NATIONAL QUALITY FORUM

+ + + + +

RESOURCE USE PROJECT:

PHASE II CANCER TECHNICAL ADVISORY PANEL

+ + + + +

TUESDAY

JUNE 28, 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., David Penson, Chair, presiding.

PRESENT:

DAVID PENSON, MD, MPH, Vanderbilt University Medical Center, Chair

ROHIT BORKER, PhD, GlaxoSmithKline

STEVEN CHEN, MD, MBA, University of California-Davis

TIMOTHY GILLIGAN, MD, Cleveland Clinic Taussig Cancer Center

DWIGHT KLOTH, PharmD, Fox Chase Cancer Center

LOUIS POTTERS, MD, FACR, North Shore-Long Island Jewish Health System

JAY SCHUKMAN, MD, Anthem Blue Cross and Blue Shield

JOHN SKIBBER, MD, University of Texas-MD Anderson Cancer Center

LOUISE WALTER, MD, University of California-San Francisco

NQF STAFF:

TAROON AMIN

HEIDI BOSSLEY, MSN, MBA

LAURALEI DORIAN

SARAH FANTA

CAMILLE PRESBURY

SALLY TURBYVILLE, MA, MS

ASHLIE WILBON, MPH, BSN

CARLOS ALZOLA, NQF Statistical Consultant

ALSO PRESENT:

TODD LEE, PharmD, PhD, American Board of Medical Specialties (ABMS) (via phone)

ROBIN WAGNER, ABMS (via phone)

KEVIN WEISS, MD, ABMS (via phone)

TABLE OF CONTENTS

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

8:39 a.m.

MS. TURBYVILLE: Welcome, everyone. We are going to start this morning with just some welcome and introductions, and then we will go into the disclosure of interests.

So we are so pleased that you are here today. We are really thrilled with the work that we have done with you all so far, and we are really looking forward to evaluating and moving through the measures today and hearing what your thoughts are on the measures that have been submitted, and capturing your ratings as we move through the meeting.

I want to give Heidi Bossley, who is our Vice President, Performance Measurement, an opportunity to welcome you as well, and then I will turn it over to your Chair of this Technical Advisory Panel. You don't want to --

MS. BOSSLEY: No. You have to tell me what I am supposed to do disclosures, too. But we are thrilled to have you here, truly appreciate the amount of work we are asking you to do. We have recognized that, and it is truly appreciated by NQF. Thank you.

MS. TURBYVILLE: So, David, if you wouldn't mind.

CHAIRMAN PENSON: Yes, sure. Thanks, Sally. I think the way to do this to begin with is probably to do introductions, and then maybe do the disclosures first, if there are any. Then we can talk a little bit about our goals today, and sort of the process.

So why don't we around the table and introduce ourselves. I will start. My name is David Penson. I am at Vanderbilt University, Nashville, Tennessee. I am a urologic oncologist who also does health services research, and I run our Center for

Surgical Quality and Outcomes Research at Vandy.

DR. WALTER: I am Louise Walter. I am from the University of California-San Francisco. I am a geriatrician, and I am also a health services researcher at the San Francisco VA.

DR. SKIBBER: I am John Skibber. I am a surgeon at MD Anderson Cancer Center, and I am the Chief Surgical Quality Officer there.

DR. POTTERS: I am Louis Potters. I am a radiation oncologist. I chair Radiation Medicine for the North Shore LIJ Health System on Long Island.

DR. BORKER: I am Rohit Borker. I work for GlaxoSmithKline. I am in the U.S. Health Outcomes Group, Director, Oncology. Are we also doing disclosures?

MS. TURBYVILLE: We will do that after. We have instructions. Thanks.

DR. SCHUKMAN: Good morning. I am

Jay Schukman, Senior Medical Director of Anthem Blue Cross and Blue Shield here in Virginia and a Regional Vice President with WellPoint for the East Region.

MS. BOSSLEY: Heidi Bossley, VP of Performance Measures in charge of the CDP process.

MS. DORIAN: And I am Lauralei Dorian. I have recently started here at NQF, and I am happy to be working on this project.

DR. KLOTH: Dwight Kloth, Director of Pharmacy, Fox Chase Cancer Center and Secretary of the Pharmacy and Therapeutics Committee which has a lot of linkages to Joint Commission, compliance, quality and so forth.

DR. CHEN: I am Steve Chen. I am oncologist at UC-Davis for a few more days, and then I will be at City of Hope, and a health services researcher as well.

DR. GILLIGAN: I am Tim Gilligan. I am a medical oncologist at the Cleveland Clinic.

MS. WILBON: Good morning, everyone. I am Ashlie Wilbon. You have all received a lot of emails from me probably. It is nice to see you all, and I am a Project Manager for this project.

MR. AMIN: Hi, everybody. Taroon Amin. I am a Senior Director here working on both work with the CDP and on the Measures Application Partnership.

MS. TURBYVILLE: And I am Sally Turbyville. I am the Senior Director on this project in performance measures, and again just thrilled to have you here, and also want to make sure I try to take a moment to thank the team for all their hard work in getting us ready for this meeting today.

CHAIRMAN PENSON: Great. Thank you, everyone. I will echo what the NQF folks have said. Really, everyone did a yeoman's effort getting the reviews done in advance. I think that will help us today, I hope.

Before we get into that, we

probably should do the disclosures now.

MS. BOSSLEY: Okay. This normally is done by our General Counsel, and I have been given a script just to make sure I do it correctly.

As you may remember, we asked you to fill out a disclosure form a while ago, and we are asking you to orally disclose. You don't need to include everything that you put on your form. I would just include those things that are relevant to this project. So if you have any grants, receive any speaking engagements with any organization that would be relevant to cancer care, I would probably just include that.

I would also remind you, you all are here as individuals. You are not necessarily here on behalf of your organization, and we do ask you to represent yourself with your expertise.

We are going to have you go around. You all have introduced yourselves.

So I would just say whether you have any disclosures or not, and then we will actually follow up and make sure if anyone has any questions for anyone who has disclosed something, we will give you an opportunity to do that as well.

DR. KLOTH: We should reiterate what we had previously submitted electronically?

MS. BOSSLEY: If it is relevant to the project, yes. Yes. So why don't we start.

DR. WALTER: I have no disclosures.

DR. SKIBBER: John Skibber disclosure is I am on the NCCN Executive Committee for their colorectal cancer database.

MS. WILBON: I'm sorry. Could you just say your name before you have your disclosures for our transcription, so we can associate whose disclosure goes with who.

Thank you.

DR. SKIBBER: That was John Skibber.

DR. POTTERS: Lou Potters. I have no disclosures.

DR. BORKER: Rohit Borker. I work for GlaxoSmithKline, and I have stock ownership in GlaxoSmithKline, Amgen, and other pharmaceutical companies.

DR. SCHUKMAN: Jay Schukman, no disclosures.

DR. KLOTH: Dwight Kloth. I am on two guidelines panels for NCCN, antiemetics and myeloid growth factors. I have done speaking or consulting for Amgen, SI, Novartis. I think that would cover the relevant parameters.

DR. CHEN: I am Steve Chen. I have a few disclosures. I do have a grant for the California Breast Cancer Research Program. As far as industry disclosures, I have a research contract with Agendia, Inc. With

Genomic Health I have a pending research contract, and with LifeCell, I have a protocol under review.

DR. GILLIGAN: Tim Gilligan. A couple of things to disclose: I chaired a panel for the American Society of Clinical Oncology on the use of tumor markers for germ cell tumors. I am on the NCI PDQ Adult Treatment Editorial Board that writes treatment summaries for all adult cancers.

I am oncologist at Cleveland Clinic, who has a big stake in this, and I in 2010 received a one-time fee for speaking from Pfizer.

CHAIRMAN PENSON: And this is David Penson. I have pharmaceutical disclosures or industry disclosures. I do service as the Vice Chair for Health Policy for the American Urological Association and, as such, I am a paid consultant to the Board of Directors for the AUA.

MS. BOSSLEY: Okay. Does anyone

have any questions for any of their colleagues on anything they have disclosed? Thank you very much.

MS. TURBYVILLE: Are there any panel members on the telephone as of yet?

CHAIRMAN PENSON: All right. A few people are in the room but not at the table. If you all would like to introduce yourselves, that would be great.

(Audience introductions.)

CHAIRMAN PENSON: Thank you. Is anyone on the phone? All right, I think it is just us.

So let me just say a few words about this process. I also double as a member of the Steering Committee. So I get double the work and half the fun, I would say.

So my goal today, I was telling Sally beforehand, is to get through all four measures, so we don't have to do a painful conference call. That will be the reward and, as such, what we are trying to do here -- We

will try to stay as close to the agenda as possible.

I would like everyone to feel free to speak their mind, but at the same time, I would like to encourage people not to repeat what others have said, just for the sake of time.

These measures, compared to some of the others we have seen in this arena, are relatively straightforward. So I hope that the discussion will be relatively streamlined, and the work that everyone did in advance was really helpful; but again, that being said, I want to ensure that everyone has a chance to say what is on their mind and their thoughts.

Going through the agenda, we will start -- we are actually running a little bit ahead -- with the colon cancer measures this morning. We will go through the colonoscopy measure first, and we have left about an hour and a half for that.

We will take a break, and then we

will do the episode of care for treatment of localized colon cancer at 11:15. We will have some time at 12:25 for members and public comment, folks on the phone, if anyone joins us. We will take a lunch break at 12:30 and start again at 1:00 with the breast cancer measures.

We have two breast cancer measures to do. I think everyone has been through these. One is episode of care around a case of newly diagnosed breast cancer, an episode of care around a breast biopsy, and then again we will have time for public comment and any member comment, and then we will wrap up.

The way this works, my understanding of it -- Sally, correct me if I misunderstand -- is that as the TAP we go through these measures and provide our opinion as to whether it meets the criteria, high, medium or low or insufficient, as it were, and these recommendations then go up to the Steering Committee and then go up to even a

committee above that for final approval.

My suspicion is that what this TAP says will probably swing the Steering Committee. So it is important, as you put your votes through, to understand that what you say here will determine whether or not this gets NQF endorsement.

With that in mind, I want to just say a few words about how these measures are evaluated. As part of the Steering Committee, this was a new realm for NQF. I don't know how many of you all have served on NQF panels before. Having served on patient outcome measures and other quality measures, those are easy compared to these.

I think the Steering Committee really had a hard time wrapping themselves around some of these issues. For example, because they are cost measures, there wasn't the level of evidence required for a quality measure, because it is fairly generic. A dollar is a dollar, within reason, obviously.

So as you go through this, that level of evidence isn't required.

Another key point that I think, as a Steering Committee member, those of us on the committee who are clinicians had a hard time accepting was the concept of we are just looking at dollars, and we are not looking at quality, and we really want to get efficiency and value.

One of the key points, one of the key principles of NQF, is that these should not be used in isolation. They need to be used with a quality measure, but I want, for lack of a better way to put it, folks to suspend disbelief today and accept that. It is hard to do. I am with you on that.

It was really -- There was a very long discussion in the Steering Committee, and I think a lot of us, myself included, were very uncomfortable and still are uncomfortable, but have gone from very uncomfortable to the only way we are going to

do this is if we accept that. I think that, if we can get this piece done, we can start talking about efficiency and value in the next iteration.

As far as the criteria, I won't go through them, because when I saw the responses that people gave -- I had originally thought may be I would send out an email to everyone until I saw your responses. I think people really do understand what we are trying to get at here.

In some ways, I was telling Sally, we took the measure -- the criteria that -- the existing criteria that they used for the quality measures and put them into the cost measures. To some degree, it is a round peg in a square hole, and to some degree it was sort of taking an octagon and knocking it into a circle.

I think the Steering Committee did as best a job as they could without completely reinventing the wheel. So we really have

these four areas to deal with: The importance to measure and use, with its four subcriteria, and I won't go through it unless there are questions or discussions that people want to have with it.

There is the acceptability, which I think is going to be the majority of our discussion. I don't think people are going to argue a whole lot about importance. There may be some discussion, but I think the issue really comes up around scientific acceptability, and to some degree usability and feasibility, but it really is in the scientific acceptability; risk adjustment, does this make sense; accountability issues.

So I think we are going to spend the majority of our discussions with each measure on criteria number two, and I think that is appropriate.

The other thing I would say -- and, Sally, correct me -- we will obviously go through each measure and each criteria and

each subcriteria in turn. Do you want us to vote on each subcriteria at the end of a discussion?

MS. TURBYVILLE: Yes. So what we have done in the past, and it worked pretty well, is to have all of you using the clickers that were handed out, rate each subcriteria once the discussion of the criteria is complete, so that we can keep moving through the whole process.

So we are going to do a little recap on some presentations, have people talk a little bit about how the voting works following, but as we go through, so that we don't have to wait until the end of the day and then go back and recap.

One of the things that staff will do as you move through measures is try and make sure we don't have rating creep, so that we are not getting harder or easier as we move forward in other measures, and we will kind of remind you what you thought before. Doesn't

mean you can't go back and say, well, actually, now we need to change that. It is just to try and help you find that consistency throughout the day.

CHAIRMAN PENSON: To that end, I sort of -- This is a personal opinion. The way I approached the high, medium, and low or insufficient, which I think is another criteria, was I looked and I said, in my mind using sort of the NIH criteria for a grant review, if it is high, there are really almost no weaknesses or just it's fine.

If it was moderate, my feeling was that the strengths outweighed the weaknesses. The weaknesses were moderate at most, but it wasn't enough to kill it, in my mind. It was, when you got to low or insufficient low, where it was just basically, this ain't going to fly, it doesn't pass the smell test, or insufficient, I just wasn't convinced, but perhaps it can be addressed.

I think that, if you sort of keep

to that mentality -- and I think everyone did when they were going through this at home -- I think we will be okay.

So I am trying to think if there are any other issues that need to be addressed. We do have a report from our statistical consultant, and I think that will also be very helpful, who just walked in the room. So do you want to introduce yourself?

MR. ALZOLA: I am Carlos Alzola. I have been involved with this process for a few number of measures now, and I met with Sally and some other people here.

CHAIRMAN PENSON: Thank you. The only other thing I will say as far as -- and then I will turn it over to Sally and the team: You know, we have this Excel spreadsheet in front of us, and I can see it, and I think everyone else can see what everyone else voted on.

I don't want to call people out. So in other words, looking at some of the

results, there are places where everyone said high or medium or one person said low, and I will just basically, stead of saying, you know, so and so, why did you rate it low, I will let you guys speak out, because it may be as the discussion goes on, you change your opinion, and if you had changed your opinion, that is good. We are trying to arrive at some sort of general consensus.

MR. AMIN: As we sort of walk through the discussion, if there are elements that are rated more low or insufficient, the more detail that we can get on the rationale for that scoring process, the better we can provide that feedback to the measure developers.

CHAIRMAN PENSON: Okay. Sally, go ahead.

MS. TURBYVILLE: So I am going to hand it over to Ashlie. We want to spend a little bit of time making sure we are all on the same page. So I know some of you have

seen this before. We might speed up and slow down as we go through, when are hitting on something new, try to slow it down. That said, if we are going too fast or too slow, feel free to signal to us, please.

DR. SCHUKMAN: I just had a question on the resource use, how that will ultimately tie to the quality measures, and how granular can you get tying those two together at a high level. I am just trying to understand what the work product is going to look like or if we know that yet.

MS. TURBYVILLE: We don't know that yet, and I appreciate David's perspective in the Steering Committee. I think there is a very strong guiding principle, both from the Steering Committee in resource use and, certainly, the TAP is welcome to echo that sentiment if it is agreed upon, and then there is also a principle from NQF that quality should be a part of this.

What we learned as we went into

developing the criteria and thinking through evaluating measures is that where we are now today, both in the industry as well as our conceptually evaluating, we needed to at least get resource use measures evaluated as a building block as we move toward value, and knowing that we have over 600 quality measures currently endorsed.

So I don't know that we know what the exact match is, and I don't think all the developers know this yet either. So, hopefully, this will continue to encourage us to get there. So it is really to kind of push us to keep on thinking about value and efficiency, I would say.

CHAIRMAN PENSON: The only thing I would add to that, because, Jay, I am with you 110 percent on this -- How long did we spend on this? A good hour, and it is heated. The one thing that gives me a little bit of relief in this committee is that, where cancer is concerned, there are a lot of quality measures

specifically tied to colon and breast cancer.

So you could see the start of a value ratio developing here. It is much more problematic, speaking as a Steering committee meeting, where we are looking at non-condition specific measures. You know, general episode groupers for care in a managed care population over the course of a year, it is easy to measure what resource use is there, but what is the denominator there, and that is a bigger problem.

So it is no consolation. Again, I will ask you to suspend disbelief to some degree.

While folks are getting set up, there is one other thing as I look through my papers. The NQF team had assigned the panel sort of lead reviewers for different elements. If people are comfortable with that, I will ask each person as we go through that to sort of lay out their thoughts in two or three minutes, and then have an open discussion.

That may facilitate things a little bit. Go ahead. Thank you, guys.

MS. WILBON: Good morning, everyone. We just have a few presentation kind of introduction slides, as Sally mentioned, to kind of get us all on the same page. We are going to go over a little bit just a couple of slides on the consensus development process. You can figure out and kind of visualize where this TAP fits in the overall process for these measures.

Then we will go into an overview of the criteria and the subcriteria, and then a little bit about the meeting process toward the end.

So for the meeting, we do hope that in evaluating the measures -- you guys are the second TAP to meet -- that we will be able to, hopefully, at the end draw out some lessons learned and be able to move that forward through the other TAPs that are going to be meeting. So any feedback that you have

throughout the process for how we can improve things and, hopefully, make things more efficient, we are welcome and open to that.

So the consensus development process is approximately an eight step process. We have already done the first two steps that are in gray, which is for a call for nominations, which is what got you all here, and the call for candidate standards.

We are now in the consensus standards review process, which for this project we will go into a little bit later. It is a little bit longer and more elaborate than most projects.

Then after we have recommendations from your ratings and then the recommendations from the Steering Committee, we put those out, and the measures back out for public and member comment. Those comments are usually integrated back -- or sent back to the developer for any improvements or changes that are needed, and then we put the measures out

for member voting.

All of that goes on to our CSAC, and the CSAC is our Consensus Standards Approval Committee. They are an oversight body that we have here at NQF that oversees all the projects, the consensus development projects, and ensures that the process was followed and that the recommendations that were made were in alignment with the criteria.

The CSAC decision is then passed on to the Board for ratification, and then after -- and at that point the measures would be endorsed. After the endorsement process, we do actually -- endorsement stamp of approval -- we do actually have a 30-day appeals period where developers or public can send in any concerns for how the measures or the outcome of the project.

So this is just a pictorial of how the process works. obviously, what is in yellow is kind of where the Technical Advisory Panels feed into the process.

So this is just a brief slide to kind of -- so everyone understands what is the definition we ended up with, particularly through our work with the Steering committee in the first year of this project, that resource use measures are broadly applicable measures that compare health service counts in terms of units or dollars, that can be applied to a population or event, and are broadly defined to include diagnoses, procedures or encounters.

Those counts can be the frequency of defined health system resources. some may further apply dollar amount, allowable charges, etcetera, to each unit of resource.

So for this particular project, as I said, we have a little bit more of an elaborate process or standards review process. We realize that because these were new measures, the first time we have ever evaluated them, that we wanted to give ourselves a little bit more lead time with the

first effort, which was with our cardiovascular diabetes TAP.

They have already met, and we are still working through those measures, and that was what we kind of coined as Cycle 1, and with a little bit of lead time on Cycle 2 where we would, hopefully, be able to kind of feed in any lessons learned for the remaining TAPs that we have started now.

So you guys are the first TAP in Cycle 2. You will be followed by the bone joint and the pulmonary TAP in the next few weeks.

So these are just -- I won't spend any time on this, but it is just kind of a high level timeline of how we hope to get through each of the cycles for each of the consensus development process steps.

So this is the process, the general review process that we aimed to use for this process, and we do do a brief staff review when the measures come in to make sure

that they are complete, to the best of our ability. We try to do follow-up with the developer for any empty fields or things that don't make sense, cut and past errors.

We don't catch everything, but we do try to do a little bit of clean-up before they go out to the panel, and then to send in to the statistical consultant, who is Carlos, and send that feedback to you guys so that you can, as you begin your review, have that additional -- any guidance around the scientific acceptability to help you further evaluate the measures.

So for those measures that are condition specific, they go to the TAP first, and then, as Dr. Penson mentioned, your review will then be forwarded to the Steering committee for final recommendations.

So the role of the TAP is really to evaluate each of the subcriteria, identify strengths and weaknesses of the measures, and particularly focus on the clinical construct

of the measure.

Our Steering Committee was built with a little bit of different expertise. So our TAPs are specifically built with clinical experts and methodologists to really give the measure -- to ensure that the measures have that technical review.

So what we are planning on doing today in terms of the review and with the TAPs is to do a systematic criteria by criteria evaluation. We will start with importance, and then move sequentially through each of the criteria and subcriteria, and to really be focused on how well the developer has demonstrated that the criteria has been met.

You will rate the criterion, and we will kind of go into, in some later slides, about how the electronic voting tool works.

So I will hand it over to Sally.

MS. TURBYVILLE: So what we are going to do today is finalize the evaluation and rating of the four submitted measures

using the subcriteria that we sent out earlier to all of you, was finalized last year with guidance from the Steering Committee, as David mentioned, and we do ask the TAPs to focus on the subcriteria, which then will help the Steering Committee, as Ashlie mentioned, make determinations about the overall criteria.

So as you know, there are four major criteria. They are: The importance to measure and report -- specifically, is the topical area that is being selected an area that is of interest to think about measurement of resource use; certainly, the scientific acceptability, which has the reliability and validity component of your evaluation, as well as thinking about the risk adjustment approach, and other kind of components of the measure construct; how usable is the measure itself, you know, the final results; and then feasibility.

I already briefly talked about the measure to report, how important it is. We

are looking at four subcriteria: A national goal -- is it a problem area with opportunities for improvement? Is the purpose and objective, as submitted by the measure developer, clear to you as the reviewers of the measure, and do the resource units that they are capturing for the resource use measure make sense to you as experts in this area?

So it is feasible that the area would be important, but the resources that they are requesting be captured as part of the measure potentially don't jive. So we ask you to think about that.

Scientific acceptability, as we talked about, reliability, mainly: Are the results consistent? Can they be reproduced or are they not -- have they not been able to demonstrate some kind of reliability or generalizability as far as it being able to be something that we would run the specifications and see the results as they should be?

How credible is it? There are a couple of areas in validity, both for the measure itself, and then when we tie it back to importance -- Importance, remember, is a topical area. Validity is the opportunity where you say, well, in how they constructed the measure and the area that is presumably important to measure, are those tied together? So that we will ask you to talk about that, and disparities, if and when applicable. We can talk about that a little bit more as we get to that discussion.

So as David mentioned, when we are thinking about the rating, we just wanted to provide you a little bit of guidance on high and moderate and low, mainly focusing on the liability and validity, but I think you can see some of the gradations between high, moderate and low, or when we are thinking about high reliability, there should be both clear specifications that could be implemented in a standard way. So they would have to be

unambiguous.

There would have to be someone that a user could take and then implement the measure, and that there is empirical evidence of the reliability of the measure, both for the data elements -- so the data that are being captured -- as well as the measure score. So can the score that the measure is producing -- do they demonstrate that it is reliable?

Validity has the same kind of both things happening at the same time. So are the measure, as I have said, specifications consistent with what they described important, and then is there empirical evidence of the validity, both for the data elements that are required in order to specify the measure, as well as the measure, and then whether flex to validity are empirically assessed. That is a high.

So as David said, really, you don't feel there are any weaknesses in the

measure as it is submitted. The testing that they submit is relevant, makes sense, and is complete.

Moderate would be that the specifications are unambiguous, but for reliability they may empirically demonstrate only that one of the two, the data elements are reliable or the score. So we are looking for one or two, and then validity similarly. You have the empirical evidence of validity, but again there is this "or" about the score or the data element itself. Still, there should be some assessment of flex.

Then low: Really, this is when the specifications perhaps have some ambiguity or requires some improvement, and that the empirical evidence doesn't demonstrate reliability. Same with validity, it doesn't reflect the evidence or the intent as described in the measure importance, doesn't jive with all of you, or the empirical evidence is demonstrating that the measures

are actually not valid or flex and validity have not been assessed.

As David mentioned, there is also insufficient evidence. If it is just an inappropriate method or scope of reliability testing, and you want these developers to either take the time in this project or in future projects to do different testing, that will be a signal to that, as well as the validity.

Any questions about that before we move on? I just want to make sure we are all on the same page as you move through your voting of the measures. I think everybody understands that pretty well, from the expressions.

So I just want to briefly touch on how we thought about resource use measures. We did want to make sure that we were able to capture different types. So we have episode based measures, procedure measures, population based measures. I believe all four of these

measures are episode based measures. So under the project, though, we did collect other types of measures as well.

I don't want to spend too much time on this, because all of you did do the preliminary review, and as David said, based on the notes we got, it seemed like there was a good grasping on what they are.

We had four areas that we asked developers to submit guidance or specifications on. So we had data protocol: What data are needed? Data cleaning steps. Clinical logic has to be specifications, has to be unambiguous. As Ashlie said, certainly, as a good group of clinicians here in cancer care, we really want you to take a deep dive into the clinical logic of the measures.

The construction logic, which are those steps that are beyond the clinical logic: Sometimes there aren't any steps beyond the clinical logic, but sometimes there are, that are just parameters around the

measure, time periods, etcetera, that need to be applied. Then certainly, the risk adjustment and costing methods, and then reporting guidelines. How do you attribute the results of the measure? How do you define the peer group in order to create your benchmark for resource use?

This is just an illustration of what we were talking about, and I don't want to spend too much more time on this. You can see that, where the general methods were submitted, that was just to help us understand. It is not the part that is subject to your evaluation. However, it does inform you so that you can then evaluate the data protocol, the resource units, the clinical logic, and the construction logic, the adjustments for comparability, and then the reporting.

So we already talked about the reliability and the validity. As you can see, there are two subcriteria for reliability, and

there are six subcriteria for validity, and we will walk through this.

What is going to come in real handy -- could I borrow this for a second? -- is as we are walking through the measure, if you just pull this out as we are evaluating it and have it in front of you, in our experience it is really helpful, because we will use it to guide us through the rating process.

Again, there are these six criteria, subcriteria, for validity. Disparity: Just to note, one of the TAPs before talked about disparities, in particular race, ethnicity, socioeconomic status, and to us typically, we don't want these types of things to be risk adjusted away. We want to reveal these kinds of differences, but we may ask developers to stratify by the population so that action can be taken.

What we have had discussions about so far, and we are going to ask the Steering Committee to further discuss it, and they

started their conversations about it, is whether it really makes sense for resource use at this time. The evidence isn't really clear. These resource use measures are not measures of appropriateness, so something that we would certainly welcome your expert input as well, once we get there, and what your thoughts are on that.

Then usability. I want to talk a little bit about usability here. So this is the spectrum of what we think about when we want to -- when we are talking about measures. You can see all the way on the left is benchmarking. When we are saying benchmarking, we are talking about those internal quality improvement measures.

Those are the type of measures that NQF doesn't endorse, because they don't need that national standardization for implementation. They are within a system or within a network, and it is just being used for quality improvement and tracking.

As we move over to the right for accountability, these are the measures that we are interested in endorsing and making sure that there is a standard specification, so that people can implement them in the same way, compare results, etcetera.

So as you can see, when we talk about usability and we talk about public reporting, we are not just talking about the very end, which will be public reporting to all, but we are also talking about these accountability models s well.

I don't want to spend too much time on this, because I think David did a fantastic job of recapping both the Steering Committee's sentiment about what this project is doing and how it fits into this concept of efficiency and value as well as NQF's position.

Also, you can see here, if we think about our integrated measurement framework, certainly, we realize that the

resource use measures and cost can be examining slices of a clinical episode or it can be looking at the entire episode or trajectories from the episode. So just a contextual illustration. We don't have to spend too much time on this.

I think we have enough complicated things to think about.

CHAIRMAN PENSON: I love that cartoon.

MS. TURBYVILLE: Is it a snail or is it a -- I'm not sure -- caterpillar perhaps? Okay.

Then feasibility: We certainly want to make sure any measure that is endorsed by NQF is something that can be implemented today. So while we understand that there are measures that we certainly are very interested in but the data perhaps aren't quite available, I don't think that is an issue typically for the measures that we have seen come through, because they have all been

claims based.

So just as a reminder, we will pause here and there to make sure that we are opening it up for public and member comment as well as the public and members who are here in the audience, and we may also at times try -- you know, if we think we are getting too off time, to try and avoid conference calls, we might try and see if there are any other inputs, and I know David is certainly going to help guide us through that as well.

We will ask the measure developers to briefly introduce the measure. They should be available to ask questions, and I think they might be on the phone now. Then we will ask you to weight this criteria.

I think we went through the evaluation's process. We start with importance and then move -- We will probably have feasibility go pretty quickly, because it is administrative claims data, and then dive into the areas where there really should be

some distinct differences measure by measure.

These are the measures we are looking at. Thank you again for all the reviewers on these measures, and then I think at this point I am going to hand it over to Sarah to make sure you get the instructions that you need in order to rate the measures, because it is an application.

MS. FANTA: Good morning, everyone. All right. So each of you should have received a little remote. That was particularly assigned to you, so we know which way everyone voted. We will be using these as we go through and rate each subcriteria.

I actually have the receptor on my computer. So as you vote, if you want to just point in my direction, I can pick up the votes very easily.

As you can see, the keypad is numbered zero through nine, and there is a caution symbol in case you want to delete your entry, change your rating. Then if you just

press Send after you make your selection, and point at me, then I will get the rating.

Then here is just an example of the types of things you will be voting on. So one, Yes, and two, No. You would just press it on your keypad, and point to me. This is just an example.

If you want to modify your response, you can just press the Caution button, press the number you meant to send, and then press Send. You will have 60 seconds to vote, and there will be a live tally, letting you know how much time is left, and then once everyone has voted, the voting results will be displayed on the screen.

MS. WILBON: There is also a little handout in your folder that gives you instructions on how the -- if you kind of get mixed up, if you want to refer to it before we vote, you can do that as well.

MS. FANTA: All right. Here is just our next steps and upcoming dates. We

actually have a Steering Committee meeting tomorrow and on Thursday. So we have been very busy. I will be discussing the CV diabetes measures as well as non-Commission specific measures, and as far as this meeting goes, NQF staff will serve as the liaison between this TAP and the measure developers.

So if there is any follow-up needed or any questions that need to be answered, we will definitely be communicating with them and then reporting back to you any progress that has been made.

Your final TAP ratings will also be sent to the Steering Committee to help inform their decision as they go through the cancer measures, and we will schedule any follow-up calls as needed, but hopefully, we will get everything done today.

CHAIRMAN PENSON: I am optimistic. I really am. All right. So far, so good; don't jinx it.

All right. So I guess we really

should start. The first measure we are going to look at is 1583, which is episode of care for a 21-day period around colonoscopy.

Is someone on the phone from ABMS Foundation?

DR. WEISS: Yes. Kevin Weiss is here. We also have Robin Wagner and Todd Lee.

CHAIRMAN PENSON: Great. Good to hear from you, Kevin. It's Dave Penson. What I will ask you to do, as Sally had mentioned, is maybe spend just a minute or two or three just discussing -- introducing the measure to the panel, and giving us sort of a broad overview.

I think everyone in the room has poured through the materials pretty well. So I don't think you have to go into great detail, just the high points to get our discussion started, if you would.

DR. WEISS: Sure. That would be great, and maybe give a framing. Also, just a note, which I appreciate that you noted that

this was coming from the ABMS Research and Education Foundation, not from the ABMS per se.

This project was funded by the RWJ. We looked at types of care measures, and the philosophy of the project. You will be looking at several measures today from the project. So I can maybe give you that general, and won't have to repeat it later.

It was to develop measures that were physician led around the cost of care, and in doing so, we looked at the areas that will have priority, and we were pleased that this particular measure and several others met the needs of NQF in that prioritization.

We have -- In looking at this measure, we have looked at one or two types of measures we present. One would be high frequency measures that are very common across the health care system, but maybe not large in individual cost, but recognized that there was a perception of variability.

To develop these measures, we brought together a multi-disciplinary team of physicians and non-physicians to have us vet what they thought was an important measure that could be linked to a quality measure at some point, recognizing that the cost of care measures by themselves do not provide enough information and could actually provide wrong information if not paired with quality measures.

