored in one database. And in answering the question that you posed, some of these issues have not been fully sussed out, but that's more perhaps standardized questions applied to a bunch of different models as opposed to a problem with the model itself.

CHAIR WEINSTEIN: There are some very specific things and we are not piling on here, but I think that the notion is is that I actually think this is an easier measure potentially to use. They exclude some things like the pharmacy benefits or exclude patients without pharmacy benefits, which is really a positive.

But I find this -- you know, most people could use this. They wouldn't have to buy the Ingenix tool, which I think we are going to get to that, you know, later on, which is a big issue, because the CMS site allows this kind of use for everybody.

So there is some usability issues here that are very significant and I wouldn't want to get lost in them feeling criticized inappropriately. So just to echo the comment.

MS. WILBON: So let's go ahead and take like maybe a 10 minute break. I know originally we had 15, but we're a little bit-- we're not that far behind, but about 15

minutes.

CHAIR WEINSTEIN: We'll catch up.

MS. WILBON: We'll catch up. So we are going to come back and finish usability and feasibility for this measure and then move on to the Ingenix measure.

So for those on the phone, a 10 minute break. Thank you.

(Whereupon, at 11:16 a.m. a recess until 11:30 a.m.)

CHAIR WEINSTEIN: Are the measure performance results reported or suitable to report to the public at-large in national or community reporting programs? Is there evidence that the measure performance results are available?

So this is two separate questions in some ways. I guess we have one answer for both, which is hard, because right now, they are not available. And they need some work. They could be available for Part B. For A of Part 3(a)(1), are the results reported in

public? They are not.

So do we again go with insufficient or are we going to -- how are people interpreting this differently than me?

MS. O'NEILL: It seems like insufficient is the appropriate thing, because the other ones seem like we are judging how well they are doing this. And they aren't doing it, so -- and it's part of that sort of general signal that this is a measure in development.

CHAIR WEINSTEIN: Right.

DR. RATLIFF: And the developers know they've got Robert Wood Johnson funds for their ongoing development and this is a developing process.

MS. O'NEILL: Right.

DR. RATLIFF: So they are just not there yet. I think it is sufficient probably just, you know, making that point.

CHAIR WEINSTEIN: Do you have any comment, Taroon?

MR. AMIN: I think the only comment that would be made here is recognizing that the process of where resource measures are in development broadly, the expectation that it would be reported to the public at-large is not necessarily --

CHAIR WEINSTEIN: I think if you had the question, are the measure performance results expected to be reported, you know, at some point? Yes. But that's not the question.

MR. AMIN: Yes.

CHAIR WEINSTEIN: So we can't really say anything but insufficient. But I just want you to understand that.

MR. AMIN: Right.

CHAIR WEINSTEIN: Yes. I hate to say that we haven't voted, but -- there were six insufficient.

DR. RATLIFF: Yes, six insufficient, sir.

CHAIR WEINSTEIN: So did the

submitted information demonstrate that results produced by the measure are meaningful, understandable, useful for quality improvement and public reporting or was a credible rationale presented? Discussion by the group? I don't want to lead this one, because I'll say the wrong thing.

MR. AMIN: It's being evaluated, right?

CHAIR WEINSTEIN: Any other comments? Anybody else? Patsy, anything? No. Okay.

DR. RATLIFF: It's two moderate and four insufficient.

CHAIR WEINSTEIN: Okay. Are the data and result details maintained such that the resource use measure, including the clinical and construction logic for a defined unit of measurement can be decomposed, interesting word, to facilitate transparency and understanding?

MS. WILBON: I'm sorry. I just

want to go back before we get into this one. Can you just give me an idea of why the insufficient for whether or not -- for 3(b), whether or not the measure is meaningful, understandable and the results are useful? Is that based on some of the issues you had with the scientific acceptability and the reporting of the measure scores? Could you just give me --

DR. RATLIFF: I voted moderate, because I was giving them the benefit of the doubt that as they developed this per their -- they are probably going to get there. I could easily see voting insufficient, though, with the idea being that this is under development and we don't know where they are going to bring that train into the station.

CHAIR WEINSTEIN: Yes.

MS. WILBON: Okay.

MS. O'NEILL: I think, you know, that some of the questions even that Carlos raised about the observed versus expected and

what those numbers were, we can't tell until it has sort of been run through the drill whether or not you are going to get a usable result that would change practice patterns, because we can't tell yet really if those are different numbers, you know, with the confidence intervals.

CHAIR WEINSTEIN: The danger of a priority accepting something without the evidence would not be in our best interest, at this time.

MR. AMIN: Any time.

CHAIR WEINSTEIN: Are the data and result details maintained such that resource use measure, this particular measure, including the clinical and construction logic for a defined unit of measurement can be decomposed, I guess disassembled, to facilitate transparency and understanding?

So if you broke this down, this measure, could people really understand it? I would change the word decompose, but

questions by our colleagues about this?

DR. RATLIFF: So I guess just logistically, is this referring to the observe versus expected ratio that is being developed by additional practitioners or is this the more overall data set that is being developed in evaluating each patient's episode?

MS. WILBON: It's more about the construction of the measure. So in the way that it is specified, so how they have constructed the episode, how they are assigning and attributing, you know, the cost of the physician.

DR. RATLIFF: Not just the end result, but the entire spectrum?

MS. WILBON: The entire measure. Could somebody kind of take it apart and say oh, okay, I understand how they are attributing physicians. I understand how the time -- you know, the --

CHAIR WEINSTEIN: As to the construction of this --

MS. WILBON: -- how it is risk adjusted, right.

CHAIR WEINSTEIN: -- model, is it understandable?

MS. WILBON: Right. The different pieces of it, you know.

MS. O'NEILL: And could you build it with their --

MS. WILBON: Right.

MS. O'NEILL: -- based on their definitions?

MR. AMIN: That's two high and four moderate.

CHAIR WEINSTEIN: Next question. Does the measure meet NQF -

MS. WILBON: So that's for overall.

CHAIR WEINSTEIN: Oh, sorry. Does the --

MS. WILBON: Yes, no, that's --

CHAIR WEINSTEIN: We don't do that.

CHAIR WEINSTEIN: Are the required data elements routinely generated and used during care delivery?

MS. WILBON: So 4A and 4B, are those two feasibility criteria that I was telling you about, that because these measures are based on admin data and admin data are generally created during care delivery, and as is 4B, which refers to whether or not the data elements needed to run the measure are available electronically, which they are.

So we can just do a -- if everyone is okay with that --

CHAIR WEINSTEIN: Can I argue though?

MS. WILBON: Sure.

CHAIR WEINSTEIN: Because they are not all available. The preference issue, which is talked about here, it's not in their model, but NQF would want it. So do we --

MS. WILBON: But it's not --

CHAIR WEINSTEIN: Specified --

MS. WILBON: They haven't --

CHAIR WEINSTEIN: -- in their model.

MS. WILBON: It's not specified in their measure.

CHAIR WEINSTEIN: Okay.

MS. WILBON: So, as written, you wouldn't need it to run their measure, as specified.

CHAIR WEINSTEIN: Correct. Thank you.

MS. WILBON: Right.

DR. RATLIFF: So working within their model --

MS. WILBON: Right.

DR. RATLIFF: -- the data elements they are looking at in their model, are we recording that already? Can they get that from an EHR?

MS. WILBON: Right.

MR. AMIN: That's six high.

CHAIR WEINSTEIN: Are the required

data elements available in electronic health records or other electronic sources? Is that the same thing?

MS. TURBYVILLE: Yes, it should be high.

CHAIR WEINSTEIN: So it's just asking the same question a different way?

MS. TURBYVILLE: Yes.

DR. RATLIFF: Actually, A is just saying that you are measuring it and that's a sign that you are putting that measure into an EHR, I guess. I misspoke, but I'm saying EHR.

MS. WILBON: Not just EHR.

DR. RATLIFF: Yes.

MS. WILBON: This is claims data.

DR. RATLIFF: Or claims data.

MS. WILBON: Yes.

DR. RATLIFF: Some administrative database.

MR. AMIN: That's six high.

CHAIR WEINSTEIN: Are the -- are susceptibilities to inaccuracies, errors, or

unintended consequences and the ability to audit the data items to detect such problems identified? Comments by the group? I'm not sure that they addressed this. Anybody?

MS. O'NEILL: Starting with your first point about, you know, what kind of inputs there are to coding, I mean, not that that's an easy thing for anybody to do, but that would be a source of error that is not--

CHAIR WEINSTEIN: But it wouldn't be an error from their model, because they are just taking the claims codes.

MS. WILBON: Right.

CHAIR WEINSTEIN: That would be an error -- a step from the UB-92 forms or something.

MS. WILBON: Yes.

CHAIR WEINSTEIN: Yes. Any other comments?

DR. RATLIFF: I think we have noted them multiple times the potential sources for bias in that.

MR. AMIN: That's two high, three moderate and one low.

CHAIR WEINSTEIN: Yes, sir. Can the data collection strategy be implemented? Is the measure already in operational use or did testing demonstrate that it is ready to put into operational use?

Any comments or questions? My sense of this is just they haven't made a model of this to be industrial. They have just been doing their own testing of it, at this point. So I don't know if it is ready.

Does anybody feel differently?

DR. RATLIFF: I mean, we discussed whether or not they looked outside of MarketScan or looked to a more generalized approach and the answer was no. So I don't know if this has been explored yet.

I think the general concept though probably --

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: -- is very valid.

CHAIR WEINSTEIN: Yes, I'm sure.

DR. RATLIFF: Or it could be.

CHAIR WEINSTEIN: It's just they haven't done it. My sense is compared to Ingenix, it's got a product out there that they are testing. This is not. That's not a problem, it's just not there. But am I misinterpreting for the group?

MS. WILBON: So, again, let me just read the full criteria here to help -- hopefully this will help clarify.

So it is asking whether or not the data collection measurement strategy can be implemented as demonstrated by operational use and external reporting programs or that testing did not identify barriers to operational use.

MS. SINNOTT: So in this case, it has neither external operating -- reporting activities nor has testing been done.

DR. RATLIFF: But are you asking us to speculate could it be done? Do we see

any barriers to applying this measure to say another provider database?

MS. WILBON: Right.

MS. O'NEILL: I mean, so the fact that they are just -- they are using standard administrative data, I mean on a very basic level, could another system get at their system standard administrative data? That simple answer would be yes. But has it been vetted? I guess that answer is no, so far.

But are we really looking at are the data elements that -- or the inputs to the measure standardly available?

MS. WILBON: We're asking more so about how feasible is it or how easy is it for a user to pick this up and implement it? Is it implementable, I guess, if that's a word. And are there barriers to doing that, you know?

Right. So examples would include, you know, data availability, timing, frequency, you know, complex sampling required

to run the measure, patient confidentiality issues or fees for use of proprietary specifications.

So those are some of the things that would, you know, hinder or limit the feasibility of running or implementing the measure.

CHAIR WEINSTEIN: Well, but, you know, you and I talked on the phone even for the Ingenix thing, we are going to -- we would have to pay a fee to be a user. We don't know anything about this one.

MS. WILBON: Yes, it's -- it would be open to the public. It's a -- it would be free.

CHAIR WEINSTEIN: As opposed to Ingenix, which wouldn't?

MS. WILBON: Which would not. Which we will get to, obviously, when we discuss that.

CHAIR WEINSTEIN: Yes, yes, yes, gotcha.

MS. WILBON: Yes.

CHAIR WEINSTEIN: I just want to be clear in my own mind.

MS. WILBON: Yes.

CHAIR WEINSTEIN: But, yes, I just don't know that it is ready. I mean, it's exciting. I'm struggling with the answer to this question. Maybe it's I'm making too much of it. Anybody else?

MS. TURBYVILLE: Jim, could you provide some examples of the barriers that you are seeing to it being feasible right now, just for clarity sake?

CHAIR WEINSTEIN: Well, I just don't know. I mean, my sense is if this gets validated and it works, are they going to commercialize it? I mean, I don't know what is going to happen. Are they guaranteeing us that this will just be a public measure and they are going to give us the software free for every place in the country?

MS. WILBON: So we have them on

the phone, so we can clarify. But my understanding is that it would be available publicly, that there wouldn't be any funding for it. We do have a process with all the measure developers that submit measures to us, they have to tell us whether or not they will be charging for it. And this -- any measure that gets endorsed should be available publicly in the specification.

So, essentially, what would happen with this measure, as with other measures that are not proprietary with fees, which is a little bit different than what we are going to see with Ingenix, but for this particular measure, the specifications would be available publicly.

The developer -- if someone wanted to use this measure, they could email the developer and say hey, I want to run this measure. They would take the specifications back to their house or whatever system they are in and have a programmer program it and

they would use it however they intend to use it in their system.

CHAIR WEINSTEIN: So this would be Microsoft Resource Utilization Version 1 that I could have for free?

MS. WILBON: Yes.

CHAIR WEINSTEIN: And install on my computer system?

MS. WILBON: Right. Obviously with some programming. But it would be a per system implementation.

CHAIR WEINSTEIN: And ABMS has no intent of trying to regain their cost, even though I know they have been funded by RWJ in some way. Is that --

MS. WILBON: Yes.

CHAIR WEINSTEIN: And we ask them?

MS. WILBON: Yes.

CHAIR WEINSTEIN: We are asking you.

DR. MANHEIM: There is no intention in bringing anything proprietary.

CHAIR WEINSTEIN: So you imagine that if the University of North Dakota -- I said I wanted to use your tool, I could go to the website at ABMS, download it and I could be in business? And you --

DR. MANHEIM: Yes, it would require some programming on your part.

CHAIR WEINSTEIN: Yes. And if there was a problem with it, you would have a 1-800 I have a problem number?

DR. MANHEIM: Todd, do you know the answer to that?

DR. LEE: It wouldn't be a software application that would be available. It would be the specifications and the technical appendices that would be available that users would need to translate into a software application, whether it is, you know, a vast programming language or some other application that they could use to run their data through our algorithm.

CHAIR WEINSTEIN: Yes. So my

sense is sometimes that is not so easy. And so those were my questions. Sorry.

MS. SINNOTT: And also, a programmer isn't a programmer and that that kind of translation doesn't necessarily happen in a valid way.

CHAIR WEINSTEIN: That's right. That's what I was asking. They are not going to have technical support though.

DR. RATLIFF: But if the NQF adopts this measure, does the NQF then popularize it or are you just going to say hey, this is a good measure?

MS. WILBON: So, no. NQF - once they are endorsed, they are just out there. We do -- we are looking to -- we will have a database available hopefully later this year that will provide like a central housing for all of our measures that are endorsed and give access to the public. Give the public access to the measures and to contact information to developers to ask questions.

But it is common that a lot of developers don't have -- you know, they offer support, I guess, as they are contacted, but I'm not sure --

CHAIR WEINSTEIN: But what happens often times is, you know, the SF-36 is a good example now, now it's bought by Ingenix and we can't really use it, you know. Or by United, I should say.

So I just -- that's my question. It's not anything more than that. It's a long way from knowing that answer for me.

MS. BOSSLEY: Right. Just to clarify, there is no requirement that developers have an 800 number or anything. The specifications need to be updated and on their website and available for individuals or maybe not on the website, but can be accessible.

CHAIR WEINSTEIN: And you know, the American Board of Medical Specialty is a wonderful group, but it's not really

commercial. I mean, they are trying to do some commercial things, I know, but they are the certification board for specialties, that this isn't.

MS. WILBON: Yes.

CHAIR WEINSTEIN: You know, so I just worry that they are going to be able to sustain this over a long period of time and that's just the reality. So that's all. I don't want to belabor it. Thank you for answering. I'm sorry, go ahead.

MR. AMIN: I just want to quickly clarify and ABMS might want to clarify this also. This is from the Research and Education Foundation, which is separate from the ABMS credentialing group.

DR. MANHEIM: And part of the purpose of this was to provide a non-proprietary clear specifications with the positive and negatives, I guess. It's non-proprietary. You don't have as much support. And, yes, this was done under the ABMS

Research and Education Foundation.

DR. LEE: And yet, I should note that neither Willy nor I work for ABMS. We are both academic researchers that were part of the development team of this project.

MS. SINNOTT: Just to add a little more to we don't know, once something like this became public and freely available, there would be nothing to restrict anybody else from adopting it and commercializing it in some way, either by providing support or something with a feedback to ABMS or something.

So you would get the algorithm and access to it for free, but your support you would have to pay for, for example. So and if I were in the business of creating measures and this got endorsed by NQF, the first thing I would do is integrate it into my measurement software program.

MS. TURBYVILLE: Which a lot of them do.

MS. SINNOTT: Which a lot of them

do, Sally says.

CHAIR WEINSTEIN: So can we vote?

MR. AMIN: That's four moderate and two low.

MS. WILBON: Great.

CHAIR WEINSTEIN: Is there another question?

MS. WILBON: Well, that completes your first measure. Three to go.

CHAIR WEINSTEIN: Yes. The next three are going to go fast.

