NATIONAL QUALITY FORUM
+ + + + +
RESOURCE USE PROJECT:

PHASE II

PULMONARY TECHNICAL ADVISORY PANEL

MEETING

+ + + + +
TUESDAY
JULY 19, 2011

+ + + + +

The Meeting was held at the National Quality Forum, Suite 600 North, 601 13th Street, N.W., Washington, D.C., at 8:30 a.m., Kurtis Elward and Janet Maurer, Co-Chairs, presiding.
PRESENT:

KURTIS ELWARD, MD, MPH, Co-Chair
JANET MAURER, MD, MBA, Co-Chair
GERENE BAULDOFF, PhD, RN
KATHRYN BLAKE, PharmD
DALE BRATZLER, DO, MPH
ZAB MOSENIFAR, MD
LINUS SANTO TOMAS, MD, MS

MICHAEL SCHATZ, MD, MS
RICHARD STANFORD, PharmD, MS
NQF STAFF PRESENT:
CARLOS ALZOLA, Statistical Consultant to NQF
TAROON AMIN, MPA, MA, Senior Director

HEIDI BOSSLEY, MSN, MBA, Vice President of Performance Measures
LAURALEI DORIAN, Project Manager

SARAH FANTA, Research Analyst

ANN HAMMERSMITH, NQF General Counsel

SALLY TURBYVILLE, MA, MS, Consultant to NQF

ASHLIE WILBON, MPH, BSN, Senior Project Manager

ALSO PRESENT:

BENJAMIN N. HAMLIN, MPH, NCQA Director of Performance Measures

SHEILA HEITZIG, Director of Practice and Policy, American Academy of Allergy, Asthma and Immunology

THOMAS LYNN, MD, Ingenix

CHERI ZIELINSKI, Ingenix

I-N-D-E-X

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

8:35 a.m.

MS. DORIAN: Good morning, everybody. I think we're going to get started now. I'm Lauralei Dorian, and I think I've met most of you this morning actually, and if not, you've probably received some emails from me over the last couple of weeks.

I'm actually quite new to NQF and I'll be working as a project manager on the remainder of the resource use program of work.

And I would like to say thank you so much on behalf of everybody at NQF and my team members for being here today. We really appreciate the time you've taken out of your busy schedules to be here and provide your expert opinions.

It's a really crucial part of this project and we're really appreciative of the work that you've put in and that you will put in today. So thank you for that.

And I'd also like to say a special

thank you to our two chairs, Dr. Elward and Dr. Maurer. Thank you. Thanks for the work you've put in and for the leadership you will provide today.

So I think the first thing that we'll do is I'll have the rest of our team at NQF introduce ourselves and then I'll hand it over to Ann Hammersmith, who's our general counsel at NQF, and she'll go around and have each of you introduce yourselves and do your disclosures of interest, which is just a regular part of what we do, particularly when committees are meeting for the first time. And then we'll go through a brief PowerPoint presentation that we've prepared for you and then we'll hand it over to your co-chairs to do their welcomes and thoughts for the day and take it over from there and lead the day.

So I'll have Ashlie start, I guess.

MS. WILBON: Good morning, everyone. Thanks for joining us. And again,

I'm the Senior Project Manager on the project and happy to have everyone here this morning.

MR. AMIN: Hi, good morning, everyone. My name is Taroon Amin. I'm a senior director here at NQF.

Thank you for all your hard work. I know there is a lot of work that went into reviewing these measures, very complex, and we look forward to this morning's discussion.

MS. FANTA: Good morning, everyone. I'm Sarah Fanta, Project Analyst. Looking forward to working with all of you today.

MS. TURBYVILLE: Good morning, I'm Sally Turbyville. I was previously working with NQF on this project for the past year. I've transitioned into a consultant role.

And in complete agreement, we're very appreciative of all the hard work and looking forward today to your opinions as we work through the measures.

MS. BOSSLEY: I timed this just

right, didn't I? Heidi Bossley, Vice President of Performance Measures, and again to reiterate what staff has said before, but we truly appreciate all of the hard work. And we know it's a lot of work we've asked you to do. So thank you very much.

MS. HAMMERSMITH: Good morning, everyone. I'm Ann Hammersmith, NQF's general counsel. What we'll do now is combine introductions with the conflict of interest disclosures.

If you recall, several months ago we sent you a form that we asked you to fill out, which you did, and then we reviewed them in great detail.

And what we ask you to do today is to orally disclose anything that you revealed on the form or that has happened since you filled out the form that you believe is relevant to your service on this committee.

Just because you disclose something does not mean you have a conflict of

interest. The idea here is openness, transparency with each other and so on.

We don't expect you to recount your CV to us. We know that you're all extremely capable people -- that's why you're on the committee -- but just note things of relevance to your service on the committee.

Excuse me. I'm glad you're pulmonary doctors because I seem to be having a little difficulty here.

We are particularly interested in your disclosure of research support, consulting relationships, and grant funding that's relevant to what's before the committee. I also want to remind you that you sit on the committee as individuals. You are not a representative of your employer or of any organization that might have nominated you for service on the committee.

So with that I'll ask each of you to identify yourselves, tell us who you're with, and then if you have any disclosures.

So I think I'll start right here.

CO-CHAIR ELWARD: I'm Kurt Elward. I am in practice in Charlottesville and I also have a clinical professorship at VCU in Richmond and a research appointment at UVA.

I'm involved in a couple of evaluation projects where we're trying to help physicians implement quality improvement guidelines, and those are funded in part by NIH and in part by some pharmaceutical support through a foundation, to make sure that there's no direct handling of money and things like that. So, is that it? Yes, thanks.