In asking them to look at the issue of colon cancer and colonoscopy opt-out; and the measure that we present to you today here looks at the concept of an episode of care around colonoscopy, it was felt that there was enough variability in the process of colonoscopy, both in types of approach in terms of -- specifically, in terms of whether anesthesia was used and also in terms of numbers of biopsies and approach to pathology within that process, so that all those collectively warranted enough perceived

variation that they wanted to, and did, develop the measure which we present today.

The episode begins before -- just before the colonoscopy, recognizing that there may be some preparatory work which is involved, and then follows up through a period that was felt to be appropriate in terms of when the acute complications might arise, and correctly follow from colonoscopy.

I will stop there, just to check and see if Todd Lee, who is our health economist, would like to add anything.

DR. LEE: No, Kevin. I think you have captured all of the important factors.

DR. WEISS: Great. Probably the only other thing that is worth noting is that we will be presenting in this model a risk adjustment model for trying to manage the issue of the complexity of individual patients, and that model is one that is consistent to other models. So you may or may not decide to spend extra time on that part of

the discussion, because otherwise, you know, if you repeat it to other measures you will see today.

Was that the type of overview that would be helpful?

CHAIRMAN PENSON: Yes, I think that is very helpful, and I just -- I will ask the people in the room, is everyone comfortable with the measure, their understanding of the measure? Any questions right off the bat for the ABMS Foundation folks? Okay. Well, I have got noes around the table, Kevin. So I think everyone is with us here on the same page.

So I guess we will start the discussion around the measure. This measure specifically, when we talk about distinguishing this, in the afternoon, from the breast cancer measures, is accountable to the level of the provider, and I think that is going to shape the discussion, and I think the risk adjustment piece -- we are going to spend

probably some time on that.

Let's start with the importance to measure and report. The first subcriteria, 1a, deals with whether or not this is addressed as a national health goal or priority or is a high impact aspect of health care.

Pretty much across the board, everyone rated this as high. There was one or two people who said it was moderate. I will open up the floor to any comments. I think we all know this is a fairly common condition. Colonoscopy is very widely used. It runs up a bill in many places.

Do other folks have things to add to that?

DR. GILLIGAN: Just briefly, I think we probably get the high, just based on what we know walking into the room. The application itself here actually didn't provide any evidence that it is high. I think most of us just believe it is high.

The first paragraph talks about breast cancer, which was probably a typo, and then the rest of it talks about colon cancer treatment costs, and colon cancer treatment costs have very little to do with colonoscopy costs.

So I think we kind of threw them a bone on this one and said, well, we know it is important, even though you haven't actually shown us any evidence that it is important.

CHAIRMAN PENSON: Yes, I think you are right about that, Tim, and I do think that comes up in some of the other criteria. I notice you and I kind of hit on the same thing, but I think in the overall impact, I am getting the impression that everyone felt pretty unanimously that that was high.

MS. TURBYVILLE: Can I ask a question? So given that these are the experts here, and this would be potentially an application that would go out to the general user, is the missing information because there

isn't as much literature and evidence of that, and so the experts here are able to say, even though the literature provided was more downstream, as people in the field we know there is variation; or is there an opportunity for there to be evidence cited, not clinical evidence but evidence of variation, to support the importance of the measure a little bit more directly?

CHAIRMAN PENSON: Well, Sally, let's -- The variation piece comes up in the next subcriteria. So let's table that for just a minute, because I have an issue there. Let's just start right into very high impact piece of it. Rohit?

DR. BORKER: Hi. This is Rohit again. I'm sorry. I wanted to -- We agree with what Tim and David were saying. I think, in terms of high impact, the literature that would be really needed that will help us is it is not just the cost of colonoscopy, but what proportion of total colorectal costs are --

colonoscopy costs; because if they are minimal, then doing anything in colonoscopy is not going to help reduce that total burden. So that is the missing link, to me.

CHAIRMAN PENSON: And I do think we are going to come to that in the next piece.

So just to keep us moving along -- Sorry, Jay, go ahead.

DR. SCHUKMAN: I was just -- You had mentioned something earlier. Is the denominator going to be at the individual level? Is that correct?

CHAIRMAN PENSON: So if you look in the report, what you basically see is reported out by provider, and my understanding is I interpreted the measure -- and we can ask the folks on ABMS Foundation on the phone, but it is actually probably going to be attributed to the provider who performs a colonoscopy, would be my guess.

DR. SCHUKMAN: Okay. Is that

going to be generally true for all these measures?

CHAIRMAN PENSON: No, sir.

DR. SCHUKMAN: Okay. I just wanted to be clear on that.

CHAIRMAN PENSON: Yes. So this is just for this measure. it will be accountable to the -- most likely, the gastroenterologist who performs the colonoscopy or the colorectal surgeon.

DR. SCHUKMAN: The other question I had, to follow up on that, is how we are going to gain access to the administrative data to do this.

CHAIRMAN PENSON: Well, I will ask Kevin and the team from ABMS Foundation on the phone. I am assuming it is probably publicly available data like Medicare, etcetera. But, Kevin?

DR. WEISS: I don't -- Let me ask.

CHAIRMAN PENSON: I'm sorry?

DR. WEISS: I don't understand the

nature of --

CHAIRMAN PENSON: So the nature of the question was -- and we are sort of getting ahead of ourselves here, because we will come to this a little later, but we might as well just address it now.

As far as administrative data used to populate the measure, can this be done from publicly available data like Medicare or Ingenix, etcetera?

DR. WEISS: Correct. It can be gotten from administrative data available that would either come from a private or public payer. It could come from a health system, if they have got comprehensive use of those administrative data or a health system that has all the different data elements.

CHAIRMAN PENSON: Jay, does that answer your question?

DR. SCHUKMAN: Yes, it does. Thank you.

CHAIRMAN PENSON: So the question

is, if you -- I think that we will keep moving along, because there are going to be other issues. Do you want to vote now on each or later at the end? Okay, great.

So let's move on to 1b, and b is the demonstration of resource use or cost problems in opting for improvements. So this is really where you say there are issues with variation in delivery of care across providers or groups.

There is a little bit more variation here. I probably was one of the people who was most sort of had problems with this, and I would share with you my thoughts on this and then open it up to the floor.

There is no doubt in my mind that there is going to be variation in care around colorectal cancer geographically by provider, etcetera, but understand that the inciting episode here is colonoscopy, and you going around a 21 day period around colonoscopy.

I wonder if there is going to be

variation in costs based by colonoscopy, because everyone is going to have had a colonoscopy. What I think is going to drive it, as Dr. Weiss implied, are things like differences in anesthesia use, potentially how many biopsies are taken, and in my mind I am not sure that is meaningful.

So at least in my preliminary read, I read that as a low, because I wonder if, in fact, there may be appropriate variation, and we are not going to capture that. That was why I was sort of less enthusiastic. I will open the floor now.

DR. POTTERS: I had a similar concern. I didn't see any evidence presented that there was variation. I can believe that there probably is, and again I have doubts about how significant that variation is going to be. I would like to see more data on that.

CHAIRMAN PENSON: Other folks? Jay?

DR. SCHUKMAN: I just had a

question. It is going to look at a 21-day segment, but in terms of variation, say, over a period of a year -- I am just asking, you know. There is going to be those who do it probably more frequently than they should, and I am just curious how that would be captured.

CHAIRMAN PENSON: Well, it won't, because the inciting event is colonoscopy. So each episode is colonoscopy and the 21-day period around it. Louise?

DR. WALTER: I guess it just depends on what the important question is. If it is really about anesthesia and complications, that sounds like that is what they are trying to go after versus who actually -- you know, the disparities that exist with colonoscopy use.

CHAIRMAN PENSON: So is that meaningful? That is the question.

DR. GILLIGAN: I guess, in their defense, it is meaningful in the sense that this is a screening procedure. So it is done

to millions of people. So even a small difference has a huge multiplier here. So small differences in cost could have a big impact on dollars spent.

DR. SKIBBER: That is a good point. I am not clear exactly from the developers that they are only restricting themselves to screening procedures, and that is going to have a -- as evidenced by their references and by the attached article, that is going to have a significant impact, I think, on the variation in cost between providers.

If you look at their article at the end, inpatient costs for surgeons to do colonoscopy, in one of their slides, is five times that of other providers, and I can tell you why that is. It is because they are sick patients who are having a clearly diagnostic or therapeutic colonoscopy, and they are in a high risk group for perforation and bleeding.

That is not accounted for anywhere

in this measure. Now if they want to say we want to study the costs associated with screening colonoscopy in appropriate age groups, or something, that is going to be a much cleaner question.

I think this overall suffers in the way it is going to impact on providers' sense of do I need to improve, because they don't make that designation between a screening procedure and a therapeutic or diagnostic procedure.

CHAIRMAN PENSON: So let me just interrupt for one minute, because I think we have the measure developer on the phone. So let's just confirm that that is the case.

Kevin, is there any possibility or are we misinterpreting it? There is no exclusion criteria for a prior diagnosis of colon cancer?

DR. WEISS: Todd Lee is around --

CHAIRMAN PENSON: We can't hear you. I'm sorry. You are sort of breaking up.

We are really losing you here. Maybe take it off speaker for a minute.

DR. LEE: This is Todd Lee. I can try and answer the question. They are asked whether or not the episode includes prior colon cancer.

CHAIRMAN PENSON: Well, the question becomes, basically, is this a screening -- or what you are hearing from the panel is concern that, if this is a screening colonoscopy versus a surveillance colonoscopy, you are going to have major differences in cost. So one thing that could really limit it to screening colonoscopy might be an exclusion of patients who have a prior diagnosis of colon cancer.

So the question is: Does this measure only apply for screening colonoscopy, surveillance? Have you thought about this at all?

DR. LEE: It actually -- So the current measure captured --

CHAIRMAN PENSON: Are you on a speaker phone? You kind of keep going in and out.

DR. LEE: Sorry. It is a handset.

CHAIRMAN PENSON: We can hear you pretty good right now.

DR. LEE: It is intended to capture both groups, the way the current measure is currently specified. We realize that there might be differences in cost. However, it is difficult within an administrative dataset to differentiate with exact clarity whether or not an individual would have had a previous diagnosis of colon cancer versus a screening colonoscopy.

Our work group felt that it was best to include the entire spectrum of colonoscopies as events in this episode.

CHAIRMAN PENSON: Is there any way to at least capture that in the risk adjustment?

DR. LEE: Well, what I don't know

is how good the administrative data are and being able to differentiate the different screening versus a colonoscopy done in a person who has had previous colon cancer.

CHAIRMAN PENSON: Okay. Steven, question, comment?

DR. CHEN: I think my comment would be I actually gave this a better score here on 1b, which is that there are opportunities for improvement. However, I would be much more concerned when we get to the scientific validity of this, that I think that the way that they have constructed this, particularly things like there is going to be a huge cost difference between people who get a colonoscopy with no biopsies and a colonoscopy with biopsies that is going to engender pathology reports and what-not.

That may or may not have anything to do with the provider as opposed to the person who they are screening. On top of that, people who are lower cost may actually

be providing a lower quality colonoscopy, because they are not finding the polyps, to begin with. So you say, wow, I am really cheap.

CHAIRMAN PENSON: So what I am hearing -- Lou, go ahead.

DR. POTTERS: You know, there is also in the broader context the patient who gets screened on an assigned frequency versus the patient who never gets screened and then presents with a disaster.

So, you know, if you are going to try to understand the cost of somebody who gets screened on a regular basis against the inpatient who has, obviously, got an advanced, previously unscreened lesion, that that creates conflict.

So in the broader context of trying to understand how they teased out the data, which I guess we will get into, is by provider and not by patient, it creates that conflict. I don't know how they could get

that data or tease it down, and it doesn't sound like they can, and that is one of the issues.

CHAIRMAN PENSON: So what I am hearing, just to reiterate, because I think there are some concerns here, is that there may be opportunities for improvement here, but the problem becomes the way the measure is developed, because we can't look at screening versus surveillance colonoscopy, and there are some issues with risk adjustment, that any variation seen may actually be appropriate variation and that, in fact, we may be rewarding providers who provide worse care, as Steven pointed out.

I personally am not convinced that there really is evidence of variation around these costs with colonoscopy specifically, but that being said, if there is evidence of variation and with the proper methodology, there may be some opportunity for improvement here. Is that a fair assessment? Okay.

DR. GILLIGAN: Can I ask on question, because one of the things that came up for me here was -- I mean, if this was a grant application, we expect them to make the case, not us to have to make the case for them. So that kind of surprised me here.

CHAIRMAN PENSON: Yes, and I think that that is a good point. I will cut them some slack, only because they have a form they have to fill out, and there is an element of trying to be sort of terse. I think it is fair for us to bring our knowledge in. We don't have to act in a vacuum, as if we were in a study section.

I think that reflects our response to 1a, which was I think everyone in the room felt that was important, even though we didn't get our five-page diatribe on why. I think we all know that. But here, you know, Tim, I agree with you. I would have liked to have seen evidence that there is variation, because I am not feeling it, for lack of a better way

to put it.

Okay. If people are comfortable with 1b, I will move on to 1c. The purpose and objective of the resource use measure and the construct is clearly described. I think that, for the most part, everyone felt that was a true statement or a moderately true statement. I thought it was fairly straightforward myself. I will open it up to discussion.

All right. I am not hearing anyone. So I think we all can agree that probably was the intent, and the description was reasonable.

The last piece in the importance to measure realm is the resource use service categories that are included are consistent with and represent the conceptual construct of the measure.

So just to start the discussion here, I thought that this -- When I was considering this measure, 1d and 1b tracked

very closely, because my gut feeling was that the variation was going to be related to exactly what everyone was saying, which was use of anesthesia, use of pathology services and, to some degree, complications.

So the question is, was that meaningful? I wasn't convinced, for the same reason that I wasn't completely convinced with 1b. I may have been a little harsh, as people are talking, but I am curious to get other people's opinion on the service use categories -- the resource use service categories. Did folks think that the -- I think it was eight categories, that the costs were divided into were reasonable?

DR. BORKER: I have a opinion on that. So I am not a clinician, but it does seem, based on what the developer submitted, that anesthesia is used without any sufficient evidence that it works, but especially in the light of colonoscopy.

So, to me, that is appropriate to

look at those resource use and at some point connect it to the outcome. I know we are not doing it at this forum, but --

CHAIRMAN PENSON: Yes. I completely agree with you. Again, I think at some point this becomes a numerator of a value ratio, and I think we just have to sort of take NQF at its word. I am looking at Sally. I am blaming you for everything Carolyn Clancy has ever done, for better, for worse. But I think we have to take them at their word.

Like I said, the only thing that gives me some sort of relief here is that, with colon and breast cancer, there is just a litany of quality measures out there, but I agree with you that you need to consider both. John?

DR. SKIBBER: I think that that gets back to that issue, though, of not separating screening colonoscopy from a therapeutic one. It gets into all these issues, and it does even get down to the point

of attribution.

You know, when you read the developer's submission, it implies that the anesthesia is some sort of a patient preference or they ask repeatedly if the patients would pay out of pocket.

That may or may not be, and it is not clear to me that the evidence or even the rationale for that statement is correct. It may be a patient preference issue, and it may be a provider issue or it may be clearly indicated, and their rationale, to me, seemed like it went back and forth, and there is not a clear statement about that.

Again, this speaks more probably when we get into the attribution.

CHAIRMAN PENSON: But let me stand up for a moment just to sort of reiterate what you are saying so that that way the measure developer gets feedback, because it is resonating with me, as you say.

I think other people around the

table are shaking their heads, which is, you know, that inability to break up a screening versus a surveillance colonoscopy is really problematic. Is that a fair statement?

DR. SKIBBER: It is not even clearly the issue of surveillance. it also speaks to the issue brought up by one of the other members, which is there are a lot of instances, actually, where a patient is undergoing a colonoscopy without a diagnosis of cancer preceding that. However, they have symptoms. They have obstructive symptoms or whatever, and those are not going to be in the same risk category as a patient who is turns 50 and shows up for an asymptomatic colonoscopy.

I think several times now we are getting back to that issue.

CHAIRMAN PENSON: And that may have an impact on this anesthesia question.

R. POTTERS: I think one positive attribute is that we are probably going to see

geographic variation as a significant driver. You know, practicing in the northeast, I can tell you that the expectation is that you are going to get anesthesia with your colonoscopy, and it is interesting in the context of just talking about this with some colleagues and laypeople.

At least on Long Island and Manhattan, you are going to see a huge geographic variation against this data.

CHAIRMAN PENSON: You know, the other point I would add to that is, thinking about this as someone who is approaching the age to have a screening colonoscopy, if you are going all the way to the right colon, then I sure as heck would like some anesthesia, but it may be that, yes, there is less anesthesia used and the colonoscopy could potentially only look at the descending colon, and it gets to the point where we do need to think about quality here as well at some point, but I think you are going to see geographic

variation. It is provider and patient preference and community standards, but that is value. That is value, documenting that. Steven?

DR. CHEN: Again, I think I would renew my comments about 1b, which is to say that I had less problem with this as far as how they divided up the categories than I do with how they are going to use the categories. I will reserve the rest of my comments for when we get there.

CHAIRMAN PENSON: So I think, again -- Tim, do you have a comment?

DR. GILLIGAN: Well, no, just very briefly. I think that, if there are practice patterns and people on Long Island like general anesthesia -- my experience in Boston was that we did it with conscious sedation and didn't have a lot of discomfort, and it worked just fine.

That is kind of the sort of thing that we do want to find, because if we are

spending a lot of money on anesthesia and we can do just as good a job without it, then we should know that and reward people for more efficient care.

CHAIRMAN PENSON: So I think it is important. There are a couple of points I want to reiterate to the group, because what I am hearing from everyone is that this actually goes back to the 1b discussion we were having before, that even though there is not a lot of evidence that there is variation, if there is, we would like to know that. So that has some value.

I think Steven's comment about breaking up 1d and 1b is critical. When you vote on 1d, what you are really voting on is the -- I hate to use the term validity, because it comes into number 2, but from a clinical component, do the service categories resonate with you; whereas, the opportunity for improvement is where you want to vote on 1b. So I would urge you to sort of break

those two up.

What I am hearing is, actually, more positive than negative here, if that is a fair statement. I do hear the concerns about screening versus -- I don't want to keep harping on that, John, but screening versus surveillance, that it is symptoms when people come in the door, and perhaps that is a discussion that we can get to when we do risk adjustment, because I think that will be a big discussion coming down the pike.

Any other comments on importance, the criteria 1 in general? Steven?

DR. CHEN: Just one other thing, and again I am having trouble sorting out whether these go to the ones or the twos, but I did have one concern as far as how they did their categories, where they include things like inpatient stays and things and other imaging that may or may not relate to the colonoscopy at all.

They are excluding people once

they hit two days before their colectomy, but what day you get your colectomy is heavily probably dependent on what day your surgeon is available as opposed to how long it takes for real. So if you are doing other preoperative staging studies, that may or may not fall within the two days before surgery or outside the two days before surgery, and that has almost nothing to do probably with what the gastroenterologist did in their colonoscopy.

CHAIRMAN PENSON: Yes. And I think that may actually fall into the 1d criteria here. So I think it is reasonable to consider that when you put in your vote for 1d, and I think that is a well taken point, and we will get into accountability. I mean why does the gastroenterologist get potentially penalized for what the colorectal surgeon does? Maybe that is just the nature of the beast.

Okay. I think I am not seeing anyone looking to holler anymore. So let's

vote on criteria 1.

MS. TURBYVILLE: So if you look at the flat screen right there as a reminder, we are looking at high impact. One on your clicker is high; two is moderate; three is low, and four is insufficient. Point don't Sarah.

CHAIRMAN PENSON: Don't forget to hit Send. That is what I just learned.

MS. TURBYVILLE: We are missing one.

CHAIRMAN PENSON: We've got nine responses. Is that everyone?

MS. TURBYVILLE: There you go.

CHAIRMAN PENSON: Good. Great. So I think the majority of the panel felt this was high. A few folks felt it was moderate. That is reasonable.

Let's move on to 1b, and the next one we are going to talk about is the performance gap. This is demonstration that there was variation over or less than optimum

performance across providers or population groups. So let's go ahead and vote on that. Remember to hit Send.

MS. TURBYVILLE: And point toward Sarah. There we go.

CHAIRMAN PENSON: Okay, great. So this is interesting. The group sort of split up. I'm sorry? People migrated down, for better, for worse. So there is some discussion here.

Let's move on to 1c. Let's go ahead and vote on that. This is the purpose and objective of the measure and the construct being clearly described.

One person voted low on this, and I don't want to call people out, but I just want to make sure you meant to vote low on this. So this was basically the purpose of the resource use measure and construct are clearly described. If you voted low, I didn't really feel that there was a lot of negativity here. So whoever voted low on that --

DR. KLOTH: I was on the fence, to be quite honest.

CHAIRMAN PENSON: Okay. And that is not a problem, voting. I don't want to influence your vote. I just want to just have for the record why, so the staff can report back to the -- You vote the way you want to vote. All right? But I just -- and it can be because you didn't like the color of the paper. That is okay, but I just want to have it on record so that the NQF team can report back to the measurers. It is a matter of public record. So if you don't mind.

DR. KLOTH: Well, I was really ambivalent about voting 2 versus 3, and I was very persuaded by some of the discussion about the level of documentation and justification. Sitting on research review for so many years, I perhaps get too fussy at times and too picky.

CHAIRMAN PENSON: So what you are saying is you were concerned when you heard

the discussion regarding documentation and how it is defined, that basically you felt that perhaps this wasn't clearly defined enough. Is that what you were sort of --

DR. KLOTH: That's right, and we had a lot of discussion about screening versus diagnostic and all the variables, and really parsing that out so that we are really measuring what we think we are measuring in the way that we really do intend to measure it.

CHAIRMAN PENSON: Okay. Thank you. Again, I don't mean to call you to the table, but we just didn't have that discussion at this point. I want to make sure that we are properly documented.

Let's move on to 1d. So this is the service categories that are included are consistent with and representative of the conceptual construct represented by the measure. So this breaks down to the various A categories, etcetera, and I will open it up

for voting. All right, we are all done there.

I think this is consistent with the discussion we had where we had sort of a spread. Most of us felt it was moderate. Okay, next. I think that's it. so we don't have a yes/no for this one at all. Okay, great.

So we are making pretty good time, relatively. We are going to now move on to the scientific acceptability of the measure, and for this one we do have assigned folks. We will go through each of these and have a discussion.

The first one, 2a: Was the measure well defined and precisely specified so that it can be implemented consistently within and across organizations for comparability? I was the assigned reviewer for 2a.

I felt it was. When I reviewed the measure itself, it made a lot of sense to me. Someone has done a fair amount of

research in administrative databases that I could identify patients undergoing colonoscopy, that I could define the seven-day period beforehand, the 14-day period afterward, etcetera.

With regard to inclusion and exclusion criteria, I have to say that that was my only sort of hook on this, was that -- and I think Lou Potters noticed this either on this or one of the other measures, that the concept of having two years continuous coverage in some respects is problematic, because how was that chosen? What does that mean?

On the other hand, if you don't have it, it is a problem as well. I think it is more critical for some of the other longer measures as opposed to this one, which is only two or three weeks, but I do think that, if you know someone is about to lose their insurance in a few months, it is going to affect the way you do business as a provider.

So I think it is an imperfect but reasonable exclusion criteria.

So that is why I actually voted this high. I will throw it open to the floor. Tim?

DR. GILLIGAN: I just had a brief comment. I am not an expert on this, but I think the reason we need that two-year role in this is to make sure that you know who the patient is and whether or not they have exclusion criteria and stuff like that. I think that is why they have it so long. Maybe others know more than I do about this.

CHAIRMAN PENSON: No, and that is exactly right. You need to have some sort of background information for the risk adjustment. You need to -- There are a number of reasons why you want to have that. It is imperfect, as I think Lou mentioned, but it is making the best of a bad situation.

DR. GILLIGAN: Right. I mean, there is no way around it. So it is just

stating -- It is sort of just stating a fact, that that is what it is.

CHAIRMAN PENSON: Yes. I'm sorry, Louise.

DR. WALTER: Yes. I also wanted to say, on the specifications, actually, the other reason I rated it high is because it actually based the claims and specifications on an Annals of Internal Medicine article by Warren et. al, which I thought that was an impressive part of that, too.

CHAIRMAN PENSON: This works, in my assessment, and anyone -- I don't want to say anyone disagrees. It sounds so heavyhanded. But any negative comments on that? I thought this was straightforward, but I have been wrong before.

DR. BORKER: A quick comment. It is nothing negative about this particular approach, but one of the issues -- and I am not a statistician either; so you have a statistician here. But whenever we are

analyzing the costs, there is obviously that issue with being able to compare one cost or point estimate of a cost to another point estimate, because of all the non-delaying distributions which are not pretty straightforward.

So what I would like to see additionally is -- and those are finally -- This is going to get used to compare costs at regional level or even at a wider level -- some level of developer's recommendation in terms of what adjustments we need to make, in addition to just assigning one dollar to all the zero dollar claims.

CHAIRMAN PENSON: I think I can put your mind at ease, because I do use standardized costs and, in fact, in this measure they are actually risk adjusted costs. We will talk about whether or not that particular methodology is reasonable in a little while, but I think that the standardized cost and methodology is pretty

accepted out there.

DR. BORKER: Right. My issue is not around the standardized costing. So the standardized costing will assign a unit cost to a particular resource use. It is comparing provider one to provider two. Let's assume provider one comes up at $10,000; provider two comes up with $13,000. To me, that comparison needs to be more explicitly stated.

CHAIRMAN PENSON: So I will ask you to table that, because I think that comes up in the risk adjustment.

DR. BORKER: Okay.

CHAIRMAN PENSON: Because I think one of -- I don't want to bias the discussion. So I will just say that I think we need to talk about the risk adjustment. So I will ask you to table that, because I think it comes later. We are just talking about A-1 right now. Any other comments about A-1; that is, just sort of specifications and sort of the basic sort of guts of the measure? All right.

Hearing none, I am going to move on to A-2,k which is reliability testing. Reliability testing, it says, is also repeatable and produce the same results a high proportion of the time. Lou Potters, you were the primary reviewer for that.

DR. POTTERS: So I voted high, but then this weekend I sort of looked at it, and I guess I agree with the consensus of more of a moderate.

These are commercial claims that are then filtered down by provider. They have a single reference that was noted on a publication, but the reliability of this across, I guess, the dataset, the market scanned dataset, is a compilation, and that is going to be used as the base. But it is not completely clear that this is going to work across the board.

CHAIRMAN PENSON: So would you change your sort of view from high to moderate or high to low?

DR. POTTERS: Yes, from high to moderate.

CHAIRMAN PENSON: Before the panel chimes in, I will ask Carlos to just give his sense on this, particularly your thoughts on reliability and validity. But let's start with reliability.

DR. ALZOLA: Reliability?

CHAIRMAN PENSON: Yes.

DR. ALZOLA: When I looked at these measures and I look at the data derivation process and how, as someone who is going to implement the measure, going to do it, they have two kinds of measures. Some of them apply a commercial software, and those have been tested, and using administrative data they can be reproduced. No question about it, and essentially true in this case.

So the only problem that I find is that this depends a lot on who is going to implement it. So the specifications are clear, but there is always room for reading

something a little different than the way the measure developers did.

So in that respect, there is -- Even though I consider it reliable, but there is some room for not being able to reproduce things exactly the same way that the developers did.

In terms of the data, the data itself -- you know, it is claims data. It is not always very reliable, but there is not much anyone can do about it.

DR. POTTERS: You know, it was an iterative process that they went through, and to me, when you go to write software code to get your result, if that formula isn't formulaic in the sense that that is what you have really proven, people can write that formula multiple different ways and get different answers.

CHAIRMAN PENSON: Steve?

DR. ALZOLA: And most -- More than the formula, usually where the differences can

happen is when someone is extracting the data from the database. Interpretations that look very crisp and clear on paper, when one goes to put it in computer code, the difference is going -- different interpretations can happen.

CHAIRMAN PENSON: Let me ask one question, Steve. I can't find this. I was looking. In the one article that they talk about, was that Joan Warren? Was that Medicare data? So, remember, the Medicare is one dataset, and Healthscan, Ingenix is another dataset, and they may or may not work equivalent.

I have a lot of faith in Joan and the NCI team where Medicare is concerned, but it is Medicare. They are the maestros at that. Whether or not it works in Ingenix, I don't know. Steven?

DR. CHEN: Yes. The question that is asked is, is this reproducible; and I think the answer is yes. the one thing that I would have as far as reliability is that they allow

each place to do their own data cleaning, and there is no discussion as to how they should do this data cleaning, which allows you the potential for gaming with garbage in these situations. So I will probably switch my vote down, too, as well.

DR. POTTERS: And I guess that is a different question, which is how much of that is reliability versus validity in a sense. If they tell you what to do, but then you do something different, which category does that fall into? It is a semantic.

CHAIRMAN PENSON: I am not sure it falls into either one. I mean, that is just dishonesty.

DR. POTTERS: No, I am not saying they are dishonest. I am just saying they just don't care.

DR. CHEN: And I would point out that you could be highly reliable and totally wrong.

CHAIRMAN PENSON: Let's focus on

the reliable piece for now, and what I am hearing is that there are some concerns here that this may not always be reproducible each time, even when someone is doing their very best effort to do so.

The question I would ask is: Does anyone want to disagree with that statement and say, no, it is, or does anyone want to say, actually, that is true, and it so bad that it is a problem and that we should be voting low, because I am hearing this sort of consensus in the middle. Rohit?

DR. BORKER: So I put low here, only because it hasn't been demonstrated that this is -- that this method is reproducible in another dataset. I mean, to me, they did a very good process with it being a data process, but that process was applied to the same database again and again. There is no evidence of it being reproducible somewhere else.

CHAIRMAN PENSON: So I would just

advise you that that insufficient as opposed to low. In other words, it is not that the evidence speaks against it. It is just that there is not enough evidence. So if that is your opinion, then you would vote insufficient.

DR. BORKER: Okay.

DR. TURBYVILLE: And just to remind everyone, and not to sway what your rating here is, that NQF doesn't prescribe to the developer how to do the reliability testing. So, clearly, your input is important, but we do allow, especially at the first time of endorsement, to think about the ability for developers to test it.

You know, real life or out in multiple databases, we acknowledge, is limited. I think ratings probably should reflect that, if it is something that you are very concerned about, and we would capture that. In follow-up, we would hope that it would have been tested in more expansive data

sources.

CHAIRMAN PENSON: Other discussions with regard to reliability subcriteria 2a(2)? So just to sort of again recapitulate the discussion I have heard, since I know also the ABMS Foundation team is on the phone, is that what I am hearing is that, for the most part, people are feeling that this probably is reliable, but they are not 100 percent swayed.

There are some concerns, mostly of a moderate nature, both that the formulas are created from a good process, and there may be room for error there, and also it has not been adequately tested in other datasets. For some people, that may be okay, and for others they may view that as real major problem, and that will be reflected in the voting.

Did I miss anyone's thoughts, comments? Excellent.

MS. TURBYVILLE: You could vote on the two reliability criteria.

CHAIRMAN PENSON: That would be fine. So let's do the 2a variables. So let's start with 2a(1), which is: Is the measure precisely specified so it can be implemented consistently, and again it is high, moderate, low or insufficient. Looks like we got that there.

So I think most people felt that was high or moderate.

Let's move on to the 2a(2), which is reliability: Are the results repeatable, produce the same things?

Let's just vote again. We are down one person. We got it. Okay, well, consensus. So moderate. That's good. Consensus is excellent.

So then we have to have an overall reliability. I didn't know we had to do this. I apologize. So, basically, based on those two subcriteria, what is your overall feeling about reliability, based on this discussion? Okay, we have everyone. You got to like that.

It is all about consensus. Very good. All right, excellent.

Is that what I wanted? It is not what I want you to do. What I want is just to send a -- If we send a consistent message, I think that is useful for both the ABMS Foundation folks, and it is also very useful for the Steering Committee. Excellent.

So now we are going to move on to the 2b issues, which include validity an risk adjustment. The first one is 2b(1), and that is that -- We are not going to vote. So don't worry -- the measure specifications are consistent with the evidence presented to support the focus of the measurement under the criterion. The measure is specified to capture the most inclusive target population indicated by the evidence.

I think Tim's comment before about the fact that there is not a lot of evidence here -- let's remember that a lot of us in the room are content experts, and if you feel that

it meets your take of the evidence, I think that is reasonable.

So the reviewer for 2b(1) was me, and I felt that this did sort of capture what they intended to capture. They were specifically focused on costs around colonoscopy, not around colon cancer treatment, and I think that is a key thing.

If you accept the argument and the evidence, you are going to have to make a leap of faith, but I think we have made that leap of faith already. There may be an opportunity for room for improvement here, and this is something worth studying.

I did feel that the data elements captured what they were trying to capture and the evidence, and with that I will open it to the floor. Okay, I am not hearing anyone argue on that, and I guess that is good.