MS. WILBON: Yes. So Ingenix, let's go ahead and do the next measure and then we will see how far we can get before lunch.

Yes, the next measure is 1609. It's an ETG-based Hip/Knee Replacement measure by Ingenix.

CHAIR WEINSTEIN: Yes.

MS. WILBON: So do we have someone from Ingenix on the phone?

DR. DUNN: Hi, yes, this is Dan

Dunn and I'm on the phone and I'll try to do my best here. Also Howard Tarko, who is our medical director for -- one of the medical directors for the methodology. Just as a note, that the lead clinician, Tom Lin, had a family emergency and we'll do our best to answer your questions.

If there is anything you would like us to follow-up on, we are happy to do that.

MS. WILBON: Thank you, Dan.

DR. DUNN: You're welcome.

MS. ZIELINSKI: Hi, Operator, this is Cheri Zielinski, I'm on the line.

MS. WILBON: Okay. Hey, Cheri, glad you guys were able to make it. If you could just give us a brief intro to the measure and then we will pass it back to the TAP. Thanks.

DR. DUNN: So this is Dan. I can do that. Okay. This is a hip and knee replacement, correct?

MS. WILBON: Yes, that's correct.

DR. DUNN: Yes, okay. So the measure focuses on resources used to episodes of care for patients who have undergone a hip or knee replacement. The methodology itself is based on the episode treatment group and procedure episode group methodologies developed and maintained by Ingenix used broadly in the industry.

The procedure episodes identify a unique procedure event, as well as the related sets of actions performed before and after the procedure. That includes work, often therapy, prior to the procedure, the procedure itself, including the inpatient stay and other surgeons work, et cetera, as well as post-op activities, such as any repeated surgery, outpatient follow-up, physical therapy.

The methodology is included that assigns a severity level to each episode. And so the results would be, you can think of it as, a hip replacement episode with a severity

level, a knee replacement episode with a severity level. And if you were going to do measurement, you would, you know, take into account the fact that you have a different episode for hip replacement, different episode for knee replacement with different levels of severity. Those together define, if you will, the risk values of the measurement.

There are a number of resource use category numerators, if you will, included with the measure. The total cost of care, care by -- cost by type of service, as well as some utilization measures for specific types of care.

MS. WILBON: Okay. Thank you.

DR. DUNN: You're welcome.

MS. PAXTON: This is Liz Paxton. I was wondering how you are handling laterality, especially in terms of total knee replacement.

DR. DUNN: That's a good point. So the question is if there is a bilateral?

MS. PAXTON: Oh, or a subsequent knee replacement, not necessarily a simultaneous bilateral procedure, but --

CHAIR WEINSTEIN: Do you record right and left? Do you record right and left in your data system?

DR. DUNN: Yes. So there is right and left, if they are indicated on the administrative data, that's captured. If it's bilateral in the same event, both. I mean, it's indicated by the procedure code modifier, that is captured.

If there is a knee replacement, that episode, for example, and then say within the time period defined to cover the -- you know, say one knee replacement episode, that means kind of overlaps within the episode, that's also recorded as the fact that there is overlapping knee replacement episodes of care.

In the case of the bilateral, you know, that would be something that someone would control for or exclude, if they decided

that those are going to be more work. If there is overlapping, usually people treat that as an episode that wouldn't likely be included, you know, just difficult to have a complete picture of what went on.

CHAIR WEINSTEIN: One of the things that -- I'm still not sure that -- so what you said the answer to that question was is when it is available, you get it? So it's sometimes available, right versus left? It's not a required data field in your administrative data set?

DR. DUNN: Yes, that's correct. I'm assuming that procedure code wouldn't give you that alone, that that would --

CHAIR WEINSTEIN: So it's not.

DR. DUNN: Right. It would show up on the modifier.

CHAIR WEINSTEIN: Yes, yes. Most people don't have that. I think, you know, you guys have done some tremendous work, like ABMS. And we appreciate that, number one,

because it's fairly complicated.

The thing I run into in this particular diagnosis is preference. The rates of procedures even in your write-up are quite variable. You talk about, you know, Wisconsin and other places with rates varying from 162 per 100,000 to almost 300, so there is at least a twofold variation in the rates of these procedures and the cost continue to climb.

And I know just from my own work that the rates of these procedures go up for a number of reasons, just the aging population, plus people are doing them in younger populations than they have done before.

And there is no preference. And so the indications for this like back surgery, get to be a little blurred, although, you know, there is a clear, you know, x-ray changes in the studies out of Canada that you are probably familiar with where patients were

actually given choice.

There was only about 16 percent of patients when given a choice in Canada actually wanted the procedure, which then doesn't get dealt with here.

And so the issue is it's a very effective - cost-effect procedure. People really get good relief of pain and become very functional. And we're going to get into this in disparities. There is quite a difference in rates of these procedures in non-whites, which we can talk about, which I think are not talked about in your write-up.

But how do you address this preference issue, if at all, in your data systems? I'm just curious, because it really is an underlying problem for preference-based decisions.

DR. DUNN: And another great point. We don't deal with it in this measure, so the assumption here is that the -- a decision was made to go forward with the knee

replacement, for example, and then, you know, given that, measure the cost associated with it.

We also, you know, have -- the later discussion is joint degeneration episodes that you can then, you know, look at rates of surgery within those. But within this episode itself, knee replacement or hip replacement and the decision for surgery has been made.

CHAIR WEINSTEIN: Yes, it's just for NQF. To me, this is a major issue around quality. And just because something can be done and has a good result, doesn't mean that it should be done. And a well-informed patient might choose differently.

And I don't know how that gets addressed, but I think it is significantly important. And, of course, there is no outcome data here. And, you know, the readmission rates, complication rates, these are fairly high in some of these things that

are very costly.

It's a great procedure. I'm an orthopedic surgeon. I understand it, but I worry about the ever-increasing rates without those kind of things being measured. And it's no function -- no reflection on Ingenix. They have nothing to do with that, but the notion is, I think, if NQF is going to be a quality measure place, those things need to be addressed in the episode, if we are going to talk about the usability of these things.

MS. O'NEILL: So the way the reporting comes at the end of this measure is on a, you know, per physician measured against their peers. So the decision to do the procedure has already been made. So basically, the measure compares resource utilization once the decision is made.

But to your point, we are not measuring the quality of the decision making or the process of the decision making. And we would have to probably look at some type of

defined population for rates and maybe even cohorts from different age groups what you might consider a somewhat appropriate rate for people on different age cohorts within a given population, how that would be managed by the system.

But this is just after the decision is made.

DR. RATLIFF: So as I see the difference for lumbar radiculopathy, for low back pain, most of the time you are treating those conservatively. For a fractured hip, most of the time you are going to surgery.

Here is the one where there probably are a lot of different conservative treatment options that we are ignoring and going straight to the subset of patients that are having surgery.

So going back to the lost work, you may be losing a lot of healthcare expenditures with regards to this conservative treatment by focusing on the subset of

patients that are going into the operating room.

But again, that's not really what this measure is looking at. It's not looking at the larger set of patients.

CHAIR WEINSTEIN: And you really need to look at that article by Gillian Hawker and Jim Wright and others that was done years ago from Canada and Ontario. And you can argue whether it is right or wrong, but I think NQF's obligation as a quality group giving the nation measures in these kinds of preference-based decisions needs to get into patient preferences somehow, whether it is through shared decision making or some other methodology, because this is a great procedure for the right person.

But the complications can be significant and the cost huge. And when you start to do this in people that, you know, it gets back into Windberg's work originally on tonsillectomy and hysterectomy, you know, if

people don't have problems, they do pretty well. But should they really be done, you know? So we need to at least underline that. At least I would like to as a Committee Member.

MR. AMIN: We will make sure that is in the report.

CHAIR WEINSTEIN: Any other opening comments, by any of our other colleagues? Can we get to the work?

So does this measure focus address a specific national goal? So is this an important condition? I think most of us would say with the increasing rates of these procedures and the cost issues in an aging population, the answer would likely be yes, but we should all make that decision.

MR. AMIN: That's five moderate or five high, one moderate.

CHAIR WEINSTEIN: Was the data submitted that demonstrated considerable variation or overall less than optimal

performance across providers or population groups, disparities in care?

Comments from the group?

DR. RATLIFF: It's a lot of data presented in their submission, but not so much data about hip and knee replacements. More generalized like data about patients who are sick and seeing a doctor for some reason. So I don't know that specifically that relates back to a patient choosing to undergo this elective orthopedic procedure.

I mean, I know it does. It's just that data is not really in their submission.

CHAIR WEINSTEIN: And I didn't -- I guess it didn't specify a cost measure, to me. It gave guidelines, but no recommendation. The process, to me, was very complex and hard for me to follow or explain.

The rankings are slightly confusing. In some cases, your lowest number was the strongest association and in some cases your highest number was the strongest

association.

And you assume coding is consistent between facilities and it not necessarily is, it's common. And you did not address specific resource utilization within a procedure or E&M visits type of provider, et cetera, and you did not address non-billable activities in these processes.

So those were things I found problematic in the performance gap.

MS. TURBYVILLE: Can I just --

CHAIR WEINSTEIN: Please.

MS. TURBYVILLE: -- note? For this performance gap, what you want to keep in context is does the -- in this particular measure, focus area, it's not whether the measure is constructed as doing these things.

CHAIR WEINSTEIN: I'm sorry.

MS. TURBYVILLE: So for the importance criteria, try to keep the kind of thinking about the area in which it is examining. So did they provide literature or

did they give you some distribution information indicating that there is an issue there, whether it is high variation or the variation is --

CHAIR WEINSTEIN: Yes. They did.

MS. TURBYVILLE: Right.

CHAIR WEINSTEIN: But they didn't give a preference issue, yes.

DR. RATLIFF: Where is the variation data?

CHAIR WEINSTEIN: On page -- early in their discussion about the procedure, they had some data. It's in this page here where they talk about OA accounts for 55 percent of all arthritis, da, da, da, hip/knee joint procedures accounted for 35 percent of the procedures from 1990 to 2000, age-adjusted rates of total knees in Wisconsin increased 81 percent from 160 per 100,000 to 294 per 1,000.

Rates increased among young patients. Cost -- they had some rate data and some references.

DR. RATLIFF: That's a given. But what's the variation between facilities and the variation between practitioners?

CHAIR WEINSTEIN: Oh, no.

DR. RATLIFF: With regards to this procedure.

CHAIR WEINSTEIN: No.

MS. WILBON: So --

CHAIR WEINSTEIN: No. Sorry.

MS. WILBON: So just as a reference using the table, so the submission items, if you are looking at a submission form, that this information should be reflected in are the two, so the IM-2, 2.1, 2.2, 2.3, 2.4 and 2.5. So within --

MS. O'NEILL: So I think that --

MS. WILBON: -- that section is kind of where you should find whether or not they demonstrated that or not.

MS. O'NEILL: So they quoted the variation and rate of the procedure being done over time and in different locales, as Jim

pointed out, but the actual measure, as it's structured, is comparing the utilization of resources between people that are doing the procedure.

So how much variation is there in length of stay, drugs, endoprosthesis, utilization. You know, I mean, that's really what the end reporting is about. So I think that's the conflict.

DR. RATLIFF: Okay. So their point here is now going back to Dartmouth and talking about different utilization of the procedures. You are already taking a subset of patients having the procedure. Where are you showing the variation within that subset when they don't get -

CHAIR WEINSTEIN: They don't address this, but I know it is happening. And I know they must have it in their data. Is there a reason you didn't address it?

DR. DUNN: This is Dan. I'm looking at the slide. We missed the mark on

that specific point. We could follow-up if that's allowed, but you're right, we didn't answer the question.

DR. RATLIFF: So as we discussed earlier. This was cut and pasted from other Ingenix things, where they just took this out and like stuck it into this document, because they didn't to, you know, frankly put forth the work to like look up these citations.

And we all know that data is out there. They just are not presenting it to us.

CHAIR WEINSTEIN: Yes, but like you said, I mean, it's in their database. They have these various providers across these organizations.

DR. RATLIFF: Right.

CHAIR WEINSTEIN: And they have -- they probably have some of the best data in the world on this. Yes. So we can vote. That was a great discussion. Thank you, everybody. It's very helpful.

MR. AMIN: That's one moderate and

five low.

CHAIR WEINSTEIN: Is the purpose objective of the resource use measure and the construct for resource use/cost, over-cost, clearly described? Discussion?

MS. SINNOTT: I would just highlight, and I think I'm on page 3þ, Purpose, they list four items: Payment program, public reporting, quality improvement internal to the specific organization and quality improvement with benchmarking with no further description or narrative about testing or where the research is being -- I mean, it looks to me like these are ideas thrown out rather than reporting on their use.

MS. WILBON: I don't want to sound like a broken record, but again, keeping in mind the importance criteria is about the area that is being measured. So did they describe the purpose of the measure? And then later on, when you get into the details of the measure construction and how it is reported,

that's more the scientific validity, usability.

So this whole section of importance is, again, are they picking up an area that demonstrates a resource use problem? Are they describing what the objective of their measure is, which is potentially to measure the resource use of X condition or surgery? And so that's it.

CHAIR WEINSTEIN: Other comments? We can vote. Do you have to wait for that clock to go down? No? Okay.

MR. AMIN: That's four moderate, one low and one insufficient.

CHAIR WEINSTEIN: Next. Are the resource use service categories included in the resource use measure consistent with the representative conceptual construct represented by the measure?

So do they have the right categories within this measure for this procedure? Any comments by the group? Where

is that? What's the number? So the resource -- what page is that on? I'm sorry.

MS. WILBON: Two of the PDF.

CHAIR WEINSTEIN: Yes, so they have admissions, discharges, outpatient, emergency department, pharmacy evaluation and management, procedures, surgery, imaging, diagnostic and lab. So are those the -- did they include all the right categories? Did they leave something out?

The one thing that happens with a lot of these patients is they go to rehab facilities post-procedure and I didn't see that here.

MS. WILBON: I don't think that was on our list.

CHAIR WEINSTEIN: But it's an important one, because these patients often they try to get them out of the hospital really quick to a rehab facility and it's a transfer of cost. And those are big costs that we need to consider in the management of

these patients.

DR. DUNN: Jim, this is Dan. That is part of our resource use.

CHAIR WEINSTEIN: What's it under in the list then? Is it outpatient facilities?

DR. DUNN: Under -- yes. We have inpatient facility broken up into acute and non-acute.

CHAIR WEINSTEIN: Page 5.

MS. O'NEILL: And then the DME is captured, I saw, in another list. Is that correct?

DR. DUNN: Right. That's not broken out as a separate category, but it's included as part of the cost under a larger category.

MS. O'NEILL: Yes, okay. Thank you.

CHAIR WEINSTEIN: I'm sorry, where did you see the rehab on page 5?

MS. SINNOTT: There is a couple

places.

CHAIR WEINSTEIN: Page 12. I'm sorry.

MS. SINNOTT: And this is a question for Ingenix. Where are the rehab therapies on the outpatient basis?

DR. DUNN: The physical therapy for example.

MS. SINNOTT: And OT?

CHAIR WEINSTEIN: I got it.

MS. SINNOTT: Correct?

DR. DUNN: Yes, that's -- I'm not sure what page this is on, but it's under S-9.7. S-9.7 has both -- itemization of all the resource use categories we included, but that -- physical therapy and OT are broken out as a separate measure category.

MS. SINNOTT: Okay. Thank you.

MS. WILBON: 25.

DR. RATLIFF: Yes, from review of the Excel sheets that you provided, I mean, it seems like a pretty wide net.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: So I think you are capturing what you need to capture.

CHAIR WEINSTEIN: I thought I had read it, but I didn't see it.

DR. RATLIFF: I think it's in here, yes.

CHAIR WEINSTEIN: It is. It is. Thank you. Okay. Do you have everybody set? Good.

MR. AMIN: That's two high and four low.

MS. WILBON: Moderate.

MR. AMIN: Oh, and moderate, four moderate.

CHAIR WEINSTEIN: Is the measure precisely specified so it can be implemented consistently? Any discussion on this?

MS. SINNOTT: This is Patsy. There is a discussion about an eligibility table and the strength of the clinical relationships and assignment to diagnostic

classes specific or not, but all of that is not detailed to be repeated by anyone other than Ingenix.

The clinical logic that goes into tying events to events to create an episode is not described. It is not even described as a consensus process among physicians or a consensus process or a research into the data to see how things link up.

So that's my concern.

CHAIR WEINSTEIN: Any other comments?

MR. AMIN: That's two moderate and four low.

CHAIR WEINSTEIN: Does reliability testing demonstrate that the results are repeatable, producing the same result a high proportion of the time when assessed in the same population in the same period of time and/or that measure score is precise?

Any comments?

MS. WILBON: So again, before you

guys move on, if we could just get a little bit more, yes, explanation of the lows? I know Patsy talked a little bit about it, but is the -- did you feel like the way that they were written, that they weren't clear or that you feel like it is only Ingenix can repeat or can actually use? I guess I'm just looking for a little bit more, I guess, to that.