CO-CHAIR MAURER: Hi, I'm Jan Maurer and I'm a pulmonologist. I live in Phoenix, Arizona, and I work for a company called Health Dialogue, which is a disease management company that creates and implements programs for people to better manage chronic diseases, of which asthma and COPD are two.

I also have a clinical professorship at the University of Arizona in

Phoenix, and I am on the Quality Improvement Committee of the American College of Chest Physicians and I sit on the Board of Trustees of the Chest Foundation. I think that about covers it.

DR. MOSENIFAR: I am Zab Mosenifar. I'm an academic pulmonologist at Cedars Sinai in UCLA. I've been there for about 33 years or so.

I have a large fellowship program, actually the largest in the country perhaps, and I mainly do research in the COPD area. My research is funded by NIH and some pharmaceutical industries.

My research right now involves mainly lung volume reduction surgery in non-invasive forms via stents and various devices, and I also work on use of growth hormone on patients with COPD as well. I have no conflict.

MR. ALZOLA: I'm Carlos Alzola, and I'm an independent statistical consultant.

And I was hired for this project and I put together these assessment worksheets to help you evaluate the measures.

DR. BLAKE: I'm Kathryn Blake from Nemours Children's Clinic in Jacksonville, Florida. I've been doing asthma research for 25 years.

I've research support from no pharmaceutical companies, only from the American Lung Association and NIH, and I consult for a project in Missouri looking at pharmacy and Medicaid claims. That's it.

MR. BRATZLER: I'm Dale Bratzler. My contact information has actually changed from what's in the materials. I'm actually with the University of Oklahoma in the College of Public Health. I'm a professor and associate dean there.

I have no conflicts to report. My experience with pneumonia, particularly over the past almost 12 years now, has been coordinating the National Pneumonia Project

for the Centers for Medicare and Medicaid Services.

DR. BAULDOFF: Hi, I'm Gerene Bauldoff. I'm a clinical professor at Ohio State in the College of Nursing. My focus is COPD and pulmonary rehabilitation.

I have worked as a lung transplant coordinator in Pittsburgh. I've worked as the rehab coordinator at Pittsburgh on the NET Trial, and I'm on the board of directors for the American Association of Cardiovascular and Pulmonary Rehabilitation.

DR. SANTO TOMAS: I'm Linus Santo Tomas. I'm from Milwaukee, Medical College of Wisconsin. I'm a clinician educator and a pulmonary as well as an intensive care specialist, and no conflict of interest.

DR. STANFORD: I'm Richard Stanford. I'm actually at GlaxoSmithKline. I'm in the Department of Health Outcomes.

My background is, I have a clinical pharmacy degree as well as a Masters

in preventive medicine. I have been doing observational research for the last decade mainly in asthma and COPD, so I have a lot of experience I do believe in large databases such as what we've looked at.

I don't have any conflicts outside of the fact that I work for pharma.

DR. SCHATZ: I'm Michael Schatz. I'm Chief of the Department of Allergy at Kaiser Permanente San Diego.

I serve as the Co-Chair of the Joint Task Force on Quality Measures, which is a joint task force of the American Academy of Allergy, Asthma and Immunology and the American College of Allergy, Asthma and Immunology.

I've been involved in database research at Kaiser, including some specific research looking at quality measures, although not resource measures. And some of our research has federal funding, but I do have funding from GSK, Merck, Genentech and

Aeroquin, in terms of research funding, and serve as a research consultant for Merck, Amgen and GSK.

MS. HAMMERSMITH: Okay, thank you. And there's no one on the phone, Sarah? Okay.

All right, thank you for those disclosures. Do any of you have any questions or anything you would like to discuss with each other based upon the disclosures this morning?

(No response.)

Okay, thank you. Have a good meeting.

MS. DORIAN: Great. Thank you, Ann. Thanks, everybody, we appreciate that. Now what I'll do is just briefly go over some of the logistics for today. I'll also actually have the people in the back introduce yourselves, if that's all right.

MR. HAMLIN: Good morning, everybody. My name is Ben Hamlin. I'm Director of Performance Measurement at NCQA.

MS. HEITZIG: Good morning, I'm Sheila Heitzig and I'm the Director of Practice and Policy for the American Academy of Allergy, Asthma and Immunology, and the staff of the Joint Task Force on Quality Measures for the American Academy of Allergy, Asthma and Immunology and the American College of Allergy, Asthma and Immunology.

MS. DORIAN: Thank you. So as I was saying, I'll go over some of the logistics for today and then just do a brief PowerPoint presentation to sort of situate you to where we are in our CDP process.

Then I'll hand it over to Ashlie at that point, who will briefly touch upon some of the subcriteria that you'll be discussing. And if at any time you have any questions, please feel free to jump in.

This is a new process for us, particularly in relation to resource use measures so we want to learn from your feedback all the time and we're really open to

listening to what you have to say about the process thus far.

So as you know, we have five measures that we're discussing today, three from Ingenix and two from NCQA. And we're fortunate that we have our NCQA team here in person and we'll have Ingenix on the phone, so that resource is there for you to use.

Usually what happens is at the beginning of each measure discussion, the measure developers will give a very brief introduction of the measure as to what the measure is about to remind you. And then they'll be there throughout your conversation to answer any questions that you might have, so definitely use that. It's there for you to use throughout the day.

We also have Carlos here for the first half of the day, I believe, and he's provided his statistical analysis already and he will give a brief overview as well if you'd like to him to, and he'll be really important

particularly during the scientific acceptability component of our conversation.

And just a reminder that it's important for you to speak into your mics because it is being recorded. And Sarah or maybe Ashlie later will go over the voting device that we'll be using.