Now we will move on to 2b(2), which is the validity testing demonstrates that the data elements are correct and that

the measure score correctly reflects the cost of care.

What I will do is I will ask Lou, who is responsible for that, to comment. Then I will ask Carlos to add his thoughts. So, Louis?

DR. POTTERS: So again this was like the weekend switch. In terms of validity testing, I thought they did a better job than the reliability testing, because the data just is what it is, and they have the publication. They have gone through their breakdown in terms of standard deviation and how they fractionated everything, and I thought it actually worked out okay.

CHAIRMAN PENSON: Carlos?

DR. ALZOLA: Yes. I don't have much to add to that. I just look at how the costs were distributed in terms of what lines of service were responsible for the major costs. They seem to make sense to me. Of course, you clinicians know better than I, but

they all seem consistent with what you would be expecting.

CHAIRMAN PENSON: All right. And Steven, we will open up to the floor.

DR. CHEN: So this is where I actually start to have issues with precisely what is being included, and I think that is where this goes. Right?

For instance, I have concerns about in the seven-day run-up to the colonoscopy, which is included, including things like ER visit. I can't imagine how scheduling a colonoscopy generates an ER visit that is attributable to that provider. If anything, it is the opposite.

I will renew my concern about the cancer work-up, things like CTs and MRIs, after colonoscopy. It is unlikely that that is what you are using to look for free air. You may use a CT scan at some point, but an MRI of the liver is almost certainly not part of the colonoscopy. That is the colon cancer

work-up.

Then again, my concern about including as a mix people who have biopsies and therapeutic polyp excision with people who just had a straight screening colonoscopy, are two radically different populations that I don't think risk adjustment handles well. I think that does need to be a stratification, not a risk adjustment.

CHAIRMAN PENSON: So what I would say to the last comment is let's -- that last piece, let's table that and discuss that in the exclusions, the next subcriteria, because what I am hearing is perhaps these patients either should be substratified or excluded altogether. Is that a fair statement? But I think the two other points are worth discussing further, because I think they do fall into this subcategory.

Is it valid to include that information there? It is going to be attributed to the individual performing the

colonoscopy, but that may not be a valid cost associated with a colonoscopy per se. It may be a valid cost associated with the diagnosis, is what you are saying.

What I would ask you is if you feel that that is such a major problem that it really is going to cause you to vote low, for lack of a better way to put it, because I think it makes a difference.

DR. CHEN: I think as it is currently constructed, I would probably vote low, but its remedy-able. You know, it is able to be remedied.

CHAIRMAN PENSON: So what I am hearing is that that, as currently written, is a major problem that causes you to really question the validity of the measure, but it could be fixed. Can you just comment again for the NQF team how you would think about fixing it?

DR. CHEN: Reexamining that -- For instance, an MRI of the abdomen probably has

fairly similar cost to the professional fee of the colonoscopy, if not more. So it could quite easily be imaging aspect could overwhelm the entire episode of colonoscopy of an otherwise uncomplicated colonoscopy that is done without anesthesia.

Similarly, I worry about attribution. While the gastroenterologist may choose to do anesthesia, how the anesthesia is performed may or may not be within their purview as to which drugs and what-not as well. So again, I worry about attribution as it pertains to the validity, because if the validity has to do who it is being attributed to, then there are things out of their control.

CHAIRMAN PENSON: So I wonder if -- - One of the key things that I am hearing here as you are speaking and I am thinking about this, and you are certainly making me think about this, is that you would be able to capture a new diagnosis of colon cancer

immediately following the colonoscopy, and that certainly is something that would be worth stratifying by, because that is going to affect things.

If someone does have a diagnosis of colon cancer, it is going to affect the way they are imaged. It is going to affect other resource utilization, even up to a colectomy. I am not a colon cancer surgeon or deal much with colon cancer in the clinical setting. Is that a fair statement?

DR. SKIBBER: Again, it speaks to an issue that I have a little bit that is similar, is the fact that even in the attached article, they don't account for -- They account for inpatient costs as part of this model. However, inpatient costs that may be generated even before the patient has the colonoscopy would be captured.

What I am referring to there is a patient admitted for a symptomatic problem, then undergoes a colonoscopy. I think that is

going to distort some of the provider attribution, and it is a similar point.

CHAIRMAN PENSON: Do you think that is something that could be fixed in a risk adjustment model? Steven's comment is you would just have to stratify. There is no way to risk adjust.

DR. SKIBBER: I would have to ask the developer on that. I don't know. Did they have any -- Does the developer have any idea if this was addressed by their working group?

CHAIRMAN PENSON: So I would ask Dr. Weiss and the ABMS Foundation team on the phone. Are you guys there?

DR. WEISS: We are. However, the phone is in and out. So if you could repeat the question.

CHAIRMAN PENSON: So the question was: There was concern regarding attribution of costs, that certain patients are going to have more imaging. Some patients are going to

come into a hospital symptomatic, and there is no real indication to us in reviewing the measure that that was considered by the working group.

Was that considered by the working group in the discussion? Is there any sort of risk adjustment for that? Is there discussion about stratifying analyses? Where was that with regard to your working group?

DR. WEISS: So I will start and, hopefully, Todd can jump in, in terms of the risk adjustment model which he is intimately familiar with.

The working group recognized that the older someone -- of these colonoscopies would be for screening activity, that there would be a small portion that would be done for other reasons, and that -- and that in here that that should be relatively balanced, and that more severe patients require -- risk adjustment. So they did not want to, didn't feel the need to, recognizing that --

CHAIRMAN PENSON: Kevin, you are breaking up a little bit there. I'm sorry. Could you repeat that last part? It may be a tech problem on our end. We just don't know. We don't think so. Maybe if everyone here could turn off their microphones, maybe that will help.

DR. WEISS: Oh, that sounds better at my end, for sure. How much of what I said was heard?

CHAIRMAN PENSON: We sort of lost the last minute of it. I think we have identified the problem. It has got to do with microphones on On. So if you want to start from the beginning, that is fine, too.

DR. WEISS: That sounds painful, although it wasn't a very long answer.

The working group very much focused on the fact that what we were really capturing for the majority -- well, the overwhelming preponderance is really screening, and that there would be some non-

screening activity that goes in here, but that the very complex patients who have lots of comorbidity would be picked up in the risk model adjusted for, and that hence the comparisons would likely happen within peer group, that there would probably be a tendency to a more apparent random directional bias -- not random, but just directional bias in terms of any variability in terms of patient severity, recognizing that it was an imperfect way to solve it, but that is the best possible solution that was worked out.

To the risk model, we have got Todd Lee on the phone, and I don't know, Todd, if you want to -- if you can come off Mute and talk a little bit about that as well.

CHAIRMAN PENSON: Before Todd goes ahead, I just want to let you know, because you can't see the visual cues. I am going to state right up front that some of the clinical experts in the room, when you mentioned the distribution of screening versus surveillance

colonoscopy, were shaking their heads.

So it may be necessary at some point for you to provide evidence that the majority of colonoscopy is screening and that the impact of a clinical surveillance colonoscopy or colonoscopy for severe symptoms or even moderate symptoms due to colon cancer is minimal. Just to let you know, that is what I am seeing here.

DR. WEISS: No, that is helpful.

CHAIRMAN PENSON: Todd, if you want to mention a few words specifically as to whether or not the risk adjustment piece captures severity of symptoms, because we will talk, I think, in great detail about the risk adjustment model in a few minutes. But does the risk adjustment model capture at all severity of symptoms or, for that matter, new diagnosis of colon cancer after the colonoscopy?

DR. LEE: No, it only considers information preceding the episode and is not

symptom based. Rather, it is going to risk adjust for other coexisting conditions, and that is really the focus of the risk adjustment model, and looking at severity of patients in terms of other concomitant medical conditions.

CHAIRMAN PENSON: Okay, that is great. Thank you. So I am going to sort of keep us moving along for the sake of time. I'm sorry, Louise.

DR. WALTER: No, no, the only other thing I want to put on the validity is I think the lack of testing in people 65 and older is a concern for me. I put my geriatrician hat on, and that will actually come to some of my points in the exclusion criteria.

CHAIRMAN PENSON: I think that is a major concern. I agree with you.

So just to reiterate what I have heard in the validity discussion, I am hearing some real concerns here that are related, to

some degree, to the risk adjustment piece and the inability to sort of capture: Is it valid to capture, say, imaging in a patient who has a new diagnosis of colon cancer that is going to be attributed to a gastroenterologist that, in fact, the gastroenterologist had nothing to do with it?

I am mostly hearing some concerns about symptoms at presentation, whether or not that is going to affect the validity of the measure. We will get to the inclusion criteria, but these are all sort of intimately related, for lack of a better way to put it.

Are there other comments or concerns regarding validity?

DR. GILLIGAN: I am wondering if I have interpreted this differently, because I notice on this measure and also on 1584, I am the only person who voted low.

When I looked at Carlos' evaluation on 2b(2), it seems like all the standards that were suggested there were not

met, to me. So has the data been compared to other authoritative data sources? No. Has the data integrity been checked? No. Is the data representative of the target population? Not really, because the target population is over 65.

So that left me feeling like we didn't really meet the criteria, but everyone else felt like it did. So I am wondering if I am using a different ruler here.

CHAIRMAN PENSON: No, I don't think you are. I think it is a matter of interpretation. I think that I could see why you would vote that way, and it is just a matter of reference setting. That is all. But I think the comments you are raising are quite reasonable, and I think that the discussion reflects that.

MS. TURBYVILLE: And just to add for your thought process, the minimum threshold that we require for validity, especially in consideration that this is the

first endorsement of these type of measures, is face validity.

So that would be a minimum threshold, that they demonstrated that face validity in a systematic manner was met.

CHAIRMAN PENSON: Go ahead, Rohit.

DR. BORKER: I had a good comment. So I completely agree with what Steven was saying in terms of what is attributable, what is not, but to what developers have produced here, if the end was to look at absolute cost of care for a given provider, definitely those are valid concerns. But should we think about what the instrument is getting used for or the measure is getting used for? It is to compare across providers.

So if one can assume that within a certain setting, say community setting, if one provider had a similar case mix of patients as another, then shouldn't those differences get neutralized, if you may?

CHAIRMAN PENSON: Well, I -- Go

ahead, Steven. I'm sorry.

DR. CHEN: I guess I would say to that, there's two issues. One is small sample size for a lot of providers. So you are going to have case mix variation, just by pure randomness.

The other is systematic bias. For instance, colorectal surgeons do, I think, 10 percent of the colonoscopy or eight percent of colonoscopy. They have a radically different case mix than people who are doing screening colonoscopy.

So to the extent that you are using inter-specialty provider comparison as well, that is radically different.

CHAIRMAN PENSON: Other comments? I think we have beat validity to death, but we are not done yet, because I think many of the remaining points are going to come back to us. So let's move on to 2b(3) which is the exclusions, that basically these are supported by the clinical evidence, and the

specifications for scoring include computing exclusions and, finally, if patient preference is a basis, there has to be evidence for that.

For this one, with the exclusions 2b(3), John, you were the primary.

DR. SKIBBER: I felt that the exclusions expressed by the developer were reasonable. They appear to be specified. The exclusion of age less than 40 is arbitrary but reasonable. The high cost condition exclusion appears well stated and obvious. The issue about colectomy for cancer within two days of colonoscopy is reasonable.

You know, that is interesting, though. You are going to lose a small number of complications related to that, if one of the objectives is to assess provider resource use based on complications.

If you go back and look at their article, the number that you are probably going to exclude based on that is low, is very low, in fact. So I think that is fine as an

exclusion.

The inflammatory bowel disease exclusion is a reasonable one. We get back to that insurance exclusion, which is their major one. When you look at their article, they exclude over 50 percent of the colonoscopies in the database.

Just in an overall sense, I realize that that is what they have to have in order to record comorbidities and have administrative data, but I have to tell you, that completely excludes the ability to address disparity. To me, that is not even an issue, because of the way that they create that exclusion. That cannot be addressed in this at all.

CHAIRMAN PENSON: So I will just ask you on that, because I think that is a very good point, but there is a specific subcriteria for disparities. So let's come back to that, and remember to bring that up when we get to that.

DR. SKIBBER: That was all I had.

CHAIRMAN PENSON: Are there other comments on exclusions? Steven?

DR. CHEN: So I have covered a fair number of them. I do have some questions. One was the exclusion includes people who don't have coverage at least 320 days before and after for a 21-day measure.

I understand for the year before you are using it for the comorbidity adjustment, but for the year after I am not exactly clear on why that would have any effect, say, after, say, a few months after.

So going back to your question of we are eliminating all these people, we end up with a very small sample size. To the extent of improving the sample size, it may be worthwhile to consider cutting the back end, because I don't think the back end has a lot.

The other comment I had, had to do with -- Again, I was looking for ways to consider excluding some people who I thought

were not the same, and then also re-including some people that maybe aren't under active treatment. So there is not really a reason to think that their colonoscopy would be performed dramatically differently, but the issue has to then be whether that could be risk adjusted away; because you will notice that the number one medication given is anti-lipid agent, which probably had nothing to do with their colonoscopy cost. I am just throwing that out there.

So you could do similar sorts of things with something like HIV/AIDS. I don't think we would necessarily perform their colonoscopy any differently on face validity, but their costs would be significantly different.

So if you are going to include one thing but not the other, why not CHF then as a high cost. CHF is an incredibly costly thing, to the point where they qualify for hospice. So that's just my thoughts.

DR. WALTER: Building on that, I had big trouble with the cost exclusions, as you were talking about. Excluding people for renal disease, HIV, and transplant might be appropriate in younger people, but that would completely not capture high use in older people, which is more congestive heart failure, dementia. There's a lot of other high cost things. So I don't think that is very inclusive.

CHAIRMAN PENSON: So I will put out to the group, to Steven and Louise and John who started this discussion, in your mind, is that -- I mean, obviously, it is remediable, but is that a major limitation that affects your feeling that this is appropriate. Would it lead you to vote low, is basically what I am asking?

DR. WALKER: Yes.

CHAIRMAN PENSON: All right.

DR. SKIBBER: Again, a lot of that concern -- and I agree with both the members -

- is this issue of not identifying screening, because if you look at their article again, their cost, their phenomenal amount of costs are related to management of cardiovascular conditions, which again is just going to muddy any interpretation of provider attribution and may, in fact, muddy your ability to draw any distinctions based between even institutions, if not regions. It is a concern.

CHAIRMAN PENSON: We are going to get to this in a minute, but my question to everyone in the room -- and we are talking about exclusions, but we are sort of bouncing back and forth -- does that affect the validity, because you have included these costs which may have nothing to do with the colonoscopy that are related to CHF, and by the same token, does the risk adjustment methodology, which does include CHF, if I am not mistaken, as one of the covariates, adequately address that to the point where your concerns are mitigated, I guess is the

term I want to use -- mollified?

DR. WALTER: Well, I guess I would like better explanation for why they chose certain things for exclusions and certain things to risk adjust, because it just seems like the exclusions are not appropriate for older adults.

CHAIRMAN PENSON: So I think, on that theme, what I would like to do -- and I think the folks -- I am sure that Todd and Kevin on the phone are hearing that discussion specifically justifying why certain things were excluded and why they were included in the risk adjustment model or why they weren't included at all.

I want to call the question or move us on to the risk adjustment, and then what I think we should do is vote on these four, because they are related.

So we have talked about the exclusion, and to reiterate, I think there are some serious concerns there. Louis, I'm

sorry.

DR. POTTERS: Well, I had the disparities. There is really not much to say about disparities in the context of the application, because they reference a section that has like two sentences in it. It basically just says that it is age related.

So it does get back to what John was saying in terms of the insurance issue and the filtering down. So we may want to take that whole block.

CHAIRMAN PENSON: Oh, you mean all the way down to disparities? I just think that, just so we can remember what we did, because it is a lot of information, although I am open if people want to keep running. But I think probably we can -- I think it is better if we vote, because we will vote in disparities in 10 minutes probably. But before you vote now, I think we do want to do the risk adjustment piece. That is the last one, if people are okay with that.

So this one is 2b(4), and it is the risk adjustment: Is there an evidence based risk adjustment strategy specified or is there a rationale that it does not need to be included?

I think this is where you can ask the question, does the risk adjustment methodology pass the smell test, not just is it there. I think that I will ask, just for the sake of time -- we don't have to repeat a lot of what we have already discussed regarding things like screening versus surveillance and CHF.

So with that, I will turn it over to John, who is the primary reviewer.

DR. SKIBBER: The risk adjustment use was the CMS version of eight ccs for assignment of comorbid conditions based on ICD-9 coding during the preceding 12 months.

This was adapted from a total cost model to an episode based model. I have a concern that this risk adjustment model treats

colon-specific comorbidities similarly to non-colon related conditions, and that I am not sure I see that this has been validated, for all the reasons we just talked about.

I believe that the statistician, who can speak for himself, has concerns about the validity of the risk adjustment that should be addressed.

The other issue, I think, included in this measure was the stratification. That was based on a task force recommendation of stratification of 40-75 years or greater than 75 years. Again, I am not sure that this is totally valid.

One interesting point is they based some of this on that Warren article, and if you read that, that article is for outpatient colonoscopy. It is not for all colonoscopy, and I think it is going to speak to the issues brought up by one of the members.

CHAIRMAN PENSON: I will ask

Carlos to say a word or two about the risk adjustment, then throw it open to the masses, as it were.

DR. CHEN: Thank you. The main thing about the risk adjustment is that there is a lack of information here about how the model was fit and some statistics on how we -- especially how it calibrates to the population, how the predictive relates to the observed, and also are there measures of goodness of fit.

The other that seems curious to me is that, if you look at the model, it is adjusted for lots of conditions, and apparently they intend to select those risk factors. They just used some statistical significant criteria.

I am not sure that all these really should pass the test in terms of clinical validity. So that is something to reconsider.

Mostly, I felt a lot today that

people with cardiovascular conditions would have really severe costs, much higher costs. However, the coefficient for CHF is only $88. So it is not having any of these conditions. It is only -- Most of them are less than $100.

So if you take a base without any comorbidities, it is going to cost -- The risk adjustment is going to say $1131. If you add, if anybody is going to have may two or three comorbidities, so their estimated cost, their expected cost, is going to be just 1500, no more than that, and yet, when we look at the distribution that they observe, we have costs in the upper range, over $2,000.

So it seems to me that that is -- The model does not seem to be capturing those high costs. It could be that those high costs are not really appropriate, but I really would need to see some evidence of calibration for this model.

CHAIRMAN PENSON: Louise?

DR. WALTER: Well, probably

because it was in younger people versus older people.

CHAIRMAN PENSON: Steven?

DR. CHEN: I would say, to what Carlos said, I think the methodology they used to create the risk adjustment is reasonable. What we don't understand is they don't give us an R-squared or some sort of idea of how well does the model work.

In particular, risk adjustment tends to really fall down at the outliers, because a model will say, you know what, I am just going to ignore that outlier, because, well, we are just going to get that one wrong. But in fact, what we are looking for in a resource use thing is we are focused on analyzing outliers.

So without understanding how the fit is toward the edges, even a scattergram of their the residuals of the first 75 percent test set would be incredibly helpful in this situation. So it is almost an insufficient as

much as anything.

CHAIRMAN PENSON: Yes. I am with you. I actually felt it was insufficient. When I looked at it before sort of Carlos' review, I sort of -- you know, having played in this space, and looking right off and saying, well, they used a Delphi process, well, you know, that is just 10 people's opinion. I am fine with that. There's a lot of people smarter than me. But the fact of the matter is you are going to test that, and I didn't feel like they had done a good job.

They have a very complicated equation that has been -- they have done some statistical maneuvering, but I just in my gut didn't feel right. Then when Carlos as a statistician sort of -- looking at his review sort of confirmed that.

My feeling is that there may be something here, but they haven't documented that for me. Other comments? Wow, I thought that was going to be a much longer discussion.

So I think that what I am hearing is significant concerns from Carlos regarding the risk adjustment, all the things we have discussed already today vis a vis age, vis a vis why the colonoscopy is performed, issues with other comorbid conditions, attributing one thing to another, attributing CHF costs to colonoscopy and potentially not capturing that.

I think that there was some understanding of the risk adjustment methodology and some sort of acceptance of it, but by the same token, what I am hearing in the room was that the evidence presented just didn't get the ball over the goal line.

So I am going to sort of now ask us to vote on these four criteria. Before I call the vote for 2b(1) through 2b(4), are there any other comments or thoughts people want to add?

All right. We are down one panel member. Should we just go ahead? I am

getting the "sure, why not." Sally, Louis is out of the room. I think we still have a quorum.

MS. TURBYVILLE: See if the developer wants to provide any input.

CHAIRMAN PENSON: Okay. That is a reasonable idea. Kevin and Todd, I think you have been listening to the comments. Before we take a vote -- we are waiting for one of the panel members to come back in -- are there any thoughts that you might want to add, any comments, to what has been said?

DR. LEE: Sure. This is Todd Lee. I will take a couple of minutes to address a few issues that I heard, one around exclusions.

We have standard exclusions that are applied across all of our measures. So if we consider this measure in a suite of an additional measure -- I know you are not evaluating it as a suite, but we felt that there was value in having some consistency

across all of our measures, so that if folks wanted to implement these as a suite of measures, it would reduce the possibility of errors.

For that reason, we have the standardized time window, the one year before and the one year after. Realizing that we may not need the one year after to capture all the resource use related to a colonoscopy episode, we wanted to go with the strategy that we wanted to keep this as consistent as possible across all of our measures.

A similar sort of issue applies to selecting these, quote/unquote, "high" cost exclusions, HIV/AIDS, active cancer, renal failure, transplant status, and again this is a standard across all of our measures.

That is consistent with several other resource use measures, and again it was an acknowledgment on our behalf to try and have some consistency across the measures to reduce, again, the likelihood of errors when

individuals or groups might be implementing these measures. It is essentially a standard set of code that you could use across all of our measures to exclude individuals.

I fully acknowledge what Dr. Walter said about this probably being more applicable to a younger population than it is to an older population. There is a big caveat around all of the information that you are seeing in that this is a commercially insured population in which we tested these episodes.

So we don't have the evidence on those people that are older than 65, and we would fully acknowledge that.

The final thing that I will touch on is the risk adjustment issues that you all just talked about, and think that you captured the information very accurately.

We certainly have fit information in terms of observed predicteds and how each of these 12 different models that we evaluated with our work groups fit.

We went through a process of having them say, hey, what is clinically important, and then we went through another process of saying, okay, what if we used some statistical fitting methods and seeing how those two things differ.

What we found out is we ended up falling on this risk adjustment model that used whether or not things were associated significantly with resource use.

We know that we are not going to predict -- or be able to predict all of the variation in the observed -- and we don't want to, because some of that, as was noted, might be due to difference in practice patterns and not differences in coexisting conditions.

It is not surprising to me, given that we are focusing on a 21-day period, that these coefficients are small. The coefficients that you see with other chronic conditions in our risk adjustment model are small, simply because we are talking at a very

constrained time period.

Finally, we have all of the information that you might want around risk adjustment model performance. We just did not submit that as part of our initial documentation, and again as we have done for other measures, we could submit that as part of the response to all of your comments.

CHAIRMAN PENSON: All right. That is very helpful, and I think, for the sake of argument, we will move on to voting, if that is okay.

MS. TURBYVILLE: If they could email it to Ashlie, we might be able to keep things moving so that you can adequately rate today, if that makes sense to you.

CHAIRMAN PENSON: So I guess -- Why not? So what I would say is the sooner you can get us that information specific to this measure, because I think there is some concern about the risk adjustment that matter.

We are going to call the vote now,

but certainly, we could circulate this either through a follow-up phone call or even today by the end of the day. So if you could email that to us, that would be great, Todd.

With that, I am going to ask -- We are all back in the room. So let's start voting. We are going to start with item 2b(1), which is: Are the measures specified consistent with the evidence or, if nothing else, our clinical take on the evidence? If everyone could vote now, that would be great.

So I think most people felt this was high or moderate, and I think that reflects the comments of the room.

Next is validity testing. This is: Does the validity testing demonstrate the measure elements are correct and the measure score correctly reflects the cost of care and resources provided? I do think there was some discussion here. So let's go ahead and vote here.

I think this reflects the

discussion fairly well.

The next one is 2b(2) and probably 2b(3). This is the exclusion question, and I think there were some real discussion here regarding age and other issues. So go ahead and vote.

Okay, we got everyone.

Then the last, and certainly not least, was the discussion around risk adjustment, 2b(4), and again there was discussion here as well.

I do think these scores reflect the discussion as well. Okay.

Let's move on. I think we are now officially behind, but hopefully, for the remainder, I think these are a little bit less controversial. We will go through what is left in the scientific accuracy, and then we will go through usability and feasibility, hopefully, relatively quickly.

The next one is 2b(5), which is differences: Data analysis demonstrates for

scoring that the differences -- you can identify statistically significant and clinically meaningful differences in performance or there is evidence of over or less than optimal performance. The primary reviewer for this was John, 2b(5).

DR. SKIBBER: The data will be provided in summary reports on each provider and will be expressed as the proportion of observed expected -- risk adjusted expected ratios that are above the 75 percentile for the peer group and overall.

The stewards proposed that this is going to contribute to controlling for case mix. Again, we spoke about that. It doesn't appear that that is tested at all.

The main driver of this cost difference is likely to be anesthesia use, and this will be shown with -- I don't know how this is going to directly impact on our ability to truly identify clinically meaningful differences between providers. It

is sort of like it is a fait accompli. They already know that is what is going to show.

I do think that there is an issue in anything like this of small numbers of individual cases for individual providers that is going to be profoundly affected by their case mix. The steward does acknowledge this in their submission. Other than acknowledging it, they don't necessarily state that they are going to account for it in any other fashion.

CHAIRMAN PENSON: Okay. Other comments? I agree with you, John. I just -- Looking at this personally, I don't know what to make what the differences is. First of all, I am not sure what the differences are going to be. I am sure they are going to be. They always are, but I don't know how to interpret the differences.

I didn't know whether that was meaningful and whether I could interpret it, which, if I could interpret it, I don't -- I didn't know whether to score this low or

insufficient, but the bottom line is I don't know what to make of it, and it is worrisome to me.

CHAIRMAN PENSON: On 2b(5)? Okay, we will move on to 2b(6). This is the one that looks at multiple data sources, and is there a demonstration that they can produce comparable results? I think, John, you win again.

DR. SKIBBER: This doesn't appear to be fully addressed in the submission. An additional concern is that the method they describe in their description for getting their resource use data only accounts for DRG facilities, and there will be some facilities that are currently at least DRG exempt, and how their data is going to translate exactly into the comparable data, while they propose a model for that, I don't -- Maybe others might know if that is a valid model.

MS. TURBYVILLE: I just want to add something for your deliberations before

you get further into this conversation. Typically, the way we think about this is, if a developer is requesting for it to be endorsed for use in different types of data -- for example, they are using both -- they tell a user they could use both administrative data and clinical enriched data, or differences in that type of manner -- then we would want them to demonstrate that, by specifying these different data sources, that they have tested them to see if the measure performs reliably in that manner.

I think, certainly -- and we have heard the conversation here that there are issues with administrative claims data, that that is, I think, going to be an issue throughout the measure specifications and concerns that administrative data itself, even if it is just commercial, as they are looking at it here, may have challenges. But for this particular subcriteria, we are really looking at when they have specified these distinct

data sources, and I do believe that they are just looking at commercial administrative data.

So we would endorse the measure if it were recommended for that to be implemented only in commercial data administrative populations.

DR. WALTER: The problem is there are many types of claims databases, like they did not validate it in anything other than an employer based, versus Medicare claims. So I didn't know whether to rate this insufficient or not applicable.

CHAIRMAN PENSON: Yes, I am with you on that, because the question isn't -- You don't have to meld Ingenix with Medicare to do it, which is what you are getting at, Sally, but what they have looked at is they have basically looked at a commercial payer dataset. I think it was Healthscan or Ingenix, but does it work in Medicare?

It probably does, but there is no

evidence there to support it. So my gut feeling was I would have liked to have seen that evidence.

MS. TURBYVILLE: And just -- I don't want to beat this up too much. We would not endorse this measure if it were recommended for implementation in Medicare data. It would have to be endorsed for commercial populations only.

I realize the measure goes 65 and over, and that is a whole 'nother conversation for this committee to have, but when it is tested in a certain database, a commercial population, that is all we can endorse it. What you said, you know, there are variations in commercial data. That is a question of how representative the market scan that they used, which is quite large -- whether that helps address some of those concerns of all of you.

CHAIRMAN PENSON: So I guess what I am hearing from you, Sally, for the panel is that, if this was endorsed, it would only be

endorsed in the population -- It wouldn't be used in Medicare, because it hasn't been tested in that dataset. It would just be Market Scan.

So in that respect, if you were just looking at Market Scan, I guess it sort of changes today for me. Others?

DR. WALTER: I guess I missed the point of the question then, because I thought it was about has this been tested in multiple data sources, and it has only been tested in one. Right?

CHAIRMAN PENSON: Right. So what they are basically saying -- I know what you are saying, Louise, and I think what Sally is telling -- What I am hearing is, yes, if they were looking at -- If they wanted to get it endorsed for Medicare and for Healthscan, then we would have to see them both, but it is only going to get approved for the one.

DR. GILLIGAN: So this becomes not applicable then?

CHAIRMAN PENSON: Yes, I guess it becomes unapplicable. I think that is reasonable.

DR. BORKER: Okay. So that answers my -- My confusion was generalizability versus testing.

CHAIRMAN PENSON: Okay. So I guess, if it is not applicable, we can move on. Lovely. So the next to the last one in this set is disparities: If disparities in care have been identified, did the measure specification, scoring and analysis allow for identification; that is, by race, ethnicity, status and gender.

This is 2b(6), and John, this one you have as well, with disparities.

DR. POTTERS: I had disparities.

CHAIRMAN PENSON: Oh, I am sorry. I apologize. Sorry, Louis.

DR. POTTERS: There is not much to say, because they really don't address it. I interpreted that as -- which is why I raised

the issue under exclusions and risk adjustment, because it gets back to the -- it really gets back to the unknown denominator, and we just said from an N/A -- you know, not applicable -- for the previous section, against other data sources.

Whatever the denominator is out there is out there, and whatever is in, I guess, this Med Scan data is going to subdivide whatever is in there.

CHAIRMAN PENSON: But what I would ask specifically, as I interpret it, was can the current Market Scan data, and can the current measure be used to address disparities by gender, by race, by age?

When I read it, I didn't see any evidence one way or the other.

DR. POTTERS: Right. Well, they didn't really address it.

CHAIRMAN PENSON: Right. So that, in my mind, is insufficient.

DR. POTTERS: Right.

CHAIRMAN PENSON: Other comments? That is what I am hearing you are saying, too, Louis.

DR. POTTERS: I went through the -- I am going through this thing. I am like where is the discussion on this, and there was really nothing. So they don't look at it, unless they want to comment.

CHAIRMAN PENSON: Kevin, Todd, did you do any analyses, disparities analyses, by gender, by race? Any evidence to help guide us?

DR. WEISS: Yes. We have not done anything by gender. We cannot in this dataset do anything by race, not captured as part of the Market Scan data. We did do some age based analyses within the population.

Honestly, I don't have the data in front of me to be able to tell you what differences there were. I don't think we found anything in, again, an under-65 population, but before I commit to that, I

want to dig up some data and make sure.

CHAIRMAN PENSON: Okay. I am going to call the question, but I do want to raise one point to the panel that you just made, which is that you cannot do a stratification by race or by socioeconomic status. So if people feel that that is an important piece to have here, then you would want to have that reflected in your vote.

With that in mind, let's wrap up the scientific piece. So let's do 2b(5), which is the differences in performance. Are they statistically significant, practically and clinically meaningful differences in performance? Let's go ahead and vote on that.

Okay. So just for the folks on the phone, we have three votes for moderate, four for low, and two for insufficient for differences in performance.

2b(6) we are going to skip over, because we felt that was not applicable, and 2c is disparities. Just to reiterate -- Oh,

I apologize. Yes, we have to do overall validity.

So this basically encompasses the evidence, the validity testing, the exclusion criteria, the risk adjustment, the difference in performance scores. I will ask everyone to vote on that now.

We have everyone, it looks like. So just again for the folks on the phone, and I think this is reflected in the comments, for overall validity we have five votes for moderate and four votes for low.

Now we will move on to disparities, which is the 2c: If disparities in care have been identified, do the measures specify scoring, etcetera, allow for identification of disparities through stratification or results? Let's go ahead and vote now.

Everyone, re-vote. There we go. So this one for disparities, we had two who said it was moderate, two who said it was low,

and five who said it was insufficient evidence.