MS. SINNOTT: For a measure that is supposed to be fully transparent, all of that clinical logic should be there.

MS. WILBON: Okay.

MS. SINNOTT: And it's not.

MS. WILBON: So that was kind of a general, everyone kind of agreed with that?

CHAIR WEINSTEIN: I think they give guidelines. They are not as clear. It is not obvious to the reader what they are actually using. They have got very sophisticated formulas, hard to interpret to the novice and it's not clear what clinical information they have included within these to

make these determinations.

MS. O'NEILL: And if it's a proprietary measure, there is no -- I mean, the expectation is that they wouldn't fully divulge exactly how they get where they are going, right?

MS. WILBON: Not necessarily. So I may just have Dan talk a little bit about this, but, for our process, we do ask that -- you know, in order to enter the process, they do have to submit the specifications such that they can be -- you know, that a Committee, such as yourself, would be able to evaluate the strength of the measure.

And in doing so, they should be submitting it clear enough in a way that you feel like you would be able to duplicate it. However, there are some proprietary issues with this particular measure that actually operationally doing that, there are some limitations to that.

DR. DUNN: Yes, this is Dan. Yes,

the intent wasn't to hide anything. And, you know, the intent is to make it transparent in a way that was described. You know, proprietary or not, you know, what is being measured and using the measurement need to understand fully what is being done. So that wasn't the intent.

CHAIR WEINSTEIN: But you mentioned software, so there is an assumption that there is a system that does the mapping and the signing of all these diagnoses and the procedures. It's not a manual process, but -- and we understand people are using it, but it's not apparent.

DR. DUNN: Yes. One is obviously the software is following specifications, which is what, you know, the intent here was to describe that at a level that could be interpreted by you folks and others.

And then there is a set of software that, you know, embeds that logic and people apply it against administrative claims

data and returns results.

Now, as a note, you know, there is now no one who starts from the specification and goes up and tries to recode the logic themselves. It has just been easier, you know, for folks to use this software, rather than do that.

DR. TARKO: This is Howard Tarko. Could I make a comment and just a point of clarification? Is the issue that you are not exactly sure how these eligibility tables were generated? Is that the question?

MS. SINNOTT: That's part of it.

DR. TARKO: There is a -- we have a physician review panel that reviews all of these relationships one by one and so there is no automatic process that was used in creating these tables.

There is currently a process going on right now where all of the diagnostic ETG relationships are being reviewed by a panel of specialists. So this is not done

automatically at all. It is done by physicians using clinical judgment.

CHAIR WEINSTEIN: Yes, we just can't tell that from this.

DR. TARKO: Okay.

CHAIR WEINSTEIN: So --

DR. TARKO: All right.

CHAIR WEINSTEIN: I mean, is that a fair statement or do you think it's different?

DR. TARKO: No, I think that's a fair statement.

CHAIR WEINSTEIN: Yes. Does the reliability testing demonstrate that the results are repeatable, producing the same results a high proportion of the time when assessed in the same population?

MS. PAXTON: I was wondering if the developer could comment on the internal consistency measure? They did a great job explaining how the measure could be reproduced in different populations. But also mentioned

regression models. Could you explain that process?

DR. DUNN: This is Dan. So the question is related to our internal testing of the ability of the measure to be, I guess, both matched in a validation perspective, as well as, you know, being applied to the same set of data multiple times and getting the same results?

MS. PAXTON: Exactly. The reliability issue.

DR. DUNN: Right. So many as a note, you know, given the software application, you run the same set of data through, you know, multiple times and you will get the same answer every time.

As a related point, if you -- you know, one of the steps is that you need to validate that the software and the measure are working appropriately. As part of that, we will parallel code against the software using SAS, for example, at the end test and to

result that alignment with a 99.9 percent accuracy, you know, with claim lines being the -- but the measure matching on claim lines at 99.9 percent accuracy.

MS. SINNOTT: What do you mean matching?

DR. DUNN: Meaning you get two different processes. One is the software use, what people would use in practice. And the second is a parallel interpretation of the specification by someone who isn't involved in the process, who is writing code. And then the match is if you have 10 million claim lines, and if you compare the results from Approach 1 versus Approach 2, the match rate has to be at 99.9 percent or higher.

And also, you know, when we evaluate the differences, they are determined to be, you know, random in nature, that there is nothing to be concerned about.

MS. SINNOTT: So when you say matching, you are saying that it matches the

number of orphan claim lines or --

DR. DUNN: SAS.

MS. SINNOTT: Is that right?

DR. DUNN: It matches, yes, exactly the episode that it was assigned to.

MS. SINNOTT: Okay. So you are talking about episode attribution across the entire data set. So not specifically for the total joint replacement?

DR. DUNN: Correct, correct. Although, one of the assessments is doing that calculation separately by ETG and it has the same level of required of matching.

MS. SINNOTT: Okay. And but when your -- you are saying that when you do these two methods to run the data, run all the claim lines through, you are getting the same grouped episodes, the same number of episodes, the same number of orphan claim lines, the same attribution for physician for an episode, the same outliers are excluded and the cost assessments for the episodes are the same?

DR. DUNN: The -- on the first part of the metric, I was quoting you, is based on the grouping of SAS and attribution and then it will end up into a physician's score or different component. That actually goes to the same process or that same level.

But I was talking about the actual grouping of the information, the two different, again, approaches. And if you look at every single claim record, what episode of -- what unique episode went to what ETG was assigned to that episode, what risks or severity level was assigned, so on and so on.

But that was the matching I was describing.

MS. SINNOTT: But you haven't included a narrative about the physician scoring, right?

DR. DUNN: Right. And that was our attribution adjusted. Well, actually on this one, attribution is forwarded as the primary surgeon of the hip or the knee

replacement.

And on the scoring itself, we described, you know, the approach that was used. That's maybe a different question relative to measures.

MS. SINNOTT: Right. But you haven't talked about the reliability of the physician measurement. In other words, that there is -- that the physician -- in repeated samples, one physician would end up with, approximately, the same score.

DR. DUNN: Right. So repeated samples of the same episodes.

MS. SINNOTT: Yes.

DR. DUNN: Repeated iteration of the same episode. Yes, so that's the same type of testing and reliability that is done with that same threshold.

MS. SINNOTT: But you haven't reported on the physician part of it in this response, as I understand it.

DR. DUNN: Well, the quote is --

or the 99.9 percent is based on the assessment of the grouping itself. You're right.

MS. SINNOTT: So you are saying that not the -- the 99.9 percent of the time, the physician gets the same efficiency score in repeated samples of the same data set?

DR. DUNN: And by samples, again, I think just to be clear, it's, you know, if you run 100 episodes attributed to Dr. Smith through one -- whatever, the software approach and then are those same 100 episodes attributed to Dr. Smith from beginning to end through the SAS coded prototype parallel process, you will get that match rate.

MS. SINNOTT: Okay.

DR. DUNN: The 99.9 percent. That's sort of a standard threshold we used matching 100 percent by the time we are done almost across the board.

CHAIR WEINSTEIN: And can I just question that? It just seems like that's not possible, because every physician has their

own variability, you know, within these measures on any given patient.

And to think that the utilization of resources is the same --

MS. SINNOTT: Well, but what they are saying is if they have a cash of data and they run it simultaneously through the SAS setup and through the group, they are going to get the same results.

MS. WILBON: On the same case.

MS. SINNOTT: On the same patient population.

CHAIR WEINSTEIN: Sort of a bootstrapping. I understand that part being reliable.

MS. SINNOTT: Right. But it's the year-to-year reliability that isn't reported here. In other words, how reliable is a physician's score based on the population of episodes that goes into the scoring mechanism?

DR. DUNN: Well, that's -- I think you are accurate in describing what we are

reporting on, which is, you know, that those-- if you take the same set of data, run it through the measure, is it going to give you the same --

MS. SINNOTT: Simultaneously.

DR. DUNN: -- result, is accurate. But we weren't really responding to the question of -- which you could take a whole bunch of different ways, you know, but that bootstrapping, you know, the 100 episodes and you are pulling them out 20 at a time in repeated sampling or the year-over-year, I didn't think that was the point of this question.

But, you're right, we didn't address that.

MS. SINNOTT: Yes.

DR. DUNN: I don't think that was the question.

MS. PAXTON: Right. I do think that's critical to address all the measures on the concept that a software program is

reproducing and is not, you know, reliability. So I think that needs to be considered in all the measures.

DR. DUNN: And just as a note, you know, in responding to the template, you know, we had asked that question and reliability wasn't --

CHAIR WEINSTEIN: I think the issue is that wasn't part of the --

MS. SINNOTT: Question.

CHAIR WEINSTEIN: -- requirements of NQF for the organization to provide. So I think had it been, they would have done it, if I'm not misunderstanding.

DR. DUNN: No, that's accurate. Thank you.

CHAIR WEINSTEIN: Yes.

MS. PAXTON: It should be considered in future projects to request that, because it is really critical that these measures are sound in terms of applying them, especially to the physician level.

DR. RATLIFF: But I think they answered the question that was asked or can I ask them in a different question now, and our question is important, too, but that really wasn't posed by the NQF when they sent out this request.

CHAIR WEINSTEIN: So I think we can vote.

MR. AMIN: That's two high and four moderate.

CHAIR WEINSTEIN: What is the level of overall reliability testing precise specifications and reliability testing based on what we have just talked about?

MR. AMIN: That's two high and four moderate.

CHAIR WEINSTEIN: What is the next question? Validity. Okay. Does everybody want to have a break for lunch? Do we have to vote on this? Can we vote? High. Okay.

MS. WILBON: So let's take -- let's do a brief public comment. I know we

have got someone here in the room and some people on the phone, so we will start with those on the phone.

Is there anyone on the phone who would like to make any comments or ask any questions? Okay. I'm taking silence as a no.

Anyone in the room?

MR. MARTIN: I just wanted to thank the panel for taking time out of their busy schedules to work on this. It is incredibly important to our members at the American Academy of Orthopedic Surgeons and so I congratulate you and applaud you on your efforts.

MS. WILBON: Thank you. And on that note, we will take a few minutes. Okay. So it looks like we are going to do a working lunch.

CHAIR WEINSTEIN: Yes.

MS. WILBON: So we will break for about 10, 15 minutes to get food and come back and then we will pick up with food in about 15

minutes. Thanks.

(Whereupon, the meeting was recessed at 12:37 p.m. to reconvene at 1:00 p.m. the same day.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

1:00 p.m.

CHAIR WEINSTEIN: Okay. So are the measures -- we are on total knee still.

Are the measure specifications consistent with the evidence? Any discussion from the group?

MS. SINNOTT: I just had a question about why the low cost outliers are excluded and the high cost outliers are winsorized?

I wonder if the Ingenix folks could respond, if they are there?

DR. DUNN: Sure. This is Dan. That has pretty much been a convention around all of these measures. Logic being is low cost outliers may be some indication of, you know, missing data, missing services. You know, the episode doesn't make sense as to logic on the low end.

On the high end, you don't want to exclude them, because you, you know, are

potentially giving an advantage to someone being measured who has a lot of high cost outliers.

So the idea is you winsorize them, so you are measuring up to some dollar threshold, but still including those episodes in the measurement.

MS. SINNOTT: But there is no test for the low cost. I mean, is it measured at a comparison to the mean or is it just the bottom two get thrown out?

DR. DUNN: The bottom -- yes, there is a threshold. I'm sorry if this isn't the question. There is a threshold which defines, yes, the dollar amount that a low outlier is defined as the same thing on the high side.

And, you know, the argument is you exclude the low outliers from the measurement, so that they are put aside and not included in the creation of the physician score with the logic being that those episodes probably have

some other issues related to data.

CHAIR WEINSTEIN: I mean, the easiest thing would be to include them and see if it changes the result. So do you get the same result when you include those lower expense or did you find out that -- from a sensitivity analysis or something, and I know it's not required, but you didn't arbitrarily eliminate X numbers of people because of their cost or did you or what was your methodology, I think, is the question for making that determination?

DR. DUNN: Yes. The methodology for determining what the low cost outlier threshold is based on, you know, distribution of statistics, like the bottom 2.5 percent.

CHAIR WEINSTEIN: So you had a frequency distribution and you took two standard deviations and you said, you know, at three, they are out or something?

DR. DUNN: Yes. And that's to determine that dollar amount that is kind of

applied as a standard. So that is repeated every time you run this for a certain population. That's usually done as -- even though some customers do recreate their own outlier thresholds, we include outlier threshold as part of the methodology.

And then the next step was to look and see what are those episodes that got excluded? Do they make any sense? You know, in this case, you know, do they have -- you would expect the hospital stay and, you know, the surgeons and so on. And in a lot of the cases, those outliers in -- you know, are below that threshold.

CHAIR WEINSTEIN: So it just practically didn't make clinical sense when you had your consensus panel look at the data and said this doesn't make sense. How could they only be in this hospital 10 hours and not have an x-ray, whatever the reasons were?

DR. DUNN: Right. Or a $5,000, you know, knee replacement doesn't make any

sense. Exactly right.

CHAIR WEINSTEIN: Yes.

DR. DUNN: And on the high side, the high side is more atypical. Not -- maybe a good signal on how well a physician is doing.

CHAIR WEINSTEIN: But, Patsy, are you just asking for the methods by which they made those determinations?

MS. SINNOTT: Well, yes. I mean, the inclusion of the high cost outlier, the problem is that that's not going to be equally distributed across all surgeons. And so if you have one high cost case, and you've only got 30 cases that you are being measured on, that's going to affect your quite comparative score if nobody else has a high cost.

CHAIR WEINSTEIN: And even by location, you could have a high cost place where all the docs are high cost. You would want to understand that as not representative of the sample across.

MS. SINNOTT: Geographics.

CHAIR WEINSTEIN: Right.

DR. DUNN: But again, you know, the --

CHAIR WEINSTEIN: But you didn't see that kind of distribution or you didn't look for it?

DR. DUNN: The logic was you wouldn't want to throw them out to say the threshold was $50,000. You know, throwing out cases at $55,000, but there is a number of surgeons who have cases at $49,000 and it wasn't fair. So the compromise is let's only measure the first $50,000 of these costs.

You're right, some surgeons may have more outlier cases. You can count them up, you know, as part of the investigation in the results. But I guess I would argue you don't want to throw them out.

CHAIR WEINSTEIN: Yes. Well, it's so rich to actually look at this and understand it, I think, is Patsy's point. And

that information could be incredibly valuable in starting to understand episodes.

So there is not a criticism. It's just the value of not including it or the value of including it becomes an important discussion. Other comments?

DR. DUNN: I --

CHAIR WEINSTEIN: Do you understand?

DR. DUNN: Yes, I agree.

CHAIR WEINSTEIN: Yes.

MR. AMIN: And we will also take a note of that in our minutes.

CHAIR WEINSTEIN: Yes. Other comments? Oh, sorry, Craig?

DR. RUBIN: It's a different question. Do you have the ability to report the resource use in those between 63 and 75 versus 75 and 96, refer to age rates of 63 to 96?

CHAIR WEINSTEIN: They're segmented. That's how segmented population

looks. Do you have the ability to do that? I'm sure the answer is yes.

DR. DUNN: Yes. And any patient or episode, clinical attribute, you can, you know, process the data and upset at -- that's usually part of the investigation people do to, you know, get behind the overall results.

CHAIR WEINSTEIN: But I think that it's a significant point. Yes, please, go ahead, Craig.

DR. RUBIN: Well, just I didn't see that and certainly if you are comparing populations, you know, there is a lot of comorbidities in the -- that's a 30-year plus range and I just didn't see that in the materials where that was being looked at. But could be a major finding of importance, depending upon the makeup of your patient population.

CHAIR WEINSTEIN: Did you adjust for comorbidities?

DR. DUNN: Right. So --

CHAIR WEINSTEIN: Yes.

DR. DUNN: -- age, gender, comorbidities and condition status factors.

CHAIR WEINSTEIN: Yes, yes.

DR. RATLIFF: How do you adjust for comorbidities? We're kind of getting ahead of ourselves though.

CHAIR WEINSTEIN: Yes, yes, yes.

DR. RATLIFF: Yes, but there are methods of weighing the different comorbidities. They can tell us how they did it.

CHAIR WEINSTEIN: Any other comments on this one? Okay.

MR. AMIN: We have to vote again.

CHAIR WEINSTEIN: We have to vote again. Do you need me or somebody?

MR. AMIN: It's two high and four moderate.

CHAIR WEINSTEIN: Validity. Does the validity testing demonstrate that the measured data elements are correct and/or the

measures score correctly reflects the cost of care or resources provided adequately distinguishing high and lower cost or resource use?

I wasn't actually sure about the specificity of the cost measures. I mean, you didn't give real recommendations there.

DR. DUNN: I'm sorry, whether there are any recommendations on what the cost measures were or what the measure of cost was?

CHAIR WEINSTEIN: You didn't specify them.