DR. STANFORD: Can I ask a question?

MS. DORIAN: Yes, sure.

DR. STANFORD: Have we narrowed it down to these four? Is that why we're only going to concentrate on these four?

MS. DORIAN: There are five.

DR. STANFORD: Excuse me, five. Pneumonia and then the two COPD and so that's þ

MS. WILBON: So yes, we were going to discuss those in the presentation but now's a time better than ever.

So we initially had, with the other TAPS, we've actually only gotten through

about five or six measures in one day.

These measures, as you already know, are really intense and they take a long time to get through. So we already kind of knew at the in-person meeting that we wanted to narrow the agenda down to a subset of the measures. Subsequent to developing the agenda, ABMS had informed us, actually just a couple days ago, that they were withdrawing their measures from the project. They've been submitting measures in other topic areas.

As you know, this is the fourth TAP that we've gathered for this project. We had a cancer TAP, a bone/joint TAP and a cardiovascular TAP and then obviously you guys are the pulmonary TAP.

And the evaluation process for those measures, they identified several issues and they felt kind of through the process that it was probably best that they withdraw their measures and focus on refining them and

putting some more effort into testing them.

So the feedback that we've gotten up to this point from the experts is that everyone is really excited about the measures, that they have got great concepts and they're a great start on the measures but they did need a little bit more work.

That said, we're focusing today's meeting on the Ingenix and NCQA measures which we're going to actually evaluate. We will have a follow-up conference call on the 2nd. We had already scheduled two additional conference calls.

We're going to cancel the one on the 17th, but the call on the 2nd we're going to keep and we're just going to use that call to discuss the ABMS measures, even though we won't be formally evaluating them and rating them and putting them through the process.

We do think it's valuable to have your input to forward on to them so they can use that in their refinement process and

improving the measures going forward.

Our hope is that they would use this feedback to submit them back to NQF at a time when they feel like their testing and their measure specifications are a little bit more well-refined.

So we do still want your input on those measures, and the analysis that you've done on them thus far, we will be compiling that and forwarding it on to the development team at ABMS.

But the timing of everything and the process thus far -- just that's kind of where we are at this point. Your work was not in vain, we will be using it and actually devoting a whole conference call to kind of feedback and put on those measures.

So does anyone have any questions about that? But thank you for asking. Okay.

MS. DORIAN: Okay, this is just a brief slide to remind you where we sort of are in our consensus development process.

We have completed our first two steps and we're on to Step 3 which includes recommendations from the Steering Committee with -- of course based upon your expert input today.

And then there are many other steps to follow that, that can potentially follow it, including public and member comment, member voting, our CSAC decision, which is our Consensus Standards Approval Committee, which is a standing committee that's responsible for the oversight of NQF processes and procedures, and so on and so forth.

So that's just to give you sort of an idea about where we are in this process so far. And this is just a representation, so that you can see that as well.

Now although resource use measures have been around for a long time and in many cases have been around a lot longer than quality measures, they are new to many of you,

to most of us and certainly to our NQF process.

And so it was really important at the beginning of this project that we have a really clear understanding and definition around what for this project we're considering resource use measures.

So this is what the Steering Committee, is that right? -- the agreement that they came to at that time that you see in front of you.

And so we have, as you may know, we have completed Cycle 1 of our CDP already, which included non-condition specific and cardiovascular and diabetes measures.

And we're now on to Cycle 2, which of course includes pulmonary measures and cancer and bone/joint measures. So we've been quite busy. We've been moving along at a pretty rapid pace.

We had a two-day Steering Committee conference a couple of weeks ago

followed by a phone call and then we've also had a cancer TAP meeting and a bone/joint TAP meeting, and so here we are today. It's just our timeline.

Another reminder that in terms of the measure review process, we're the first ones to receive those measures and we just review them for completeness and adherence to our guidelines, and then they go to Carlos, our consultant, for review, and then on to you.

And I won't spend too much time on this because I think you already know what your role is and I know that you had a conference call earlier to talk about this as well.

But just to remind you that you've been individually selected for your expertise, and so it's really important that today you do that deep dive, based upon your expertise, into each of the subcriteria, because that will be extremely important into feeding into

what the Steering Committee eventually decides about endorsement for a measure.

So I'll stop here and see if anybody has any questions about logistics for the day, any other questions about what I've gone over, anything about timelines moving forward, and if not, I'll hand it over to Ashlie.

MR. AMIN: Could I just add something, Ashlie, before you move on?

I just wanted to make a note for the TAP that really a deep dive in the scientific acceptability portion of the measure is really valued as the measure moves on to the Steering Committee, mainly because you're the clinical experts in this area and you're the methodologists in this area.

So the Steering Committee will weigh heavily on the discussions, the deliberations and the voting across the measure, but particularly in scientific acceptability.

So I just wanted to highlight what Lauralei mentioned. I just wanted to make sure that was highlighted in the discussion.

MS. WILBON: So everyone has at this point has hopefully looked at their measures and our criteria and is somewhat familiar with it.

But I just want to highlight a couple things. Now this is our fourth TAP meeting and we're getting pretty good at this. Unfortunately, it's our last one, we don't keep going, but we've got all this wisdom to pass onto you.

So I'll just kind of focus on the areas where we tend to have the most questions and issues in going through the criteria.

Our discussions today will be very systematic to the best of our ability, and we'll go through each of the subcriteria individually and have a hopefully structured kind of conversation focused on the intent of that criteria.

And as you know, there's four criteria, importance to measure and report, which is really focused on determining whether or not the topic that they've chosen is important to measure.

And what we've found actually up to this point is that the call for measures for this project and the scope of this project was very specific as it was. We selected the conditions.