So scientific accuracy. We are running behind. My inclination is to plow through this measure and then take a shortened break, if that works for people. I am getting yeses around the room. So that means everyone's bladder is holding out, says the urologist. Okay, great. Who invited the urologist to a colonoscopy meeting, huh?

So next we are going to talk about usability, and I think that, to some degree, these are some of the quicker discussions, I hope.

The first one is 3a, which is: Are the performance results, measure performance results, reportable to the public at large in national and community reporting programs at the time of endorsement or at least by the time of endorsement maintenance review?

So I will ask Tim, who is the

assigned reviewer, to discuss that.

DR. GILLIGAN: Yes. I don't think there is a whole lot to say. What is interesting here is that we are all over the map on how we scored this, among the four who rated it as insufficient, because it seemed to me that this wasn't really addressed in the proposal.

The standard here is whether the measure performance was also reported to the public at large in national or community reporting programs, and they are not, really, at this point. So I just got insufficient. I don't know if there are people who voted higher who will explain why they would give it a high score on this, because I didn't think I had any evidence or data to give it a score other than insufficient.

CHAIRMAN PENSON: My question, though, addressed to Sally and the NQF team, is this is currently being sort of tested by RWJ under the RWJ contract. So is this an

applicable issue now or is it something that waits for endorsement maintenance review in three years?

MS. TURBYVILLE: Very good question. So what I would ask is to call your attention to some revised language that we have for usability. During the time that this project was rolling out, concurrently the NQF was reexamining some of the criteria, including reliability, and it gets to exactly what you are talking about.

So now it is: Does the submitted information -- That is not right either. Which one is right?

MS. BOSSLEY: Sally, why don't you just -- The slide from before, can you put it back?

MS. TURBYVILLE: I will. Thank you.

MS. BOSSLEY: Let me just provide a little background, too. So we are currently looking at the usability criteria. So this is

a work in progress. So, hopefully, you all will work with us as we go through this.

We have a task force that is convening in the next month to take a look at, as we advance and move more into public reporting, actually different -- That is a small piece in the spectrum, like Sally showed, of accountability.

So how do we as NQF anticipate, when we first have measures submitted to us, what is the level of reporting that we expect or accountability? Where should it fall within that spectrum? It may not, and I think that is what you are seeing with these measures.

With these, I would also -- You are looking at a piece that falls within the efficiency framework. So I think we want to move very cautiously in recommending that these measures be put out there without that context of the quality piece.

So I think you all need to balance

this with it is the first time we are looking at these. These measures were just developed. They are tested. They have been looked at in many ways. They are being implemented within some community. So your ratings may be low, but it is in part because of just the state of where we are with everything. David, did you want to add anything?

CHAIRMAN PENSON: Yes. I don't know how, given what you have said, that we can vote anything other than insufficient, because we just don't have any information from the measure developer; and frankly, because it is in flux with NQF, it is going to make it very difficulty for us to make any conclusions.

I would argue, perhaps tomorrow or the next day or on a phone conference with the Steering Committee, that I wouldn't dismiss these measures just because we don't have this level of evidence.

MS. BOSSLEY: And I think part of

what you will do in framing in the Steering Committee will be part of this, as well as framing that. This, again, is the first step, and we, again, don't know what the implications are, how this will work out, and I think it is part of a broader look in the way of resource use combined with quality, and then in general.

This is something that most of our steering committees struggle with when they see new measures, because there is no information most often on how they are being used and how useful they are, how understandable they are.

So you are not the first, and you won't be the last to struggle with this. So we do have a committee looking at this specifically and, as we have more information, we will provide it back to everyone.

CHAIRMAN PENSON: So I am going to perhaps make a jump and wonder if anyone feels that they would be able to vote anything other

than insufficient here. If so, that is okay, but just holler out now. Steven?

DR. CHEN: I am fine with insufficient. I do want to throw out something when we get to 3b for the folks to consider.

CHAIRMAN PENSON: Oh, absolutely. We are going to go through each one individually. I am not suggesting that we blanket it, but at least for 3a, which is useful to the public, we just don't know at this point. Okay.

Then I will move on to 3b, which is that the results are meaningful and understandable and useful to the intended audience for both reporting and quality improvement.

DR. GILLIGAN; Yes. So I think, for me, actually the discussion we have had kind of sums up where I fall on this issue, which is we have had a lot of debate about how we interpret this and what the meaning is, and

my reading of the debate is it is really not clear to us what the meaning of it is.

So I rated it low for that reason, that it is just not clear yet what the outcome of this is going to mean to anyone.

CHAIRMAN PENSON: So would that be a low or insufficient?

DR. GILLIGAN: For me, that is a low.

CHAIRMAN PENSON: Okay. I think that is very reasonable. Steven, you had a comment with 3b specifically.

DR. CHEN: Yes. It is somewhat of what Tim was just saying. To the extent that I have issues with validity, I also have issues with portraying to the public a thing that I consider potentially invalid. But on top of that, I do want to throw out something with standardized pricing as it pertains to public reporting.

Standardized pricing, to me, makes a lot of sense as a researcher, because I want

to equalize resource utilization. What the public might actually care about, though, is what they are going to be charged.

So if some hospital has an extraordinarily high charge for something that some other hospitals are an extraordinarily low charge for, from the public's perspective the resource utilization is dramatically different, even if they do an identical thing.

DR. KLOTH: If I can inject some additional thoughts on that question, that very important question, as we move forward -- presently and as we move forward, there is a nationwide recurring, ongoing problem with drug shortages of a panorama of drugs which are involved in the care of these patients, including propofol.

So shortages and those related issues and dealing with shortages can have wildly fluctuating impacts on what the costs are to the provider and, therefore, the cost to get passed on to the patient.

In some cases, providers have to go to what I would call the black market. It is more generally called the gray market. You could call it scalpers, but the question is, when there is instability in the supply chain of medications, it can have significant impacts.

CHAIRMAN PENSON: I think those are good points. I also wonder if, to some degree, we have captured that already in the validity piece with the standardized pricing, but again, the question now is: Given that, is it useful for public reporting?

What I am hearing is concerns. The questions is whether or not those concerns warrant a low vote or whether they warrant just more evidence, and that is a matter of personal opinion.

Other thoughts on reporting and quality improvement vis a vis whether or not these are meaningful? I would add, and I rate insufficient, I think as a consumer, and a

consumer looks at what is a risk adjusted score for a provider, what they cost, even if you put quality on it, I am not sure a consumer -- and I am not condescending to the consumer, because I am a health care consumer, too. I am not sure what to make of that.

If Dr. Smith versus Dr. Jones' ratio is 1.2 versus 1.1, how do I use that? Maybe there will be more evidence to tell me how to do that. You know, I will get my Consumer Report circle or something down the road. That is where I think we are going to end up, frankly.

That being said, for me I just -- I don't know what to do with this. So I am between low and insufficient, leaning more toward insufficient.

DR. POTTERS: I think it would be unfair to -- this is just my opinion. It would be unfair to vote insufficient, given the fact that, for Section 2, we had a hard time on the validity part of it.

So given the fact that we understand this, to a large degree, more so than the general public, to just say it is insufficient and yet the validity doesn't count or doesn't show at least a high vote would not be right.

CHAIRMAN PENSON: I know exactly what you are saying, and the question becomes, did you capture that in the earlier vote or do you want to -- I don't know a better way to put it -- double jeopardy? I don't know. I think in the end the whole thing is a Gestalt. Whether it is low or insufficient, as Sally alluded to when we started, if it is not moderate or higher, the Steering Committee is going to have a hard time running with it. I will tell you that up front. Steven?

DR. CHEN: In distinguishing between the two, I guess the way I meant to think about it is I tend to vote insufficient if I think the part has validity, and if they just give me more information, I could vote

high. I vote low when I think this is probably not fixable. Even if you manage to fix the validity, I'm still not comfortable that this should be released to the public.

CHAIRMAN PENSON: Yes, and that is kind of where I am playing out in my mind, too.

MS. TURBYVILLE: I just want -- Before you vote, that it is important to remember that what we are talking about is various types of public reporting as demonstrated in this slide. So certainly, it is important to think about the individual consumer, but there are potential other uses for the measure for you to deliberate as you think about your potential ratings for this measure.

CHAIRMAN PENSON: All right. Let's keep plowing through the threes and fours and get to a break, because even I am getting tired now.

So let's talk about 3c which is

the clinical and construction logic, and that is basically, that the data and result detail are maintained such that the resource use measure, including in the clinical and construction logic, defined in the measure can be decomposed to facilitate -- and here is the magic word when I reviewed this -- transparency and understanding.

Again, I will throw this to Tim.

DR. GILLIGAN: Yes. For me, that is the key word at the end there. Most people rated this moderate or high, and some curious people still feel that way about it.

I think this O to E ratio and what it means was a concern I had, which is why I gave it a moderate, but I did think that with more evidence -- I think that does have meaning. It just needs to be spelled out a little bit more as to what that is going to mean, but I don't want to confuse 3b and 3c, because I think we raised a lot of issues on 3b, and 3c is a separate issue, and we should

look at that cleanly.

I think, for my money, it deserved a higher score than 3b, because I think the problems there can be fixed with more data.

CHAIRMAN PENSON: So I was the outlier here, looking at everyone's scores. The reason I did was -- and perhaps I am not being fair about this -- was something caught my eye in looking at the provider report, which was that you had reporting by specialty type, and then the magic word, peer group; and peer group was not well defined.

I think, as a clinician, I get very nervous when something can be turned on the -- maybe I shouldn't wear my clinician hat, but I think that is why we are all -- those of us who are clinicians are here.

I think that needs further discussion, because in the end, when you are going to have accountability and comparisons between providers, and you are going to have potentially physician tiering, whatever it is,

you need to be completely transparent, and it wasn't, to me. But I was probably a little harsh. No getting around that.

Other comments with 3c? All right, I will keep moving along to 3d. 3d -- what's that? It's N/A.

MS. TURBYVILLE: Right. So we didn't ask the developers to try and harmonize at this point. If we get in the process and they need to, we will work with them to do that.

CHAIRMAN PENSON: Terrific. All right. So let's then vote on the threes now, and then we will, hopefully, get through the feasibility quickly and take a break.

So the 3a is regarding measure performance, that there would potentially be reported to the public at large and national community reporting programs by the time of endorsement maintenance review, and discussion about exceptions.

This was the one where we stopped

and discussed whether or not there was no evidence presented and that NQF is still sort of moving on this. I think the general feeling -- I don't want to sway -- I do want to sway the vote. We sort of said it was going to be insufficient. So let's go ahead and vote. We stopped the discussion based on it, and it's okay.

See, someone is messing with me, okay? That's fine. As long as you meant to do that, I'm fine with that. Dr. Potters -- it's always the radiation oncologist messing with the urologist. I'm teasing you.

All right, let's move on to 3b. So just for the folks on the phone, one person voted moderate. The rest of the panel voted insufficient. I would argue that I don't think that is a negative reflection on the measure.

3b is usability: Did the measure results -- Are they meaningful, understandable, and useful to the intended

audience, for public reporting, quality improvement? We could have discussion here. Go ahead and vote.

There we go. So this one, 3b, we had six votes for low and three for insufficient.

3c in usability is around clinical construction logic. The resource use measure, including the clinical construction logic for defined measurement, can be decomposed to facilitate transparency and understanding.

So here we have seven votes for moderate and two for insufficient. And, obviously, 3d was not applicable.

So let's see if we can't run through the feasibility measures. I think they will be relatively quick, and then we can take a break.

We are going to 4a, which is regarding the byproduct of care. For clinical measures, the required data elements are routinely generated and used during care

delivery. I don't think we have an assigned reviewer here.

For the most part, looking -- It's me.

MS. WILBON: So, David, the 4a and 4b, because it is administrative claims data, they are kind of --

CHAIRMAN PENSON: They are assumed?

MS. WILBON: Yes, you guys can still vote, obviously, but --

CHAIRMAN PENSON: Basically, that is what I was going to say, is I don't think there is a lot of discussion, that these are routinely capture. We will vote at the very end. And administrative data, I don't think anyone is arguing with that

With 4b, again obviously, the data elements come from administrative data claims. So they are going to be there. I don't think anyone is arguing with that.

4c is susceptibility to

inaccuracies, errors or unintended consequences related to measurement, and these are judge to be either inconsequential or minimized or, alternatively, they can be identified and avoided. So, Tim?

DR. GILLIGAN: Yes. I didn't see any issues here, except those that are inherent working with administrative datasets, which are imperfect by definition. So I didn't see any way they could do more than they had done, honestly, on either 4c or 4d. I think there are limitations to this methodology, but you have to work with these datasets to get the data.

CHAIRMAN PENSON: Anyone have anything to add to that, or disagree? All right, excellent.

Then we will just briefly discuss 4d. The data collection management strategy can be implemented as demonstrated by operational use in external reporting programs or testing did not identify barriers for

operational use.

So, basically, can it be done? Can it be operationalized in a reporting program?

DR. GILLIGAN: I didn't see any reason why it couldn't. There are always going to be cost and manpower issues, but other than that, I didn't see anything exceptional here.

CHAIRMAN PENSON: And I think our votes reflect that. Anyone want to add to that? Okay, why don't we vote on this, and then we can take a 10-minute break.

So 4a: This is basically required data elements generated and used during care delivery in an administrative dataset. I am tempted just to vote by acclamation here.

I think we will do these two by acclamation from now on. Oh, there we go. Okay. Insufficient -- I don't want to put anyone's feet to the fire, but given the discussion, did someone vote insufficient by

accident? If not, again, the person who voted insufficient -- we certainly need to include some sort of comment. So could someone take credit for that?

DR. POTTERS: I voted insufficient. It wasn't clear based on what the intent was, you know, by Sally's comment.

CHAIRMAN PENSON: All right. So, basically, just again, the idea here is that these are sort of very feasibility issues. So in other words, do the data in the administrative datasets -- are all the required data elements routinely generated in an administrative dataset and used for care delivery?

So if that is unclear to you, if it is insufficient -- is that how you interpreted it? It's clear. So --

DR. POTTERS: I would go with the consensus.

CHAIRMAN PENSON: Okay. So again, you don't have to, Lou. So can we just

reflect that? So it will be eight votes for high, and one vote for moderate.

For 4b, I am going to vote by acclamation. Can everyone agree that it is a high probability that the required elements are available in an administrative dataset? Okay. So we voted -- all nine voted for high.

Now this is where I think we do have to have a vote, which is 4c. Susceptibility to inaccuracies, errors, or unintended consequences related to measurement are judged to be inconsequential or can be minimized or can be monitored and detected.

There we go. We have nine. So we have four votes for high and five votes for moderate.

The last one is barriers to use, that the data collection and measurement strategy can be implemented as demonstrated by operational use and external reporting programs or testing did not identify barriers to operational use. Let's go ahead and vote

on that.

We have all nine. So we have four votes for high and five votes for moderate. Then I think we have to vote for the overall feasibility, if I am not mistaken. No, we don't? Okay, great.

So let's do this. We are running a little bit behind, about 15 minutes behind schedule, as it is. Let's take a 10-minute break and reconvene at 20 of 12:00, with hopes of going through the next one somewhat quicker, although perhaps not much.

(Whereupon, the foregoing matter went off the record at 11:31 a.m. and went back on the record at 11:46 a.m.)

CHAIR PENSON: So we'll get started again. With any sort of luck this will go a little quicker. The reason I think this will go quicker is because we'll sort of, to some degree, try to truncate our comments.

I'll ask people to really focus on Criteria 2, the scientific piece of it. I

think that we can spend a lot less time on the importance part and certainly much less time on usability and feasibility, because this measure, in many respects, sort of recapitulates the earlier measure.

And with that in mind I will ask, are the ABMS Foundation folks still on the phone? Have we scared you guys away? They finally said we've had enough of those people. Okay, very good. Well that makes it move along even faster.

So the next measure we're going to talk about is episode of care around treatment of localized colon cancer. I just need to get my notes up here. But basically I think that this should be an interesting discussion.

In many respects it's very similar to the last measure. I'm actually trying to get the measure up here, so forgive me. But I'll start the discussion while I'm getting it up.

With regard to Measure 1a, which

is the does this address a goal priority identified by the partnership, is this a high impact issue? I think most people voted yes on that. Is there any real discussion, anyone feel this isn't a high-impact topic?

I sort of figured we'd end up there, I think everyone voted this as high. The next one is 1b, A Demonstration of Resource Use or Cost Problem and an Opportunity for Improvement. Is there data demonstrating variation, delivery of care, cross providers and population groups.

And this, to me, even though I don't remember seeing a whole lot of evidence presented that there's opportunity for improvement here.

It was sort of my take on this that there probably is real variation in the way colon cancer is treated across the United States, and by extension there's going to be variation in cost and there is room for improvement here.

I'll open it up to the floor. I think some people, actually there were one or two people who voted low here, and so I wonder if there are people who have different thoughts?

DR. WEISS: Just to note on the telephone, with apologies. Kevin Weiss here, we were on the line but we were not able to speak, they had us muted.

So we're available to you, it seems that you're well on the way so we're available to you as you go forward.

CHAIR PENSON: Okay. Great, before you comment I'll just invite the measure developers if you want to add anything specifically about this measure now would be a good time.

DR. WEISS: Probably the only thing I would say is that as you can imagine this is, as the working group was deliberating this issue of cancer the stages of cancer became very much he issue of clinical concern

right up-front.

And the way that this was addressed by the Work Group in an obvious way is to find localized colon cancer by way of treatment exceptions and that's how this got built. For the members of the group I'll just let you know I'm an internist not an oncologist nor a surgical oncologist.

So I can't give you any of the nuances with it to how this was decided, but that was the intent of this measure and it's we did not look at the more advanced cancer related to colon cancer because of the inability entirely to manage the staging question.

CHAIR PENSON: So I'll just, you kind of went in and out, but I'll just repeat for the group as what I heard in you saying was basically that this was limited localized disease as best you could because you felt that that would make it more comparable among patients, is that a fair statement?

DR. WEISS: Yes.

CHAIR PENSON: Okay. Louise, you were going to say something?

DR. WALTER: I guess one quick thing on the stage, it sounds like they're using colectomy as a proxy for stage, but you could still have really quite advanced disease and get a colectomy.

So actually my main, the reason I voted this low was I didn't see how this was going to identify, without the very important clinical characteristics like histology or stage, how it would identify meaningful differences or, you know, inappropriate variation versus variation based on how sick the person was, as far as their disease.

CHAIR PENSON: And I think that's a very well taken point and I know it's going to come up again when we talk about risk adjustment.

And I guess, just to interpret what you're saying, Louise, is that you think

that's such an overwhelming problem that it actually affects your ability to look among differences in use. Other folks want to add to that?

Okay. So I'll move on, again, to 1c which is the purpose or objective of the resource measure are clearly described. I think in this respect I think the purpose and the objective is clearly described.

And I think most people felt that way, but not everyone, so I'll throw it open to the group here.

DR. WALTER: I guess my main question, because I just didn't understand, is this just going to look at variation of chemotherapy or what is it that they're actually, they don't actually have a hypothesis unlike the colonoscopy one which actually said we think anesthesia is going to be different, we think complications are going to be different.

I didn't seen any hypothesis

around what they thought was going to be the driver of costs.

CHAIR PENSON: So I'll put that to the measure developers on the phone that one of our panel members is having some real concerns with this, which I think to some degree I share.

And I think others on the panel may as well, that it's really hard without adjusting for things like stage and grade to interpret these things.

And so what are your thoughts on this vis-a-vis, what do you expect to find here?

DR. WEISS: So the workgroup was again driven by oncologists on the group and the general surgeon input was, general and colorectal surgeon input, was that they felt that there was a lot of variability in complications associated with the surgery, some of which appeared not really post op, but actually took a number of months to evolve.

And that it was important to capture a long episode cross associated with it and that would be associated with the care that would encompass the ability to look at the various types of complications that were fixable over a short period. And major complications that actually led to severe issues of patient compromise.

It's not a quality measure, clearly, but they understood it associated with initial procedure and that they would see the variation in resource use attached to this.

CHAIR PENSON: So let me sort of summarize what I hear you saying, Kevin, for the group. And sort of help us to sort of focus on the importance piece. And I think this helps you, Louise.

Because what I hear you saying is basically by focusing on patients who have had colectomies, specifically, you think that you're probably selecting for patients with

localized disease, which I think many of the people in the room might disagree with you on.

The surgeons in the room, I think, definitely would. But that being said, I think it's a rough proxy to some degree, maybe.

But what you're really hypothesizing here is that you're going to capture differences in cost primarily related to surgical complications after colectomy.

Either immediately, post operatively or afterwards. And that certainly you have some stratification, like chemotherapy that may or may not help. So is that a fair summary, Kevin?

DR. WEISS: Correct. And the fact that the costs associated with care, particularly if there is a complication, really would rest outside of the usual, if you did it on a short-term morbidity, a 30 or 60 day window you'd miss a lot of the care complications that may end up in long-term

care or long-term care needs in some fashion or another.

So that's why the episode was built around that long duration. The term of the episode was built because that is the treatment term for a localized colon cancer.

Now in addition it's recognized that if a person had just localized -- and by way of my listening to the working group I don't want to say that I don't have the clinical expertise in this area.

That if they have localized colon cancer without chemotherapy that would really bespeak the this is probably an advanced disease.

Although there are some times when one would do a rescue, kind of some sort of a salvage operation to a person with colon cancer, who has very advanced, to do a colectomy. But to do that without chemo would bespeak a localized process.

CHAIR PENSON: Okay. So I think

I'm going to sort of discuss with the panel a little bit about how to interpret this. Because I think they're going to be -- I see you, Steven, so give me a second.

I think there are going to be issues here with validity and risk adjustment we're going to get into. But I think that when we're looking at this first criteria, which is sort of the importance to measure and report, we really should sort of focus on the importance.

And what I would say is based on what Dr. Weiss has told us that this is really focused on looking at sort of sequelae of colectomy in colon cancer patients. That's how we're going to kind of end up interpreting this, that's the de facto what this is.

Is that going to be important and meaningful? So that would be my comment. And, Steven, you had your hand raised.

DR. CHEN: Yes, I was going to say as far as 1b is concerned I do think that

there's a big opportunity here because I do think there is a fair bit of variation. As far as 1c I think that discussion just reflects exactly why I put moderate.

Because I didn't have a good sense of what were they attacking. Because one of the big variations is actually do Stage II colon cancers get chemotherapy or not.

Somewhere around a third of them do. And then about a third of people who are Stage III who are supposed to get chemo don't get chemo.

So the stratification on that leads you to big issues. And then, not to mention Stage IV people who get colectomies anyway. But having said that I would say 1b for me was a high and 1c was a moderate best because I don't thing there's clarity here.

CHAIR PENSON: So again, what I will tell you here is I hear what you're saying, I agree with you 110 percent, but wonder if that's a question for the next

section on validity and risk adjustment. But maybe not.

DR. WALTER: Well I guess my main feedback was there's nothing about this being sequelae for colectomy in anything on the section on impact importance. So the feedback to the developers would be put that in there, because there's nothing about that.

CHAIR PENSON: Right. And I would add that it's also about given the long-term followup it's about sequelae of chemotherapy as well. And if you look at the accountability piece, the surgeons are only held accountable for the first six weeks and the rest goes to the medical oncologist and, again, I think we'll get into this discussion a little later, but that's a concern.

So with that being said I am hearing some concerns with 1c specifically and perhaps a little bit with 1b.

Let's just quickly go through 1d which is the resource use service categories

that are included are consistent with and represent a conceptual construct represented by the measure.

And I think that for the most part people were on board with that, if I'm not mistaken. The people thought that the categories made sense. Any disagreement on that? And most of the scores were high or moderate.

So again, not to truncate discussion, I think to summarize where we've been with this I think in the room people feel that this probably is important but there are some concerns. Because what exactly are we measuring here. And we will get into that with validity for sure.

But let's just accept the fact that there's no discussion about the sequelae of treatment, whether it's surgical or potentially chemotherapy and it's not really discussed well there. And I think that gave people some pause. Anything else to add to

that, the summary?

Okay. So let's vote on this then. So the first one is on the impact. Does the measure focus on a specific national health goal/priority, or is there evidence to support that it's high impact, or in our clinical experience is this high impact issue. So let's vote.

I have to say I don't love this voting system. We're missing two. We're missing one, one. Nine, we've got them all, good. So not surprisingly everyone voted this was a high impact issue.

Let's move on to 1b, which was demonstration of resource use, or cost problems and opportunity for improvement. So in other words is there a performance gap or is there variation?

It's always that last one. There you go. And this was, again, fairly acceptable. Five people voted high for 1b, and four people voted moderate.

1c, was the purpose/objective of the resource use measure, including its components, and the construct are clearly described.

And I think let's just vote on that. There we go, we have all nine. And so we have two who voted high, six who voted moderate and one who voted insufficient.

And I think the person who voted insufficient I think we've had that discussion if you feel that, I felt it sort of makes sense.

So let's do the last one which is 1d, the resource use service categories included are representative of the conceptual construct.

Well let's just vote on it, it's quicker. There we go, we have all nine. And so we had five who were high, three were moderate and one who was insufficient.

I'm going to call out the person who was insufficient, just felt that the

comments that we already discussed cover that or if they want to add anything else.

(Off microphone discussion.)

CHAIR PENSON: Okay. All right. Why don't we just re-vote so it reflects it properly?

Should end up five to four now. I like it when that happens. So five voted high and four voted moderate. Okay, excellent. So now let's move on to the scientific criteria.

And we're going to start with the first one, which is the measure is well defined and precisely specified so it can be implemented consistently within and across organizations. And the assigned reviewer for this was John.

DR. SKIBBER: Okay. I felt that the measure would benefit from a more explicit statement of the meaning of localized, because that's not clear. Also it's going to be important to eliminate any costs related to disease surveillance that may develop during

that first year.

It's very common that patients are going to followup for surveillance during the first year, in fact it's in all the guidelines. There's no accommodation for eliminating those costs.

The one thing that's not mentioned, and it may be an assumption by the developer, is that from my experience of looking at administrative databases that differentiate colon from rectal cancer, that is ill defined at best. They don't mention it at all.

And I think they at least need to acknowledge that the treatments are different between those and that rectal cancer should not be included.

The general approach is clearly stated. The target population, that note was not filled out on their submission sheet, but I would assume is going to be fine. Data inclusion and exclusion criteria are clear.

CPT coding for a colectomy is extensive, to say the least, and includes a number of procedures that are not commonly done for colon cancer.

And I would say that when you look at the way that they created an exclusion for inflammatory bowel disease in the initial, the colonoscopy set, and then you compare it with this they completely ignored that.

So what that means is there is a number of those, actually relatively high-cost procedures that are done for either IBD, polyposis, a variety of things that they don't limit their colectomy definition to.

And that muddies the patient population somewhat. I'd like to either, they should at least either acknowledge this or recreate this in some fashion.

CHAIR PENSON: So maybe I'm misinterpreting this. But are you saying that basically there are going to be patients who are included in this measure who did not have

a colon cancer diagnosis?

DR. SKIBBER: They might have had it, however the setting in which this cancer was found and treated is going to be substantially different than, what I think is the purpose and a very worthwhile one, which is to look at the patient who's undergoing a routine colectomy. So that might be a consideration.

CHAIR PENSON: So let me just state this so I understand and also so that the measure reviewers, your concern is that a patient comes in with say IBD and is having a total colectomy for symptoms and they find a unexpected colon cancer and that patient's included?

DR. SKIBBER: Yes.

CHAIR PENSON: Okay.

DR. SKIBBER: And you know, the procedures that they have described some of them are very high-end technical procedures for patients with unusual conditions, like you

just described. I'm sure their working group is well aware of this and probably felt that it wasn't important, but --

CHAIR PENSON: And, John, how important do you think it is? I mean is it a fatal flaw, is it a minor point?

DR. SKIBBER: No it's a fixable, I would consider it's a fixable thing. If I was trying to do something that was really going to be resource efficient and really look at the issue of localized colon cancer I would adjust that.

But just to carry on and not belabor that, the way that they capture costs appears to be reasonable and clear. To get into the risk adjustment it's a similar, for me at least, it's a similar set of problems.

CHAIR PENSON: Right.

DR. SKIBBER: To the colonoscopy issue. Probably more important here, the comorbidity is, the way this adjustment is used, again the statistician can address his

concerns, I'm not aware of any valid way to take into account the way that they created their risk adjustment strategy, as said before.

CHAIR PENSON: So just for the sake of process, because I think we're going to have that discussion when we get to the specific risk adjustment sub-criteria, but let's table that.

And by the way it's very similar to the last measure and I think we're going to come up with a lot of the same things.

Let's just focus just on that 2a1 to start with, which really is, if you feel that it's well defined and precisely specified. And I'm hearing that it's not perfect but it's okay?

DR. SKIBBER: Right.

CHAIR PENSON: Other comments?

DR. KLOTH: A couple of questions. I just can't figure out if these patients that are collected as part of this measure will be

patients receiving adjuvant chemotherapy, yes or no?

And if so I can't find if they're going to stratify per KRAS testing, which is a vital component of determining what is optimal therapy. I just couldn't, if those portions are there I just couldn't see them.

CHAIR PENSON: So they're not, as best I could tell. And I'll let the folks on the phone correct me if I'm wrong. They do control for adjuvant chemotherapy in that they stratify their analysis by date the patient received chemotherapy or not afterwards.

Now whether or not the people in this room feel that's adequate is another discussion which I think we're going to have in a little while.

But as far as KRAS testing, you know, I mean I suppose that could be captured in an administrative data set, so I'll ask the folks on the phone, did you guys attempt to capture that at all?

DR. WEISS: Let me see if Todd Lee is available. He was on mute and I'm not sure he's been able to get off mute.

DR. LEE: So can you hear me now?

CHAIR PENSON: Yes, we can.

DR. LEE: Okay. Good. I didn't mean to be a Verizon commercial, sorry. Yes, all of the testing as part of the standard care of the patient undergoing a colectomy is intended to be captured.

So the KRAS testing will be captured as long as it's captured in one of our CPT codes that's listed here or it has a eligible ICD-9 associated with that claim. We do not stratify the population by receipt of a KRAS test.

CHAIR PENSON: All right. I think that answers Dwight's question, thanks Todd. Steven?

DR. CHEN: All right, since this has to do mostly with implementability and reproducibility, I think my concern, one of

them harkens with John was saying as far as the reliability of people who, they get a colectomy for a large sessile polyp or whatnot, partially colectomy and then it's found to be cancer later, whether they're going to be included or not.

Because the measure says that they have to have ICD-9 code and the operation on the same claim-line. Now some institutions will go back and change the ICD-9 code and bill it properly. Some institutions quite frankly are too lazy to do that. So I have worries about reproducibility as it pertains to that.

The other issue I have goes back to the colon and rectal issue. They're radically different diseases, I think for reproducibility and for reliability sake you'd want to separate them.

But the more you separate them the less reproducible it's going to get. And that's a conundrum I don't think is --

CHAIR PENSON: Well I actually think so we will get to sort of the reproducing and reliability piece. But just stick on this first part, this 2a1, I think that it critical.

Because if you don't feel that you can group colon versus or not the measure specified the right pop rectal patients together that speaks to whether population on the measure.

And I'll defer to both you, Steven, and you, John, as colorectal surgeons whether or not you feel that's an appropriate grouping.

And if you say no, it fatally flaws it, I think that's something that the panel needs to know. We rely on you guys for your clinical opinion on that.

DR. SKIBBER: I think there's two points there. One is that the ICD-9 coding is different. So within the limitations of that that's fine.

The other thing is that, and I think I got this from the initial presentation by the developers, was that their working group decided, as many of the working groups do on this, decided to say it's really almost an intention to treat.

If they were treated as a colon cancer then they were, you know, then that's the way it's going to work out for this measure, which is fine. I don't have a big problem. But there definitely will be some overlap.

One thing that might be helpful though is to exclude any radiation therapy. I don't see that addressed. Commonly rectal cancer is treated with that and that's a big difference than colon.

And so if that could be either considered as an exclusion that might be helpful in clarifying that.

DR. CHEN: Yes, I do think they need to be separated. In every paper I've

every written about colon, I've written about colon specifically. In every paper I've written about rectal I've written about rectal specifically.

The papers that have combined the two I've criticized heavily at meetings to say that you're mixing apples and oranges.

CHAIR PENSON: Okay, and can that be done with administrative data comfortably for you guys? So I wonder and, Sally, are we better off voting after each discussion with the sub-criteria or taking it on whole do you think?

MS. TURBYVILLE: There's a benefit of moving through maybe 2a2, but if you think the conversation is so lengthy that you want to capture it quickly we're fine with that. But it may be worthwhile to go 2a1, 2a2. And then we can dive into validity after that.