DR. DUNN: Well, there is the resource measure is what, you know, cost overall -- you know, cost by type of service. And we --

CHAIR WEINSTEIN: Yes.

DR. DUNN: Go ahead.

CHAIR WEINSTEIN: No, you go ahead. Sorry.

DR. DUNN: And then we weren't specific, if this is what you are getting at,

on, you know, whether you use standard price costs versus, you know, allowed amounts, for example.

CHAIR WEINSTEIN: Exactly.

DR. DUNN: Okay. And I know there are people who do both and actually compare them and some that do one or the other.

CHAIR WEINSTEIN: Do you have a preference or how did you actually do it?

DR. DUNN: I think -- well, when we shared the -- some results with you as part of the submission, that was based on the standard price, because our benchmark data, you know, needs to be standard priced to be able to put things together, the cost, the different contributors.

But my preference is actually for both. I think if you do standard cost, you know, it does get around that question of being able to look at utilization and treatment decisions, practice patterns, but the real cost and -- you know, does reflect

many times choice of facility, choice of device, you know, things like that.

CHAIR WEINSTEIN: The kind of things Mary Kay was talking about earlier.

DR. DUNN: Yes.

CHAIR WEINSTEIN: Any other comments?

DR. RATLIFF: I'll bring up one point. A lot of your kind of final results go to the individual surgeon performing the procedure that your PEG is like associated around.

You have the near and further, I believe, arms for preoperative evaluation with the further being six months. So then your surgeon is going to have attributed to him cost accrued by the patient in the six month period prior to the procedure being performed.

So that seems, to me, to be a more valid or more representative of the efficiency of say, a health care system or a local practice environment, not so much the

individual surgeon whose outcome measure is going to be influence by that further assessment.

DR. DUNN: The FCI argument is from their side that window on the beginning part is too long. And this is a no given, the -- given the way the logic works, you know, there is a concept called the Close Windows and then the Further Windows.

And the Close Window -- I need to look this up quick, but I believe that's 14 days before.

DR. TARKO: That is correct.

DR. DUNN: Thanks, Howard. And then the Further Windows on the beginning side, it has to be a specific procedure code that makes sense relative to the surgery. So you really don't get a lot of -- unless it is something that, obviously, would be related, like an MRI or some other test to inform the decision on the procedure itself, it isn't likely going to find any services that relate

here.

But the things within the 14 days, I think, we probably agree make sense. So it's a, you know, valid point that the beginning part is likely going to -- the things that are happening to the patient may be out of the control of the surgeon, but the way the logic is constructed, it's pretty -- whatever. It's pretty exclusive on the types of services that actually become part of the episode during that, you know, longer preperiod.

MS. SINNOTT: So are you saying that primary care management or PMnR management prior to the referral to surgery and then physical therapy or occupational therapy would likely not be attributed to the surgical event or the surgical episode?

DR. DUNN: Physical therapy would, that's one of the targets, there is a target procedure code and then physical therapy is a target procedure code. Pain management would,

MRIs would, x-rays would.

DR. RATLIFF: So E&M visits, physical therapy, MRIs, probably injections. E&M Codes, once the patient has his as a diagnosis code, it's going to show up on every single E&M they have from the PCP. So it's probably going to be tagged and pulled out, all of which is going to be attributed to this PEG if you are doing the further preoperative evaluation metric.

And I'm just saying again, not -- I'm just saying that there may be a lot of variation there that has little to do with the procedural efficiency itself.

DR. DUNN: Yes, and that -- E&Ms actually would not be applied to that 14 day window, but some of the other examples you mentioned would be.

CHAIR WEINSTEIN: Yes. The issue here is as you accept or don't accept this methodology for the episode. The system's efficiency or inefficiency in getting the

patient to treatment, should they want it, in a timely way with things that matter, I mean, you could be on all kinds of medications that are extremely expensive and that's a burnup period, have lots of images that have no real impact on the then surgical procedure and then the follow-up.

So the attribution model -- I don't know how to get around this, because this is what happens. But I'm thinking out loud with you, which probably deserves more discussion.

You know, when you get to this data, you want to sort of get to what is the ideal efficiency and effective episode for the average patient. And you sort of laid out a structure for that given what you perceive is the average, not necessarily the best. Is that fair?

DR. DUNN: The average and average meaning that's the timing and the --

CHAIR WEINSTEIN: Yes.

DR. DUNN: That's the timing. That's fair, yes.

CHAIR WEINSTEIN: Yes.

DR. DUNN: And then that logic piece was designed, this is probably not average here, but try to focus on what makes sense to include differently depending on the timing.

CHAIR WEINSTEIN: Yes, and with no real outcome data, you don't actually have some measure of effectiveness or value at this point.

DR. RATLIFF: So again, what I think you are commenting on is the system's efficiency.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: Not the procedural efficiency.

CHAIR WEINSTEIN: Or not the doc, not the surgeon's efficiency, necessarily.

DR. RATLIFF: Yes. I agree with that.

CHAIR WEINSTEIN: Yes. Is that how you guys see it?

DR. DUNN: Yes, on average, at least a small percent of the dollar is conducting that type of --

CHAIR WEINSTEIN: Yes. And most of the dollars are going to be to the device.

DR. DUNN: Yes.

CHAIR WEINSTEIN: The length of stay, the operating time.

DR. DUNN: And then things that you don't want to happen, that happen on the back end.

CHAIR WEINSTEIN: Yes.

DR. DUNN: Right.

CHAIR WEINSTEIN: But in the average case, it's going to -- the big costs are the length of stay, the device and the time in surgery. There's no question about it.

DR. DUNN: Okay, yes.

CHAIR WEINSTEIN: Unless you -- do

you have different results?

DR. DUNN: No, you're right. It's probably close to 90 percent of the --

CHAIR WEINSTEIN: Yes.

DR. DUNN: It depends on other things, at least a dozen here.

CHAIR WEINSTEIN: Yes.

DR. DUNN: Not one.

CHAIR WEINSTEIN: Okay. Can we answer this question?

MR. AMIN: That's one high, four moderate and one low.

CHAIR WEINSTEIN: Next. So you guys never ask questions about when we have something high. You only ask questions --

MR. AMIN: Well, I was hesitating on that one. In fact, that was -- I'm not sure that -- but --

CHAIR WEINSTEIN: That's okay. We will keep going.

MR. AMIN: Yes.

CHAIR WEINSTEIN: But feel free

to. Are exclusions supported by the clinical evidence or analysis of frequency and distribution?

Do I understand that question? Are exclusions supported by the clinical evidence or analysis of frequency and distribution? Is information about the impact of exclusions for patient preference transparent?

It's not there. You don't have that information, Part B of that or Part 2 of that. So the upper part of that question are exclusions supported? Any comments on that?

MS. SINNOTT: Only that we are back to the kind of diagnostic classification and the, you know, black box in this of the whole system and how the diagnostic hierarchies work. I mean, granted this is a procedure-based episode definition, but we still don't know how, you know, this episode -- let's say we have a total joint replacement and the patient gets pneumonia, is that in or

out of the episode? Do we know that?

CHAIR WEINSTEIN: I think it is in.

I think it is in. You guys should -- can you comment on that, the designers?

DR. DUNN: Yes, sure. This is Dan. I'll let Howard add to this. So the service -- if you think of the way the logic is working, it is creating a condition episode, which is a joint degeneration episode, the way this one works. And then it is looking at the procedure episode within the context of that condition.

So only things that group to that condition episode are going to be, you know, on their way into the total knee or the total hip replacement. So the pneumonia would not be included, unless, you know, it happened during the course of the inpatient stay and made them, you know, stay in the hospital longer, for example.

CHAIR WEINSTEIN: Only in

hospital? There is not like a 30 day window? You don't have a window?

DR. DUNN: A window?

CHAIR WEINSTEIN: Because this episode goes beyond the hospitalization.

DR. DUNN: Right. But the service of this is that happened, you know, within -- part of the windows are only those services that relate to the condition itself.

CHAIR WEINSTEIN: Yes, but they didn't have pneumonia when they came in to get their total knee replacement. They developed it post-op, which could be possible. They could have aspirated or something. I don't know.

DR. TARKO: May I comment on that?

CHAIR WEINSTEIN: Yes.

DR. TARKO: Maybe -- what would happen in the ETG methodology is there would be a separate episode from the pneumonia that would be created. It would be considered a comorbidity of the procedure and would

contribute to the severity model in that sense, because comorbidities can't cross episodes.

DR. RATLIFF: Let me ask that a different way. A patient gets a knee replacement and gets a post-operative pneumonia. On post-op day 6 and has to be readmitted to the hospital for inpatient treatment of their pneumonia after they have had, say, a hip replacement.

How does that factor into your model for increasing the cost of that index procedure, the hip replacement?

DR. DUNN: Yes. This is Dan. And unfortunately Tom Lin would be the best person here. We can follow-up on this. My interpretation is that if that admission is for pneumonia, it would not be included in the replacement. The cost of that admission would not be included in the replacement episode.

DR. RATLIFF: As a proceduralist, let me say sweet as not responsible for any

post-operative medical complications. That's wonderful.

DR. RUBIN: Okay. I have a different question. A little bit cleaner maybe. So the patient comes back in three days later with a pulmonary embolism, how is that handled?

DR. TARKO: In the methodology, there is a -- I'm not sure if that was in our presentation, but there is the concept of a consignment and the consignment is associated with an episode and that would be included within the consignment, even though it would create another episode.

CHAIR WEINSTEIN: Let me try to help out here and tell me if I'm wrong about this. But a lot of the payers, maybe United, they are thinking of the DVT pulmonary embolism or infection as a new episode potentially.

But the severity adjustment, which he was started to allude to, might take that

into account. On the other hand, if you organizationally said I'm going to do total knees and take a bundle payment and go at risk for any readmissions, then that's a different story.

DR. DUNN: It --

CHAIR WEINSTEIN: Because I think there is sort of apples and oranges here. The surgeon might say okay, that DVT had to be related to the pulmonary embolism and had to be related to my hospitalization for my knee. I don't think that's the question in this grouper design, but can you guys explain that?

DR. DUNN: Yes. I think the point that you are on, just as background, is they have a discussion that took place out in California through IAK where they were looking at both the knee and hip replacement specifications we are looking at. And, you know, some of those readmissions if they are not, you know, obviously attached to a reoperation or, you know, something with a hip

or a knee, the lead diagnosis, you know, would -- it has to be part of other episodes.

The discussion around readmission actually became what else do we want to add to this, either as an outcome measure or as, you know, part of the cost of the episode itself. So it is -- if it's not, obviously, clinically related, it becomes a new episode.

MS. SINNOTT: I guess we are struggling with not obviously clinically-related as a concept or at least I am. That, you know, if I have a total knee replacement and I get a DVT, I think that's clinically-related. And the payers will.

So here is another question. First, it is now five months after surgery and I have a 30 degree knee flexion contracture and I need to go back in to have it manipulated. And how is that gathered or not into the surgery episode?

DR. DUNN: That would be Howard. Would that be under physical therapy as a

follow-up procedure?

DR. TARKO: Or a --

MS. SINNOTT: It's a surgical procedure.

DR. TARKO: -- separate procedure like a release. I believe it would start a new episode. If it were a manipulation, it would be a target procedure and go to the original episode.

CHAIR WEINSTEIN: I don't think we are going to answer all these just as a point of interest, but because, you know, for the hospital they would like to start a new episode and have a new payment. From the -- well, they could say the patient wasn't compliant with their exercises and, therefore, that's why they got the contracture.

They could say they didn't mobilize, they didn't take their coumadin. I mean, who knows the reasons. So these get sort of -- yes?

DR. RUBIN: Critically important,

too.

CHAIR WEINSTEIN: Yes. But speak up.

DR. RUBIN: Well, no, I think it is extremely important, because you have two hospital systems and one has a low rate of these complications that are clearly related and the other doesn't. There are interventions that you can develop and patients that have choices.

Besides, I know this tends to be search eccentric, but, you know, there is other consequences in terms of the costs of this problem that need to be described. And while I have the mike, I guess we are talking about a lot of things that need to be adjudicated.

And it's not clear to me who actually makes these decisions in terms of looking at a finding and saying well, it is or is not linked to the prior hospitalization. Is that an individualized decision or is there

certain training, so that it is done somewhat uniformly?

CHAIR WEINSTEIN: There is --

MS. SINNOTT: I believe it is built into the software.

DR. RUBIN: Okay. There is a lot of references in terms of, you know, it will--

DR. RATLIFF: It's not clear.

DR. RUBIN: -- it seems to be by choice. And computers don't usually, you know, make the calls.

MS. SINNOTT: Choice of what?

DR. RUBIN: Well, for example, you know, the description of somebody with complication, deeming it related or not related.

MS. SINNOTT: Oh. I think that this leads to kind of a larger question, for me, which is is there a place where all these relationships are delineated, so that a surgeon could go or a user of the methodology of the software could go say, okay, I

understand this clinical logic.

Pneumonia is or is not part of the episode. DVT is or is not. So that it's not just -- I mean, even you folks on the phone are not 100 percent clear how the episode logic is being built. And I think that is -- what is interesting, to me, is as someone evaluating for a public use methodology that all this clinical logic should be accessible in some way.

DR. DUNN: Yes. And I apologize for not having the right person, Dr. Tom Lin would be the right person to help you understand this. And, you know, what I would go to is to go to the code sets which were submitted and it's either diagnoses that would map to the joint degeneration episodes, which then drive what ends up in the knee and hip episodes.

So pneumonia isn't one of those diagnoses. You know, some musculoskeletal-related diagnosis for hip and the knee would

be. It's pretty clear in the set of code tables what is in and what is not. I think our challenges, off the top of my head, are -- I would say without Tom on the phone, I'm not able to tell you exactly.

DR. RUBIN: So I could not find in your -- I may be looking in the wrong location. The S-5 joint degeneration hip/knee codes for PE, acute MI, post-op, wound infection, pneumonia, all common complications from these procedures. So it would be helpful to know if they are there somewhere.

DR. TARKO: They are not part of the code set for those particular -- for this particular measure, because they would be codes that would begin a separate episode for pneumonia.

CHAIR WEINSTEIN: So just touching in again, are you --

DR. TARKO: That's just the way the methodology works.

DR. RUBIN: Well, except in one of

the papers you referenced, as background, quotes "those complications as being common complications for these procedures." And so it would seem that that should be part of the coding for an episode to capture that, because those are, you know, modifiable risks.

DR. TARKO: Yes, they do affect the risk in that they are comorbidities of the procedure. It's in a separate table. And they will contribute to the severity of the episode through the severity model.

DR. RATLIFF: And just for the Ingenix commentators, before you mentioned earlier that these were outcome measures. They are not outcome measures. They are resource use measures. And if you are not capturing the most common perioperative complications that are driving up resource use, then you are missing something.

That may not be part of your PEG model, but that probably reflects more on a weakness of your model, not necessarily a

weakness or interpretation of it.

CHAIR WEINSTEIN: Yes, let me defend them a little bit, not that I disagree with anything that has been said, but what they are trying to create is an episode grouper for their routine average total knee.

The rates of DVTs could be as high as 24 percent, whether they are clinically relevant or not, it's a high rate of DVTs. The rate of PEs is fairly low. It would be certainly higher than a back surgery.

So in the episode, to include complications or not, and to me this gets into more of a contracting issue, because as you were stating, when you look at the volume outcomes data, people who do high volumes of these things tend to have less complications, less -- lower lengths of stay, et cetera, et cetera, et cetera. And there is lots of data on this.

I'm just -- I wonder if we are, and I would appreciate NQF's help here,

overstepping the episode a little bit, but I don't want to be the adjudicator of this decision. What I want to do is try to answer the question that is being raised by the Committee.

Anybody want to help us out here?

MS. O'NEILL: Well, if we are looking at the resources that can be used affiliated with these diagnostic -- I guess, the surgical treatment of these diagnostic categories, and in a significant proportion of people under going this treatment, those resources are used, the treatment of the DVT, the treatment of the pneumonia or whatever, then those elements need to be captured if that is what, in fact, we are measuring.

And now, I think a lot of our filtering on these things ends up being are we blaming the surgeon for all the things that are happening or are we really trying to get our hands around what it costs to take care of folks with these conditions?

And if we are trying to get our hands around the resource use/cost of these conditions, then these common complications need to be measured.

CHAIR WEINSTEIN: But there are strategies, whether it is anticoagulation, you know, extubation, early mobilization, there is best practices that --

MS. O'NEILL: Sure.

CHAIR WEINSTEIN: -- limit these complications in good organizations. But I don't know that this is satisfactory for the group or for the process. So I want to try to get to some place that is satisfactory and I'm not sure how far to go.

DR. RATLIFF: Two more points and I'll shut up. For your clinical severity levels, you model the severity of procedure based very, very simply MSDRGs and whether or not you have an MCC.

So you have like these severity levels for the procedure you are performing.

Now, so going into the procedure, you have got a stratification for how much you think it is going to cost. But coming out of the procedure, how are you capturing the increased risk of perioperative adverse events that are going to occur, presumably, in your higher clinical severity patients? I'm not hearing that.