Just by virtue of the way we set the project up the measures are important, so that discussion is generally pretty brief. We try to keep it as brief as possible.

We're going to have the co-chairs lead the TAP through kind of a discussion of each of the subcriteria and get any overall input, but really reserve the bulk of the discussion and our time in terms of time management in the scientific acceptability section. So we will be discussing importance but it'll be an abbreviated discussion.

The scientific acceptability criteria, the goal is to determine whether or not you can make valid conclusions about resource use, whether or not the measure's reliable and valid.

The usability is focused on how usable the measure is. Are the results and the information that you get out of the measure, are they usable for the intended audience?

And then feasibility goes to how much burden there is with -- it's more focused on implementation. So what would it take to implement the measure? Is it realistic?

Is it feasible for someone to pick up the specifications as written in their system or whatever level of analysis that it's specified at and implement the measure? So that's what feasibility is focused on.

So within importance, I'm just going to do a very high level skim here, we're going to be determining whether or not the

measure focus or the topic of the measure addresses a national health goal or priority area and whether it's high impact, whether or not they've identified a problem area with opportunity for improvement, whether or not the purpose and objective of the measure or the intent of the measure that they have identified is clear and whether or not the resource use categories they've selected makes sense.

So based on the topic they've selected and the service categories that they've selected, does it make sense based on what they say they're measuring?

Scientific acceptability again is focused on determining how reliable and valid the measure is.

And then we do have kind of a dangling subcriteria out there on disparities. What we found in our discussions with disparities up to this point with other TAPs and the Steering Committee is that obviously

disparities are really important.

NQF has done, we're actually gathering a committee now to kind of talk about how disparities can be measured and how we should be supporting disparity measurement in quality improvement and so forth.

But, particularly because these measures use administrative data, we've found that there are some limitations in measuring disparities and reporting that out.

So to this point we've been kind of framing this disparities discussion around, is the measure structured or constructed in such a way that if the data was available that they would be able to report out stratified disparity information?

And again, it may be framed, depending on the condition, particularly for cardiovascular conditions you may feel like, our TAP felt like it was very important that, you know, disparities can be a really important of how resource use is distributed.

So again, there may be some condition-specific things around how disparities are reported for resource use measures, but in general we've found that there are limitations inherently in the way the data is gathered for these types of measures that makes reporting disparities difficult. So just kind of keep that in mind in your general context.

So again, our discussion around scientific acceptability is focused on identifying whether the measure is reliable and valid.

And we just wanted to kind of go over in a little bit more detail about what your high, moderate and low ratings actually mean when you're saying that you're evaluating the measure.

For a high rating of reliability, it means that all the measure specifications are unambiguous and likely to consistently identify who's excluded from the target

population, that it's clear how the resources and costs are being measured and how to compute the score.

That they have evidence of reliability of both the data elements, so in this case it would be the administrative claims data or the resource use categories and the measure score.

And that the measure score, that they've identified an appropriate method and scope of reliability testing for that. So it's a pretty high standard; it's pretty unambiguous. They've done everything they could possibly do to make sure and demonstrate that the measure's reliable.

For validity, again, very high standard; that the measure specifications are consistent with the intent; that they describe any importance to measure criteria; and again, evidence of validity at both the data element and the measure score level; again, that the measure score is appropriate; the measure

score validity testing is appropriate method, scope and within acceptable norms, and that they've identified any threats to validity and have addressed them in some way.

So a moderate rating for reliability, one step down, again the specifications need to be unambiguous, but that they can demonstrate reliability of either the measure score or at the data element level. So it's an either/or for the medium or moderate score.

And similarly for validity, the specifications need to reflect the intent and the focus of the measure and they can demonstrate validity either at the measure score level or the data element level, or that they've done systematic assessment of face validity.

So face validity is the kind of threshold that we ask for in terms of validity testing when they submit measures.

For a low rating of reliability,

you may find that some of the specifications are ambiguous. There's potential for confusion in identifying who's included and excluded.

It's not totally clear how the resources and costs are being measured or how to compute the score, and that they use an inappropriate measure of reliability or that the testing that they've done, you don't feel that the measure is totally reliable.

For validity, again the specifications don't support the evidence that they cited for the intent of the measure and that empirical testing that they've done to demonstrate validity is either not sufficient or they have not done measure score, demonstrated validity at the measure score or the data element level and they have not identified threats to validity. An insufficient rating, I think we tend to get a little bit of confusion around a low score and insufficient. Insufficient I would say, and

I'll open this up to my colleagues to help me on this one, insufficient we found is either they haven't submitted anything for you to be able to demonstrate whether it's high, moderate or low.

Maybe you're missing information, there is a particular statistic missing that you would like to see and you don't have the information you need to determine whether it's high, moderate or low.

MR. AMIN: Or I would just add that the method they used was inappropriate in the view of the TAP, so it's either not provided or is actually inappropriate. So that would be the sort of threshold for insufficient.

MS. WILBON: Any questions about that? Okay.

MR. AMIN: Can I just add something extra?

MS. WILBON: Sure.

MR. AMIN: Just another thing that

I would just add to keep in mind. It's really important to keep in mind that face validity is an acceptable standard in this process.

I know that sometimes we want to move beyond that for understandable reasons, but that is an acceptable standard for where we are with resource use measures broadly. All the measures will go into a three-year evaluation. All the measures at NQF go through a three-year maintenance process where measure developers will have to demonstrate how the measure's been used and sort of the validity of the measure as it's been in use for three years.

So I just wanted to keep that in mind that face validity is an acceptable standard for this process right now.