CHAIR PENSON: Okay. I'm fine with that. So what I'm hearing, to wrap up on 2a1, is that on the one hand the measure is

well defined, it is specified and it can be implemented consistently.

But I'm also hearing some real concerns from the clinical experts that the way it's defined and the way things are grouped may not be acceptable.

There's another way to put it. And I'm hearing real concern, basically no, I can't do it from Steven and at least moderate concern from John.

And so I think that that's something worth considering when we take into votes. Certainly if you accept that you can do colon and rectal together I think it's certainly specified and clear.

Let's talk a little bit about reliability and repeatability. I think many of the issues here are going to be the same as what we saw in the earlier colonoscopy discussion.

So what I'll ask to do is I'll ask Jay, who has been assigned to talk about

reliability, to add anything. And if you think it's the same as the earlier measure just say so and then ask Carlos to add.

DR. SCHUKMAN: I think it's primarily the same as the earlier measure there. You know there was a comment, something earlier on in here and you referenced it earlier, is the issue between the surgeons and the oncologists and splitting them up and looking at it.

Because attribution is always an issue. Always an issue going forward and I want to put that out there.

CHAIR PENSON: Yes, I agree. I think this is a real concern here. You know the question becomes at least, and I think we'll come into that particularly when we get to reporting in 3. But can you reliably and reproducibly assign to a surgeon and a medical oncologist? I worry about that.

DR. SCHUKMAN: Yes, I do too.

CHAIR PENSON: I really do.

Carlos, your comments and your review on the reliability piece?

MR. ALZOLA: No I feel that the same comments as for the other measures apply. In the sense that how reproducible it is for someone who wants to implement the measure. But I guess I agree that the attribution is going to be a problem.

CHAIR PENSON: Okay. Other comments on reproducible or reliability? Dwight.

DR. KLOTH: A question. This is stimulated by a discussion we had at Med Onc. Faculty meeting just yesterday. And the pending mandatory conversion to ICD-10 coding rather than ICD-9.

What I see in the document is ICD-9 and I can't find any reference to ICD-10 and if or how they're going to address that.

CHAIR PENSON: You know this is where my grandmother would look and go "Oy" so --

MS. WILBON: So if I can, actually, I did a project here in NQF last year on ICD-10 coding. We actually have gathered a group to think about how we might be converting our whole portfolio of measures at this point that are based on ICD-9 codes, use 9 codes and we're going to have to convert them at some point all to ICD-10.

So we actually have some processes that we're putting in place now to work with measure developers to get those measures converted. And at this point just where this project happens to fall we won't be requiring them to submit ICD-10 codes until October.

So we'll be working with them through maintenance and annual updates processes that we have imbedded in our NQF review process.

CHAIR PENSON: So yes, Sally says don't worry about it now. Which I kind of am, but is it safe to say, because Dwight I think you raised an important point. Not just for

this panel but unfortunately for everything we do.

But is it safe to say that we can proceed on ICD-9 only and in the maintenance, assuming it's endorsed in review, they'll deal with the ICD-10 question, is that a fair statement?

MS. WILBON: Yes, that's fair.

CHAIR PENSON: So I'll ask you, I mean, Dwight, it's a great question. I have no clue how I see it, God knows, oy. But let's just go on the assumption that ICD-9 is okay with this, if that's all right with you all.

(Off microphone discussion.)

CHAIR PENSON: Yes, it's going to be ugly, it's going to be really ugly. Other questions about reliability/reproducibility, comments? All right. Not having heard any.

So what I'm hearing, we talked a little bit about, we're going to vote now on 2a1 and 2a2. I'm not hearing overwhelming

concerns about reliability/reproducibility, but going back there are some that Carlos raised earlier with the colonoscopy measure, and I think there are some here as well.

And I think are reflected in an earlier discussion. So let's go ahead and vote. First on 2a1, which is the specifications.

And again, this is the one where basically I think, if I can summarize. If you feel comfortable grouping colon and rectal together then you probably are okay with this and if you feel that's a fatal flaw you want to reflect your vote here.

DR. LEE: Can I, this is Todd Lee from ABMS.

CHAIR PENSON: Yes, sure.

DR. LEE: Can I ask a clarifying question there? Because our measure focuses solely on colon cancer from a diagnostic code standpoint. The intent was to focus on colon cancer.

We realize folks with rectal cancer may get in. So am I hearing that that would have been the preference that we would have explicitly excluded people with a rectal cancer, ICD-9 code?

Because right now that group is not included unless they also have a colon cancer diagnostic code.

CHAIR PENSON: So, Todd, I'm glad you jumped in there. Your timing is very, very good. So looking at the colorectal surgeons they would have preferred that those were specifically excluded.

The question becomes is, if a patient has both a colon cancer code and rectal cancer code, because if they just have a rectal cancer code what you're saying, Todd, is that they will not be included in the measure, correct?

DR. LEE: That's exactly correct.

CHAIR PENSON: Okay. So my question to the content experts in the room

is, all of the patients have a colon cancer code. Some of them will also have a rectal cancer code as well. What do you think about that, Steven?

DR. CHEN: Yes, that's still going to be a problem. Because you're still going to treat their rectal cancer with radiation oncology.

They have much higher incidence of using chemo at lower stages. It's a dramatic difference in their resource utilization to have any rectal cancer component.

CHAIR PENSON: And if you either excluded them or at least stratified them would that make you feel better about the world? So Steven said yes. Just so Todd and Kevin you hear that. John do you agree with that?

DR. SKIBBER: Yes, I think if there's an explicit exclusion that's clear it's fine.

CHAIR PENSON: Okay. So I think

with those comments we can vote now. So for the benefit of the folks on the phone, seven people voted moderate and two people voted low.

And I would strongly suggest that, I think I know who voted where. I would strongly suggest that you consider the comments of the two people who voted low.

And next we'll vote on 2a2, which is reliability. Does reliability testing demonstrate the results are repeatable and do they get the same results the high proportion of the time when assessed in the same population and the same period.

Let's go ahead and vote on that. Okay. And so you have eight votes for moderate and one vote for insufficient. And I think that's based on prior discussions which have been had. We'll keep moving along then.

(Off microphone discussion.)

CHAIR PENSON: Oh, sorry. We

have to -- thank you, vote on overall reliability. Was the overall reliability testing both based on the two prior.

And here we have nine people who voted moderate, so I guess we consensus. Not what I expected, I don't know why, but okay. Very good, I guess that's a good straw for that.

So let's keep moving along. At some point I'll get hungry and then get cranky and then we'll really be in bad shape.

(Off microphone discussion.)

MS. TURBYVILLE: Did you try and plow through some scientific acceptability and then --

CHAIR PENSON: Yes, let's try to, let's see where we're at, but I think what we'll probably do is we'll plow through the scientific acceptability. Break for five minutes to get lunch and then do a working lunch to hopefully catch up.

So let's look at the evidence

question that measure specifications are consistent with the evidence presented, or at least your clinical spin on the evidence, to support the focus of measurement.

And that the measure is specified to capture the most inclusive target population indicated by the evidence.

And so this may be another area where we can talk about that colon versus colon versus rectal piece. And this will be John.

DR. SKIBBER: I basically said my piece on this before. So if they take that into account I am happy with where they include patients. You know this also works down to the issue of the, well, I think it's fine basically.

CHAIR PENSON: All right, did others have new information to add? I think we've covered a lot of this. I'm sorry, go ahead.

DR. CHEN: Is this where we talked

about the stage issue, here? Or should I hold that for validity testing?

CHAIR PENSON: I would hold that for validity or even risk adjustment, I think that's a critical piece. I wonder if that's not in the risk adjustment piece.

But in the end, because it can't be fixed in the risk adjustment, it may be a validity issue.

But here it's basically that the evidence presented, you know, supports the focus of the measurement and it captures an inclusive population. So I think the staging piece should probably wait. Other comments?

DR. KLOTH: A question.

CHAIR PENSON: Yes, sir.

DR. KLOTH: And maybe there's other portions where this would be applicable but I'll ask it now. And if the authors are on the line.

How will they track chemotherapy costs delivered in the hospital outpatient

department. Because that's not, I see reflected ambulatory, take-home tablet costs.

That's by NDC code. But I don't see where they're going to track the actual chemotherapy resource utilization?

CHAIR PENSON: So I will refer that to the folks on the phone. To some degree it goes back to something we discussed earlier with the service categories. But Todd or Kevin, could you answer that question?

DR. WEISS: Yes, we have a long list of J codes in the specification that are intended to capture those chemotherapies that are delivered in an outpatient chemo unit or something like that.

CHAIR PENSON: Okay. Thanks, I appreciate that, Todd. So I think we've discussed 2b1 regarding the specifications being consistent with evidence.

I think we should move on to 2b2, which is validity. Where we're going to have, I think, a more sort of interesting

discussion.

Now to remind you validity testing demonstrates that the measured data elements are correct and that the measure score correctly reflects the cost of care and resource provided. So in other words are we capturing what we say we're capturing.

And I don't think is where we have a discussion about staging, which I think comes in sort of, but maybe it does come here as well, it really comes into risk adjustment. But I'll throw open the floor and I'll ask Carlos to comment on this.

Let me actually do this the same way we've done it. Dwight, you were the primary reviewer for this. Do you have anything to add? And then I'll ask Carlos to comment.

DR. KLOTH: I'm going to disclose that I didn't get to, I did not complete my homework assignment.

CHAIR PENSON: You're a good man

to admit it and you are forgiven. You are smarter than me. Jay?

DR. SCHUKMAN: I just have a comment here on J codes. Particularly when you collect J codes. I mean there are a lot of administrative data limitations around J codes.

Because you're not going to capture all of them, particularly in hospital outpatient facilities. Because of the revenue codes. You just simply aren't going to capture those J codes. So I already see and issue with that here.

The other thing I noticed here is that despite the robust database that they had our there, you know, there were only 1,843 episodes that qualified, which I think is a limitation right there.

That's a fairly limited number of episodes that met the inclusion criteria. The other thing that's interesting, as you might imagine, the chemotherapy group was much

higher in cost that the non-chemotherapy group overall. So while the average was 65,000 they're way out there on the tails.

CHAIR PENSON: So those are really key points and that's right where we are now with this validity discussion. Because what you're basically saying is that capturing, we're not measuring what we think we're measuring. Or we're not doing a good job doing it.

So I think those are very important points. Let me just, before we, let me just ask Carlos to chime in on his thoughts on validity here.

MR. ALZOLA: Well, again, for validity the same thing I look at is whether the distribution of the cost along the different lines of service made sense.

And again, I think we found that what made sense that the chemotherapy group was, in the cost, was a lot higher than for the other group.

And within those lines, I think that, again, for the non-chemotherapy group the majority of the costs were attributed to the inpatient stays and the colectomy qualifying.

So again, in terms of tests what seemed clinically reasonable in terms of which are the components of the cost, that made sense to me.

CHAIR PENSON: Other comments on validity? Louise.

DR. WALTER: Just a quick thing about the vast majority of people who have colon cancer are over 65 and, again, this was not at all looked at in that population which I think is a big problem.

CHAIR PENSON: Yes, I think that's a very key point. That, you know, validity goes with generalizability and can you generalize and so I think that's a key point. Steven?

DR. CHEN: Two things on validity.

One and then full disclosure. I do have research contracts with some of the genomic assaying companies, but they're very tests and to the extent that they start getting you're probably going to want to track that. Because they're going to run about $3/$4,000 a pop.

The other thing though, I'm very worried about validity in the terms of what's a Simpson's paradox for people. Their case mix is going to highly determine where these patients fall in.

And so you could have someone who's cheaper on both the chemo patients and cheaper on the non-chemo patients and look worse overall.

And you could also be incenting poor care in the sense of you have someone who really looks expensive, they're Stage II, if you give them chemo you can kick them into the higher expensive group because you're stratifying based on chemo.

And so again it lends towards

promoting inappropriate care here.

CHAIR PENSON: So that's a problem we're always going to end up with here. I mean you've got to remember that it's, I think it's NQF's belief that these should not be used in isolation, that they need to be coupled with a quality measure down the road. You know because obviously sometimes cheaper care is worse care.

So we have to go on the assumption that there will be some sort of quality tied to it. But I think the points you raise are very valid and need to be considered. So other comments with validity testing?

So what I'm hearing, you know, we talked about some issues with 2b1 with the evidence but I'm really hearing concerns about validity.

I'm hearing concerns about generalizability. Jay's comments about whether or not we actually are capturing all the costs are really in the sweet spot for

this criteria.

And the comment that you have such a small number of players who are ultimately in there makes you wonder if perhaps something is being missed. I think Steven's comments are well taken as well.

So I am hearing some concerns here and I'm not sure they're completely addressable. Rohit.

DR. BORKER: One question. And this point of -- so this might apply generally to all the specific measures is the enrollment, the company's enrollment, that by definition these people have to be in the data system for a certain time.

So are we excluding patients who are more severe or have a more aggressive disease. They may not be metastatic when they are seen but they could have, you know, severe disease that they die and most likely to be more expensive. Are we excluding them? And it applies to this --

CHAIR PENSON: I think, you know, when you look at a one year period, without being able to control for stage, which I don't want to give this away, but to me that's the Achilles tendon here and it's not fixable in my opinion.

You know, but the fact of the matter is is that if you have a really advanced patient who has a colectomy, turns out they have a really advanced disease and they die three months later your efficiency measure is going to look great, because he didn't run up any costs for nine months.

So you're going to penalize a doctor either for his case mix or potentially even for being, you know, not giving a treatment when he should if you don't have a quality measure to go with.

As far as the continuous, two years continuous coverage, I think what we discussed earlier applies here as well. So I'm going to keep us moving because I am

getting hungry which is always a bad sign.

The next one was exclusions and there are some patients excluded. We talked a little bit about the colon versus colon and rectal and I don't think we have to visit that again.

For the most part I think people were okay with the exclusions in the preliminary. Although I think that may change now that we've had this discussion about colon tumors versus colon and rectal tumors. Do other people want to discuss exclusion criteria here?

I'm seeing no's. And Sally's rolling over here to say something to me.

MS. TURBYVILLE: Not necessarily great news. So we're still waiting for part of the lunch and it should arrive at 1:00.

CHAIR PENSON: Okay. That's perfect, we're going to tie it up perfectly. They're going to time it perfectly.

MS. TURBYVILLE: We don't want you

starving.

CHAIR PENSON: It's all good. It's all really good. As long as I keep drinking coffee we're in great shape. It's all right. I've done worse.

DR. WALTER: Actually I'll jump in one other quick thing. Because they did mention in their specifications that they wished they could have looked back more than a year to exclude people with colon cancer, because obviously you're still going to include people that had colon cancer two years ago, three years ago and is that a big problem.

CHAIR PENSON: I think that's a very reasonable point. Steven?

DR. CHEN: I would also throw out the possibility for them to consider excluding people who by the nature of their treatment appear to have metastatic disease.

For instance they get a liver resection within the year. That's someone who

has metastatic disease.

CHAIR PENSON: But you know the question with that becomes, you know, it's almost it's post-hoc, so someone has a colectomy and then have a liver resection three months later. But you know, you're looking backwards, it's hard to do that.

I mean I think in the end we're going to get into the risk assessment discussion and, simply put, it's hard to do this unless you control for stage. And I don't grade in colon cancer as well. But certainly stage. You know, node stage, nodal status and metastatic status is almost meaningless to me without that.

So I think on that note I'm going to keep us moving along because I think we need to have that discussion now. We're up to 2b4, which is the risk adjustment piece.

And basically to remind everyone here, do outcome measures have an evidence based risk adjustment strategy? Is it based

on patient clinical factors that influence the measured outcome?

Or they should have rationale for not having risk adjustment. And I'll start the discussion to repeat what I just said.

When I read this measure, without being able to control for stage of presentation, it's meaningless. It's just not fair.

And one could argue that well you're going to have, if you're going to accountability then you're going to have, Dr. Smith and Dr. Jones are going to have random assignment of metastatic patients that it's going to be the same across the board.

But John's shaking his head and I don't buy that either. I mean there are docs out there who attract the worst cases and you could easily risk adjust that and say SEER-Medicare.

But you can't do it in Market Scan, you don't have stage information. And

to me I see it as a fatal flaw and I'll open up to the floor.

DR. GILLIGAN: And the problem is here, Medicare is just only part of the country so that's not a national data.

CHAIR PENSON: I don't think anyone has used SEER-Medicare in this setting for public reporting. I'm not aware of that.

DR. WALTER: And from my reading they didn't do any risk adjustment, right. So I always laugh at measures that look at the same kind of care for people 18 to 85. I mean that seems a little --

CHAIR PENSON: So no, there is, I think there was an error in the submitted form as opposed to what's online. And Kevin and Todd jump in here. They used the same risk adjustment for this measure that they did for the colonoscopy measure. That's correct isn't it, Todd?

DR. LEE: The intent is to use the same measure. But because, I think, as you as

you all have noted, we had such a small sample size here our risk adjustment models would have been very questionable simply because we're only talking about just over 1,800 cases.

So our measure specifications are developed so that we will do further risk adjustment testing in additional populations. We don't have any risk adjustment data that was presented as part of this submission.

CHAIR PENSON: Okay. That helps. Because when you go online you talk about that HCC model again.

DR. LEE: Correct. And that was just to illustrate the process through which we will do risk adjustment for this episode.

We did not test it in the, or test our risk adjustment methodology in the Market Scan data because of the small sample size that we were dealing with.

CHAIR PENSON: Okay. So that clears things up, although it doesn't really

help. In that I think --

DR. WEISS: Kevin here real quick.

CHAIR PENSON: Go ahead, Kevin.

DR. WEISS: This is very helpful to us and again I really appreciate your input back to us as measure developers.

One of the things that kind of, that I'm trying to understand, for the experts in the room. To have a colectomy for colon cancer and not treat with anything else, is that the, because I heard that that concern was that we may not be capturing people who have nodes and other stuff.

Would that ever be not treated if it was beyond, you know, would they ever take someone with an advanced stage and not treat them?

CHAIR PENSON: Steven's shaking his head, I'll defer to him.

DR. CHEN: Yes, about a third of people how have Stage III colon cancer, which the consensus guideline is pretty clear that

they should all get chemo barring some sort of, you know, some other comorbidity that prevents them from doing so, about a third of people don't.

And then when you get to Stage IV, there are a fair number of people who may chose something else because they have overwhelming disease. They may chose hospice or no further treatment.

CHAIR PENSON: You know my problem with it is, just to jump in. I don't mean to interrupt you, John, but the problem is it's circular reasoning. Your outcome, which is cost, is also your, you know, an independent variable in a model which is controlling for stage. So that's not how you do an analysis in my mind.

You can't have your proxy for stage also be an outcome measure, it doesn't make any sense.

DR. GILLIGAN: Also it sets up a scenario, like you said, where you reward

people for giving poor care.

CHAIR PENSON: Yes, I mean to me, I'll say up front that I don't want to be too heavy handed but I'm a voting member as well.

That on the one hand I would say insufficient if I believed that the HCC model that you've put together for, well I shouldn't say you, I'm not speaking directly to you Kevin or Todd.

I just, I would feel comfortable if I felt that the HCC model ultimately could risk adjust here then I say insufficient evidence. But the fact of the matter is is that without stage I just don't think it's doable. So to me it's done before we even start.

DR. WEISS: That's very helpful to hear your reflections of that. The working group, of course, went through the same consideration of concern.

I mean I think they carried much of the same concerns and felt that this was

way to manage this. But that's actually very good to hear your reflections here.

CHAIR PENSON: I would defer to John and Steven. I'm a urologic oncologist so I'm probably talking out of school, but I suspect I'm not wrong.

DR. SKIBBER: You know that's why my first comment was that they need a definition of localized. Because the significant number of patients that present with metastatic disease there's no individual treatment for Stage IV disease. There are guidelines but those patients tend to have a variety of presentations.

Some of which require a colectomy, many of which, frankly, and this is supported in the literature, is that there's a significant number of patients that go on to have chemotherapy without colectomy.

And so until you get to the point where at least say it's Stages I through III versus a Stage IV patient you really have an

ill defined population.

DR. CHEN: Yes. So I think I would have less concern if say this was 60 day episode where I can say, you know, we'll have you do the surgery, whatever.

But once you start to get into how you treat them over the next 14/15 months it becomes a huge problem.

And if they can't use their own cancer registry or the NCDB or something like that you've got issues that I think, as you say, are basically unsolvable.

And the other thing is just to commend to you the idea that if you were looking at large regions you might be able to say well it washes away because it's okay.

But your average surgeon in the country does fewer than ten colon cancers a year. And so there's no ability to get a large enough sample unless you use their whole career.

CHAIR PENSON: We'll get into that

region issue this afternoon, with the breast measures. Tim.

DR. GILLIGAN: All right, I was just thinking the breast measures that we're going to talk about this afternoon, they stratify people based on which chemotherapy they got.

And that would be one, at least, step in the direction of trying to balance against this bias this talking about if we don't control first.

CHAIR PENSON: Yes and they do talk about seraphying by receipt of chem versus not. Although they don't do by type. And the question becomes do you think that's a reasonable proxy.

Other comments on risk adjustment? I've think we've beaten this one deep into the ground. Carlos.

MR. ALZOLA: Yes, one comment that I just didn't realize until today is that how many resources they incur depends a lot on how

long they live. So all those patients who die quickly, or soon, are, again, reflect a low amount of resource use only because they die.

So I think that should whatever risk adjustment approach they take in the end they need to consider that. I mean you can use some kind of sensoring method that are kinds for sensoring or some kind of exclusions but I think that's an important thing to consider.

CHAIR PENSON: I think that's a well taken point. It's really problematic, and even if you had stage you'd still have other issues as well. So again other issues with risk adjustment?

So I think we're all hearing a lot of concerns here. Why don't we do 2b1 through 2b4 now? Only because I think they tend to fall together.

So the first one, 2b1, is regarding, is validity and do the measures, are they consistent with the evidence. And

just again to summarize, I think there was some discussion here and some concerns about whether or not this fell in line.

There we go and so this one split with five people voting moderate, three people voting low and one person voting insufficient. And I think our comments certainly reflect that.

Next is 2b2, which is validity testing. That the testing demonstrates that the elements are correct and they measure the score correctly and they measure resource use correctly.

And there were a number of comments raised here as well. And here we had one person who voted moderate and eight people who voted low.

Next we'd go to 2b3 which were the exclusions. Any exclusions are supported by clinical evidence otherwise they're et cetera, all that stuff. Are the exclusions appropriate is basically the bottom line here.

CHAIR PENSON: Yes, it does doesn't it. And here we had six people who felt that the exclusions moderately met the criteria and three felt low, it did not.

And let's, finally, to risk adjustment. And obviously we had a long discussion about this. There we go. And so we had seven who said low and two who said insufficient. And I, again, think that reflects our discussion nicely.

Let's keep moving along, we're actually picking up a little time now which is good. We'll move on. We'll finish up the scientific piece. We're on to 2b5 which is the looking at differences.

So basically does the data analysis demonstrate that the methods for analysis and scoring allow for identification of statistically significant or clinically meaningful differences in performance or is there evidence that overall less that optimal performance.

And basically I actually wrote not applicable and I'm not exactly sure why I did that. I'll throw it open to the floor.

DR. WALTER: I mean I didn't think they addressed this. So I voted insufficient. But I never know whether it should be low or insufficient, but.

CHAIR PENSON: Go ahead, Steven.

DR. CHEN: I voted low last time for the same rationale I did for the previous one, which is to say I don't think this is fixable, as choosing that over insufficient.

CHAIR PENSON: Of the folks in the room. I mean basically, obviously you can generate statistically significant differences in resource use, but how that get interpreted, particularly in the absence of adequate risk adjustment is problematic. I don't know why I wrote not applicable.

DR. GILLIGAN: I think in the report, they just said that they weren't addressed. And I guess, I don't know, I put

not applicable too. I think that's probably what I was --

CHAIR PENSON: So I guess the question becomes probably not that it's not applicable but it just would then be insufficient. Or if you don't believe that this is going to make for meaningful differences you'd vote low.

Is there anyone who wants to speak on a positive note? So I mean I think I was in the same place with this. I don't think --

MS. TURBYVILLE: Let me see. Let me just make sure.

CHAIR PENSON: So it'll be the S12's. We're looking at 1584. It's a lot of paperwork.

MS. TURBYVILLE: It is.

CHAIR PENSON: Sure it is.

MS. TURBYVILLE: So here, presumably is a sample report.

CHAIR PENSON: That's the provider report, so that's --

MS. TURBYVILLE: That's not helpful. And then SA reliability and validity. It's fine if you didn't find it, I just want to make sure.

CHAIR PENSON: I know where I'll find it.

MS. TURBYVILLE: So starting with S11.6. I know, it's harder to think when you're actually, at least for me, navigating.

CHAIR PENSON: Here we go. So we're looking at S12's is basically it.

MS. TURBYVILLE: So if we had the sample report. Let me see.

CHAIR PENSON: So type and score, so here it is right here. It writes type of score is a ratio. They have a sample report, interpretation is an O to E.

And again if you're going to have an E you have to have risk adjustment that works. So at this point it remains to be seen if that's the case.

There is a detailed score

estimation of how it's put together. But again, the E depends on the risk adjustment model. And so basically it's just O to E ratios.

DR. WALTER: Isn't this also a fatal flaw in that they're trying to do these artificial windows of attributing costs to the oncologists and surgeon. Like somehow they hand off versus doing it at the same time. So I think that would also be a problem with this.

CHAIR PENSON: So basically what I'm hearing from the group based on this is that all we have is a sample report, an O to E ratio with a risk adjustment that we don't know anything about.

There are issues of accountability which will come up again, I think, in criteria 3, but probably are worth mentioning here as well.

That it may not be, it affects a meaningful ability to interpret the scores.

Other comments where this is concerned? Okay. We get now to multiple data sources. And just for the sake of time we're not combining data sources here.

This goes the same as before and I think we said not applicable before. So let's just sort of avoid that discussion.

The last discussion here would be in disparities in care. If disparities in care have been identified measure specifications, scoring analysis allow for identification of disparities through stratification of results.

This is by race, ethnicity, SES or gender. Or alternatively they justify why it's not there. The stratification we're seeing here is by receipt of chemotherapy.

And I think this does get back to were we were before. You can obviously stratify it by gender in the Market Scan data, but you can't stratify by race or SES.

And I think the question is is

this important? I would argue it's got some meaning and some importance. So certainly you can't do it, so other comments?

MS. TURBYVILLE: I think how we scored it last time.

DR. KLOTH: Just to repeat an earlier comment that KRAS testing is critical and if that's not well described.

CHAIR PENSON: Well yes. That's going to be, I don't know when they talk about disparities I don't think they're talking about stratification by KRAS testing but more population characteristics, like race, stage.

So I think we can probably go through 2b, finish up 2. Let's do differences in performance.

This is again, just to go through this, this is basically looking at are these differences measurable, statistically significant and clinically meaningful.

So let's go ahead and vote on

that.

Yes, I think everyone's getting hungry. Three more. So the majority of the people voted low, three people voted moderate, six people voted low. I think the comments reflect this. Specifically what I think Louise said.

Next is multiple data sources, this is not applicable so we can switch over this one. We have to do overall validity and then we'll do disparities.

So what is your overall gestalt for the validity testing, this includes risk adjustment, it includes the ability to catch what we're measuring, et cetera.

And I guess it does include stratification for disparities. That shouldn't be there, okay. So let's go ahead and vote on what we talked about.

So it's actually we have consensus. The panel feels low, everyone voted low validity. And unfortunately I think

that will be reflected by the steering committee with this measure.

Let's do disparities next. So we talked a little bit about disparities of care and stratifying by various groups.

So we actually have a spread here. One person voted high, two people voted moderate, three people voted low and three voted insufficient.

I'm wondering again, I don't mean to call people out, but for the person who voted high do you mind just saying why?

DR. BORKER: Sure this is one thing, the reason why I voted high is because even the fact that the database lacks race information or socioeconomic status, it's a limitation of the database.

But to me if that data is available there's nothing preventing this measure from using that data to stratify. That's the reason I --

CHAIR PENSON: Okay. So I think

that the message there is, if I can interpret that a little bit just for the comment, that basically much of what is there is acceptable to you and the pieces that are missing that is a flaw that should be addressed in future work but you can live with it as is?

DR. BORKER: Right. And this only applies to the stratification, not necessarily the validity part of it, so absolutely.

CHAIR PENSON: Okay. Terrific. Let's keep plowing along. I think we can do this part relatively quickly. I think we can be done by ten after one if I'm lucky today. So next we're going to talk about usability.

And I'm the primary reviewer here and I think many of the issues that we dealt with with the colonoscopy are going to be the same issues here.

Frankly what we end up here, the first one, 3a, is this reportable to public at large in the community reporting programs.

And at this point I felt it was

indeterminate. This is being tested through the RWJ contract and this is changing in NQF and I think, just like before, it's probably the same issues and you should vote more or less the same way you did for the breast biopsy. I think it's insufficient. But defer to others.

Okay. This is where it does get a little bit more interesting with 3b which is the performance results are meaningful, understandable and useful to intended audiences both for public reporting and quality improvement.

And here initially I thought it was sort of indeterminate, that there wasn't a lot of evidence there.

But now as I start to get a better understanding of what's going on with the risk adjustment and these O to E ratios I'm somewhat between low and indeterminate.

Because I have no evidence that the risk adjustment works. So in some

respects I should just leave it as if you gave me evidence I'd feel better. But then I know full well that the risk adjustment method is not going to control for stage and I think that'll kill it.

So in the end I don't know how we can interpret this because of those limitations. Comments?

And then 3c which is the clinical and construction logic. You know, this is where you have transparency and understanding and this is where I sort of got hung up on the accountability piece.

That how can you accrue costs in the first 42 days after surgery to the surgeon and everything after that to the medical oncologist?

As a surgeon if I get a parastomal hernia it's going to show up six months later and woe be my poor medical oncologist who's now going to pay for my poor technical skills. Not that she would ever blame me in a million

years because she loves me so much.

But the point is to me that's where the construction falls apart. Everything else I could live with. But that is where the accountability piece came in. Comments?

DR. SKIBBER: One small part of that is the other side of the coin, which is that their period for the accountability for the surgeon is 30 days before the colectomy.

And frankly a large part of the staging work up on a diagnosed patient may occur during that time, which really may or may not actually be under the control of the surgeon.

Those things are often patient comes with all their X-rays and whatever tests anybody else thinks they ought to have before they show up in the office. And I just think that speaks to the same issue you've brought up.

CHAIR PENSON: Other comments?

Steven.

DR. CHEN: I think going on further with that that also as we get back to rectal cancer there's a lot of neoadjuvant chemoradiation is happening.

CHAIR PENSON: So I think there are issues there. Other comments? So why don't we vote on the usability piece. Obviously harmonization is not applicable here. So let's start with useful to the public.

Are these results reported to the public at large, is there evidence that these are available. And again this is what we talked about before, they're with RWJ and NQF is still sort of all over the place here. Should I not say that for the public record? I apologize, I'm hungry, I'm punchy.

MS. TURBYVILLE: It's your interpretation.

CHAIR PENSON: The steering committee is on part of that, I just threw

myself under the bus. So we had one vote for moderate, one vote for low and seven for insufficient.

And now we'll move on to usability, 3b, did the information demonstrate that the results produced by the measure are meaningful, understandable and useful for information for QI and public reporting. Or if not was a credible rationale presented?

Let's go ahead and vote on that. And this one we have eight votes for low and one for insufficient. Next is for the clinical and construction logic.

The data and result detail are maintained such that resource use measure, including the clinical constructional logic for the defined unit of measure, are transparent and facilitate understanding. And here we have eight votes for low and one for insufficient.

And obviously we're passing on the harmonizing one. So let's quickly go through

feasibility and then we'll break for lunch. I'm hypoglycemic at this point, let's put it that way.

So basically with feasibility with 4a and 4b, just to refresh everyone's memory, this is more about administrative data claims and I would actually see if I can't push us through for acclimation for A and B.

So for A this uses administrative aid and the required data elements are routinely generated and used during care delivery. So blood pressure, lab tests, diagnosis and medication order.

So for things like genetic testing and KRAS testing, I think it would be captured in the administrative data set, is that a fair statement? Steven? Dwight?

If it gets paid for it would be captured, right?

DR. CHEN: Sorry. The genomic information almost certainly would be because it's a significant bill. I don't know if KRAS

would show as specifically KRAS or just some sort of extra staining, depending whether they do it by dish or not.

DR. KLOTH: I think that's an appropriate concern. Or if it was done gratis for some reason by a lab and not billed.

CHAIR PENSON: Well I think anything that's done gratis that doesn't get billed we miss. I mean that's a limitation in administrative data. But it will get captured if they want to get paid, correct?