And when you go from that to relative risk modeling, where you go through a relatively long explication of your risk adjustment method on page 31 of your PDF, and then that kind of disappears, I don't see where that risk adjustment is brought back in to either your modeling of your clinical severity or of your individual physician's output, for lack of a better word, in terms of limiting perioperative adverse events, having better outcomes in terms of we're losing those adverse events.

So again, going back to my point, I just don't think we are coming to an answer

for this question.

MS. O'NEILL: And I would just say that if you are capturing complications as increasing the risk, that becomes somewhat circular, particularly if the incidents of these complications varies by quality of system.

So we don't want that to be pushed into the risk. We want it to be pushed into the resource use relative to the episode.

CHAIR WEINSTEIN: And what it gets to is --

DR. DUNN: Yes.

CHAIR WEINSTEIN: -- is this the provider level issue that you are comparing? Is it an organizational level issue? Because at the organ -- what you know from all of the volume outcome studies, it's the system and the process. It's not the individual surgeon often times who creates the issue.

So these are all the right questions. I just don't know how to

adjudicate this.

DR. DUNN: Yes. And this is Dan. Maybe to state it simply what we are doing, because I think we are kind of answering it in different ways here. So one is the services that end up grouping to the knee replacement episode as we have defined it, are those that are found in a joint degeneration condition episode.

For things like the pneumonia, some others, you know, vascular complications, those would not group this episode. So those are separate. So those complications, some of them that were mentioned, unless they were, you know, something related to the orthopedic condition itself, would not be included.

The second point is which some noted is correct is the only risk severity adjustment that is done here is based on the MSDRG. So those complications do not drive the risk of the knee replacement episode. Although, they may have use of risk

adjustments for some other episodes downstream, but not for this one.

CHAIR WEINSTEIN: There is another way to say this. For example, to me, please, correct me if I'm wrong, if you had a vascular event in doing a total knee replacement, which happens rarely, that is not part of the episode. You don't get paid for that.

So the hospital length of stay is going to get longer, more procedures are going to be done. The organization is going to have to eat that cost, basically, in that, because it's not part of the bundled payment episode issue, because it's not supposed to occur most of the time. You know, 99.99 percent of the time.

On the other hand, you know, if you get a DVT peri-op, it might be the same issue. And if that happens three months later, it's a new episode because now they have a PE or something that -- it's not supposed to happen in a well-organized, well-

running system.

So I think this is a circular argument a little bit, but I think organizations will worry about what is included and not included, because their payment will be affected by readmission or not, which is, you know, Steve Janks work 70 percent readmission from CMS, you know, big issue.

You know, it's a couple billion dollars. The bigger issue is the chronic conditions, the hospitalizations.

Quite frankly, I think we should go on with the questions. We have had some good discussion. Whatever our answers are will be our answers. And there is no right answer, so unless somebody disagrees, could we go forward, please? Okay.

MR. AMIN: That's two moderate, three low and one insufficient.

CHAIR WEINSTEIN: Next question. For outcome measures, which we don't have any,

is there any evidence-based risk adjustment strategy or rational data support no risk adjustment stratification?

I don't know if outcome measures is the right term there. I think the question they are really asking is there an evidence-based risk adjustment strategy?

MS. WILBON: Right.

DR. RATLIFF: Can I ask a question of the developers? How do your clinical superiority levels relate to your relative risk adjustment or your risk adjustment methodology referring specifically to S-10.1 from page 31 of your PDF?

DR. DUNN: Sure. So the -- I touched on this a bit before this. The severity of risk levels that are assigned to the episode, are they simply on the MSDRG for the admission that the replacement happened within?

And each of those MSDRGs map to a -- let me open exactly that table. I said,

for example, a major joint replacement, an MSDRG for a major joint replacement, a reattachment of the lower extremity without a major complication, comorbidity, is assigned to Severity Level 1. No episodes with that MSDRG for the inpatient stay. They got a Severity Level 1.

On the other end of the spectrum, a bilateral multiple major joint procedure of lower extremity with major complications and comorbidity go to Severity Level 4. And then the other DRGs fell in between.

So the DRG will -- assignment will trigger the severity of the episode and then that will give it, you know, a Level 1, 2, 3 or 4. And then that's what you will see on that S-10 table. Did that help?

DR. RATLIFF: So understood.

DR. DUNN: Okay.

DR. RATLIFF: What's your risk adjustment then? Maybe there was a lot of cut and pasting from like a CHF model on your risk

adjustment method. How does that actually go into like your output with regards to your procedural efficiency? I don't see how these two things relate at all.

DR. DUNN: Yes, so the, you know, general approach to creating an overall risk adjusted measure, so think of the assignment of severity level to risk assessment. So taking the results of that severity level assignment using observe to expected ratio approach, that's where the risk adjustment is happening.

So the expected results for a physician is based on their mix of knee replacement episodes and hip replacement episodes by severity level, as well as the experience of their peers.

DR. RATLIFF: The problem with what you have is, you know, if you were more specific of doing a knee with some comorbidity adjustments, this all seems to be for another project. It's not to criticize you, but it is

a little bit not addressing specifically the knee in the dialogue here with comorbidities.

Obviously, knee patients can have CHF or diabetes and all those kinds of things, but the text does not read as if it was done for this particular diagnostic group.

MS. PAXTON: Well, it seems like there is a lot of opportunity to apply more sophisticated risk adjusted model considering the work that has been done in this area.

CHAIR WEINSTEIN: Other comments? Okay. We will take a vote here.

MR. AMIN: That's five low and one insufficient.

CHAIR WEINSTEIN: Next question. Are performance results reported? Do they identify differences in performance or overall less than optimal performance? Some discussion?

MS. WILBON: So just a quick -- again, this one is about whether or not they have demonstrated that the methods for scoring

an analysis of the measure identify statistically significant and practically meaningful differences.

CHAIR WEINSTEIN: I'll just say I found this very complicated. Just it's hard to follow and even harder for me to explain. So no offense, but I believe it's fantastic work, but I found it very complicated, personally. Other comments? Are you waiting for somebody? Okay. Still waiting?

MS. WILBON: So is the sentiment here that the complexity of it makes it difficult to discern whether or not the score would -- are discerning meaningful -- or you are able to discern meaningful differences based on what is submitted? Is that kind of-- does that reflect the scoring?

MS. O'NEILL: Well, just to try to read what the feedback would be to the physician and figure out what is clinically significant, you know, I mean, it tells something about utilization measurement, but

it doesn't really tell you in that given case, given a particular outcome, that you have applied the right resources. It just kind of counts resources.

CHAIR WEINSTEIN: Next. How are you doing? Okay?

MS. WILBON: We're doing okay. We're doing okay.

CHAIR WEINSTEIN: This is on multiple data sets. Again, the resources by which they use to get their analysis, basically, they used their own data, which is large and quite varied, I'm sure, with 50 million lives or whatever. So that's the question, correct?

MS. WILBON: This is one of those that ends up being a not applicable, because they are actually only -- yes, they are actually only suggesting or specifying the use of one type of data, which is administrative claims data.

So if they were suggesting like

chart review and admin claims data and clinical data, then this would be kind of the multiple data source thing, so that would be NNA. Yes.

CHAIR WEINSTEIN: Next. Validity, what is the overall level of validity from the things we have talked about, specifications, validity testing, risk adjustment, identification, statistically significant meaningful differences and for getting the multiple data sources?

MR. AMIN: That's one moderate and five low.

CHAIR WEINSTEIN: Disparities of care. Do you want to clarify this for us again, because -- how we should be interpreting this? Because I don't think a lot of these things are done either, so, but I may be misinterpreting those.

MR. AMIN: The intent of this criteria is to say that if there are disparities that are identified in this

particular area of focus for the measure, we want to ensure that those disparities are not simply risk adjusted away, but they are actually stratified, so that's the intent of what this criteria is looking to measure.

MS. WILBON: So to provide a little bit more guidance, if you look in -- under the importance criteria, specifically submission items IM-2.4 and 2.5, if in that section they are saying there are disparities with this particular focus area, but then when they go and develop the measure and you get measure results, they are not addressing them, you know, kind of to make that connection.

If you are saying there are disparities, but why aren't you -- or how are you addressing those if you have identified them is the --

MR. AMIN: And just to add a little bit more on that, keep in mind the last portion of this criteria which says that if there is some data justification for why the

stratification is not necessary or feasible, then just keep that in mind that it's not actually possible considering the data that is available.

CHAIR WEINSTEIN: Yes, let me suggest that there are references to disparity with stratified populations, whether it is Hispanic or non-whites and I don't think they had stratified or addressed it.

So I'm just -- that's my own opinion, but others should speak up.

MS. SINNOTT: The data is not there. There is no race/culture data generally available in commercial data.

CHAIR WEINSTEIN: In their data, correct.

MS. SINNOTT: In any commercial data.

MS. O'NEILL: Also, I don't know if this is entire true, but I think it's largely true that the disparities come on the point of surgical decision making, not on the

resource use after the surgery -- surgical decision has been made.

Although, there is some variation in pain treatment, but I think most of the disparities would be evident prior to getting to this PEG episode.

CHAIR WEINSTEIN: From the federal data, Medicare data, there is data that we have published multiple times on disparities and I think you are right. Once you get to that, the rates are different across different groups, ethnic groups. The rates are very different.

And I think in fairness to them, they don't -- they didn't stratify it. They don't have it, as you suggest.

DR. RUBIN: It did include a reference, not in the document, but he reference refers to that box of higher rate mortality, readmission and wound infection effort prior to major knee replacement compared to whites.

MS. WILBON: So just to --

DR. RUBIN: And the statistical analysis, NA.

CHAIR WEINSTEIN: Yes.

DR. RUBIN: You know, so.

MS. WILBON: Yes.

CHAIR WEINSTEIN: I just didn't think they addressed it, so I -- but if somebody thinks they did, please, speak up, because I think it's important. I would change my mind then, because if I'm misinterpreting this or Ingenix speak up.

Did you do this and we are missing it?

DR. DUNN: Well, it's not part of the measure methodology, so there is no risk adjustment. I think someone had mentioned the factors which recognize race or ethnicity or some other attribute like that.

If the user wanted to stratify to do analysis by that and they had that information, they could do that, but, you

know, there was no intent to include that as a risk factor and adjust it out of the measure.

CHAIR WEINSTEIN: So, NQF, are you satisfied that they didn't do it, because they didn't think they needed to?

MS. WILBON: So the criteria allows for them to either build it into their measure or provide a rationale for why it is not feasible. So if that's the case, if they -- if it's not in their data, then they should provide a rationale for why it's not in the data.

CHAIR WEINSTEIN: Could I ask, is it in your data or not, just for my clarification?

DR. DUNN: No. It's not in. I think someone mentioned it's not usually available as part of the information of commercial health problems.

CHAIR WEINSTEIN: Just for the record, that's why they didn't do it.

DR. DUNN: Actually, back to the point is maybe I'm splitting hairs here, I think if you risk adjust -- include it in the risk adjustment, then the ability to assess disparities goes away. I'm not sure you want to do that.

CHAIR WEINSTEIN: About stratifying and not risk adjusting.

DR. DUNN: Yes. Okay. So looking at the results, that way if someone had that information, they could certainly do that using this measure.

CHAIR WEINSTEIN: Which we have done on CMS data, yes.

DR. DUNN: Okay.

MS. O'NEILL: So we're saying, one, because we have a rationale for not doing it?

DR. RATLIFF: Microphone.

MS. O'NEILL: Are we saying that this is high because we have a rationale for not doing it?

MR. AMIN: Okay. 2C references S-10.2 and so if we -- if you felt that there should be justification, the justification was provided in 10.2 or from what we have heard today. So if you believe the justification is sufficient, I would vote as such.

MS. BOSSLEY: You want them to go back at some point and put it in the form, so it's there? Yes.

MR. AMIN: We have one high, three low and two insufficient.

CHAIR WEINSTEIN: They have been testing this at various places, so people are using it, just FYI. Does that make it good or bad? I don't know.

Available to the public, is that happening, Ingenix? At this point, I assume.

DR. DUNN: I'm sorry, available to the public in terms of the actual reports?

CHAIR WEINSTEIN: Yes. Performance results is what the question is asking.

DR. DUNN: There is one organization who uses these procedure episodes for knee and hip who do designate surgeons and I believe that information is available to members of that health plan, you know, the --

CHAIR WEINSTEIN: Is that the -- who is that or you're not allowed to say or what?

DR. DUNN: I would rather not without asking their --

CHAIR WEINSTEIN: Okay.

DR. DUNN: -- permission here.

CHAIR WEINSTEIN: Fine. So but it is available in some way, so that helps us answer the question.

DR. DUNN: Right. That's at least one instance.

CHAIR WEINSTEIN: Okay.

DR. DUNN: Correct, yes.

MS. SINNOTT: Would you clarify -- oops, sorry. I'm looking at page 40. You list a long list of users of ETGs and the

ERGs. Are any -- is any one of these using this particular measure as a stand-alone measure?

MS. ZIELINSKI: This is Cheri. The answer to that question is as a stand-alone measure, no. I mean, our --

CHAIR WEINSTEIN: We didn't get the vote, I don't think. Okay.

MR. AMIN: That's four moderate and two low.

CHAIR WEINSTEIN: Did submitted information demonstrate that results produced by the measure are meaningful, understandable and useful for information for quality improvement and public reporting or credible rationale presented?

MS. SINNOTT: And I want to clarify again that this measure, as a stand-alone measure, has not been used for any quality improvement activities, correct?

DR. DUNN: Well, maybe defined internal, you know, quality improvements, for

example, looking at the results and reaching out to a physician or a group of physicians for discussion. Is that -- would that qualify as quality improvement?

MS. SINNOTT: Yes, it would. But I'm referring to this measure as a stand-alone measure, not as part of a performance profile for a physician.

DR. DUNN: And so that's just a composite, for example.

CHAIR WEINSTEIN: What was that response?

MS. WILBON: As a composite.

DR. RATLIFF: Microphone.

MR. AMIN: Can you repeat that, please?

CHAIR WEINSTEIN: Can you repeat your answer, please?

DR. DUNN: Oh, sure. I was actually trying to clarify the question. I may have answered it at the same time. So there are organizations who will take results

for orthopedic surgeons, for example, and talk with physicians who are, you know, somewhat different than the norm, based on these measures on resource use.

And some of that discussion could be triggered by an overall result looking across all the episodes, these episodes and others, that are included in that provided overall result.

But that discussion will -- or that provider will get to the level of looking at individual episodes, like knee replacements and hip replacements for discussions around, you know, opportunities.

Is it all only -- is the whole discussion only focused on these episodes? I would say probably rarely. It's probably part of a general discussion and performance around these episodes will surface during that.

CHAIR WEINSTEIN: Thank you.

MR. AMIN: That's three moderate and three low.

CHAIR WEINSTEIN: Are the data and result details maintained such that the resource use measure, including clinical construction logic, for defined unit of measurement can be, I hate this one, broken down to facilitate transparency and understanding?

MR. AMIN: It's two moderate and four low.

CHAIR WEINSTEIN: Next. Are the required data that is -- routinely generated and used during care delivery? Do you want to tell us something?

MS. WILBON: Yes. So again, these next two criteria are, again, remember we are just talking about admin data and the ability for them to be generated in routine care and whether or not they are electronic, which is the following criteria, available electronically.

MR. AMIN: That's five high and one moderate.

CHAIR WEINSTEIN: Are the required data elements available in electronic health records or electronic sources, which is claims data? Is what you mean here. If not, is it credible -- one of the things I want to understand is the claims data from United, in this case, versus other claims.

So would CIGNA have the same or is this unique to them or something?

MS. WILBON: I don't think so.

CHAIR WEINSTEIN: Yes, I don't either, but I'm just asking for clarification.

DR. DUNN: Yes. This is Dan.

CHAIR WEINSTEIN: You didn't see anything you need that would exclude others from using this. Kind of -- I have --

MS. O'NEILL: No. I think -- I mean, the only -- there is -- what you don't even want to know about is the platform behavior. I mean, there is stuff that happens in organizations, based on their own quirky software and historical evolution of their IT

systems. But it is all pretty standard.

CHAIR WEINSTEIN: That would be important in usability, which we are not to yet, but you think it's feasible?

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: Yes.

MS. O'NEILL: And I think that we are using it.

CHAIR WEINSTEIN: Okay.

MS. O'NEILL: By the way.

CHAIR WEINSTEIN: Did we all vote?

MS. WILBON: No.

CHAIR WEINSTEIN: We will vote again?

MS. WILBON: Yes.

MR. AMIN: That's six high.

CHAIR WEINSTEIN: Are susceptibilities to inaccuracies, et cetera, unintended consequences due to inaccuracies, errors and the ability to audit the data items to detect such problems identified?