MS. WILBON: So some of the thinking that the Steering Committee did in the first year of this project was around how to, one, construct our submission form so that we could receive or take in the information

and the specifications on resource use measures in the right format so that it could be evaluated by all of you.

That in doing that they identified kind of five modules, if you will, or five kind of sections of a resource use measure. And you should have actually seen this reflected in the submission form.

So we've divided the resource use measure into these different categories. In your discussion of scientific acceptability that you would kind of look at how they specified each of these sections to determine how well the measure is specified, if it's clear and if they're actually measuring what they say they're measuring.

Those five modules that we identified was the data protocol section, so these are the beginning steps of how to get the data ready for implementing the measure, the clinical logic, which is where obviously we're hoping that you guys will take a deep

dive.

This is where they're identifying the actual patient population, how they're including and excluding patients by coding, ICD-9 codes and so forth.

The construction logic, this is usually any of their temporal mechanisms, so when the episode, for instance, would begin or end, age ranges, different kind of -- not necessarily around the actual patients' clinical characteristics, but more so around how the measure and how the risk adjustment interacts, how the clinical hierarchies and stuff, how all that interacts together to form the measure.

In the adjustments for comparability, we've actually pulled out how any stratification methods, the risk adjustment method and the costing method. So again, areas that we're hoping that you would also take a look at.

The reporting guidelines are, for

resource use measures we identified, we realized early on that uniquely for resource use measures it's very important how the measure is going to be reported out, particularly in benchmarking and selecting peer groups and assignment of peer groups and the attribution role. So that is going to be in the reporting section.

We did early on identify this as a section along with the data protocol that could be submitted as guidelines or specifications. If it is guidelines, you should have seen in the measure evaluation form, it would say Guidelines and then colon.

What it means if it's Guidelines is that there is some flexibility for the user, so it's not baked into the measure, particularly for attribution.

For instance, if someone wanted to pick it up and they have a different physician population that they want to use a variation on the attribution rules that they would have

the ability to do that, but that the specifications are actually the core of the measure and that part does not change.

So just kind of keep that in mind as you're reviewing the measure. Those things that are guidelines are considered flexible for the user. Any additions on that?

MR. AMIN: I would just add in the process of feedback to the developers, if the TAP does feel that the guidelines that were submitted are inappropriate, additional guidance would be welcome, I'm sure, to the measure developers on proper or other attribution rules that seem more appropriate.

MS. WILBON: So I'm going to just quickly skim through some of the subcriteria for scientific acceptability and then we'll go ahead and hand it over to the co-chairs. I know you guys are eager to get started and stop hearing me talk.

So 2a1 of scientific acceptability focuses on whether or not the specifications

are clear so that it can be implemented consistently.

So if I gave the specifications to a system over here and a system over here, if they, given that they had the same data available to them, is the measure specified in such a way that they would both implement it in the same way? It's clear that it's general understanding of how the measure should work.

2a2 focuses on whether or not the reliability testing has demonstrated that the measure is repeatable and that the measure score is precise.

2b1, and all the 2bs focus on the validity of the measure, 2b1 is focused on whether or not the information they've submitted is consistent with the intent of the measure and submitted importance and whether or not the population that they have selected is supported by their opportunities for improvement and so forth.

2b2 is focused on whether or not

the data elements are correct and the score reflects the cost of care and resources, and that the validity testing has demonstrated that the score and the results of the measure, you can adequately distinguish higher and lower resource use.

2b3 is focused on exclusions, whether or not the exclusions they've selected are accurate.

2b4 looks at whether or not the risk adjustment approach adequately demonstrates discrimination and calibration.

And 2b5 is focused on whether or not there are differences in performance and that they are statistically significant and practically and clinically meaningful.

2b6 focuses on whether or not, particularly for measures that specify different data sources, these measures all use administrative claims data. So we're going to kind of take that one out of the pot for these discussions and kind of make it an n/a.

There's not really a reason to kind of discuss over that when there's only one data source that's being specified.

And then disparities we talked about. For usability, we're asking in these subcriteria whether or not the results are reported to the public. And if they're not currently reported, if they have a plan to report them or how it's currently being used by the public in some way.

3b asks whether or not the results that they've demonstrated for the use of their measure are meaningful, understandable and they that they will be useful to the public in quality improvement and public reporting.

And 3c asks you to determine whether or not the specifications are transparent and whether or not they're understandable. So that's obviously a clear one, important one. I'm not going to spend time on that.

Okay, going to kind of skim over

some of these. Feasibility, this is another discussion that we're going to kind of let the co-chairs lead the group through. It can generally be a pretty high level discussion.

We've also found that sometimes usability and also the feasibility discussion, the same discussion will generally apply to all the measures from an individual developer.

From NCQA or from Ingenix, their underlying methodology and construct of the measure's pretty consistent throughout all the measures. So if you find one measure feasible it will probably apply to the other measures from that developer.

So where possible, where we can find efficiencies in our time and so forth. If you feel like that discussion you had about a previous measure applies we can kind of say okay, is there anything different about this measure that warrants discussion on feasibility for any of these items? And we'll kind of keep it going in that way.

4a and 4b, generally we tend to skip over a little bit or just do a very quick rating of. It asks whether or not the data elements are routinely generated. Obviously, admin claims data are routinely generated. And 4b asks whether or not the required data elements are available electronically, which admin claims data are. So we won't spend time on those two.

4c asks you to think about whether or not there are errors or unintended consequences by implementing the measure and whether or not the developer has identified any of those and identified ways to monitor or minimize those for the implementer.

And then 4d is focused on whether or not there is a feasible data collection strategy that can be implemented by using the measure. So is it feasible for someone to do all of the tasks around collecting the data and implementing the measure in their particular environment?