DR. KLOTH: I would tend to think so. And I'm sorry for continuing to harp on this. But KRAS is such a critical issue. When cetuximab first came out and then panitumumab.

Initially it was that EGFR testing was critical. And then we realized, and it was presented ASCO 2008, that what really, really makes a difference is KRAS testing. And that is just so fundamental.

CHAIR PENSON: Well, but we're

capturing, I understand your concerns and I think we've discussed it in the risk adjustment piece. But the question here is do we capture if it's billed as a resource use. And I think the answer is yes from what you're telling me.

DR. KLOTH: Well if their database capture technique is all encompassing of everything that would ever be billing for that patient from time point A to time point B then presumably the answer would be yes.

CHAIR PENSON: And I think we have to go on that assumption. Because it's all about the Benjamins here. So is it fair, would everyone agree that this is probably high?

DR. KLOTH: Sorry, me again. There is one other thing to consider. They may or may not capture variable drug costs if they're primarily at J codes. Because J codes is a function of a CMS billing unit.

It's not a function of what the

hospital or health care provider paid for Leucovorin, generic, versus levoleucovorin or if they had to switch because of shortages from one brand of a drug to another brand, et cetera.

CHAIR PENSON: And I think we've sort of captured that in the validity discussion. The question is if they got some sort of agent, whether it was generic or brand name, is that captured in the administrative data? For commercial --

DR. KLOTH: Well it wouldn't be reflected in the J code. Because J code is what Medicare is willing to pay.

CHAIR PENSON: Okay, so why don't we vote on it then. I think that's probably the best way to do this. Let's vote on 4a, are the required data elements routinely generated and used during care delivery. So we had six people who voted high and three people who voted moderate.

Okay. On to 4b, which is the

required data elements are available in the electronic health records. And if they're not in the electronic health records there's a credible near-term path. So I think our earlier discussion applies here as well.

We can just vote on this as well. And here we have eight for high and one for moderate. Now let's just have a relatively quick discussion about 4c and 4d and then we can all eat.

So 4c is about errors and inaccuracies. Susceptibility to inaccuracies and errors and unintended consequences related to measurement are judged to be inconsequential or can be minimized.

And I think this is an area where potentially you could factor in these issues with J codes and different medications. Discussion before we vote?

All right, I think we can have a vote then. Hunger is an incredible motivator isn't it. So here we had four people who

voted moderate and five people who voted low.

And I think that the low votes here, for the folks on the phone, reflect people's concerns as Dwight raised and Jay raised and Steven raised about J codes and whether or not that there can be errors in there that'll lead to incorrect values.

So the last one is 4d. The data collection and measurement strategy can be implemented as demonstrated or testing did not identify barriers to operational use.

And basically in looking at this I think that if nothing else their certainly demonstrating they can collect the data and that they can measure things fairly well.

Other comments? Okay. Not hearing any let's vote on this. So here we had four people who said high, four people who said moderate and one person who said low.

I will ask the person who said low just if they can provide some rationale for the NQF team to feedback to the measure

developers. All right, now who voted low? Did someone vote low by accident? You want to try that again?

Let's try that again. Keep voting now. It's like Chicago, vote early, vote often. There we go. All right, so four voted high, five voted moderate. I think we have the summary of feasibility one and then we can take a break.

That's it? Okay. So why don't we do this. Let's take ten minutes to get our food, get ourselves comfy and then we'll start working again. Before we do that we have to ask if there's anyone from the public on the phone.

MS. TURBYVILLE: Operator, if you could open the line and see if anyone from the public line has questions or input for the technical advisory panel.

OPERATOR: And that is star, one for any public comment at this time. We have no public comment at this time.

MS. TURBYVILLE: Does anyone in the audience have questions or input for the panel at this time?

CHAIR PENSON: Okay. So let's go get some lunch and start again about 1:20, 1:25. And we'll work while we eat.

(Whereupon, the meeting went off the record for lunch at 1:14 p.m. and resumed the meeting at 1:40 p.m.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

1:40 p.m.

CHAIR PENSON: Probably get started. But my hope is that we are running about a half hour behind. But my hope is that the breast cancer issues will be a little bit easier.

The reason I think that they may be somewhat easier is that basically with these measures that one of the major differences between these measures and the morning's measures is that there is no risk adjustment whatsoever.

And so that may or may not be acceptable to you as individual panel member. Because remember that these are not going to be accountable by provider.

They're going to be reported out by region and so in this setting the rationale is that because we're doing this by region and not individual provider it will wash out.

And you may or may not think

that's true. So before we start on the measures I'm going to invite the folks from the ABMS Foundation to say a couple words about the two breast cancer measures.

We're going to start with the episode of care for the newly diagnosed cases. And then we'll get into the biopsies. So Kevin or Todd if you want to say a word or two about these measures that would be great.

DR. LEE: Sure, this is Todd. If it's okay I'll start with the biopsy one first just because it makes for me in talking about them sequentially. So the focus of the biopsy measure is to look at resource use in the 60 day period preceding a breast biopsy.

The Work Group felt that there was going to be substantial variability in the diagnostic work-up of a woman who had screened positive leading up to the biopsy for determining whether or not the person has breast cancer, and wanted to capture the resource use in this 60 day period preceding

the advent.

I should not that we also then go seven days beyond the biopsy event to capture the resource use that might happen because of claims lag and so forth.

But essentially the workgroup felt that there was going to be enough variability regionally in the actual biopsy procedures that were done and the diagnostic testing leading up to the diagnosis that would provide enough variability and resource use to look at amongst this episode. If that's okay I'll move to our breast cancer treatment measure?

CHAIR PENSON: Go ahead.

DR. LEE: Okay. So our breast cancer treatment measure looks at newly diagnosed cases of breast cancer over a 15 month period, that's following their diagnosis.

I should say in the title we say it's a 15 month period but it's actually a year and a half because we go backward in time

for three months to look at resource use leading up to the diagnosis and then 15 months following the initial diagnosis.

We use an algorithm that's been validated in the SEER-Medicare data to identify new diagnosis for finding our breast cancer cases.

We then use these patients and follow them forward, again, for 15 months capturing breast cancer related resource use, stratify the population into four groups, those that receive chemotherapy with trastuzumab.

Those that receive but chemotherapy but don't use trastuzumab. No chemo and then a neoadjuvant chemo group and look at what our workgroup defined as breast cancer related resource use during that period.

I'll make the final note that all of this, as you noted, is reported at the regional level. Our workgroups felt that it

was not necessary to use risk adjustment following, because of the regional measure and we weren't trying to attribute this at the physician level.

Largely because of the concerns about not being able to measure stage that you all talked about for our colon cancer measures. And I'll stop there and listen to you guys and answer any questions that you may have.

CHAIR PENSON: Thanks, Todd. So we're going to dive into it here. The order is the newly diagnosed breast cancer episodic care for a case of newly diagnosed breast cancer.

Do people want to do that one first? Or would people feel better starting with the biopsies first? Or if no one cares we can just keep to the agenda.

I don't see anyone really caring. So that's good, that's a good sign. Sally I think we've broken them.

MS. TURBYVILLE: You weren't supposed to say that out loud.

CHAIR PENSON: It's okay, it's so darn obvious at this point. Okay. So let's start with the episodes of care for treatment as, then that's 1579.

And so basically let's sort of go through the importance first and then we'll get into number two. And so I'll sort of lead the discussion here.

And I think that these are, I think, for the most part looking at this there was fairly high agreement here with regard to importance. Although not 100 percent by any stretch.

With 1a, almost across the board looking at whether or not this addresses a health goal priority by DHHS or the National Priorities Partners.

I think everyone sort of felt this was a high impact condition. Breast cancer is really common. There are a lot of costs

associated with it. Anyone want to add to that?

DR. WALTER: I only voted for moderate because I actually wasn't sure how important it was to look at regional levels. Especially when I wasn't really sure what regions they were looking at.

CHAIR PENSON: And I think that's a good point. I don't know where exactly that sort of comes in here. Whether it's the impact level or the opportunity for improvement or maybe it's usability and usefulness to the public. I mean what does public, even for accountability reporting issues.

How do you interpret what goes on at a regional level. And so I think that's a reasonable issue. I think we can certainly, I think we will visit that again later. But I think beyond that I'm hearing consensus that this is probably high impact.

Opportunity for improvement.

Again this is where there is demonstration of problems with resource use or cost and an opportunity for improvement. Specifically variation, delivery of care, cross providers and population groups.

I think for the most part again people thought this was a fairly high, obviously people know about geographic variation from the Dartmouth Atlas and other places.

Although one or two people voted it moderate to low, so I'll open it up to the floor and see if people still feel that way and if so why.

DR. GILLIGAN: I just, are we on 1b, am I --

CHAIR PENSON: Yes, we're on 1b.

DR. GILLIGAN: I just think the whole concept behind these measure that we can improve spending by looking at variations in spending, to me it's a hypothesis. I haven't seen a lot of evidence, honestly, to support

that. So I think that's why I put moderate down.

That I think it'd be nice to see more evidence supporting that. Because it's not, I mean in some way you could argue that the huge problem in cancer right now is that we have all these really, really expensive drugs and expensive imaging technologies. And they're approved for use and they're in the guidelines.

And if you practice the guidelines it's extraordinarily expensive if you provide standard of care. And that that's where the crisis is then variations in care are less important than the fact that we have these hugely expensive treatments and no one's controlling the costs of them.

CHAIR PENSON: Yes. Tim, I don't disagree with you but that's a bigger issue. I mean I think in the end that whether or not you agree with this or not remains to be seen but, you know, people will look and say we

probably shouldn't have huge variations between say California and Tennessee and Indiana, et cetera, that it should more or less be the same.

So I think that that's what the criteria sort of drives that, at least when you think about geography. Now you may be right that in fact the real problem is overuse or just expensive costs overall. But I think the question becomes is, you know, is there room for improvement.

Could it be that this measure could look at that and say for expensive things across the board it can be used to lower cost.

Can it be used to look at variation? So I think what you're saying is true but I think the measure addresses that. Other comments?

DR. CHEN: The only other thing, and I had marked this high initially but the more I think about it the more I think it's

more of a moderate because the issue in opportunity is well, once you report it someone has to try to act on hey, there's difference.

And the question is who's supposed to act on it if you report a really big region. I have no independent ability to fix it.

CHAIR PENSON: See and that to me is a key point. I agree with that.

DR. WALTER: Yes because, I mean, there's like cancer centers are in certain regions and you'd think they were going to have higher costs that maybe someplace in, I don't know, Wyoming, which doesn't have a -- It doesn't mean they're doing worse care.

CHAIR PENSON: Okay. So I'm hearing some concerns but I'm also hearing to some degree that they're moderate concerns on something to consider. Other comments with regard to that, if any?

So I'll move on to 1c, which is

the purpose and objective of the resource use measure, including its components. And the construct are clearly described.

And again I think here most people felt that this was relatively straight forward, that the constructs were there and the purpose was there.

There were one or two people who went moderate and even one person who went low. And anyone want to comment on that?

DR. WALTER: Again, I guess this is my, there just wasn't a lot of hypothesis about what they were expecting that was going to be driving some of this. And I guess I would have loved a little bit more of that.

DR. GILLIGAN: Yes, I had the concern. It seemed like a fishing expedition to me.

CHAIR PENSON: Well, we're sounding more like an NIH study section than a NQF TAP. I mean I think in the end what you have to look at is not, I don't think it's,

this is my opinion. But I don't think it's incumbent on the measure developers, see we can't think of them as researchers, as measure developers, to give us their sort of research questions.

But I think that it's helpful for us to see how the measure would be used. That gives it purpose. So the question becomes, is clearly they haven't give that to us.

That's all right, we can either ask them on the phone or we could say, given what I see here I could see a purpose if I was a researcher myself.

So the question is are you okay with that, Louise and Tim, or should we ask Kevin and Todd?

DR. WALTER: I guess I'd like to hear from them, because I'm still rating it moderate.

CHAIR PENSON: All right, so Kevin and Todd, I know you guys are listening to the discussion with regard to purpose here. And

I think people want to know given that you're going to measure resource use on a regional level for breast cancer care.

What are some of your thoughts on what it's going to show and what are some of the actionable items when it's done on a regional level?

DR. LEE: Yes, I think the actionable information may be a bit more difficult for at least -- This is Todd, sorry, for me to respond to. I'm not sure if we measure this. And I'll sort of give some foreshadowing to how we operationalize a region.

We've done it in four geographical regions in the U.S. as well as by state. And so if one state is higher cost than another state, what impetus or what actionable information did that provide and who makes it happen?

I think that's a little unclear from how we've designed the specifications on

how that would actually happen and potentially there's no entity that would exist right now to use this information.

However, we do feel, and our workgroup felt, that there was important opportunities to evaluate variability in care within these regions.

Because if we can identify differences that do exist, this may actually lead to some important hypotheses rather than us knowing why there might be variability.

For example, some of our workgroup felt that maybe there are differences in number of follow-up visits and intensity of follow-up visits for patients across different regions. Or the types of services that were used.

And I think all of this underlies what we are trying to do is better understand this variability. Whether or not it exists regionally. And then hopefully eventually act upon it. So I don't know if we have a lot of

A priority hypotheses going.

And we were trying to develop this to understand whether or not variability exists and then how to subsequently change that.

And I guess what I'm hearing from you all is that there may be issues on how that happens if we measure this at the regional level rather than doing it within a health plan or within a cancer center or some other entity.

CHAIR PENSON: Well let me add to that and maybe sort of help you a little bit here as I think about this. You know I think when you're looking at big regions, cutting the country into four corners, may not be that helpful.

But as you get a little more granular, I think there's some there there, because when you start to break it down by category of use, whether it's imaging, inpatient, outpatient, there may be something

there. There's no doubt there's going to be variation, I think, in resource utilization.

I just have to open up the Dartmouth Atlas and chose any condition you want and we can go from there.

But I think that there's, to me when I look at these measures, I was a little skeptical but I started to think well what it's going to tell me is that maybe the far west does more imaging whereas the south does more procedures.

And that has some value. And that may even be actionable. Because it may be that the care in the west coast imaging use is appropriate or may be overuse that's inappropriate. So I think there's something there, is my impression.

DR. WALTER: But how will you know what's appropriate or not appropriate variation?

CHAIR PENSON: That's a well taken point. But how do we ever know? I mean how

do we know it's appropriate in Dartmouth Atlas? We want to minimize variation between areas but we just don't know.

DR. LEE: I would also argue that's why these eventually need to be partnered with quality measures so that we can look at efficiency.

CHAIR PENSON: Yes, I agree with you, Todd. Other comments from the panel members regarding purpose and objective? So what I'll do before, Louis is getting his vote.

Don't vote yet, don't vote yet. We have to do the discussion about resource use service categories.

So this one has something like ten or 11 categories. And the question is are these consistent with and representative of the conceptual construct, are they appropriate? And certainly my impression was they were, they are fairly comprehensive.

And I think everyone in the room

agreed, preliminary of them voted high. Any comments, any changes to your thoughts on that? Okay. So now you can vote, Dr. Potters. So let's go through these each.

We'll start with 1a, which is high impact. Does the measure focus address a specific health national health goal priority or was data submitted to demonstrate a high impact aspect of health care. We have all nine. And this one eight people said high, one person said moderate.

Next is 1b, which is demonstration of resource use or cost problems and opportunity for improvement. So are there data to support this or at least in your clinical intuition does it support that?

DR. LEE: We can use our spidey sense?

CHAIR PENSON: Use your spidey sense, it's okay. It's a new field, so I keep telling myself. So one person said high, everyone else, all eight, said moderate.

1c, speaks to the purpose and the objective of the resource use measure. Is it clearly described? And we have consensus there, everyone in the room voted moderate.

And then last one, 1d, resource use service categories. Are the service categories included consistent with and representative of the conceptual construct represented by the measure? And here we have seven people said high and two people said moderate.

And we have a summary on this one. All right. So now let's move on to the scientific piece of this on acceptability. So for 2a, which for 1579, the primary reviewer was Steven.

(Off microphone discussion.)

CHAIR PENSON: Okay, well. Whichever one you want, go ahead.

(Off microphone discussion.)

DR. CHEN: I think in general these were reasonably well defined. I did

have some minor issues but it's the one on algorithm. I think the one thing that I have concern is that when you do look at administrative data in this circumstance sometimes people will continue to use the same ICD-9 for this as breast cancer.

Even though they had the breast cancer treated and they're in the surveillance mode and then they get a biopsy and it might come back as not cancer but that code in that association might cause unnecessary inclusion. So that's one.

The other thing was that somewhere in the exclusions it said that they wanted to exclude people who had lymph node disease, but on the other hand the measure doesn't actually say that they want to exclude lymph node disease.

I think they were looking go exclude lymphomas, but sometimes people will use that code to indicate that there's lymph node disease. And that's on page, I don't

know what page that is, I forgot to write it down.

MS. WILBON: I'm sorry I just need to interrupt really quickly. Are we talking about 1578, the breast biopsy, or 1579?

CHAIR PENSON: 1579.

MS. WILBON: Okay. Thank you.

DR. CHEN: But as far as 2a is concerned I think in general it's pretty precisely specified with those minor caveats.

CHAIR PENSON: Okay.

DR. GILLIGAN: There's one thing that I wanted to clarify and some people have brought up in their comments as well. In terms of the definition they have these high-risk and non-high-risk, definitions of what qualifies as breast cancer.

And at one point having a cancer other than breast cancer qualifies you in the non-high-risk breast cancer group, but that's also an exclusion criteria.

And so we have a criteria that is

both an inclusion and an exclusion criteria. And I was wondering if maybe the authors can clarify how to interpret this non-high-risk group?

CHAIR PENSON: Todd or Kevin, any thoughts on that?

DR. LEE: So if you look through the non-high likelihood cases, under the algorithm from Natinger, those patients, in most cases, do not get included as incident cases of breast cancer.

They immediately lump to this group that's not likely to get in as you work through the algorithm. And they do explicitly become excluded as part of our exclusion criteria.

DR. GILLIGAN: So on page 15 where we have those three categories, why is other cancer included there?

DR. LEE: So that is a part of the algorithm for identifying incident cases. It actually ends up being part of the algorithm

where if they have a diagnosis for another cancer it may not be that they are a high likelihood for an incident case of breast cancer.

So if you continue to work through he algorithm, those patients are not likely to get into the cohort.

DR. GILLIGAN: All right, but since that's an exclusion criteria it should be guaranteed that they don't, shouldn't it?

DR. LEE: It absolutely is. So that happens in the next step of our coding. So they, if one of those persons would slip through this algorithm they would be excluded explicitly because of the exclusion criteria that you see in the next step section down.

DR. BORKER: I had the same actually, the same comment. It's like if you plan to exclude them why even have them with inclusion criteria?

DR. LEE: Well this is part of us trying to apply a validated algorithm for

identification of incident cases.

DR. BORKER: Right, I understand that but then the validity data is presented on the original algorithm, then in that algorithm any other cancer is part of the non high likelihood cases.

So by having the additional exclusion criteria in your measure aren't you fiddling with that validity measure?

DR. LEE: I got you, I see what you're saying. Why have the other non-cancers if we're already getting rid of them as a non-high likelihood case. I can't answer the question to the extent that we actually ended up excluding additional people.

DR. BORKER: Okay.

CHAIR PENSON: So let me just pile on here. But it's hard for me. So how do you actually identify the data diagnosis when you get the high likelihood case? Are you able to do that, do you think?

DR. LEE: Well, yes. Sure we can

do it. But how accurate is it? I don't have a sense of how accurate the actual date of diagnosis then becomes.

We identify the date based on the information in the claims data so I'm not sure, I don't have a good sense of the accuracy of that date.

Nor do I know if there's been, or am I aware of any validation efforts to find whether or not that date is the date of diagnosis.

CHAIR PENSON: Okay. Other comments about this first criteria, 2a1, the measure being well defined and precisely specified? What I'm hearing from the group is for the most part people sound relatively comfortable with it.

But people are raising some concerns about who is included and who isn't and some of the reasoning behind the algorithm and the date of diagnosis.

And while the algorithm had been

tested in other settings it's sort of new here so I think that's what I'm hearing. Okay. With that then let's move on to -- and thanks again, Todd, that's actually really helpful to us here.

Let's talk about 2b1, pardon me, did I skip one. It must be getting late in the day. Yes, sorry. 2a2, which is reliability testing.

And we go through this a fair amount. Testing demonstrates that the results are repeatable and produce the same results with a high proportion of the time. And the reviewer for this was Rohit.

DR. BORKER: So I'll just kind of, I looked at, obviously, my comments and the comments from other reviewers here. So in terms of strength obviously it is a process that definitely adds value to the reliability of the measure.

In terms of some of the limitations there was a concern that without

controlling for the stage of the disease or the disease characteristics it's kind of difficult to interpret outcomes.

Then in terms of the data elements there's no attempt to formally look for the liability statistics, such as interrelated reliability because a lot of these diagnosis and outcomes are based on ICD-9 and DRG codes so there's no formal analysis on that.

Again, this has been brought in other measures as well as the reproducibility of this measure hasn't yet been demonstrated in another database, another commercial database to be more specific.

Complex programming has been kind of highlighted as one of the issues that could cause some reliability issues when you're trying to, you know, reproduce the same analysis in another database. So that's the general highlight of the comments on 2a2.

CHAIR PENSON: But accepting Sally's comment before that it would only be

endorsed for use in a commercial payer database, like the ones here.

Do you feel that the algorithm, which it strikes me as somewhat complex programming as well, is going to be able to produce the same results in subgroups of that Market Scan data set, or whichever?

DR. BORKER: Sure so let me answer your first question first, which is even within commercial the Market Scan database, and again, based on what the steward has submitted, is a little bit different than a lot of other commercial database in the sense that it has a longer duration of follow-up.

It's got an employer base so even if the patient changes insurance they'll still track. In other databases if you change your insurance you'll lose that data.

So the 15 month guidelines and 15 month follow-up, this could have a little bit different implications on other databases. So that's that concern. The second was, I'm

forgetting the question now.

CHAIR PENSON: The question really was regarding the programming and the consistency of the results.

DR. BORKER: Right. So as I think another panel member mentioned earlier, you know, on paper it looks fine but when actually somebody is going to go and start programming it that's when the rubber is going to meet the road.

CHAIR PENSON: So looking at Carlos' comment with the reducibility he had some concerns. I don't think he raised the concerns you raised, Rohit, which is reasonable, the question I have for you is do you feel that it's a major limitation that you're not convinced, that it's just missing data, or it's a minor limitation?

DR. BORKER: I would say it's not a major limitation by any means. It's a minor, it's a remedial issue, but nothing that cannot be handled by modifications. Yes.

CHAIR PENSON: Okay. Other comments from the panel? Okay. So I think at that point why don't we -- I'm sorry, go ahead.

MS. TURBYVILLE: And this might be something that gets reexamined at the time of maintenance. So once it's implemented amongst other commercial pairs are they encountering issues with the specifications being able to be followed, et cetera.

And NQF does ask for additional information at that time. So hopefully in the future we'll have more information.

CHAIR PENSON: Okay. So why don't we, any other comments about 2a2? So what I'm hearing, any other comments before? So what I'm hearing, just to reiterate here is that it does seem to be relatively reducible and reliable in this data set.

But there are some minor concerns and certainly it's something that could be the focus of further testing. So having

summarized that let's do 2a1, which is, is the measure precisely specified so it can be implemented consistently. And the results are four for high and five for moderate.

And we'll do 2a2 next, which is the reliability piece. Does the reliability testing demonstrate the results are repeatable, produce the same results a high proportion of the time when assessed in the same population, the same time period and that the score is precise. And here we have consensus, everyone in the room felt moderate. Excellent.

Okay. You see I always, you see you got to throw something at me. Don't let that computer restart now, whatever you do.

So let's get the overall gestalt here, what is the level of overall reliability testing. And here we have one person voted high and everyone else, the remaining eight, voted moderate.

Okay. So now we'll move on to 2b

and do these basically four pieces together and vote. So 2b1 is the item regarding the evidence.

Basically, the measure specifications are consistent with the evidence presented to support the focus of measurement, and it is specified to capture the most inclusive target population, both Steven and Tim got to do this, so gentlemen.

DR. CHEN: So for me this was a moderate. Again, the precision is basically there I have big concerns, and this bleeds into the validity issue, of despite the fact that it's population based that there's still no risk adjustment at all.

And we know that regional comorbidity burden is different. And that's going to play a part in lengths of stay, response to chemo, these are toxic agents that we're giving people and what their comorbidities are.

On top of that I think stages is

important, less important so at the population level, but still important because we do see regional variation in just stage of diagnosis. And then finally, they stratify based on chemo, no chemo, new adjuvant and with chemo trastuzumab or not.

But what that eliminates is one of the key elements in variation in cost, which is the decision making to give chemo or not.

And that is, I understand why they did it because they don't have stage data, but again, at least that's circular thinking. Of, you know, you're cheap because you don't do this, so.

DR. GILLIGAN: So actually I thought in some ways it was the opposite. Because if you're only comparing people who give trastuzumab to other people who give trastuzumab, I think you're then better able to pick up variations in practice, other than the decision to give trastuzumab.

Because that's such a huge

proportion of the costs, if you measure the difference between just trastuzumab or not then a lot of the other differences get washed out.

CHAIR PENSON: So I think we're now getting into what I suspected we would get into here. Which is we're going to break up into Protestants and Catholics in a minute, says the Jewish kid from New York.

But the fact of the matter is you're either going to buy that it's okay to have minimal or no risk adjustment and look at it on a regional level or not.

And it's, you know, I'm not sure if this is going to be discussed in the validity piece in 2b2 and the risk adjustment piece in 2b4 or if it's appropriate to discuss it here.

I would maybe say that we kind of wait, kind of put it all into 2b2 and 2b4. I mean basically what you're saying is do the specifications capture the patients in such a

way that it lets them get at looking at the resource use around treatment? And if you say no, I'm okay, but I'm just trying to sort of focus it a little bit.

DR. CHEN: And I think my answer to that is generally yes with the caveats I gave before. I think you are including some people who don't actually have cancer and you are excluding some people who have known positive cancer because you're treating them as metastatic. And those aren't metastatic.

CHAIR PENSON: Okay. Other comments? Okay. So we'll move on to 2b2 and I think that this is probably going to be the time that we need to think about this, the way this measure is structured.

And basically this is, validity testing demonstrates that the data elements are correct and they correctly reflect the cost of care or resources provided, adequately distinguishing higher and lower cost and resource use.

So I think that last line is where you really start to say, you know, does it adequately distinguish things if you're using a regional level?

And again, this is, actually it's not. Who is the primary reviewer for this? Rohit, you are.

DR. BORKER: So I give moderate to this. Because I think a lot of things that I'm going to say can be resolved to appropriate statistical techniques.

But again, in terms of strength, I think they have done a reasonable good job with that iterative process which inherently kind of increases the face validity of the measure.

But then there are a lot of concerns and some of these concerns have been raised before. Again, there has not been a formal analysis of like a known groups testing or convergent discriminate validity in terms of whether certain groups which are expected

to cost higher, are they really costing higher.

So no formal analysis has been presented so it's kind of difficult to know whether the measure is really valid. Then the implications of that 15 month requirement hasn't been discussed. Again this is the same issue that we had raised earlier.

Are excluding more severe cases? It could be in that time frame be very expensive but overall may reduce the cost because they don't cost anything, once the patient dies.

Again there's this issue with high likelihood and not high likelihood cases. And although the algorithm has been validated the implications of these two groups hasn't been studied so it would be nice to see cost and resource use among patients who fell under the high likelihood cases versus the non-high likelihood cases.

In terms of other threats to

validity hasn't been, again, tested in another database. This was raised for reliability, the same thing applies for validity as well.

So I think another issue that was raised was the differences that we are likely to see in cost and resource use we are not sure whether it is because differences in the patient mix or is because of difference in quality of care that is provided to the patients? So that's just the summary.

CHAIR PENSON: And that's a good summary, thank you. I mean I'm sort of, and I should add that Carlos isn't here, but he had some minor to moderate points with regard to validity testing that he raised that I think are worth at least reviewing and considering. Other thoughts in the room?

DR. WALTER: Just again that breast cancer is most common in older people, so again I think that's another validity issue.

CHAIR PENSON: You know on the one

hand looking at this it sort of, you know, it had face validity, it passed the smell test.

But when you start to get sort of under the cover and look into it you see a lot of these other smaller problems. For me I'm able to, I'm okay with it.

I'll say up-front, I'm okay with this regionalization business and I think when you regionalize to some degree it does take care of some of the risk adjustment.

But I don't think it takes care of all of it. But a least from a face validity standpoint I'm okay with this.

It's not generalizable to older women, there are other issues with regard to the algorithm, but I saw them as minor problems when I was reviewing it. I'm not sure about the risk adjustment thing. So we can come to that in a minute.

DR. CHEN: I will mention as far as regionalization I'm kind of more okay with the four corners thing than I am the state by

state thing. Because there are a lot of border cities where their MSA is much more relevant than what state they actually live in.

CHAIR PENSON: That's a very good point. So other comments specifically about 2b2?

DR. GILLIGAN: Does anyone ever do chart review as validity testing for database? Or is that just impossible to do?

CHAIR PENSON: It's not impossible to do and if you think that would be helpful I'm sure that Kevin and Todd on the phone are listening and will take that into account. I mean it's expensive, it's not easy, but I think that it is, it's often lost.

I mean without going into it I think a lot of the, throughout the quality improvement movement we've often failed to do many of the validation studies we need to do. Either because they're hard to do, they're expensive, there is political expediency to

getting things done.

And I think it's a very well taken point, Tim, that while these are all reasonable things, if not now certainly at the renewal point, one would hope that there will be some real validity testing.

DR. SKIBBER: Interesting that you mention that. I think there's been the development of a number of extracted databases that are coming out now that may be very helpful in doing that sort of thing.

I think one is maybe not specifically for this issue, but the NSQIP database by the College of Surgeons as well as the NCCN for specific disease entities.

And their breast database is immense and that's pretty granular data that's reasonably reliable. At least in certain disease entities. And the NSQIP is also abstracted data, so that may be helpful at some point.

CHAIR PENSON: Yes, I mean you

really could do some studies going with NSQIP and NCvB, you know STS has their data set there are ways to do this and it just hasn't been done.

DR. POTTERS: The problem is you can't crosswalk those to the claims data.

CHAIR PENSON: Not necessarily. You could if you had identifiers and everyone signed on the bottom line. Those are big ifs though for sure.

DR. CHEN: To that point I mean I did write an abstract and the paper did eventually generate where we actually linked cost accounting data to our NSQIP data. And so it can be done.

It's not as onerous as you might think. It was onerous the first 1,000 charts when I hand checked them. But after that I had made an algorithm that is sort of done with a 98/99 percent accuracy.

CHAIR PENSON: I'll focus us again. Because as much as I enjoy this

discussion there's another comment that I have which I'm squashing, stop, just for the sake of time. Yes, yes. I think that that's actually something that's is worth discussing at a higher level on NQF side. Steven.

DR. CHEN: I did want to respond to your question as far as risk adjustment. So while we're sitting here I apologize I haven't been able to collate them, but I just took 2007 SEER at the 17 SEER registries to look at their stage variation and you can have up to 2X variation, or 3X in the various stages between SEER regions.

CHAIR PENSON: And that to me, thank you for doing that because that's an important point. When I looked at this on the one hand I was okay with the regionalization. On the other hand there's a part of me that keeps saying even when you have big regions there can be differences.

And so I think that we need to discuss that, it may be a little more involved

even than just saying I buy it or not. But I'll ask you to table that for just a few more minutes. Other comments with regard to validity testing?

All right, so let's move on to exclusions. And we did have a discussion about the algorithm and about excluding other cancers. So we probably don't need to revisit that. So, Louis, you were the person looking at exclusions, any other comments?

DR. POTTERS: Yes, there's really not that much, right. There are really no exclusions in that it's a population based measure.

The exclusions that are there, you know, we talked about are either the insurance based and clinical. That gets you down to about 24 percent of the total population but they certainly seem reasonable.

Except for that oxymoron on the second cancer word sort of included and excluded but we talked about that. So it just

sort of is, it's really no opinion about it either way.

CHAIR PENSON: So what I'm hearing is that from what you're saying, there's some issues there that either are very minor or at most moderate. Is that a fair statement? Other things to add? Okay.

So now let's have the discussion that I think all these other measures get bogged down in. Well, that's appropriate though. Which is looking at the risk adjustment piece.

The question is, is an evidence based risk adjustment strategy specified, are patient clinical factors included, or if there's no risk adjustment, which is the case here, does the rationale or data support that.

So the question at hand is, and remember the stratification is not risk adjustment, is it okay not to have risk adjustment here? And I heard Steven's comment that really, I think, gives pause. I don't

know.