So I guess, to me, are these data

elements susceptible to inaccuracies? I mean, any time you are taking data from one place and putting it in the claims data, they are all susceptible to those kind of things. I don't know what the error rates are, but any other discussion about that?

MS. O'NEILL: I guess the only other concern I have is kind of in the unintended consequences arena is from their description, if a complication does occur, it is identified as an element to their risk adjustment as opposed to being kind of tracked as a complication.

And to me that's kind of washing it out. I mean, it's not having diabetes ahead of time is a risk. Having pneumonia afterwards isn't a risk. It's a complication. And from their description it sounds like it would be treated like a risk.

DR. DUNN: Yes. This is Dan. That's -- that wasn't correct. It's our fault. You know, the only sort of risk driven

elements of this are the MSDRGs. The pneumonia would not trigger additional risk for these episodes.

MS. PAXTON: Would you be able to clarify what those DRGs or those risks without multiple complications?

DR. DUNN: Correct. That is it's in one of the tables. If the -- you know, all of the major joint replacement, knee or hip replacement DRGs some without, you know, MCC, some with MCC.

MS. PAXTON: The complications could be potentially embedded within those DRGs?

DR. DUNN: I believe that those are present on admission complications. Is that correct? I think those would be beforehand.

MS. PAXTON: Admitting DRGs?

DR. DUNN: Yes. Complicating factors beforehand.

MS. O'NEILL: So one of the things

you mention is the problems with small sample size. And, in particular, when you isolate one of these conditions that orthopedic surgeons, for example, use, you refer to the fact that it is easier to make an assessment of physician performance when there are multiple conditions in a panel, rather than measuring a single condition like this measure does.

Have you a recommendation on the minimum number of episodes on which a physician should be measured or the performance measured?

DR. DUNN: Well, in terms of measurement, you know, all these measurements are based on, at least our specifications, comparisons with peers. Then the question becomes is -- you know, how do you assess whether a difference observed is statistically significant? You can, you know, put some weight on it and sample size will be, you know, part of that determination.

So our recommendation on related to sample size is to use confidence intervals to support that comparison with a benchmark or with peers. You know, and if you look at the -- whether the tradeoff between sample size and statistical significance, you know, it will vary on application, depending on the physician or the, you know, peer group you are looking at.

You know, you probably need, you know, 30 or more episodes or higher to get something that's statistically different, unless the provider is very different from their peers.

I can't give you a recommendation on precise sample size, but, you know, just in the ballpark of what ends up being, you know, sort of the typical distribution.

MS. SINNOTT: Right. I'm just thinking most of a smaller health plan. For example, you know, how many patients with total knees are done in a year in a health

plan of 200,000 or 300,000 people? And then how many of those are actually done by a single provider? That's where the question comes in.

CHAIR WEINSTEIN: And 90 percent of knees are -- people doing knee replacements do less than 10 a year.

DR. DUNN: And that is a valid comment on challenges with these measures.

MR. AMIN: That's two moderate and four low.

DR. DUNN: Can I ask a question? Okay. Are the intent of the endorsements tied to a specific unit of measurement, that, you know, individual surgeon versus practice versus delivery system? Could the answer to this question depend on, you know, the level you are applying the measure at?

The feasibility of this is difficult at the individual surgeon level. But if you start rolling up --

CHAIR WEINSTEIN: Yes. First of

all, I don't want to comment on the endorsement by NQF. They should comment themselves. But I think what we are trying to do is understand the usability -- feasibility, excuse me, across different domains.

And at the individual surgeon level, any of these things are very difficult if the person, him or her, only does, you know, five of these, how valid is the measure? But you could imagine over a few years of use, potentially, that that could get better.

There is no secret that -- you know, and I think we have just finished this study showing that you have to do -- I think people who do more than 100 tend to do much better. And that may not be the cutoff. I might have this data wrong, but it's a number like that.

And people who do less have more complications and more problems. You are going to -- you know that from your database already, quite frankly, because you have years

of data on similar providers over time.

That wasn't a requirement of this collection process to make a determination of supporting this -- your measure, to my knowledge. So --

MS. O'NEILL: Yes, also the reporting that was included was on the individual physician level, so we didn't see a sample practice or a health system report, so maybe we are making an assumption that that's what the reporting format was going to be.

DR. DUNN: And then --

MR. AMIN: I would just add --

DR. DUNN: I think --

MR. AMIN: Sorry, go ahead.

DR. DUNN: No, go ahead, Todd. Really quick, you know the early focus of our responses are on the measure itself, rather than how it would be reported.

But my assumption is it would apply at all the different levels, that makes

sense.

MR. AMIN: So the only thing --

DR. DUNN: Including -- go ahead, sorry.

MR. AMIN: I'm sorry. It's hard to read. I can't see it. So the only thing that I would add from NQF, this is true in speaking, is that the measure would be evaluated based on the level of analysis that was chosen by the measure developer. So what you chose on 11.3, the level of analysis on page 32, so the Committee should evaluate these criteria based on the multiple levels that were specified.

So this measure could be applied at multiple levels, clearly, at the facility or the health plan level or at the population level, but it is also specified for clinician at the individual level.

So it would be endorsed for use at that level. So, thus, this -- all these criteria and, you know, obviously, the more

specific the unit of analysis, the more issues of like 4C would become more important to evaluate.

So I guess the answer to the question that you had posed is that the evaluation would depend on the level of analysis that was chosen for endorsement, since it would be endorsed for use at the individual clinician level.

DR. RATLIFF: And their primary outcome measure is an individual clinician. Now, here in S-11.3, they give a level of analysis going from the individual physician to like the universe. But what they are giving us in this outcome measure is individual physician data. So I think that's what we focused on as we approached this measure.

CHAIR WEINSTEIN: You would think, John, that if they could do it at the individual level, you could roll it up at any other level. That's -- I think that's their

assumption. Yes. So, okay, can we vote?

MR. AMIN: That's one high, four moderate and one low.

MS. WILBON: So do you want to take a quick break then?

CHAIR WEINSTEIN: Sure. We will take a quick break.

MS. WILBON: So we are going to take about a 10 minute break, for those on the phone. And we will be starting with the ABMS measure for 1585, episode of care for simple non-specific lower back pain, when we come back.

CHAIR WEINSTEIN: An easy one.

MS. WILBON: At about 2:25.

(Whereupon, at 2:17 p.m. a recess until 2:23 p.m.)

CHAIR WEINSTEIN: The first question.

MS. WILBON: So let me just check. Is there anyone from ABMS on the phone? I know we are running a little bit behind.

DR. MANHEIM: Yes, Larry Manheim again.

MS. WILBON: Okay.

DR. MANHEIM: Todd is no longer here, but I'm here.

MS. WILBON: Okay. Great. Thanks, Larry. Do you mind giving us just a brief overview of the measure before we start discussion?

DR. MANHEIM: Okay. So again, it's resource use and processes shared with an episode of care for what we define as simple non-specific lower back pain. This is triggered by an initial ambulatory care visit for non-specific lower back pain defined by our diagnoses.

It is a three month episode. Again, similar to -- we talked about radiculopathy. We also include prior 14 days, not for office visit, but in case there were lab or imaging done prior to the first visit.

An episode only begins if there is

no lower back pain diagnosis, trigger diagnosis within 90 days prior to the initial visit. It has to be a 90 day gap. And also, individuals with a radiculopathy diagnosis during the measurement period or during the prior year are excluded from consideration here.

And we allocate to physicians based on the same method as I talked about for radiculopathy. It goes to a physician and has to have 70 percent E&M visits and, otherwise, it goes to more than one physician or physicians that have 30 percent or more E&M visits during the episode, otherwise, it goes to no physician.

CHAIR WEINSTEIN: Any questions by anybody? Okay. Are you ready to go, sir? The pressure is on, sir.

Is this a high impact area?

MS. WILBON: That was six high.

CHAIR WEINSTEIN: Was data submitted that demonstrated considerable

variation in delivery of care? If somebody has a comment, speak up, otherwise, we will just keep voting.

MS. WILBON: Again, six high.

CHAIR WEINSTEIN: Is the purpose objective a resource use measure in the construct for resource/cost clearly described?

MS. WILBON: We have five high and one moderate.

CHAIR WEINSTEIN: Are the resource use service categories that are included in the resource use measure consistent with and representative of the conceptual construct represented by the measure?

MS. WILBON: That's two high and four moderate.

CHAIR WEINSTEIN: Is the measure precisely specified so that it can be implemented consistently?

MS. WILBON: So we do need a little bit of discussion here, just so we have a rationale kind of where you are going with

this one.

CHAIR WEINSTEIN: Anybody want to speak up?

DR. RATLIFF: Well, this is a -- I think -- can you go back to the question, sir, please?

MS. WILBON: Yes.

CHAIR WEINSTEIN: This is a tough one to specify. It isn't that they didn't do a good job and I was just -- as I was answering that question, I was looking back at the inclusion/exclusion criteria and I think they did a really good job. I just think it is a tough one, so I was probably a little less positive, only because I know how hard it is.

I think the measure does a really good job around specificity, so I think they were precise. But any other comments?

DR. RATLIFF: I think this is a real grab bag diagnosis. And I think a lot of different pathologies get lumped into a lumbar

DDD and I think they do about as good a job as you could hope for in parsing out that patient population.

MS. O'NEILL: I guess this is a technical question.

CHAIR WEINSTEIN: Microphone.

MS. WILBON: Microphone.

MS. O'NEILL: If this is a technical question about have they specified it, then they technically specified it, is it -- maybe we are all jumping to the clinical appropriateness of the specificity.

CHAIR WEINSTEIN: Well, the problem I got into is there is other overlying diagnoses sometimes and they have all the drugs. I mean, more drugs than I can imagine, which is -- this population sees all the time. But is it the primary problem? Is it a secondary problem?

And I'm not sure that is addressed well. That was my -- I mean, --

MS. O'NEILL: Yes.

CHAIR WEINSTEIN: -- I wasn't criticizing, but we all see these patients that have secondary gain issues that have home issues, that have work issues and back pain ends up to be the diagnosis that gets them into this episode. It's not their fault, but that's how I was doing it.

MS. O'NEILL: And there were some exclusions of things that are pretty common findings radiologically, for example. There were some exclusions that, to me, wouldn't -- shouldn't be exclusions, but maybe that's a different question.

MS. WILBON: So I think that comes up probably more so in 2(b)(1), which we will get to in just a second. But here if you guys are comfortable with the way that it is written, that someone could follow it, that someone could take that piece of paper, hand it to a programmer and say, you know, program this measure for me, that that is, as it is written, clear enough to do that is basically

what we are asking.

CHAIR WEINSTEIN: I think May Kay captured it though.

MS. WILBON: Okay.

CHAIR WEINSTEIN: We are taking what is probably really clear from an implementation algorithm to say no matter how clear it is, it's going to be a problem potentially.

MS. WILBON: Okay. Okay.

CHAIR WEINSTEIN: Why? Why what?

MS. WILBON: It sounds like everyone is comfortable with the way that it's -- it's a difficult topic, but based on it being difficult, that they did a good job, but it wasn't quite high.

CHAIR WEINSTEIN: We like this.

MS. WILBON: But that some of the issues that pertain to the actual specifications will come up in 2(b)(1), which we will discuss.

CHAIR WEINSTEIN: Yes.

MS. WILBON: Does that kind of --

CHAIR WEINSTEIN: Yes.

MS. WILBON: Okay.

MS. O'NEILL: Thank you.

MS. WILBON: Thank you.

CHAIR WEINSTEIN: Does the reliability testing demonstrate that the results are repeatable producing the same results time and time again in the same time period and that the measure score is precise?

I don't know that I have that precision issue in this. Does somebody want to comment on that?

DR. RATLIFF: I think it's a similar issue to the first group.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: They didn't do reliability testing.

CHAIR WEINSTEIN: Yes. So do you understand that?

MS. WILBON: Yes.

CHAIR WEINSTEIN: Okay.

MS. SINNOTT: Only face validity is --

CHAIR WEINSTEIN: Microphone.

MS. SINNOTT: Oh, I'm sorry. Only face validity is reported. And reliability of the physician scoring isn't reported, either.

MS. WILBON: Right. So just to --

CHAIR WEINSTEIN: It's in process. This isn't done yet.

MS. SINNOTT: Right.

MS. WILBON: So just I think this is very similar, that testing information that was meant for this measure is very similar to other measures. So if everyone is comfortable with that, I don't think -- unless there is something new particular for this condition focus that would need to be brought up. I think it would be covered. Okay.

So that's one moderate, two low and three insufficient.

CHAIR WEINSTEIN: What is the level of overall reliability and testing?

Again, we run into the same issues. We thought that there was some good specifications, but the reliability testing isn't there. So that's why you are going to see the votes you are going to see, whatever they are.

MS. FANTA: So we have one moderate, three low and two insufficient.

CHAIR WEINSTEIN: And, you know, I think we all just want to congratulate the people who have been doing these things, because we are going to run out of here when we are done sometime or they will leave the phone, but we all, as a Committee, want to express our appreciation to Ingenix and ABMS for this incredible work. This is really hard work. And we applaud that.

And our comments in no way want to discredit that or be seen in any other way. So just to get that on the record.

Are the measure specifications consistent with the evidence?

MS. O'NEILL: What evidence?

MS. WILBON: And again, this is not that -- evidence should actually like the intent or the focus of the measure. So again, evidence is a little misleading. We didn't paraphrase that well.

CHAIR WEINSTEIN: Well, but what do you want us to answer?

MS. WILBON: So we are asking if the measure specifications, as the measure is written, is it consistent with what they said the intent of the measure was? And what the focus of the measure of that particular condition is and what they are intending to measure.

CHAIR WEINSTEIN: Can we revote, Sarah?

MS. FANTA: Yes, revote.

CHAIR WEINSTEIN: Start -- yes, because I --

MS. FANTA: Oh, sure, yes.

CHAIR WEINSTEIN: I used the word

consistent with the evidence.

MS. WILBON: Yes.

CHAIR WEINSTEIN: Thank you.

MS. FANTA: Go ahead.

CHAIR WEINSTEIN: Who is going to make sure that question is interpreted the way you said versus what we are answering?

MS. WILBON: It's actually on here correctly, which is what we are going by.

CHAIR WEINSTEIN: Okay.

MS. WILBON: So it's just a slight -- it's just the slide that's wrong, let's assume.

MS. FANTA: So we have one high and five moderate.

CHAIR WEINSTEIN: Does the validity testing demonstrate that the measure data elements are correct? Does validity testing -- you know, we run into the same problems again. So we can vote. Do you want to say something, Mary Kay?

MS. O'NEILL: Just, at some point

time, we have to say that in this cadre of patients, what things are called and, as you have pointed out, what the actual underlying driving diagnosis may be, it has the highest degree of variability.

Maybe not in every clinical situation, but one of the most -- I mean, it's the one area where I think if you gave a bunch of reasonably trained clinicians the same batch of patients and even coming up with the right diagnostic code, it would be a pretty big grab bag, you know. So it's hard to get the right data in here.

MS. SINNOTT: And you have to give them all the same, what's it called, billing sheet with the diagnosis at the bottom, you know.

MS. O'NEILL: Correct.

MS. SINNOTT: Or whatever.

CHAIR WEINSTEIN: Well, the problem is some people use the same code for all of these patients independent of what the

diagnosis might be. I mean, it's just -- we all understand.

MS. FANTA: The results were one moderate, three low and two insufficient.

CHAIR WEINSTEIN: Are exclusions supported by the clinical evidence or analysis of frequency and distribution? Is information about impact of exclusions for patient preference transparent?

MS. O'NEILL: Could I just --

CHAIR WEINSTEIN: Yes.

MS. O'NEILL: -- clarify in this 90 day episode, am I correct to read that any patient that has a fusion in the 90 days is excluded?

DR. MANHEIM: Any patient that has a fusion in the prior period is excluded.

MS. O'NEILL: But not in the episode?

DR. MANHEIM: Right.

MS. O'NEILL: Okay.

DR. MANHEIM: As long as they

don't have a diagnosis one -- of a radiculopathy diagnosis.

MS. O'NEILL: Okay.

DR. MANHEIM: You know, because that may be thrown out because of that.

MS. O'NEILL: Okay.

DR. RATLIFF: As some of your exclusion criteria you list active cancer, which seems pretty reasonable, because you want to look at back pain, not people that are coming in with pathologic fractures. But then you exclude melanoma, which not infrequently goes to the spine, but more importantly, prostate, which loves going to lumbar spine and is going to give you a little back pain.

So it's going to confound your data that following this exclusionary criteria, you are going to be bringing in prostate cancer meds to the spine along with your Workman's Comp patients who have like isolated low back pain episodes.

So I don't understand that aspect

of your exclusionary criteria.

DR. MANHEIM: Oh --

DR. RATLIFF: Look at that.

DR. MANHEIM: -- so what you are saying is --

DR. RATLIFF: I don't understand why you then say active cancer (excluding melanoma, skin), prostate and CLL. Like why exclude prostate? Why do you want to have prostate cancer patients included for a low back pain measure?