Anything else to add about that? Okay. So Lauralei already went over a lot of this stuff, but for the meeting today, we're going to have the developer introduce each measure before you jump into your discussion, to give you, kind of just get you warmed up to the intent of the measure, the general methodology that they've used.

We have Carlos here to kind of do a brief overview of his analysis of the scientific acceptability section, particularly around reliability and validity testing. So we can call on him, and obviously throughout that discussion as well if you have any specific questions.

The developer for Ingenix will be on the phone when we get to those discussions and they'll be available throughout the discussion.

And then the TAP obviously will be going through each of the subcriteria one by one and rating each one at the end.

So, as you know, we did distribute that table that had all the assignments for each person to kind of lead the discussion of part of the measure.

We found that because these measures are so big, rather than just assigning one whole measure to someone it can be a little bit exhausting to go through a whole measure and kind of lead that discussion, so we've tried to break that up a little bit.

And Kurt and Janet will call on you when we get to that measure and you can point out anything that's important for the TAP discussion, anything that you identified in your own analysis. You can bring in some of the feedback that was submitted in the preliminary discussions from your colleagues to help guide you through that discussion as well.

Just a little bit about your voting device. What we've found, I think with

some of the other groups is that we'll talk about, once we get to importance and you guys have wrapped up that discussion we'll pause for about five minutes and we'll go and rate each of the subcriteria with your voting device.

We'll have the voting slides displayed on this screen up here in front of us, and as you vote, your results are captured in real time and there'll be a display of the distribution of the different ratings so you can kind of see what will be passed forward to the Steering Committee.

And I think what we've done is all importance through scientific acceptability, it tends to be quite long and lengthy. So we can break that up a little bit.

Maybe you want to get through reliability and vote on those two subcriteria and then do all of validity as a chunk, or you can decide how you want to break that up depending on how the discussion goes and how

tired people are getting.

So this slide will indicate for you which button you hit. Usually it's 1 for high, 2 for moderate, 3 for low or 4 for insufficient. You shouldn't need to hit send, but if you make a mistake, before you hit send, you can hit the triangle button with the little exclamation point in the middle of it, type in your new score, and then hit send.

So if you mess up, let us know. We can walk you through that, but it's pretty simple. We'll ask you to point your remote towards the laptop here that's up on the box. That little thing sticking out is the sensor that picks up your voting device.

So kind of point towards there and it picks everything up. You'll have 60 seconds to vote, so no pressure, but by the time we get to vote everyone will be ready to vote and you'll know what you want to put in, so don't feel any pressure by the time there.

So that's it. Thanks for all of

your patience in listening to that.

CO-CHAIR ELWARD: So it's going to be a majority vote? So how does that work?

MS. WILBON: Yes. So the TAP ratings, it's more of a -- the Steering Committee will make the final vote on whether or not the measure gets recommended, but your input on how good or bad you feel about the measure is -- they really rely heavily on your input on that.

So depending on the distribution of highs, moderates, lows, that's kind of what kind of gives the Steering Committee an indication on where the discussion needs to be at their level and whether or not it should be recommended. So does that help clarify it?

CO-CHAIR ELWARD: Okay, thank you.

MS. WILBON: Okay. At this point I'll go ahead and hand it over to your co-chairs to get you guys started and we can -- I think we're starting with the NCQA measures.

DR. BAULDOFF: I have a quick

question.

MS. WILBON: Sure.

DR. BAULDOFF: Hi. On measure 1608, which is an Ingenix measure, the ETF or -- I can't. Yes, that their methodology attachment to the 1608 was actually for asthma. Is it the same or is there supposed to be a separate one for COPD? It was in the PDF that we --

MS. WILBON: Right. Was it this, I'm just trying to, this document here would be --

DR. BAULDOFF: I have it right here if you need it.

PARTICIPANT: It's included in the --

MS. WILBON: Was it that ETG like their general methodology?

DR. BAULDOFF: Yes.

MS. WILBON: Okay. Yes, that was probably an attachment mistake and I apologize; we didn't catch that. Yes, but

generally it's pretty similar for across their measures, although it may be a little bit different.

They'll be on the phone so if there were any kind of mis-attachments on that they can clarify that during the discussion.

DR. BAULDOFF: Okay.

MS. WILBON: Okay, thank you.

DR. BLAKE: I have one quick question. You may get to this in a minute but it's really for Carlos.

I was curious as to what you had access to when you did your statistical analysis. What part of the data, that sort of thing?

MR. ALZOLA: I have access to all of the same information that you have. I didn't do any specific -- I didn't do any hands-on analysis. I look at how they presented their information and evaluate whether it's sufficient, it's appropriate for the goal at hand and sufficient for us to

evaluate the goodness of measure. Thank you.

MS. WILBON: And it's perfectly acceptable during the discussion if you feel like something is missing or you don't have the information you need, to ask the developer for that. The staff is here to follow up with the developer as needed to get it.

If you feel like you need additional information before you can make a final judgment, we will facilitate that so that you can get any additional information you need throughout the process.

CO-CHAIR MAURER: Thank you, everyone. We would also like to welcome everyone to this panel meeting.

This is really exciting: to move into a new area of measures, and I think a group of measures that are going to be really important as value-based care delivery rapidly becomes dominant across the country. So what we're going to do is we're going to trade off. One of us is going to lead one measure and

then the next, and Kurt is going to start with the first measure and we're going to call on the people who are assigned as noted.

But we are going to lead the importance piece of it first and then talk about feasibility toward the end. To start, we're going to do two measures this morning. We're going to do asthma and COPD.

We'll try to make the conversation run smoothly and I'm sure you'll help us with that. To begin the first measure, I'm going to turn it over to Kurt and I think he's going to have NCQA.