DR. WALTER: I guess I'm wavering between low and insufficient evidence. I think it depends on what the region is, because I haven't heard exactly what the region is.

Or, you know, it'd be nice to have the data that Steven provided about the variation and stage or other things across whatever region they're thinking about doing.

MS. TURBYVILLE: And that's a challenge because when we asked for unit or level of analysis from the developers we just had population, region and national. We didn't ask them to specify.

So I don't know how we would, I guess it would have to be if they have guidance on that it would it have to be part of the written text? Or does that even make sense, Ashlie?

(Off microphone discussion.)

CHAIR PENSON: So let me throw

something out there that may make this a little easier. Because I'm with you, Louise, I'm having a hard time with this. And it would certainly be helpful, and I know Kevin and Todd are listening, and I'll ask you guys to comment in a minute.

It would be helpful to know exactly how you're breaking down your regions. I know you mentioned the four corners versus individual states.

But if you're going to use both methods or one or the other, but then what would make me feel a lot better is someone reviewing this measure.

Some just basic descriptive information based on your regional breakdown to provide me with some reassurance that there is no major differences between the northwest and the southeast, or for that matter Iowa versus Arizona, as an example.

Steven's comment about differences in stage by SEER registry is concerning. So

I'll ask other people in the room and then I'll give Kevin and Todd a chance to comment.

DR. GILLIGAN: I'm just wondering, and I don't know the answer to this, but since we're breaking these patients up into all these different groups.

And most of those differences in treatment are going to be based on stage, does the different chemotherapy categories act as a reasonable surrogate for some of these regional differences or not.

Because we're only comparing neoadjuvant to neoadjuvant and trastuzumab to trastuzumab and no chem to no chemo.

CHAIR PENSON: I don't know if stratification is adequate here. I mean, because the fact of the matter is also, A, it's circular again, because your stratification variable is also your outcome.

DR. GILLIGAN: I don't think it's circular here. I think it was the other time, I don't think it is here. Because you're

doing a stratification --

CHAIR PENSON: Yes, you're right in that you're not going to be comparing those who did versus those who don't. You're right about that, that's correct. But it's still, I don't know.

DR. CHEN: To the extent you're trying to predict stage, the problem with breast cancer, not so much problem, but the facts are that we start giving chemo somewhere in the middle of Stage I, and if you're at MD Anderson somewhere towards the lower end of Stage I you start getting chemo.

And so there's that decision making, particularly in that upper part of Stage I, is very discriminating as to how you use resources. And again, there's the point to disclose that I do do work with the genomic assay companies that try to predict which part that falls in.

But that decision plays a huge part in your resource utilization for

identical size and otherwise stage diseases.

CHAIR PENSON: So I'm --

DR. GILLIGAN: I'm still puzzled though. Because the difference in cost is going to be the difference in whether or not you give the chemo. But you're only comparing people who got chemo to people who got chemo.

DR. CHEN: Yes, and so I think that's my point is that one of the biggest issues in at least resource utilization when you look at provider to provider, is whether you're choosing to use chemo for identical disease.

DR. GILLIGAN: Right.

CHAIR PENSON: You know, and yes I'm hearing a lot of issues here is what I'm hearing. And think I'm kind of in the same boat that Louise is in, which is, is this a fixable problem or do I just need data to make me feel better about it?

I think if you showed me data, you know, and this wouldn't be hard to do, to some

degree to look by population and it doesn't even have to be in the necessarily the same data set.

It just would make me feel a little bit better that, you know, over these three states versus those three states on opposite side of the country, stage is more or less the same and the age distribution is in the same ballpark, I'd feel better.

Is that a fair statement? Well at least that's me. I'm wondering if others agree with me, or -

DR. GILLIGAN: Well I think in some ways the appeal of the way this measure is done to me is that I think in some ways it's unfair to the provider to guess from such abstract data whether or not giving chemo was the right decision or not. Without stage information and all that stuff it's very hard to look over their shoulder.

But to look at someone who's giving chemo and they have twice as many

complications and hospitalizations as someone else who's giving chemo maybe that tell me something about quality of care that's actually worth knowing. So I mean I'm sort of playing devil's advocate a little bit.

But I think there's strengths and weaknesses either way. Of course it'd be nice to have stage data but you just can't, that's not an option.

CHAIR PENSON: And the question becomes again because you don't have that level of granularity, because you're not looking by the provider you're looking by large region. Here I get more comfort with the concept that maybe the regions are comparable.

I mean I sure would feel better if I had some evidence. Why don't we let the measure developers say a few words on this because we're having a lot of discussion here. Kevin or Todd, any thoughts on this?

DR. WEISS: Perhaps I'll start.

And then, Todd, it'd be great if you wanted to add anything. The discussion I had is very close tied to that of our workgroup. I didn't mention it at the very beginning of today but it may be obvious but it may be not.

But the intent of our project to try, where possible, to work on attribution to the most granular place that we thought was reasonable. And so on this one everyone recognized that without clinical staging you really can't drill down very far.

On the other hand when you get up to a certain population you should have a certain randomness of effect that you would avoid miss-classification based on stages. Now having state based information on stages would help us a lot.

However we also realized in that discussion how the practice of treatment really is developing more and more network. And that these networks are pretty complex, that often go across state boundaries.

They really work across health systems and patient, because of the way that the centers are being set up and such, that it really does begin to feel like you need to start with a very high aggregate to look at this.

And then probably the only people who could be beneath that would be individuals who would be individual systems who had large enough population that they could actually do some benchmarking but that was beyond our scope.

In that context we said that we can in this case go without formal staging but rather look at these different stratum of the types of care.

Recognizing that they would have to be matched to quality indicators to make any sense of any of the other resources these measures would be.

So I hope that addresses, kind of reflection, partly it's just a reflection of

you doing the same hammering that the workgroup did, how we got there and the fact that without that clinical enhancement of the stage this may be as good as we can get. But, Todd, your thoughts here?

DR. LEE: I don't have anything to add, Kevin, I think you've touched on a lot of the issues that, again, our workgroup struggled with and that is sounds like this technical advisory panel is struggling with.

DR. WEISS: I might add that they felt that it was important to shed light on possible variations and research use, I mean they were not forced to make this measure.

They said this is an important issue from their perspective. So this is something that we have reached a high level of concurrence in terms of a topic of interest.

CHAIR PENSON: So there's, I mean, I don't think anyone's arguing with the importance of it. I think that sort of, we are all in agreement there.

It's interesting so Steven just showed me these data from SEER and you really do see between Stage I, Stage II and Stage III these ten percent or so variations by registry and registry.

Now there are some differences in timing and other things, but the question really becomes is, it would be comforting, I think that's the best word to use here, if there was more data here to sort of make the group feel better about the risk adjustment. Is that a fair statement as I look around the room?

I think what I'm seeing in the room is sort of just some concerns that, yes it's important. Yes, we appreciate why the workgroup did it, we're coming up with the same, running into the same walls they did but it may not be quite ready for prime time without some additional data. Is what I think I'm hearing from the group.

DR. WEISS: Well one question that

might be helpful to me and the rest of our team. Is SEER data is not fully population based, it is dependent upon, it's a very nice sampling frame.

But I'm not sure is it representative of something that would be equated as a state sample? And I'm not, I just don't know SEER well enough --

CHAIR PENSON: Well that I can speak to. So the answer is it's not fully population based for the United States, so it doesn't represent all 50 states.

But within the state itself, whether say it's Iowa versus Connecticut, it is completely population based for that particular state over that particular time period.

It's not a random sample. Every patient in the state of Iowa who is diagnosed with cancer is included in the tumor registry. Same story with Connecticut. Now when you get into the metropolitan areas, you know, Los

Angeles, Atlanta, that's a different animal. But when you look at state registries they are population based by law.

DR. GILLIGAN: So one thing that someone could do is take a SEER area and run this measure with and without the staging information and compare the results and see what it adds.

CHAIR PENSON: I don't think you need to have particularly extensive data here. I think you need to give people a gestalt. You can look, SEER has been broken down into region, as an example.

Or you could break it down by state and just sort of look and say when I go to region those stage differences over the same time period seem to dissipate.

And I think it would make people feel better. Because I think there is some concern in this room that even with the sort of regional comping and even without individual provider accountability you may

have confounding by differences in population by geography. Steven.

DR. CHEN: And then just beyond the cancer staging, I still am somewhat concerned about not using comorbidity at all. Because we do know that things like cardiac disease is very widely across different states. And particularly smoking rates vary widely and they have huge comorbidity effects.

CHAIR PENSON: Other comments? So like we've done before I think this is the right time for us to stop and vote on these four criteria.

So why don't we start with number 1 which is a validity measure, are the measure specifications consistent with the evidence. For measure 2b1 we have one vote for high and eight for moderate.

Now we're on to 2b2, which is the validity testing. Validity demonstrates that they measured data elements are correct or that the measure correctly reflects the cost

of care resources provided adequately distinguishing higher or lower cost of resource use. And here we had seven who said moderate, one who said low and one who said insufficient.

Next we'll move on to the exclusions, 2b3, are the exclusions supported by clinical evidence or analysis of frequency, is the information about exclusions transparent. And here we had one vote for high and eight for moderate.

And last but certainly not least in section. Basically what we're asking here is the rationale and data support no risk adjustment of stratification.

And here we had two voted for moderate, five who were low and two felt it was insufficient. And I think certainly our discussion reflects that spread. I don't think we had much to add there.

So we'll keep moving along. We're actually doing fairly well I think. Let's

talk about 2b5, this is differences in performance.

And basically do the methods for scoring and analysis specified allow for identification of statistically significant and clinically meaningful differences in performance.

Or there's evidence of overall less than optimal performance. Lewis, you were the primary reviewer on this.

DR. POTTERS: Yes, I thought all of these 2b4, 5 and 6 sort of bleed into each other. I think we've had this discussion, to some degree in terms of, you know, understanding the stratification relative to a risk adjustment type of analysis.

And, you know, the pros and cons, as per the voting, I think there's a fair amount of moderate to this criteria.

CHAIR PENSON: I hear what you're saying. Other comments? When I first looked at this I actually wrote not applicable

because I kind of couldn't figure out what does it mean if the west coast is less or more than the east coast.

I'm not sure if that means it's not applicable or whether or not it's just difficult to interpret. So in certain respects it could be not applicable, it could be low. I don't know. You know, again, it goes back to do you think it's meaningful to look at regional differences.

The more I think about it I think in some respects it is, but in some respects it isn't. Again, I don't know the answer here.

MS. TURBYVILLE: I think you did a N/A for this one.

CHAIR PENSON: You do or you don't?

MS. TURBYVILLE: Yes, you have to have these, high, low, moderate or insufficient. So they didn't give us enough information.

CHAIR PENSON: Yes, so I think I can't vote N/A, I was hoping I could. I guess not. So I guess in the end I'm soliciting help. Other thoughts?

DR. CHEN: I tended toward moderate because I do think you are finding differences here. They are real. Whether they're useful or not I guess I mentally thought I would deal with that over in usability. But I think there are differences are they're real.

DR. WALTER: I guess I won't help you because I'm leaning towards insufficient waiting for some of this other data to be presented to me to understand how meaningful it would be.

DR. BORKER: I would rate this insufficient as well. I guess -- Oh, I'm sorry you had a comment? Because to me that is again leading back to the validity issue. It's the sealed analysis that was recommended here might be really helpful there.

CHAIR PENSON: Yes, it's difficult. I mean a part of me is with Steven where I think that there is some use to this assuming it could be collected in a valid manner. I'm sure it's sitting there, don't want to heap on because we really did heap on before with the risk adjustment piece.

But the fact of the matter is on the other hand if the risk adjust doesn't work then it's not going to fly, so I'd like more information there.

So I think in some respects maybe it's insufficient. Although I think Louis' comment, I mean on the surface it's okay. It's back to whether or not you buy it's okay not to have risk adjustment it sounds like.

DR. POTTERS: It really depends what they're going to do with it. I mean if they're going to put it on the front page of the New York Times it's not sufficient. If they're going to use the data and perhaps crosswalk it to other databases then it is.

MS. TURBYVILLE: That's a very important question. So in general you should be thinking that it's not just them that would be using it, it'd be any user and we don't control.

So when you endorse a measure it's for those accountability and public reporting models that you had for use in those and feeling that they've provided enough information to meet that kind of use. So we don't really oversee the use of the measures.

CHAIR PENSON: I'm sure starting to think that it's insufficient because the fact of the matter we do have these issues with risk adjustment.

And the fact of the matter is that even if it is perfectly risk adjusted and you see this sort of ten percent increase in region A versus region B I don't know what that means.

Now it may mean something but you've got to give me more data to interpret

it. And I don't have that information at this point.

Other comments? So next is 2b6, which is multiple data sources, methods are specified, there's demonstration they produce comparable results. And, Lou, this was you as well.

DR. POTTERS: I mean this would be an N/A if there was an N/A.

CHAIR PENSON: Yes, okay. And so let's finish up with the 2b and then we can do -- actually.

(Off microphone discussion.)

CHAIR PENSON: Yes. Yes, let's do it that way, that way we get the one gestalt measure.

MS. TURBYVILLE: It's starting with 2b1.

CHAIR PENSON: Well we did 2b1 already. We're at 2b5 so don't make us go back there. Never go backwards, Sally.

MS. TURBYVILLE: Sorry.

CHAIR PENSON: It's okay. So in 2b5, this is the differences or results reported, did they identify differences in performance or overall less than optimal performance.

Okay. And so here we had two people who felt it was moderate and seven people felt it was insufficient. And I think our discussion reflects that vote.

And then as we've said before multiple data sources was not applicable so we can step right over that. So this is the overall gestalt.

What is the overall level of validity testing, taking into account the discussion we had about risk adjustment, about the differences, et cetera. And go ahead and vote.

Okay. Well interesting. We had three people, I told you this was going to be a religious discussion. We had three people that felt that the overall validity was

moderate, three people who felt it was low and three people who said it was insufficient.

And I think the way to interpret that for when we're writing the report, and I know Kevin and Todd is listening, is I think as one of the people who voted insufficient and the others can agree or disagree, I just think that there's a need for more information.

It's not necessarily and indictment that it's no good but it really, truly I just don't know. Is that a fair statement for the folks who?

(Off microphone discussion.)

CHAIR PENSON: Well for low I think people basically looking at it and they're not buying. Yes, they're not buying that the risk adjustment, you know, people felt you needed risk adjustment. I mean there really is, sort of three people said it was okay, not perfect.

Three people said it's never going

to fly and three people said I don't know, give me more information to make me vote one way or the other.

And I think that the discussion really reflects that, frankly. And it'll probably be the same on the next measure too.

(Off microphone discussion.)

CHAIR PENSON: Yes, we're getting there, we'll be done on time. No phone calls, I swear. So we'll do 2c now. Disparities in Care. If disparities in care have been identified measured specifications, scoring, and analysis allow for identification through stratification that is by race, ethnicity, socioeconomic status or gender.

Or alternatively there's a rationale on data which justifies why stratification isn't feasible. And for this one, Rohit, you were the reviewer.

DR. BORKER: So as you can see the validations are all over the place. But I gave it a high just because of the same reason

that I mentioned last time as that's the limitation of the database.

But if the database has those data elements there is nothing preventing this measure from stratifying the data on those disparity measures.

CHAIR PENSON: See I kind of, and Sally, correct me if I was wrong, but I had an N/A and the reason I said N/A was because you're doing this sort of by region. So it's hard to stratify a region by gender or by race. I mean you could look at --

MS. TURBYVILLE: Proportions.

CHAIR PENSON: -- proportions potentially, but it's still hard to do.

MS. TURBYVILLE: I think that's an appropriate question. And it could be that, it's fine for committee to feel that the stratification of disparities isn't warranted. It's an important signal to send to developers, even in commercial databases where we know it's very difficult to do, that it is

warranted if not now in the future.

So I think especially for these commercial administrative derived pieces it is kind of important to kind of provide that feedback for the measure.

CHAIR PENSON: So allow me to summarize that as I think through this. Which is if you believe that stratification, by gender, so all the women in the west versus -- well this breast biopsy.

So all the older women in the west versus all the older women in the east, if you think that's value there that you either have to think about it as insufficient or it wasn't addressed.

If on the other hand you sort of look and say well stratification isn't that important for a population based, or for a measure which is based on region, the population may be --

MS. TURBYVILLE: Well in particular this measure.

CHAIR PENSON: This measure, so then you say high because the rationale to justify why stratification isn't necessary. Thoughts?

DR. BORKER: I have a quick question on what you're saying. So in this particular case, using that database, one can stratify older versus younger women across states.

But doesn't mean the developers have done it. How do we rate that, when we know that they can do it, it just that it hasn't been reported?

CHAIR PENSON: Well, so the question becomes, and I do think it's important, if you think that it's not important then you say, I would assume, you say high and you let it go. Because in your mind there's rationale for why stratification isn't there, by age as an example.

If on the other hand you think that's an important piece, you know it can be

done so you would put it as insufficient. Or you could even say -- Go ahead.

MS. TURBYVILLE: Right, so I wouldn't use insufficient in that case. I think insufficient would be you don't have enough information to determine whether or not disparities, socioeconomic, race, ethnicity, should be addressed for this measure.

I think you might put moderate, for example. For commercial testing, because you know they can't test it but perhaps they could have recommended that users, if they have the data stratified, because the literature supports that there are some racial disparities that you would want to exposed.

You might say low if they don't address it all, but the literature does or you know it does. So --

CHAIR PENSON: And you raised a point with the slip of the tongue, because there's no race information here. Okay. So if you feel that race is an important

stratifier and that you can't do this without stratifying for race, then you would vote low. And so it's just because of matter of clinical opinion.

DR. CHEN: And to that point I would say that race is actually fairly important in breast cancer.

From molecular studies we know that African-Americans have a much higher rate of triple-negative breast cancer which is much deadlier and requires much more aggressive therapy.

CHAIR PENSON: Okay. Other comments? All right, so why don't I call the question on this and I suspect we'll be all over the map here.

So with regard to Disparities in Care, do the measures allow for identification through stratification results or do you buy that the rationale justifies that stratification is not necessary or feasible.

There you go. So three put it as

moderate, five put it as low and one put it as insufficient. All right, we're moving right along.

We're actually I think doing fine, Sally. So next we're going to talk about usability for this. And this I think will also be a somewhat interesting discussion. The primary reviewer for this was Louise.

And basically we'll start with 3a, which basically looks at whether or not the results are reported to the public or at large in national or community reporting programs by the time of endorsement maintenance review.

And again we go back to the discussion we had before. Which is this is a work in progress both with regard to NQF and this also has not been tested by RWJ, it's currently being tested.

DR. WALTER: Right so there were four highs and two insufficients and I rated it as insufficient because it's not in use. And I think that's what we've been

consistently voting for the other things if they haven't been publicly reported, and it hasn't been.

CHAIR PENSON: So I will say one thing here, because I'm with you on that. And I was the other insufficient. And I think that for the people who voted high, you know, people have been changing their votes here.

But the one thing I will say here is that there is something to consider, the useful to the public when you report at a regional level. And for those other pieces on the slide if, and I'm not suggesting you do this by the way.

But if you look at this and you say just the way it's designed, these regional reporting measures, it's just not useful for benchmarking accountability, whatever, then you would vote low here. I don't feel that way, but there may be some in the room who do.

DR. WALTER: Okay. So that was 3a. And 3b, most people thought it was

insufficient as well. There's not really enough information presented to understand whether it would actually be meaningful and understandable to the public.

And then 3c, all over the map. Two highs, two mediums and two lows. Yes, I just didn't think this would necessarily facilitate transparency and understanding at this regional level without significant clinical detail.

So I actually rated it low. But one of the reviewers said it was high because at least the data elements were clear.

CHAIR PENSON: So I'm with you on this, ironically. But I'm curious to know, for the people who felt this was high, again I don't want to call people out, but with regard to the construction logic and how it worked and whether or not it was transparent, does someone want to speak to what they thought were the strengths here?

DR. BORKER: I think I voted high

here. But then based on the clinical arguments that have been made I'm going to revise that rating definitely.

DR. POTTERS: Yes, I mean I voted high too and then that was in the context of what I thought was, in my narrow way of thinking before this discussion, an interesting way for the developers to try and make up a stratification that I think we've had a discussion on that. Perhaps it's not as valid as I had originally thought.

CHAIR PENSON: Other comments? I definitely see us losing energy here, so let's, stay with it, stay with it folks. We're almost done, another hour I promise you.

All right so I guess at that point, any other comments about usability? So I actually feel like we're actually reaching a consensus here on this, so let's go through these together.

MS. TURBYVILLE: Before you rate, just keeping in mind we know that these

measures haven't yet been implemented and we understand that, for in particular these measures, so you also want to put the mental twist as they're presented, would they be suitable as well. So kind of broadening that context.

CHAIR PENSON: Okay. So let's start with 3a, to the measure results are reported to the public, are they going to reported to the public at large, community reporting standards.

In the past we've been voting insufficient, I'm going to actually move by acclimation, is there anyone who is going to vote differently for insufficient here?

So everyone votes insufficient, that's everyone gives it a nine, which is consistent with what we've done with other things.

So for usability, 3b, which is the public reporting, the performance results are meaningful, understandable and useful to the

intended audiences both for public reporting and quality improvement. And we can actually vote on that. All right, we're good. So here we had one moderate, two low and six insufficient.

And now we'll move on to 3c, which is the clinical and construction logic. Basically this is are the data and the results maintained such that the measure including the clinical and construction logic for defined unit of measurement can be decomposed to facilitate transparency and understanding.

Go ahead and vote here. All right that's interesting. Okay, so we had five folks who said it was moderate, three who said it was low and one said it was insufficient.

Do you think that the discussion we had reflects that? I mean I sort of thought we were going to be all over the map on this one.

MS. TURBYVILLE: The moderate would be interesting. So the conversation at

least that I picked up I heard the issue with it being at the regional level and how would that work and how it would be actionable, so I'd be very interested for those who supported it at the moderate level to, for our notes, provide a little bit of rationale would be helpful to us.

CHAIR PENSON: Don't get shy now, we're all friends.

DR. CHEN: I voted moderate. I guess my feeling on it is in the narrow context of can this be deconstructed and made transparent I think the answer to that is yes.

I don't necessarily think that the things that make up the elements are the most useful things but the technical question is yes.

CHAIR PENSON: That's great. That's very helpful. And obviously the harmonization piece is not applicable. So we'll now finish up with the feasibility items. Then we'll take a five minute break

and knock off the last one and call it a day.

MS. TURBYVILLE: Maybe we could get out early.

CHAIR PENSON: I think you're overly optimistic. If we do it's only because we're all beat to heck. All right, so 4a, and we've been through this numerous times today. I don't know if have to have a lot of discussion about it.

For the clinical measure here the required data elements are routinely generated and used during care delivery. So blood pressure, so basically we're looking at administrative data here.

I don't think that the comments before about J codes and genomic testing are quite as relevant here, although I could be wrong. Thoughts?

DR. CHEN: I think they're still relevant here in the sense that we're giving chemo and there are the genomic assays, but having said that.

CHAIR PENSON: Okay. But you'd probably call out something more along the moderate way than anything else, that would be my guess. Okay. Let's move on to 4b, the required data elements are available in the EHR. And I think everyone can agree that that's probably the case. You're probably not going to miss much here.

We may have the same issue as before with the types of chemotherapy and how things get priced out. But I think they're still going to there. Are there other comments?

Okay. Susceptibility to inaccuracies, errors and unintended consequences. Fairly, looked like it was middle of the road looking at scores. And I think we've been here before with this. I mean it's still administrative data, it's not perfect. Other comments?

And finally, barriers to use collection and measurement strategy can be

implemented as demonstrated by operational use and external reporting programs. Or testing did not identify barriers to the operational use. In my mind this becomes a matter of whether or not you buy the algorithm. And if you buy the algorithm I think it's applicable. Other comments?

(Off microphone discussion.)

CHAIR PENSON: Well right, in your opinion does it, I mean my gut feeling is it's just SAS code, you know, so it's doable and implementable at other places. Whether or not it's valid we talked about already.

All right, let's vote on this and take a short break. So on 4a, are the required data elements routinely generated and used during delivery of care. And here we had six votes for high and three for moderate.

For 4b, the required data elements are all available in the electronic health records and if not there's a plan to get them. And here we had eight people vote for high and

one person vote for moderate.

4c, susceptibility to inaccuracy, errors or unintended consequences taken into account, et cetera. So here we had two highs, six moderates and one low.

At the risk of, I just wonder did whoever voted low mean to vote low? Okay. And did you vote low on all the other ones as well or? No? Do you just mind sharing comment why so the NQF folks can report that back?

DR. CHEN: I think my concern with this one goes back again to the ability to adjust for things. I think it is susceptible to inaccuracy, that's why.

CHAIR PENSON: Okay. That's reasonable. All right. And then 4d, can the data collection strategy be implemented. That is is the measure already in use or is there testing that shows that it can be put into use. So we have four high, four moderate and one low. Again I'll ask whoever voted low

just to share their comment for the record.

DR. GILLIGAN: That was me and my mistake, I meant to vote moderate.

CHAIR PENSON: Okay. Can you change that yourself? Okay. Good so it's four high and five moderate. So that takes care of the treatment one. We have the breast biopsy one left.

Why don't we take a five to ten minute break. Let's start again at 3:15 with an eye towards getting it done in about an hour. Okay? Thank you guys for hanging in there, I appreciate it.

(Whereupon, the meeting went off the record for a break at 3:06 p.m. and went back on the record at 3:18 p.m.)

CHAIR PENSON: Guys, I don't mean to be rude, but I know a lot of folks are going out to the west coast, or a few of you are and I don't want you to miss your flights. I don't want you to miss your flights and I don't want you to have to get on the phone

with me for an hour in two weeks, okay? So with that note let us continue.

So we're on to the last of the four measures, which is the episode of care for a 60-day period prior to breast biopsy. And in many respects, this is like the other one, the one regarding treatment of care, pardon me, costs around care.

So let's start with the importance issues and we can go off from there. So 1a, just to refresh your memory if you haven't gotten tired of this already.

This focuses on a national health goal/Priority Identified by DHHS, or National Priorities Partnership, and I don't think anyone is going to argue again, breast cancer is a big ticket item, and so it's high impact.

And I think most of everyone's scores reflect that, their H's and M's. Anyone who voted for an M and wants to add, be my guest.

DR. WALTER: Well I guess I wasn't

sure what the, I mean it seems to me that the big disparities are in getting follow-up of your abnormal mammogram, or abnormal lump or something like that, this measure is not going to get at that.

It truncates it at 60 days. It's not going to measure, a lot of the delays are because you don't get follow-up in 60 days. So I didn't think this was really, again it would've been helpful to have more conceptual clarity about why, you know, what they're doing is important, but it wasn't really stated in their introduction.

DR. CHEN: I think what they're trying to get at and I'm, you know, trying to do the psychic thing here it's like a friends network for them, but is that there has been a push more recently towards trying to get people to do needle biopsies instead of open biopsies.

And yet despite consensus statement saying needle biopsy is the way to

go, only like 40 percent of breast cancer is diagnosed by needle biopsy. And so that's I think where they're going.

DR. GILLIGAN: Although, I just have to weigh in. I actually voted insufficient, but I didn't get my survey in on time. But they have nothing in the section on any impact from breast biopsy whatsoever.

So I mean, yes I agree it is a high impact, but it would've been nice if they'd said something about breast biopsy as opposed to about breast cancer.

CHAIR PENSON: Yes, we go back to this where, you know, even though we all feel that it is, did they by not putting it down, should we penalize them. They're on the phone sort of shaking their heads.

But I just don't feel, Kevin and Todd you can yell at me any time, but the fact of the matter is I just, you know, you're right it's not breast cancer, but look you don't get to breast cancer without a breast

biopsy. And I think Steven's comment about what sort of biopsies are done is valuable.

But I'm not hearing anyone say low or insufficient, I'm hearing high or moderate, so.

DR. WEISS: Would it be okay for me to step in here here -

CHAIR PENSON: Oh yes, yes, absolutely, come on in now.

DR. WEISS: Yes, so I think that you're right to help us notice that we didn't give you that level of detail. The two things that the workgroups were very clear on was the type of procedure, and that is the open biopsy or not.

And the second is that there's a proliferation of different types of pathological assays that can be used here. Some very simple, some pretty complex and very expensive.

And that there's a real belief that on both those issues there's a lot of

variability, and that means that there's research issues and quality issues probably. So that's why that's there.

CHAIR PENSON: That's helpful Kevin, I appreciate that. Any other comments about 1a? So 1b is that demonstration of resource use and cost problems represent an opportunity for improvement. Excuse me. The data demonstrated that there's a variation in the delivery of care.

It's interesting because when I looked at this originally I scored it as an M. And that's because I'm not a clinician that treats breast cancer, and it never occurred to me, you know, this issue about the biopsies, that that is actually what's most important.

I mean it strikes me though, that in the document that the issue is always about imaging, and the imaging is going to drive it. But what I think I'm hearing here is that type of biopsy is important as well.

I mean are we comparing apples to

apples? When you look at open versus say, needle biopsies, I don't know the answer to that and so I'll defer to the breast cancer docs in the room.

DR. CHEN: Yes and no. There are some people who can get needle biopsies, who are getting open biopsies. I mean we give grand rounds to the community and have arguments with other physicians who say no, they don't like needle biopsies when they clearly could do them.

On the other hand, there is some element of people who have to get an open biopsy for various technical reasons.

And there's a kind of a level that you can't go below, and that's going to vary from place to place. Often also dependent on what kind of resources they have available to them. Not every place has stereotactic biopsy available to them.

CHAIR PENSON: Other comments? I mean I'm not hearing any deal breakers here by

a long shot. I'm getting the impression that everyone thinks this is either high or moderate, and I think that's good.

So next, onto purpose then. The purpose or objective of the resource use measure, and the construct for resource use and cause is clearly described.

And pretty much everyone in the room voted that as high or moderate. I don't think there were a lot of questions here. I think the purpose is pretty straight forward. Any comments?

All right. And then we'll do 1d, which is the resource use categories. And like the earlier measure, the categories are very straight forward in my mind, and they have validity for me so I didn't have any problems with that. Other thoughts?

DR. GILLIGAN: I just wanted, one very brief comment which is because this happens so much these days. On page 6 it says, "In 2009, the USPSTF issued guidelines

advising against any screening for women in their 40s." And that's just factually incorrect.

The USPSTF recommended against routine screening. They recommended the doctors counsel the women based on whether or not they want to be screened.

And it constantly gets quoted this way, as if they said, don't screen, and that's specifically not what they said. They said, don't do routine screening, talk to the woman first. So I just wanted to clarify that.

CHAIR PENSON: Kevin, did you catch that?

DR. WEISS: I did.

CHAIR PENSON: There's no comment, just you noted it and change it, and next time you submit it, it's all good. But I think, we won't spend too much time on it because everyone's really punchy.

But Tim's, but I'm not minimizing Tim's comment, I mean I do prostate cancer,

which is, you know, the ultimate in that world. And I think people make those mistakes all the time.

And it's important because a lot of times these documents get picked up in the lay press, or by patients and other people. And suddenly what was sort of a mental lapse for lack of a better way to put it, becomes an unhealthy policy.

So with that in mind, why don't we go through the importance to measure variables. So 1a, high impact, does this measure focus on a specific national health/goal priority, or is it a high impact aspect of health care? So here we have five people voting high, and four people voting moderate.

1b, the performance gap issue, demonstration of resource use or cost problems and opportunity for improvement is there, where overall there's less than optimal performance across providers, population

groups. And here we had two people vote high, and seven people vote moderate.

1c, Purpose/objective. The purpose/objective of the resource use measure, including it's components and the construct are clearly described. And here we had five people vote high, and four people vote moderate.

And finally, 1d, the resource use service categories are the categories included, consistent with and representative of the conceptual construct represented by the measure. And here we had seven people vote high, and two people vote moderate.

Does this one have a summary one, no, okay good. All right. So you'd think by the end of the day I'd have figured that one out by now, wouldn't you? Appreciate that.

Okay so next, we're going to move onto Scientific Acceptability. 2a, and 2a is basically is the measure precisely specified so it can be implemented consistently. And

the primary reviewer for this was Louise.

DR. WALTER: All right. So it looks like we had three middle's, three high's and one low. I actually voted this low because while the measures and CPT codes are clearly labeled in the table, they're actually not specific for breast biopsy.

So there's two CPT codes, the 10021, 10022, that are basically biopsies of any organ, they're finial aspirations. Normally they're paired with a breast cancer diagnosis, and then you can use it as a biopsy.

But if it's not paired with a breast cancer diagnosis which in this case it's not, then it is basically a F&A of some organ, be it your thyroid.