CHAIR WEINSTEIN: Do you understand his question? It's pretty specific. You say cancer, but you exclude some cancers.

MS. SINNOTT: Exclude from the exclusion.

MS. WILBON: Right.

CHAIR WEINSTEIN: Yes.

DR. MANHEIM: Right.

CHAIR WEINSTEIN: It doesn't make sense.

MS. WILBON: Maybe they meant including.

DR. MANHEIM: I'm looking at it and I may --

DR. RATLIFF: It's on page 13 of your PDF, Step 3 of your criteria, the first paragraph there.

DR. MANHEIM: So diagnostic codes to identify active cancer treatment.

CHAIR WEINSTEIN: But then you say excluding certain types of cancer. Why would you exclude them? I think what you are giving is examples of cancer you would include potentially.

DR. RATLIFF: Agreed.

DR. MANHEIM: Right, right. From what I'm looking at, I should be following what you have, I don't see that.

CHAIR WEINSTEIN: Yes, it's --

DR. MANHEIM: I believe that --

CHAIR WEINSTEIN: -- an error, I'm sure --

DR. MANHEIM: It's an error.

CHAIR WEINSTEIN: -- in what was written.

DR. RATLIFF: If it's an error, they consistently make it at multiple different points in the document.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: Whenever they talk about like excluding --

CHAIR WEINSTEIN: Yes. You are exactly right.

DR. RATLIFF: -- cancer, active cancer patients.

CHAIR WEINSTEIN: The other thing they did, they say patient had fusion or other back surgery or fracture. I assume that includes osteoporotic compression fractures, which are very common cause of back pain?

DR. MANHEIM: The diagnoses are listed and I would have to look at that.

CHAIR WEINSTEIN: So those are important points that we just brought up that

you need to resolve.

DR. MANHEIM: Right.

CHAIR WEINSTEIN: Okay.

MS. SINNOTT: I'm sorry, is pregnancy in here as an exclusion?

DR. MANHEIM: Pregnancy, I believe is --

CHAIR WEINSTEIN: It's not listed.

DR. MANHEIM: -- not listed. It's not in here. I know there was discussion and it was decided not to include it as an exclusion.

CHAIR WEINSTEIN: Another good point, I think. It's hard enough to do this with including those. The Committee is recommending you make the changes that we have recommended in your model or clarify that this is an error in the printed version that we have --

DR. MANHEIM: Right.

CHAIR WEINSTEIN: -- versus your model.

DR. MANHEIM: So exclude pregnancy, don't have the restrictions and on active cancer, if that's not an error, you know, just written therein. In any case, we have to correct that.

CHAIR WEINSTEIN: And compression fractures.

DR. MANHEIM: Right.

CHAIR WEINSTEIN: Which maybe it says or fracture, so I'm just not sure. But you are talking about surgeries there, so I'm not sure.

DR. MANHEIM: Yes. So we will need to check that.

MS. SINNOTT: And what about trauma?

DR. MANHEIM: Trauma is, I believe.

DR. RUBIN: It's in there. I think it's in there.

MS. SINNOTT: As expressed as an E Code. Well, the question is whether you want

to include motor vehicle accidents in the non-specific, might I say, mechanical low back pain?

CHAIR WEINSTEIN: You get into this whiplash kind of stuff, too, you know, back pain.

DR. RATLIFF: If you're going to start excluding motor vehicle accidents, why don't we exclude like Workman's Comp and other like work-generalized accidents? And I see what you are saying, but it can quickly like broaden out and suddenly your measure doesn't mean anything to your patient population.

MS. SINNOTT: Well, but if I'm a Workers Comp carrier, I want -- I don't want to exclude the Workers Comp injuries, right?

MS. O'NEILL: Yes, most of this database would not have Comp data, I don't believe, but they would have personal injury cases is what you are saying. And, obviously, they would be excluded if they were major trauma by the other exclusionary criteria, but

not minor trauma.

MS. SINNOTT: I'm just going back to the, you know, original exclusions from the back pain, the boat, which was, you know, inflammatory, spinal arth --

CHAIR WEINSTEIN: Spinal arthropathy.

MS. SINNOTT: Thank you. And motor vehicle accidents and pregnancy and cancers and things like that.

DR. MANHEIM: Whether we would actually know whether it was a motor vehicle accident or even Workman's Comp from the data, I'm not sure.

CHAIR WEINSTEIN: Well, the other thing is in your -- in the radiculopathy one--

DR. MANHEIM: Yes.

CHAIR WEINSTEIN: -- you also, we missed this, but, included the cancers there, too. So I think it's an error. And I think your list of exclusions are a little bit better around some of these things we are

talking about right now, so you might --

DR. MANHEIM: Right.

CHAIR WEINSTEIN: -- try to --

DR. MANHEIM: Look at both of them and make sure they are both correct.

CHAIR WEINSTEIN: -- look at those and make sure that they are making sense with your physician panel. And I would ask that you submit a revised list to NQF that matches your model and/or if your model has got these in them, it's a problem. So there is some work that needs to be done that NQF needs to know about by these things we are bringing up now, because --

DR. MANHEIM: All right. We will do that.

CHAIR WEINSTEIN: -- it invalidates or weakens your model by not addressing these issues, in both cases. Anybody have other comments about that? Okay.

So are exclusions supported by the clinical evidence or analysis of frequency and

distribution? Is information about impact of exclusions for patient preference then apparent? The same issues we have had before.

NQF, will you let us know that they have done that?

MS. WILBON: Yes.

CHAIR WEINSTEIN: Yes.

MR. AMIN: That's three moderate and three low.

CHAIR WEINSTEIN: Risk adjustment for resource use measures is the evidence-based risk reason based here? I assume.

MS. WILBON: Yes. Let's check the wording here. Yes, so the risk adjustment should be based on patient clinical factors or evidence about those clinical factors that influence the measured outcome of resource use. Obviously not based on factors of related disparities and care and that the risk adjustment factors are present at the start of care and have demonstrated -- that they have demonstrated adequate discrimination and

calibration of the model.

CHAIR WEINSTEIN: Any comments from the group?

DR. RATLIFF: The same content to me as the first model.

CHAIR WEINSTEIN: Say it again.

DR. RATLIFF: The same content as the initial model from ABMS. The same issues.

CHAIR WEINSTEIN: Thank you.

MS. O'NEILL: Jim, I -- this may just be completely impractical. I note in the risk adjustment model they have got some major psych diagnoses, but they don't have any, you know, anxiety disorder, any of the more normal psych diagnoses, which is a risk factor in this group. And maybe that's because the data is too hard to get.

CHAIR WEINSTEIN: Okay. Let's vote on this one.

MR. AMIN: It's three moderate and three low.

CHAIR WEINSTEIN: One of the

things I just want NQF to know is in their page 22, while the latter is straightforward around risk adjustment, caution is warranted as the risk adjustment equations were derived from a population that may be different from the population to which the measure is being applied. That's why I said low.

I don't know what that means. Can you guys explain that?

DR. MANHEIM: Right. What it means is that the coefficients were derived from existing data and an alternative to just taking the coefficients that we used is to re-estimate it, the variables we have, within someone's given population.

CHAIR WEINSTEIN: Okay. Thank you. Are performance results reported? Do they identify differences in performance or overall less than optimal performance? So we all talked about this before.

MS. WILBON: The same one?

MS. WILBON: Right. So this

criteria reads should be that the data analysis demonstrate that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically meaningful -- practically and clinically meaningful difference of performance.

CHAIR WEINSTEIN: Are you okay, Elizabeth? Do you need some more help? Are you reading the answers for us?

MS. FANTA: We have three moderate, two low and one insufficient.

CHAIR WEINSTEIN: Thank you. If multiple data sources methods specified, do analysis demonstrate that they only used, you know, the one data source? So are we going to answer this? I thought this was one we skipped.

MS. WILBON: Yes, it is.

CHAIR WEINSTEIN: Okay. Validity. What is the overall, based on the different measures, validity of this?

MS. FANTA: We have three moderate, two low and one insufficient.

CHAIR WEINSTEIN: Disparities. Is it the same issues that we have talked about before?

MS. FANTA: One high, two low and three insufficient.

CHAIR WEINSTEIN: I'm going to have these questions memorized by the end of this. Sad. Tell us when you are okay. Are you okay?

MS. WILBON: Yes. So just for the Committee's information, what I'm doing is kind of for consistency sake, I realize that over the course of a day, you know, people get tired and there is -- that we are rating kind of the same issues, the same across the measures, particularly from the same developer, so I'm just kind of checking back to ratings to make sure that they are consistent.

So they have been consistent, up

to this point. Although, I would like to -- not to call anybody out, but whoever rated this high, if they could just -- the previous rating for this same criteria for the other ABMS measure was one low and five insufficient.

So we ended up with this one with one high, two low and three insufficient. So I just kind of want to get a feel for where people were on that.

DR. RUBIN: So I was the outlier.

MS. WILBON: Okay.

DR. RUBIN: And part was the statistical analysis. I really was looking for a not applicable, I guess, and just referenced back to my initial evaluation from this. But it's not part of the risk adjustment and so maybe I should have thrown it back to four. That seems to be a marked discrepancy, but --

MS. WILBON: Yes.

CHAIR WEINSTEIN: It's okay.

Whatever you --

MS. WILBON: It's okay.

CHAIR WEINSTEIN: -- think.

DR. RUBIN: It's the only time I've been an outlier.

MS. WILBON: As long as you -- as long as we have a justification and we can kind of rationalize it, that's fine.

CHAIR WEINSTEIN: So do you want to change your vote?

DR. RUBIN: No, that's okay.

CHAIR WEINSTEIN: You're okay?

DR. RUBIN: Yes.

CHAIR WEINSTEIN: Good. Okay. Are you okay?

MS. WILBON: Yes.

CHAIR WEINSTEIN: Next. Usability. Are the measure performance results reported suitable to report to the public at-large, national, da, da, da, da. Is there evidence?

MS. FANTA: Two moderate, three

low, one insufficient.

CHAIR WEINSTEIN: Usability. Did sufficient -- did submitted information demonstrate that results produced by the measure are meaningful, understandable and useful for quality improvement, public reporting, etcetera?

MS. FANTA: The results were one moderate, four low and one insufficient.

CHAIR WEINSTEIN: That's on usability. Are you okay? Do you need something answered?

MS. WILBON: Yes.

CHAIR WEINSTEIN: Because we want to make sure you are --

MS. WILBON: I just -- so for the ABMS measure, 1586 on the lumbar radiculopathy, for this -- this is 3(a), correct?

DR. RATLIFF: Yes.

MS. WILBON: Okay. The vote was that everyone voted insufficient. So I just

kind of want to point that --

DR. RATLIFF: One thing for the discussion there --

MS. WILBON: -- point out some --

DR. RATLIFF: -- to make the same kind of -- I think I can answer your question.

MS. WILBON: Okay.

DR. RATLIFF: The thinking on a point there, the point we brought up, it was they had funding from Robert Wood Johnson. They noted the measures had been tested for usefulness or interpretabilities. When we discussed, I guess, 1586, we sort of made the point that this was a process. We didn't really have the data yet.

CHAIR WEINSTEIN: I also think this diagnosis has much more specificity to it with the right criteria than low back pain does. So I think there is a difference.

MS. WILBON: Okay.

CHAIR WEINSTEIN: That we have tried to represent in this.

MS. WILBON: Okay.

CHAIR WEINSTEIN: Do you need more help with that?

MR. AMIN: No, that's good.

MS. WILBON: That's okay.

CHAIR WEINSTEIN: Okay. Is this the next one, sir? I thought we did this one?

MS. WILBON: We did. But just clarify it. Yes, go ahead, just show it.

DR. RATLIFF: So I guess what I take home from the Committee is that even with the data, we still think this is going to be low?

MS. WILBON: I --

DR. RATLIFF: Because of the patient population, because of the diagnostic criteria?

MS. WILBON: Yes.

DR. RATLIFF: Because of the uncertainty involved.

MS. O'NEILL: And I think even more than the, you know, trying to evaluate

the resource utilization of a bunch of people doing decompression laminectomies or whatever for a peer group compression, how many resources you use to manage people that come in under these diagnostic labels is unbelievably hard to evaluate, if you don't have outcome data.

I just don't even know what you, in this group, are measuring hardly, because, I mean, maybe somebody sees somebody once and doesn't like these kind of patients and they don't see them again. Maybe that's the best thing for some of these guys, but you know what I mean?

It's just too much of a grab bag. And I think this is absolutely where you would want to have an outcome.

CHAIR WEINSTEIN: Does that answer your guys' questions, NQF personnel, who are whispering? Share your feelings with the group.

MS. O'NEILL: I'm saying this and

I'm going to go to the airport and you guys can talk about what crazy things I said.

MS. WILBON: Stir the water and then run.

MR. AMIN: So the discussion on the previous ABMS measure was around the O to E ratio and whether the information was giving you enough detail to be able to tell a difference, to be able to discern a difference between different providers. Does that sentiment carry onto this measure?

MS. O'NEILL: I would just say, I mean, when you are looking at resource use expected and observed around a procedure and you are doing it around the management or the non-surgical or conservative management or maybe not so conservative management of the people with the same group of complaints, not even the same diagnosis necessarily, but of symptom complaint, which back pain is not a diagnosis, it's a symptom complaint, and to say what you would observe versus what you

would expect, unless you are looking at really large numbers, you would need to have how many resources you need to get to a certain outcome with a cohort of patients before you can tell if you are doing enough, too little, too much.

You know, I mean, I don't know how you -- what yardstick you would be using really.

CHAIR WEINSTEIN: Is that helpful?

MR. AMIN: Yes, it is.

CHAIR WEINSTEIN: All right.

MR. AMIN: Thank you.

CHAIR WEINSTEIN: Is this the next one, Sarah?

MS. FANTA: Yes.

CHAIR WEINSTEIN: Are the data and result details maintained, such that the resource use measure, including the clinical and construction logic for a defined unit of measurement can be broken down to facilitate transparency?

MS. FANTA: The results are one

high, two moderate and two low.

CHAIR WEINSTEIN: Feasibility. Are the required data routinely generated and used during data care delivery?

MS. FANTA: Four high and one moderate.

CHAIR WEINSTEIN: Are the required data elements available in electronic records?

MS. FANTA: Five high.

CHAIR WEINSTEIN: Are susceptibilities to inaccuracies, errors or unintended consequences in the ability to audit the data items to detect such problems?

The problem with this is the specificity of these diagnoses or the lack thereof, so people tend to use different codes, maybe even for the same patient if they saw him on two different days, is one of the issues you may see in some of the responses here versus the other radiculopathy one. Still waiting?

MS. FANTA: One moderate, three

low and one insufficient.

CHAIR WEINSTEIN: Do you need some clarification, Sally?

MS. WILBON: Go ahead.

MR. AMIN: I guess the question that the team is thinking about is whether that is a concern with administrative data broadly applicable to any measure or this is particular to this topic area, because --

CHAIR WEINSTEIN: Yes, that's what I was trying to give you a clarification expecting this response.

MR. AMIN: Okay.

CHAIR WEINSTEIN: In this particular diagnosis, the Time 1, Time 2 diagnosis in the same patient may be very different, unlike the others. A hip fracture is a hip fracture. A knee replacement is a knee replacement. A disc herniation with radiculopathy is pretty clear.

MR. AMIN: Okay.

CHAIR WEINSTEIN: But back pain,

today it's back pain, tomorrow it's back pain from a different one of these codes. So it isn't that the data isn't there. It's the reliability of using the same code for the same patient at different times. Over time I think it would change.

MR. AMIN: Thank you for that clarification.

CHAIR WEINSTEIN: Is that okay, Sally?

MS. TURBYVILLE: It's interesting.

CHAIR WEINSTEIN: Yes, true, unfortunately. Unless my colleagues feel differently? No.

DR. RATLIFF: I see no nomenclature for this just we don't have a good language for describing these conditions, so you are stuck with that, with a measure like this, those administrative data for back pain.

CHAIR WEINSTEIN: Can the data collection strategy be implemented? Is this

measure already operational? So that answer is no. So it's not ready to be implemented is the way I would sort of look at it.

MS. WILBON: Right. So it doesn't -- okay, that's fine.

CHAIR WEINSTEIN: Who are we missing here?

MS. FANTA: Two low and three insufficient.

CHAIR WEINSTEIN: This next one is going to be actually pretty easy, I think. We'll see what the group thinks.

We're not taking a break.

MS. WILBON: Oh, we're not?

CHAIR WEINSTEIN: We're going to keep going.

MS. WILBON: Oh, okay. We are just taking a five minute mind break.

(Whereupon, at 3:06 p.m. a recess until 3:07 p.m.)

MS. WILBON: Is there someone from Ingenix still on the line?

CHAIR WEINSTEIN: Are you prepared to talk about the next measure, because some--

DR. TARKO: That's it. Dan Dunn will be doing that.

CHAIR WEINSTEIN: Is he there?

MS. WILBON: So the next measure--

DR. TARKO: Right here.