CO-CHAIR ELWARD: Yes, thank you very much. It's really a pleasure working with you all. I know several of you, and I appreciate the great amount of work you've done.

I've noticed a lot of people not only contributed on their assigned measures but other measures as well. Given the amount of material, that evidences a great amount of

interest and dedication so thank you very much.

So as Janet said, we'll start with 1560 first and then move through these. And as we go through the agenda, if there's adjustments that need to be made, things that can make things go smoothly, we'll be working on that and I appreciate your thoughts also.

So if we can start with 1560 perhaps we can have NCQA talk first about the overall measure and get an overview of that.

MR. HAMLIN: I'm actually going to introduce both of our measures because the majority of the methodology is identical for both measures.

Really it's the clinical condition and the denominator population that differs between the two. The NCQA measures of Relative Resource Use are a total annual so it's all services for members identified with a clinical condition. There's no attribution of specific services to the disease state

itself.

They use a standardized pricing methodology and they aren't risk-adjusted, so we use the standardized pricing based on a price list that I believe you all had access to that are generated for us annually that use a combination of different sources, using Medicare data and clinically -- I'm sorry, and adjusted using a number of other different sources from Ingenix and others to provide what we call sort of a standardized resource use that allows us to compare health plan to health plan, that we use them for health plan comparisons.

We've been collecting data on these measures now for five years. The last two years, this is Year 2 of public reporting of the results of this data.

So these measures are collected and publicly reported through our Quality Compass module. So it's available to subscribers of Quality Compass.

Specifically for asthma, the clinical population uses a very similar population of that as defined with the HEDIS asthma measure, and we only report the results alongside of that HEDIS asthma measure.

So our equation for value includes the Relative Resource Use paired with the quality measure, so we're reporting both of these results together as part of that public reporting.

I'm going to leave it there and if there're specific questions as we go through I'll try and answer them then without overwhelming you.

CO-CHAIR ELWARD: Any questions for NCQA right now?

(No response.)

Great. Well, in terms of the importance it seems like, I think the importance of the measure is extremely pertinent.

In terms of going through the

exact criteria, the overall cost of care of asthma is huge. I think it's well documented just how much opportunity there is. And so it seems like on many scales, the measure's very important. Any other thoughts about that? Well, perhaps we can go to Kathryn. Would you like to start us off on scientific acceptability?

MS. WILBON: So sorry. Real quick before we move on to importance, I mean before we move on to scientific acceptability, even though the discussion may be brief, we do actually have to vote on importance.

So if we could just have everyone, we'll just on the screen, we'll show each of the subcriteria and we'll have you guys vote high, moderate, low based on your evaluation of the measure, if you feel like they've demonstrated each of these criteria.

DR. MOSENIFAR: Are we going to vote again after we hear from the reviewer, or this is just the last vote?

MS. WILBON: No, this is just on the importance criteria. You'll have plenty of, you'll have lots of other votes left on the scientific acceptability, usability and feasibility. So this is just the first chunk.

MS. TURBYVILLE: Yes, the importance of this problem, you know, it's not the importance of the measure. It's the importance of this problem as a health care cost in the country.

DR. MOSENIFAR: Yes, thanks for correcting that.

MS. WILBON: So I also just wanted to point to you real quickly in your folders, that table that I've actually got up there will help kind of guide you through the discussion as well throughout the day so you can see which subcriteria that we're going to be talking about, this kind of two-column table.

And we'll have it up on the screen throughout the day as well to help you get an

idea of what we're evaluating.

So 1A is the first subcriteria we'll vote on. It's asking whether or not the measure focus addresses a specific national health goal or priority or demonstrates a high-impact aspect of care. So we'll have everybody go ahead and vote. Okay, that was nine high.

And we'll move on to 1B, which asks whether or not -- that the data submitted demonstrated resource use or cost problems for improvement variation in delivery in care or population groups.

MS. FANTA: So it's seven high and two moderate.

MS. WILBON: So 1C asks whether or not the purpose or objective of the Resource Use Measure and the construct for resource use costs are clearly described in the submission. So was the intent of the measure described in the submission adequately?

MS. FANTA: So we have nine high.

MS. WILBON: 1D asks whether or not the resource use service categories that were identified in the measure are consistent with the measure intent and the measure concept.

MS. FANTA: The results are nine high.

CO-CHAIR ELWARD: Thank you. Thanks very much. I'm giving Kathryn a few more minutes. I don't mind --

DR. BLAKE: For 2a2 and 2b2, did you want Mike to go first with the 2a1?

CO-CHAIR ELWARD: Yes, you're right. Well, go ahead.

DR. SCHATZ: So let's see, are we going to put up -- again, I found this very difficult to apply my clinical expertise to, but from the best I could tell from my experience and from what I could read, I didn't see any problems with, and relying on the statistical consultant, I didn't identify any issues in this sub-measure, for this measure. I'd certainly open it for other

input.

CO-CHAIR ELWARD: Any other thoughts?

(No response.)

All right. And let's go to 2a2. Kathryn?

DR. BLAKE: Mine was actually made very easy because Carlos had done a lot of the assessment on the reliability and validity.

So for 2a2, can you scroll it up for 2a2? It was the reliability testing, and when I looked at that again I also found this hard to kind of wade through.

For instance, when I was reading the description of the data and the sample, there wasn't a lot there. But then when I read more into my sections then I found out more of what the data and the sample was. So that sometimes was difficult to get through.

In terms of the analytic method, so the data and the sample, once I reviewed everything, the data and the sample

were adequately described.