So I didn't see that there was any evidence that this was a validated algorithm for identifying breast biopsy. And at least in my work doing claims coding, I know that that is not a specific code for breast biopsy.

CHAIR PENSON: Other comments in the room? I'm going to take the prerogative and ask the measure developers about Louise's comment, because I didn't capture that, not being someone whose focus is on breast cancer. And that's a major concern, are you picking up non-breast cancer or non-breast biopsies?

DR. WEISS: If Todd's here, if he can maybe address that?

DR. LEE: Yes, so these were through our workgroup, identified as an iterative process, you know, I'm not aware that, why they got through our workgroup if they're not specific to breast biopsies. I can't answer that right now.

DR. WALTER: And let me put, I was going to put one other thing because I looked at the evidence that you provided, the data, and there was a statement about, they couldn't understand why the cost was so much lower for CPT code 10022, which is the non-specific code, versus 19103, which is a specific breast

cancer biopsy code. So that actually would track with that this is not a specific for breast biopsy, based on your data.

CHAIR PENSON: Louise, is there something you could suggest here, whether it's either taking out those codes, or some sort of evidence they could provide to make you feel better?

DR. WALTER: Well I guess we're getting back to chart review. But it's a little concerning since one of the non-specific was the second most commonly used CPT code, so I don't know how often is that used for breast biopsy versus other biopsies.

CHAIR PENSON: Steven?

DR. CHEN: Yes, I actually had a concern about F&A too. In particular my concern was that it excluded you if you had an F&A followed by another breast biopsy.

Now in general outside of breast cysts, we don't do F&A for anything besides abscess and cysts, unless you don't have core

available to you, and that's where people stop doing needle biopsies. Because they just get so poor information, and so I would say that in my mind, I would probably exclude that.

And then as a triggering event, and put that into resource use, the same way that you might use an MRI or something like that. Someone is wasting their time putting a fine needle into something that didn't need a fine needle. It's just my, because the other ones are not cancer directed questions, they're abscesses and cysts.

CHAIR PENSON: So Sally, how best to handle this because what I'm hearing from the panel is, this is worrisome, probably the best way to deal with this would be to exclude these patients. But I also add as someone pointed out, that if you exclude the patients you lose half of the sample size, and I don't know what to do with that.

MS. TURBYVILLE: I think there are several possibilities. Throwing it back to

the developer to see if, for example pairing it with a diagnosis or some other code would increase it's reliability, and whether they can do that in this project.

Or if the committee votes on the measure as it's specified now, and I think it might be an opportunity to ask the measure developer if they think they can think about this a little bit more and come back, if not to this panel to the steering committee, as having addressed a concern with this TAP.

CHAIR PENSON: So, what I would then support here, or propose I should say, is this is probably insufficient evidence. And what I mean by that is that, it would be accompanied by a comment basically saying that, as currently defined this measure is problematic because of these two CPT codes. And that it would be preferable to have these removed, so that that way the cohort is more precisely defined.

However, that being the case, then

it would be useful to have the cohort rerun and see what sort of sample sizes come up. So does that seem like something that would be reasonable?

DR. GILLIGAN: Just one question. You said that if you pair the CPT with the breast cancer diagnosis, then it becomes more valid. Is that an option?

DR. WALKER: Well that's just, yes, I mean because at least then you think, gosh if it's paired with a diagnosis, they must be biopsying the breast versus without anything, yes.

CHAIR PENSON: Yes, but the problem becomes is, what do you do with negative biopsies? You don't want, I mean what you could potentially do is pair it with benign diagnoses from the breast too.

But what I'm hearing here is, is that with the inclusion of these two CPT codes it's causing real problems certainly. I'm glad you're here, I didn't even think about

that. Other comments?

So I think what I would, and just to repeat what I had said was, that I think it's probably, the best way to handle it would be to say that there is insufficient evidence as it's currently written.

It's not acceptable because the CPT code is capturing non-breast biopsies in all likelihood and we would propose, or suggest I should say, either excluding those patients, or potentially tying it to a diagnosis code that localizes it to breast, whether it's breast cancer or benign breast condition, and then --

(Off microphone discussion)

CHAIR PENSON: I guess in certain respects I would, what's that? I would say it's insufficient because the fact of the matter is, is low implies that we don't think it's going to work.

And it's not that I don't think it's going to work, it won't work in it's

current format. It may work with the modifications but we need to test that.

MS. TURBYVILLE: Right. I completely understand with what your struggling here with, I think we would like to vote on how the measure is specified now.

CHAIR PENSON: Okay.

MS. TURBYVILLE: Whether the measure developer can follow-up or chooses then to remove the measure, or continue to pursue it, you know, as it is written, we'll explore.

I mean obviously there's time line issues for those projects so we'll work with the developer to see if they can come up with an approach that would address your concerns, and how you rate the precision of the specs.

CHAIR PENSON: So I think what, so then that's what you prefer, so let me, before we talk anymore in the room, I want to go to the folks on the phone, Kevin and Todd, and make sure you guys are hearing what we're

saying.

And if you have any questions about this or disagree, let us know now so we can discuss it, because obviously this will affect things. Kevin, Todd?

DR. WEISS: Yes, so there's not a process for us to easily say to you that if it's a very clear concern of specificity of the diagnosis that is affecting your ability to go forward, that we would just take these out and work with the more specific, the smaller end.

So there's no way of us doing it on the fly like this, but and I can't speak for Todd and the rest of the group.

But I would suggest from my understanding of how they work is that, they would probably be very pleased to make sure that the specificity increased in fact if this was a big problem that would make us all feel uncomfortable.

So I don't know if that did

anything more than just reflects the fact that we're very much attuned to your feedback.

CHAIR PENSON: We sort of went in and out for the last part of that, so if you just repeat it again, I'm real sorry.

DR. WEISS: That's okay, I'm an instructor of repeat. We are very, you know, we are as concerned as you are if, in fact, that these diagnosis would lack a lot of population where there was nonspecificity and created any sort of misclassification into the population.

And so if we had a way to, a mechanism on this call to just say change it by deleting these two codes, we would. I just don't think that the process is set up to allow us to do it, either on NQF side or on our side.

But we would be very responsive to that concern if it was addressed to us. And to the extent that you could look at the rest of the measure and give us some feedback in

the context. It would be extremely helpful because it seems like a very straight forward issue to address.

CHAIR PENSON: So yes, I'm with you on that and I think everyone in the room is too. As Sally pointed out, we're sort of obligated to vote on the measure as is.

But as Sally also pointed out, it's entirely possible that you could turn around, crunch this, get it back to us fairly quickly, and either as a TAP or even as a steering committee, address it that way.

So I think that what I'm hearing from Sally is that NQF will find some sort of mechanism, because it's a very discreet request that your getting from the TAP, which is easy to deal with. All right.

DR. WEISS: That's fine.

CHAIR PENSON: Great. Thank you. Other comments in the room? Okay, so with that, let's move on from 2a1 to 2a2, which is reliability testing. And reliability testing

demonstrates the results are repeatable producing the same results a high proportion of the time when assessed in the same population. Louise?

DR. WALKER: Oh sorry, I didn't have this down for me.

CHAIR PENSON: Oh, I'm sorry, sorry, I apologize. You're right, you don't. This was Dwight, who the dog ate his homework there, so I'm just teasing. It's that point in the day. So basically I think that while looking at this, I think I was the only one who voted insufficient.

It's sort of all over the map, most people voted moderate. Rohit, you said low. So I think I'd ask Rohit just to comment on the negatives, and while I figure out why I put it insufficient again.

DR. BORKER: This is again 2a2, correct?

CHAIR PENSON: This is 2a2, yes.

DR. BORKER: So this is kind of

reflecting the same concerns I had with other measures is, we haven't really evaluated this measure in another database, and that to me is like the biggest evidence gap here.

Whatever processes that they have done has been in the same database. Until we have that data, I can not evaluate. And to your earlier comment, I should put that as a insufficient evidence rather than the low evidence.

CHAIR PENSON: Right.

DR. GILLIGAN: I have actually the same concern. I thought it was low or insufficient as well.

CHAIR PENSON: I now have found my notes. The reason I actually said it was insufficient was because of a comment that Carlos made in his, was data reproducibility assessed and he specifically said there's no evidence that the process was validated, or there was any type of QA to insure accuracy.

So he raised some concerns that, I

mean my common sense says it's probably very reproducible, but there's no, the data aren't provided and they were requested.

Other comments? Okay, so why don't we vote on 2a2 and 2, let's vote on these, whatever numbers we're up to now. It's been a long day.

2a1. So 2a1 is this is precisely defined and implemented consistency, and this was, there was a lot of discussion around whether or not we're including breast biopsies, biopsies which are not of the breast

So here we had two who said moderate, four who said low, and three who said insufficient. So I think we've had enough of a discussion here that we don't have to beat this over the horse.

And the next one is reliability. 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.

And so here we had three who said moderate, and six who said insufficient. And I think that Carlos's comments sort of addressed that, so. I'm sorry.

MS. TURBYVILLE: Overall.

CHAIR PENSON: Just looking at what Carlos said, okay. Let's move onto 2b1.

(Off microphone discussion)

CHAIR PENSON: Oh summary, gosh darn it, that's what you were saying, God I was so lost today. All right, so this is the summary for a liability. Just vote and shoot me now, okay. And here we had one that was moderate, two that were low, and six that were insufficient.

Now we can move on, good. Yes, this is the fun one, actually I think it may go a little quicker now.

So, we're now onto 2b1, which is related evidence. Evidence measure specifications are consistent with the evidence presented to support the focus of

measurement under criteria on 1b. And the measure is specified to capture the most inclusive target population indicated by the evidence.

I think to some degree this is going to get back to that issue of the breast biopsies that aren't really breast biopsies. So the reviewer for this one was Louise, so.

DR. WALKER: All right. Yes, I was going to say this tracks very well with 2a1. So in addition to it not being specific for breast biopsy, I think another question I had was a lot of the measurement choices were not necessarily justified. So why a 60-day window before biopsy?

I had a question like, well many women don't get follow-up breast biopsies within 60 days of a lot of their imaging.

So therefore, a region with lower costs in the 60 days before biopsy, maybe that's due to long delays between getting all their testing and then their biopsy. And how

would this be detected? So I guess I had some questions about that.

Also, why stratify at age 30, mammography. Generally the guidelines suggest starting at age 40. So there wasn't really a lot of rationale for that. So those were sort of my comments on specification.

CHAIR PENSON: Other comments regarding sub-criteria 2b1, whether or not the measure is consistent with the evidence? I think Louise kind of hit the nail on the head.

Now my question to you is, with regard to this, I mean does this also come back to low based on what we discussed before? I mean I think that issue about not, if it wasn't for the non-breast biopsies, I'd be a lot more comfortable here, but I really think that's a major problem here.

Okay, let's move onto 2b2, which is validity testing demonstrates measure data elements are correct, and that the score correctly reflects the cost of care. This was

assigned, I'm looking here, to Dwight.

And I think in the end, we can sort of use Carlos's review as a proxy. And I think Carlos in reading the review picked up some minor concerns, but nothing that was overwhelmingly problematic.

I mean I think that, you know, if you get, let's for lack of a better way to put it, suspend disbelief for a minute, and say that all the biopsies were done on the breast, then you are validly measuring the resources here within reason. And I think Carlos may have raised some minor concerns, but that was it. Steven?

DR. CHEN: I think the one thing that for me is of major concern is, what generated this biopsy to begin with. Is it something that is a palpable mass? Is it something that is of minor concern?

You know, because I think that that does change the resource utilization, because a palpable mass, you're much more

likely to go direct to biopsy, whereas something that maybe was a little amorphous, you may end up with an MRI somewhere along the way.

And that has nothing to do with practice variation, it has everything to do with what's actually happened.

CHAIR PENSON: Well, I understand what you're saying and it raises two points. I don't think that that affects the validity of the measure, I think that it is a risk adjustment situation. So in other words, you know, why was it done? We can risk adjust for that.

And so I don't know if it's appropriate to discuss it here. We can discuss it later, but then we'll get into the discussion later again like we had on the last one is, do you need risk adjustment when you're looking at a regional setting.

Shouldn't the number of women who have palpable breast masses in Florida be the

same as in Arizona. And I'm not saying that's true or not by the way, I don't know. But I would think if it's okay with you, this probably comes up in the risk adjustment piece.

DR. WALKER: But if you don't believe the codes, how can, I mean can you vote anything other than low?

CHAIR PENSON: Well, here's what I would say is that, I think we've already dealt with that in all fairness. So I mean as, let's now say that we're cool with, we've gotten over the fact the inclusion criteria is a problem, okay.

But now let's now say, looking at the outcome measures, you know, which is resource utilization, do they have face validity? Does that fly with you?

Because otherwise what we're going to end up doing is we're going to put low for everything. It may bounce back and then it may fix this problem, this problem.

If there's a problem with say the validity of these measures, you know, going with what we were getting at before with J codes, and this and that, I don't want that to get lost because we were so caught up in the issue with the CPT code. So we've already sort of voted on that.

Let's now pretend as if everything else is okay, so we can give additional information to the steering committee and to the measure developers, so they can make appropriate changes. So having said that, what do you think?

DR. CHEN: I'm looking at this again, and I think by and large I think they've captured most of it. I think, the one thing I'm looking for and I can't seem to find it in here, is lymph node assessment and things like that that sometimes comes after biopsy, but it's a minor point that I don't think prohibits it from going forward. And it may be in here, I just lost it in the mess.

DR. WALKER: I guess the only other thing I'd like to see for validity testing is, different intervals. Again, convince me of why 60 days is the interval that we should be using.

CHAIR PENSON: I think that's a reasonable point. Other comments? All right, so now we're onto 2b3, and this is the exclusions, did I miss one, I don't think I did. No, 2b3, okay good, I'm not completely losing my mind, only sort of at this point.

So the exclusions are supported by clinical evidence, and the measure specifications include computing exclusions so it's transparent. And the reviewer for 2b3 was Rohit, and again I'll ask you before you start, remember let's sort of, we've beat the excluding the CPT codes at this point. Just so we can give useful information.

DR. BORKER: So just to kind of point out the strengths. One of the things that you just mentioned was transparency in

terms of the exclusion criteria and it's impact that got tested on the cohort size.

But that also serves as one of the limitations as they lost pretty much 52 percent of their potential eligible patients. So you're losing half of your patients, we don't know what, you know, what impact that has on the outcome measures. That's the major point there so, and there are some minor things that I don't want to bring it up.

CHAIR PENSON: Steven?

DR. CHEN: One exclusion I just bring up for a discussion and I'm not sure how I feel about it, is they excluded unpaid claims and they zeroed those out.

But for things that have zero cost, they gave them a one and gave them the standardized price. Which would seem to imply that if you have a crummy insurance company covering your state that you might look really good on resource utilization.

CHAIR PENSON: I think that's an

interesting point. I didn't catch that, that's interesting. I was on Rohit's comment, which is you throw out half your patients because of your exclusion criteria, what does that do to your generalizability? And could that vary by region which would confound any comparisons.

I looked at it not as a major point, but a moderate point. In other words, I didn't want to throw the baby out with the bath water, but that was something there.

DR. WALKER: This is just a question. Again, it would be nice to have explanation, rather than excluding women with a prior history of breast cancer, they stratify by this, and I just didn't know why, why stratify versus exclude, and what was the hypothesis behind that?

CHAIR PENSON: That's a reasonable point, I mean maybe they just wanted to capture more interesting data. As long as they stratify I think you're dealing with it,

but I think it's, other points on 2b3, exclusions?

So what I'm hearing, just to summarize, is some minor to moderate points here. A lot of good information for a summary report and I didn't hear anything that killed it, but I definitely heard some interesting thoughts there.

Let's move onto the criteria which we always seem to spend the most time talking about, which is appropriate too obviously, is 2b4, which is the risk adjustment issue.

And again, there is no risk adjustment here, so that what you're voting on with 2b4 will be whether or not the rationale as you see it, for the lack of risk adjustment is appropriate. There is some stratification, but I don't know if I would say that's adequate, and so I'll throw it open to the floor.

DR. CHEN: I feel like I'm talking less.

CHAIR PENSON: No, it's okay.

DR. CHEN: There's a couple of things. One is, and I didn't issue this objection before, is you have this, again it's 320 days of coverage on both sides for a 60-day episode.

And that only makes sense if you're going to do risk adjustment, which they specifically said they don't want to do.

So I would say that I would leave that in and get the risk adjustment in, particularly because what kind of biopsy you choose may be very related to what their comorbidity is, in particular anti-coagulation becomes a huge deal if you want to do a needle biopsy.

CHAIR PENSON: And I'm with you on that. I think that the exclusion by, you know, two years continuous coverage is a fine point, it's a good point, but it's also a fine point. In other words, a minor point.

What I would ask, which I'm having

a hard time with, I personally am comfortable in this particular measure not to have the risk adjustment.

Only because the things we've talked about, Steven, with regards to, for example anti-coagulation, palpable breast masses, that's going to the, I would assume that the distribution is going to be random between the west coast and the east coast.

I may be wrong about that, but that's my gut feeling, that you're not going to find more people on anti-coagulation, significantly more in California than in New York, but I could be wrong.

That's what it boils down to. If you think that you're going to see this variation by region, then risk adjustment is necessary. If not, then it's not.

My gut is on this one, and I'm not a breast cancer doc so, either the folks who are say so, you know, I'm not feeling it, so.

DR. WALKER: Well I guess it just

comes down to again, trying to really understand what region it is, if it's the state, you know, maybe there are. But, you know, if it's bigger, maybe not. So I guess I wish there would've been more specification on the region.

CHAIR PENSON: Other comments? All right. Well, we are going to get done on time the way we're going now. We beat them right down. Okay, let's vote on these four like we've done before.

So the first one is 2b1, which is the evidence or the measure specification is consistent with the evidence. And here we have seven who say moderate, and two who say low.

Then let's go onto 2b2, which is validity testing. Does validity testing demonstrate the measured data elements are correct and that they correctly reflect the cost of care resources provided.

There we go. And here again, we

have the same seven who said moderate, and two who said low.

Okay. Let's move onto 2b3, which is exclusions. Are the exclusions supported by the clinical evidence or analysis of frequency and distribution, and is it transparent. And again it's seven moderate, and two low. At least we're all consistent.

Now this one I'll bet you we'll get different ones though. This is the risk adjustment question for, and basically there is no risk adjustment here. Do the data or the rationale support no risk adjustment stratification strategy.

And so here we had four moderate, three low, and two insufficient. And I think the discussion reflects the uncertainty in the room regarding this and I think that's appropriate.

All right, let's keep moving along. We're doing great and everyone's going to catch their planes, and the work's going to

get done.

So we're going to 2b5, this is differences in performance. Data analysis demonstrates that the methods for scoring an analysis of the specified measure allow for identification of statistically, and again importantly clinically meaningful differences in performance, or there is overall evidence of less than optimal performance. And 2b5 was reviewed by Rohit.

DR. BORKER: So for 2b5, one of the points that was raised was it's not clear how unwarranted variation would be determined. So I guess the, one of the people who commented was to see if, whether any changes or variation in costs is because of difference in the patient mix again, or is it because of just a different quality of care that's being offered to the patient.

And the second comment I had is again around the statistical tests. Because these are cost data of, more than likely

they're not going to be normally distributed. So performing regular difference, like T tests and all may not apply here, so those are two of the concerns.

CHAIR PENSON: Other comments? You know, when I looked at this, it's sort of the same as the last one. The big problem I have here is not the statistical piece although, that's a well taken point that I hadn't considered with regard to normal distribution.

But really, what do I do with information, you know, if the west coast is different than the east coast. And I just don' know. I don't know what we're going to do with that bit of data.

It doesn't mean that it's flawed. I just, you know, assume it's less when I'm sort of between moderate versus insufficient.

On the one hand, the more I think about it, the more I come to the conclusion that we at least got to measure it, and look

at it, and go from there. So I don't know. Other comments? All right, I can hear you breathing people.

Yes, we're almost done, we're almost done. It gets easier from here.

Okay, let's just do disparities and then we'll be done with the scientific inaccuracy piece. And again we've had this discussion before as well, with the disparities.

You know, when you look this, if disparities in care have been identified, does the measure allow scoring specifications and stratification by race, ethnicity, other issues. And the reviewer for 2c was Dwight.

And I think in the end, we're right back where we were before with this. I think a lot of people are confused by this, and people are all over the map.

The question becomes is do you think you can really only, you know, stratify it by age, and if you think that you need to

stratify it by more than that in this, then you would say low or insufficient. Where if you think that's adequate, you would go from there.

So let's vote on this, what's left, and then we'll just go through the last pieces of it. So go to 2b5 and to summarize, this was the differences in performance and like I said there, Rohit raised some issues of statistics, and I think we are again at that issue of the clinical meaningfulness of this, and so it's, let's go ahead and vote. So we had seven who said moderate, and two who said insufficient.

MS. TURBYVILLE: We've got to do validity overall.

CHAIR PENSON: All right, thank you.

We've got to do, yes, validity overall. So this is, you know, based on sort of the whole nine yards, where you see with risk adjustment, validity testing and the

specifications piece.

So I think in the end, this is where you can take into account the business with the CPT code. So we had three who said moderate, five who said low, and one who said insufficient. And I think again the comments reflect that.

And disparities, we discussed this already. Disparities in care, if they've been identified, can you measure them or is it do you need to measure them. One, two or three? So five who said moderate, three who said low, and one who said insufficient.

All right, we're in the home stretch for real now. Let's talk about usability and then we'll talk quickly about feasability because I think we've been through those already.

Usability, 3a, which is basically the results can be reported to the public at large in national community reporting programs. And for this Steven, you were the

primary reviewer.

DR. CHEN: For this I actually had a little bit less problem with reporting to the public, because it is at such an aggregate level. Without the attribution, I'm a little less concerned about how precisely accurate it is.

And I do think that this is a problem that people should be aware of, that other regions do have more needle biopsies than they do. That's just my personal opinion.

CHAIR PENSON: Other comments? Go ahead Louise, don't be shy.

DR. WALKER: I'm losing my ability to think. Well I just, again I wasn't sure if that was important to know. And also with all the caveats of the problems, I especially was wondering if it was something that I'd want to report to the public at this current stage.

CHAIR PENSON: Let's remember that with this particular item, that this was the

one that throughout the day this is being tested by the RWJ contract, and NQF is sort of redefining the way they look at this.

You know, the downside with all these regional measures is, I don't know what to do with the information myself. I don't know if consumers, if programs will know what to do with it, so I sort of, I've been voting insufficient throughout the day on this, and I'm kind of still there with it.

Steven, I'm with you on that, with regard to the accountability is less worrisome, but even so I don't know what to do with it then, so.

Okay. Let's move onto 3b, the measure results are considered meaningful, understandable, and useful to the intended audience, both for public reporting and quality improvement. And this sort of gets to what I was just saying which is, you know, what do you do with this.

And I suspect looking at the

comments that everyone else was sort of in the same place which is, what to you do with this. I think Louise you were the primary reviewer for this. Were you, maybe not, am I screwing up again? Steven, I'm sorry, Steven.

DR. CHEN: And this is why I did put insufficient here, because I don't see how it's an actionable piece of information beyond consumers saying, hey, I heard we don't do a lot of needle biopsies, am I eligible? It's the only thing that they can do.

CHAIR PENSON: Yes, but now the question becomes is it's not actionable, but insufficient supplies it with more evidence that might be actionable?

DR. CHEN: Yes, I think if they put together some sort of detail that says, this is how one would expect to use it, I might be convinced that it was useful. But it's insufficient leaning towards low in this sense.

CHAIR PENSON: All right, that's

good. Other comments? All right, let's move onto 3c then, which is the one concerning clinical and construction logic, that the data and result detailed are maintained such that the measure, including it's clinical construction logic, can be decomposed to facilitate transparency and understanding. Steven?

DR. CHEN: This one I put low mainly because it used the word construction logic, and I'm not really pleased with their construction logic. Do I think it's decomposable, it is. I'm not particularly sure I want them to decompose it because I don't really like it.

CHAIR PENSON: So with regard to the construction logic, and again I'm just trying to generate thoughtful comments for the reviewers on the steering committee.

I mean it's straight forward having a biopsy although, maybe not straight forward as to where the biopsy is performed.

So that part strikes me as easy, so what sort of issues are you having with it?

DR. CHEN: I still have trouble with F&A just in general, because it is not typically used as something diagnosed heading towards a cancer diagnosis.

And so including that logic, but then excluding them from if they have had a second biopsy which would presumably be the, I got an F&A, it looks weird, now I want another biopsy so I know what I actually was supposed to have done the first time.

Or, you know, I thought it was cyst, I aspirated it with the F&A, and it turns out now there's a mass. So I think they excluded the wrong thing.

CHAIR PENSON: So are you inclined to say this is insufficient or low?

DR. CHEN: I could live with either. It certainly could be fixable I suppose, so insufficient would be fine.

CHAIR PENSON: Okay, good. Go

ahead.

DR. BORKER: Quick comment. I'm not sure if they excluded the second biopsy. I think they called the first one the entry or incident biopsy, but then they just ignored the second biopsy. And maybe the developers can tell us more on that, if that's true.

CHAIR PENSON: So Todd and Kevin, could you just talk a little bit about what you did with patients on this measure that had first versus second biopsies?

DR. LEE: It's only the first biopsy that is identified in a measurement period that is included. If the second biopsy occurred within seven days of the event those resources would be captured, but other than that, we only identified a single biopsy per individual.

DR. BORKER: Right. The question is, so for an individual if you have the first biopsy then entered into the database, but then within seven days they had the second

biopsy, what happens to that data, post that second biopsy? Is the data excluded from the analysis, or is the patient excluded from the analysis?

DR. LEE: After seven days the data is not included as part of the resource use associated with this episode.

DR. BORKER: Right, so for a case where they enter the system, within three days they had the second biopsy, what happens to the remaining four days?

DR. LEE: They are included. The remaining four days are included, but any resource use after day seven is not included as part of the episode.

CHAIR PENSON: Right. Okay, other comments?

MS. TURBYVILLE: If it's included as a resource, you still have the second biopsy?

CHAIR PENSON: That's my understanding, yes. Steven?

DR. CHEN: I think the other thing I didn't really like about the logic part, there are some thing's that are kind of flexible as to when you might get them, whether you get an MRI before, or after for an obvious breast cancer.

And so the window seems arbitrary enough and short enough, that things might fall outside the window that should belong in the window, just because at my institution getting an MRI within seven days is actually reasonably difficult, unless you personally make a phone call.

CHAIR PENSON: Okay, but what I'm hearing from you though is, while these are major concerns, they are addressable concerns?

DR. CHEN: Yes.

CHAIR PENSON: Okay. So and obviously, other comments about 3c? So 3d is not applicable, so let's vote on the usability and then we'll wrap up on feasability, and go from there.

So let's start with 3a, the measure performance results were reported to the public at large in reporting programs, either at the time of endorsement and maintenance review. And obviously we've talked about the ongoing issues here with RWJ, et cetera. Go ahead and vote.

(Off microphone discussion)

CHAIR PENSON: What's that?

(Off microphone discussion)

CHAIR PENSON: Okay, let's vote again.

(Off microphone discussion)

CHAIR PENSON: Now we're going to get done on time, I'm a machine. There we go. Okay, we have five for moderate, and four for insufficient. Okay.

Next is 3b, the results are considered meaningful, understandable, and useful to the intended audiences for public reporting and quality improvement.

(Off microphone discussion)

CHAIR PENSON: And here we've got one moderate, three low, and five insufficient.

Next is the clinical and construction logic that data results are maintained such that the measure, including the clinical and construction logic can be decomposed to facilitate transparency and understanding. Go ahead and vote on this. And here we had four moderates, one low, and four insufficients. And I do think that the discussion reflects that, that some people were more convinced than others.

Okay, so let's do the last three measure pieces fairly quickly. I think these actually will be quick, not just because we're tired but because they're relatively straight forward in this setting.

So 4a is the feasability measures. This is the one that the required data elements are routinely generated and used during care delivery and are going to be

captured.

And I think unlike the discussions we had with colon cancer and the earlier breast cancer treatment, because this is a sort of diagnostic work-up, we're probably going to catch most of this. We're not going to have issues with J codes, et cetera. So my inclination is that this is probably high.

4b, the required data elements are available in the electronic health records. Similarly I think that this is going to be captured as well, and again high.

Why don't I just move for acclimation here, does anyone feel that it's not high, for 4a or 4b? So that we're good at.

4c and 4d are worth a little discussion before we wrap up. So for 4c, this is the element susceptibility to inaccuracies, errors, and unintended consequences.

And again, this is administrative data so we don't have to repeat that. Steven, you're the reviewer here, were there any other

issues beyond it's administrative data?

DR. CHEN: I think the only other thing is that a lot of these biopsies are being done for unknown diagnoses by definition, and so you're going to get a lot of variation in ICD-9 codes.

And so that makes it a little more susceptible to inaccuracy, but not fatally so, probably. My biggest concern were the original concerns that Louise had.

CHAIR PENSON: Okay. And 4d is the data collection strategy can be implemented as demonstrated by operational use or the testing didn't identify barriers. And my inclination was with this was fairly straight forward, programming.

So let's vote on 4c and 4d, and then we can wrap up. So let's do 4c, this is inaccuracies, errors, and unintended consequences.

MS. TURBYVILLE: One more.

CHAIR PENSON: Keep voting people.

MS. TURBYVILLE: Your thumbs will be very strong after today. Running out of time. No?

CHAIR PENSON: Oh.

MS. TURBYVILLE: Okay. It just came unplugged.

CHAIR PENSON: Let's try it again.

(Off microphone discussion)

CHAIR PENSON: Yes. There we go.

MS. TURBYVILLE: Wow.

CHAIR PENSON: So for 4c, we had seven who said moderate, and two who said low with regard to inaccuracies and errors.

And 4d. There we go, nine. And here we had five who said high, and four who said moderate. So that takes care of that.

Before we go we have to do public comment. So operator, could you open up the lines to public comment?

OPERATOR: Certainly sir, I'd be happy to. Ladies and gentlemen for public comment, please press star, one. Again that

is star, one for public comment. There's no public comment at this time.

CHAIR PENSON: Thank you. Anyone in the room, audience in the room? Everyone's afraid to say anything because everyone wants to leave. Well before I turn it over to Sally, well Kevin, Todd, do you guys have anything you want to add?

DR. WEISS: For me it's just a thank you from the committee. You've done very thoughtful and respectful review of the work that was done by this project, and the workgroup, and the staff of the team. Really appreciate your time taken out to deal with these difficult and new types of measures.

CHAIR PENSON: I think I can speak for everyone in the room, that we're really grateful you were on the phone today, it really made this easier. I'm getting a lot of head shakes.

It really is good to be able to talk to the measure developers, and work

together to sort of identify the strengths, and identify the weaknesses which often can be improved. So thank you guys for spending the day. This is probably more painful for you two, than any of us.

And I personally as the Chair want to thank everyone else in the room. This is a lot of work and I appreciate it.

DR. POTTERS: So I just want to personally and have the minutes reflect, that we thank you for running the meeting officially.

CHAIR PENSON: I guess we're on time.

(Off microphone discussion)

CHAIR PENSON: Thank you, appreciate that. So Sally, I'll turn it over to you now.

MS. TURBYVILLE: I just want to echo what David said, as well as Kevin on the phone. Thank you, all of you so much for all your time in preparing for the meeting. And

then also, today in contributing your expert input and opinion.

I think you were very clear with your rationales. I think it certainly helped the measure developers, and certainly will help us at staff as we continue to move these forward. Next steps, Ashlie if you want to quickly?

MS. WILBON: So next steps, everyone submitted their votes today on the measures. We will be following up with the developers on some of the questions that you guys had, and we'll email that back to you.

What we've found with some of the other TAPs is, it seems to be a little bit more difficult over emails for people to re-vote, but that we'll probably end up sending out the information if you guys have any like, verbal, you know, responses or statements you want to pass forward in response to what they've submitted.

And we'll move that forward to the

steering committee, rather than having you guys re-vote, and we'll put your votes here and get the context that you rated the measure as is, and that the additional information submitted by the developer, you know, here's kind of what any other additional comments as it moves on.

So no follow-up conference calls. So just look forward to some follow-up emails from us, and hopefully wrap things up.

So again, thank you David for keeping us on track and keeping us entertained and awake, so we appreciate it very much. And he's actually going to be here two more days for the steering committee meeting, so appreciate that. Thank you.

CHAIR PENSON: I mean I get to be like Lou Potters, and I get to sit there and listen and sort of let other people talk. Thank you again guys.

(Whereupon, the meeting went off the record at 4:21 p.m.)

D

E

F

W

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Resource Use Project

Before: NQF

Date: 06-28-11

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