MS. WILBON: -- we are discussing is No. 1603, which is the ET-based hip and pelvic fracture measure. If some one could just give us a brief overview of the measure before we --

CHAIR WEINSTEIN: Yes, and could you clarify? To me, this is about hip fractures, because pelvic fractures, the terminology, I just want to be clear because your literature reveals -- is talking about hip fractures pretty much.

So the word pelvic in there is interesting to me.

DR. RATLIFF: My interpretation of this is just hip.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: Why are we saying hip fracture? When I was filling this out, I don't know about anybody else --

CHAIR WEINSTEIN: Me, too, but they keep -- they have the terms here, so I just want to be clear, because pelvic fracture is a whole other ball game.

MS. WILBON: Yes.

CHAIR WEINSTEIN: Not that it doesn't occur in the elderly, but -- okay. We are ready to have you tell us.

DR. TARKO: Well, we're waiting for Dan Dunn. He'll be here in a second.

CHAIR WEINSTEIN: He will be here what?

MS. WILBON: Dan Dunn is going to be on the phone.

CHAIR WEINSTEIN: Oh.

DR. TARKO: We're getting him right now. He'll be right here.

CHAIR WEINSTEIN: Yes, we're

waiting on him not happily. Well, we do have other jobs. We're ready to go, Sarah.

MS. WILBON: We are just going to go ahead and go and then when Dan gets on the phone, we will ask him any questions as they come up.

DR. TARKO: Okay.

MS. WILBON: Okay. Thanks.

CHAIR WEINSTEIN: I just do want to clarify this is about hip fractures. And I would eliminate the word pelvic for right now, unless I hear otherwise from Ingenix.

DR. ROBERTS: Are we sure that's what they meant?

CHAIR WEINSTEIN: That's why --

DR. ROBERTS: Because if the pelvic --

DR. TARKO: That would have been--

DR. ROBERTS: -- is all the way through there that needs to be removed.

CHAIR WEINSTEIN: That's what I'm asking. Can you guys answer that question?

DR. TARKO: It was our understanding that it was hip and pelvic fractures, so I'm here in Tom Lin's stead, but that was our understanding it was including pelvic fractures as well. That was our error.

CHAIR WEINSTEIN: Yes, but did you include --

MS. WILBON: I'm sorry, can you --

CHAIR WEINSTEIN: -- codes for pelvic fractures or just hip fractures?

DR. TARKO: We did include codes for pelvic fractures.

MS. WILBON: I'm sorry, can you tell me your name? Who is talking right now?

DR. TARKO: It's Howard Tarko.

MS. WILBON: Oh.

DR. TARKO: I'm a medical director here.

MS. WILBON: Okay.

DR. TARKO: I'm here in Tom Lin's stead. He was called away on some personal emergency.

DR. RATLIFF: Bringing in like a pelvic fracture --

MS. SINNOTT: But am I correct that the ETG --

DR. RATLIFF: -- that's different.

DR. TARKO: Dan Dunn is calling in now.

DR. DUNN: Yes, hello, Dan Dunn here, also.

MS. SINNOTT: So the ETG says close fracture or dislocation by hip and pelvis, so it is an ETG classification, correct, for hip and pelvis fracture?

DR. DUNN: Yes.

CHAIR WEINSTEIN: Can you show me your inclusion criterion? Just for whatever reason -- because the codes they have here on 25, ETG does provide methodology to deal with this case where code will shift.

You know, for example, concurrent renal transplant. For hip fracture there were 26 diagnosis codes which would cause an

episode of hip/pelvic fracture to shift to an episode of joint degeneration.

So I'm confused. This is really important.

DR. RATLIFF: You know, I mean, it's cracked. I mean, if you look at their Excel spreadsheet where they go through the diagnosis codes that they are including, they are including --

CHAIR WEINSTEIN: Which page is that, John?

DR. RATLIFF: This is their Excel sheet S-5_DD, that is included in the package of information that came with 1603.

CHAIR WEINSTEIN: Oh.

DR. RATLIFF: That's fracture ilium and fracture ischium, a pelvic fracture with the disruption of pelvic circle. Closed fracture part of the pelvis. So I think we were all thinking standard hip fracture.

CHAIR WEINSTEIN: Yes.

DR. RATLIFF: But they are

including like a lot more.

DR. ROBERTS: A sacral insufficiency fracture compared with an interstroke anterior hip fracture.

CHAIR WEINSTEIN: That's why I asked the question right up front, because these are like apples and oranges and treated very differently in very different episode groupers. And so if this is -- then it is probably -- we need to decide whether we can include this or not the way it is designed.

Hip fracture is a very common, very meaningful important measure unto itself. When you get into pelvic fracture, it's a very different problem. They are usually stress fractures. They are not talking about trauma here, I'm sure, I hope. I shouldn't be sure about anything.

But unless the Committee feels differently, I think you really have to disentangle those things.

MS. SINNOTT: Am I correct that

NQF requested a measure for hip and pelvis -- felt pelvic fracture as a single entity?

MS. WILBON: It wasn't necessarily that we were asking for it in a single entity. I know the way that it was listed on the call for measures in hip/pelvic, but if they had a separate measure for hip fracture and a separate measure for pelvic fracture we would have taken that as well, I think. It was just a matter of semantics.

CHAIR WEINSTEIN: Well, this is a core question to whether we can actually answer this effectively.

DR. TARKO: The way that the measure was specified, I'm trying to find the actual statement, was there are some classifications which the episode treatment groups called condition statuses. And the condition status -- there is a condition status factor called femoral neck fracture and one for pelvic fracture.

And it was understood that that

would be the set of codes used in terms of defining the measure. The subset of the episode treatment group.

CHAIR WEINSTEIN: Are these the codes?

MS. WILBON: Yes.

CHAIR WEINSTEIN: Yes. Can I?

MS. WILBON: Sure. Go ahead.

CHAIR WEINSTEIN: Just for a second. Can you scroll? Can you scroll?

MS. WILBON: You can scroll.

MS. SINNOTT: In the beginning of the measure information --

CHAIR WEINSTEIN: You get into unspecified derangement of a joint, unspecified -- site unspecified. A lot of these codes open fracture of an acetabular. You can't compare these things. Those are night and day problems. Much more morbidity, much more complex surgery.

Hip fracture by itself has a 30 percent one year mortality uncomplicated.

MS. WILBON: So, Dan, or whoever else is on the phone, can you guys give a rationale for -- does this ETG exist in this way or was it combined in some way in response to the call or are they --

CHAIR WEINSTEIN: Open fracture.

MS. WILBON: -- separated? Maybe if you can just give us some context as to how this evolved or how you have it in your system currently?

DR. DUNN: Yes, this is Dan. I'll take a shot. And again, I apologize Tom isn't able to be here, but I'll do my best. So there is an ETG. I think if -- someone described which is called closed fracture or dislocation by hip and pelvis.

So that's the general categorization. You know, what we did is go into that ETG and identify those episodes where there was evidence of those two conditions status that somebody mentioned, fracture femoral neck and pelvic fracture.

So the episodes that find their way into the spinal measure specification are the subsets of episodes in that broader ETG where there is the indication of the, you know, fracture of femoral neck or pelvic fracture.

CHAIR WEINSTEIN: Yes, but this is a hard one. You understand the problem. A lot of your codes like the 800 codes, you know, open fracture of an acetabular is so different. Multiple open pelvic fractures with disruption of the pelvic circle, that's diastasis.

I mean, I can see the transcervical fracture, which is not a femoral neck fracture being included actually. I can see the mid-cervical fracture. I could even see an intertrochanteric fracture, a pertrochanteric fracture, but a lot of these pelvic things you can't put them in the same grouper.

DR. DUNN: Maybe that's how -- do

you have access to a table that's in the S-8--

CHAIR WEINSTEIN: Is this it?

DR. DUNN: -- spreadsheet and it's called "Condition Status to DX Code Map." I don't know if this would help the discussion, but that would give you the diagnosis code, but not for the specific subset that we are pulling out.

CHAIR WEINSTEIN: Well, this is even more confusing. Closed fracture of the shaft of the femur, closed fracture of the lower end of the femur, closed fracture of the lower epiphysis of a femur which would be in a child.

DR. DUNN: Yes. I'm sorry, the only one that I'm referring to here are the ones with fracture of the femoral neck, which is -- starts with that 70326 condition status or the 70328 coded fracture, but in the 820 range.

CHAIR WEINSTEIN: So you are including just the 70326s?

DR. DUNN: And the 70328.

CHAIR WEINSTEIN: Yes, but when you get into the 328s, you get into the pelvis fractures and acetabular fractures and ilium fractures and disruptive pelvic ring fractures. These are very different injuries.

DR. DUNN: And then the strategy then is -- maybe again clinically this -- you still -- it doesn't make sense that the risk adjustment methodology would recognize those two differently.

DR. RATLIFF: When I interpreted this, I think I just mentally read it as like a hip fracture.

CHAIR WEINSTEIN: Well, that's what I was thinking, but --

DR. RATLIFF: But you are clearly correct --

CHAIR WEINSTEIN: -- I asked the question -- yes.

DR. RATLIFF: -- it's not just a hip fracture. It's a pathologic fracture of

the femur from systemic malignancy is open pelvic fractures. It's a vast array of different injuries.

CHAIR WEINSTEIN: If this is the way it has been done, I don't think we can effectively measure this.

MS. WILBON: Okay.

DR. DUNN: Actually --

MS. BOSSLEY: So I think it would be helpful for you to at least talk through the important piece, because I'm seeing hip fractures discussed in the importance piece, but not the pelvic piece, the pelvis.

CHAIR WEINSTEIN: Yes, they are all important.

MS. BOSSLEY: So I think --

CHAIR WEINSTEIN: But the hip fracture is so common.

MS. BOSSLEY: Right.

CHAIR WEINSTEIN: There is a half a million in a year.

MS. BOSSLEY: Which is where I

think they were able to get the data on it.

CHAIR WEINSTEIN: Yes.

MS. BOSSLEY: But so I think it would be helpful to have you rate the importance piece and then I think we should probably have you, at least, discuss the scientific acceptability, because the precision of the specifications deal with this issue as well.

CHAIR WEINSTEIN: Yes, that's what I'm -- because all their writings are about hip fractures.

MS. BOSSLEY: Right. I think that's what I would do. And then maybe let's have you stop, because I don't know that you can go beyond that. What we may have to do is talk to Ingenix and make sure that this was truly the intent, because they have the person that would typically answer this question is not available.

And then if we need to get you back on a phone call to finish the discussion,

why don't we do that?

CHAIR WEINSTEIN: Perfect.

MS. BOSSLEY: Does that seem reasonable to --

CHAIR WEINSTEIN: Perfect.

MS. BOSSLEY: -- staff, too?

CHAIR WEINSTEIN: Is somebody reading this differently than me, including the Ingenix people? Because maybe we are misinterpreting what you meant to do. Okay. So let's --

DR. DUNN: This is Dan. I'm sorry. Yes, I can't help you here.

CHAIR WEINSTEIN: Don't worry about it, Dan. It's not a problem. We're going to figure it out, but we just want to make sure we do the right thing.

So I think Heidi's suggestion is the right one.

MS. BOSSLEY: Yes. So, Dan, we will give you -- they will give you some guidance. They are going to go through

importance and some of scientific acceptability, so you can talk to Tom when he is available, get back to staff and then if we need to reconvene the TAP to look at more, we will.

DR. DUNN: Okay. Thank you. Sounds good.

CHAIR WEINSTEIN: Okay. So if it is hip fracture, which can be an intertrochanteric, femoral neck, pertrochanteric, it's a big problem. You know, it's a big cost. It has a lot of issue around comorbidity issue, complications, so do you want us to actually grade this, Heidi?

MS. BOSSLEY: I think it would be helpful.

CHAIR WEINSTEIN: Okay. Okay. So from an impact point, can we all agree --

MS. WILBON: Well, I just have a procedural question.

MS. BOSSLEY: Yes.

MS. WILBON: Because I'm a little

bit confused now myself --

MS. BOSSLEY: Yes.

MS. WILBON: -- admittedly. Are they going to be evaluating this as if it is as submitted or as if they are just evaluating hip fracture? Because --

MS. BOSSLEY: No. I think they have to --

CHAIR WEINSTEIN: We're talking about the fracture.

MS. BOSSLEY: -- submit it as -- evaluate it based as it is submitted, because from the sounds of it, that's actually what they intended to do, if we are understanding them correctly.

CHAIR WEINSTEIN: The first part of this we can answer so many questions, because they are written about the variance and the issue is really written around hip fracture, all the papers they are quoting.

MS. WILBON: So but the measure --

CHAIR WEINSTEIN: But the

methodology by which they did the measure is not valid.

MS. WILBON: But the title --

DR. RATLIFF: But when they went back to do their summary --

MS. WILBON: -- I mean, let me look at the intent here quickly. Let me just see if that is -- because I think this is where it might be --

DR. RATLIFF: Their summary data answers the first question.

MS. WILBON: -- confusing is where the title says one thing, their intent says one thing, but then the specifications say another. So I just want to make sure as we go through this that we are all on the same page.

DR. RATLIFF: Even as we look at the first question, like relevance, importance of this, they talk pelvic fractures as being how they looked through their own database to get their charge discrepancy.

So again, even answering the first

question like you are asking us to do, we are still opening a grab bag --

CHAIR WEINSTEIN: You know, Heidi, could I make a suggestion?

DR. RATLIFF: -- that is filled with crackers.

MS. BOSSLEY: Sure.

CHAIR WEINSTEIN: It's just because I think it will be confusing for everybody.

We can do this by phone. This isn't a hard one to do. I would rather get the clarification and do it the right way, then start down a path that is going to get us all mixed up and not be adequate for you.

MS. BOSSLEY: That's absolutely -- we are fine with that, too.

CHAIR WEINSTEIN: Okay.

MS. BOSSLEY: And it is perfectly fine. So I think the question would be is does Ingenix have enough information to know what they need to, you know, clarify?

CHAIR WEINSTEIN: Well, we can talk to them by phone, too.

MS. BOSSLEY: Yes, yes, exactly.

CHAIR WEINSTEIN: Because the right person is not here.

MS. BOSSLEY: Right.

CHAIR WEINSTEIN: And so why don't we do that the right way? And we are familiar enough with this, you could put these questions on a monkey survey, we could all do them together or whatever.

MS. WILBON: Survey monkey, you are close. You are close. So then team, do you guys have a good idea of -- as Jim said, we can have a conversation off-line about what needs to be clarified and what maybe needs to be disentangled or what have you.

Do you have an idea about what maybe to follow-up with Tom with about, at this point?

DR. DUNN: Yes. This is Dan. Yes, thank you. And, yes, we will -- we are

probably going to need to touch back with you to clarify, but I think I know where to start now.

MS. WILBON: Okay. Great. So we will circle back with you tomorrow or Monday and kind of touch base. I'm not sure when Tom will be back, but we can touch base and figure out when to have that discussion.

DR. DUNN: Okay. That sounds good.

MS. WILBON: Okay.

DR. DUNN: Thank you.

MS. WILBON: Thanks, Dan.

DR. DUNN: Okay. Take care. Bye.

MS. WILBON: Bye.

CHAIR WEINSTEIN: Are you okay, Heidi?

MS. BOSSLEY: Oh, yes.

CHAIR WEINSTEIN: Because I really think it's the right thing to do.

MS. BOSSLEY: Yes, that's fine.

CHAIR WEINSTEIN: Okay, yes.

MS. WILBON: Okay. So that said,

that actually thus ends --

CHAIR WEINSTEIN: A record.

MS. WILBON: Yes. So we are going to open it up for public comment. We are going to open it up for public comment.

Is there anyone on the phone who would like to make a comment to the TAP before we close?

Yes, Operator, can you just make sure that all lines are open, at this point, so if anyone wants to speak, they can do so freely?

OPERATOR: Yes, all the lines are open.

MS. WILBON: Okay. Thank you. Is there anyone there who would like to make a comment? Okay. Anyone in the room?

CHAIR WEINSTEIN: Thank you for all your help. And again, I want to iterate for the Committee how much we appreciate the work of ABMS and Ingenix. This is incredible work that is really important, really, really

important, at this time.

So thank you and I hope you take our comments as just being complimentary and helpful.

MS. WILBON: And I would like to thank Dr. Weinstein for leading the group today and everyone for, you know, traveling near and far to get here to discuss the measures. And we appreciate your work.

And we were hoping to not have to do any follow-up, but, obviously, things happen, so we will communicate with you by email as much as possible. And if it warrants another phone call, we will, you know, get that arranged for some time later this summer.

So again, thanks to everyone and feel free to call me or email me with any questions or things that come up and we will keep you informed.

Great. Anyone? Okay. Thank you.

(Whereupon, the Technical Advisory Panel Meeting was concluded at 3:27 p.m.)

D

M

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Bone/Joint Technical Advisory Panel

Before: NQF

Date: 07-07-11

Place: Washington, DC

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