When I got to the analytic methods, my impression was that they didn't really provide a lot of detail on the analytic methods. It was more of an overview of what they looked for.

And I actually had a question about one part of it and I'll read the sentence because this is what was confusing. It said, did notice of public reporting of Relative Resource Use, RRU result in 2010, result in a change in the number of make-up of plans that reported or RRU in 2010?

I think, at least I think one of those was supposed to be 2009, from what I understood from what I read previously but I wasn't sure.

So that was one small question and I don't think it will change the overall assessment that I had of this measure. But overall, I agreed with the overall statements and comments that they had.

In terms of the testing results, they had a sentence in there talking about that the range in variation in both the submitted data and the final plan results were not found to be excessive nor was there a significant relationship noticed between the health plan total O/E results and plain quality results.

And it was kind of a qualitative assessment, so I found it kind of difficult to ascertain the appropriateness for the testing with that being so qualitative. Then in terms of their findings statement, I found that acceptable.

MS. DORIAN: Thanks, Kathryn. Carlos, do you think maybe you could also give us a brief overview of your findings on the scientific acceptability?

MR. ALZOLA: Sure. Thank you. Well, let me start with, how was the measure defined? Was it well defined and precisely specified? So we look at the several things

there.

We look at the clinical logic, meaning how is the measure defined and what are the clinical underpinnings that go under, in defining the measure, what are their reasoning?

They also look at the construction logic, so how to go from the database to actually arrived at the score, at the final score.

And then the risk adjustment methodology and the data derivation process. So all those things are found in various places in the document.

And so then they say that, in terms of the clinical logic, they defined what were the methods that they used to identify the conditions and the timeframe for identification and measure.

Like in most measures that we have reviewed, the conditions are identified by the diagnosis code and they have a measurement

period and an identification period.

So they are two years. One year in which they identify the condition and then the next year when they measure the resource use.

They have some exclusions that they apply to all their measures and those are cancers, end-stage renal disease, organ transplants or HIV/AIDS.

And I think even though it's not, I didn't see it specified explicitly, I think those patients are considered different and that excluding those would make the population more homogeneous because this really costly patients that could skew the resources.

MR. HAMLIN: Yes, that's correct. They're extremely high cost conditions that a small proportion of patients could skew the results using our methodology. So we exclude them from all measures.

MR. ALZOLA: Okay.

DR. BLAKE: So does that mean that you didn't exclude diseases such as

cardiovascular disease because they're more prevalent in the population so would have been less likely to skew it with a small number of patients?

MR. HAMLIN: We don't exclude them from the population. The risk adjustment does take into account comorbidities and stratifies them by their risk.

So cardiovascular conditions would be accounted for in the risk adjustment methodology but we don't actively exclude them. The four exclusions that were listed are the ones that are because of cost issues.

For the asthma population we also exclude COPD because that tends to be the ones with, again with the HEDIS quality measure, there's some factors that affect the measure on the clinical side. So we're really trying to look at just asthma alone.

One other comment about the annual analysis just to hopefully clarify some of those sentences. NCQA for Relative Resource

Use conducts an annual report, an analysis of the results that we receive.

Some other things we look at are we're looking at the types and numbers of plans that report. We look at the results from plans that are reported in past years versus plans that are first-time reporters this year to determine whether there are any differences in their status and their results primarily due to outliers.

So we restrict the results that we consider outliers to be any O/E ratios between 0.3 and 3. And we look to see if there's any differences in the results of new plans versus sort of repeat offenders, if you will, or plans that reported over a number of years.

We also do look at specific correlations between some of the service categories in the quality scores and the specific service categories in the pharmacy scores, if you will, to try and draw those correlations.

To date there have not been any that we've been able to quantify, if you will, so primarily, you know, the results are we report the results and we have not been able to show any significant correlations between any of these service categories and their quality scores.

And that annual analysis is conducted and it's a very standardized methodology for conducting that analysis and we try and look at that again year over year over year.

DR. STANFORD: I have a question. So your exclusions are occurring in the identification year as well as the measurement year, or are you just looking at exclusions within the confines of the identification year?

MR. HAMLIN: So defining the population, we use a two-year criteria based upon the HEDIS asthma measure. The measurement of resource use is only during the

measurement year which is a 12-month calendar year.

DR. STANFORD: But the exclusions are occurring in both years?

MR. HAMLIN: Right. The exclusions would affect the denominator population over either year, yes, that's correct.

DR. STANFORD: And then your outliers are based on cost outliers, is that correct?

MR. HAMLIN: Right. That doesn't meet the plan results, exactly.

CO-CHAIR ELWARD: Any other thoughts?

(No response.)

Okay, a question I had for NCQA is, it seems that you initially state you're looking at group practice level, but some of the measures seemed to be focused more on health plan or health system work. Can you address that?

MR. HAMLIN: Sure. The measures

are valid for any health plan. They're population-based measures. We use them for health plans. They have been tested and used in physician groups.

However, you do have to have a population of at least 400 members for the methodology to be valid and so it tends to be larger physician groups that can use the measures.

However, we only at this current time use them for health plans because that's what we do and that's where we found the sufficient sample size.

CO-CHAIR ELWARD: Yes, so the unit of analysis would have to be enough, if you went to a group practice level, you'd have to have enough, at least 400 patients.

MR. HAMLIN: At least 400 patients at a minimum to be able to sort of use the measure in a valid fashion.

CO-CHAIR ELWARD: Okay.

MR. HAMLIN: And for asthma this is

significant, because even at the health plan level oftentimes we end up excluding plans because of a lack of sample size. So we only include people in the analysis that have at least 400 patients.

CO-CHAIR ELWARD: Just one thing, in Step 1, I think it's Page 12, you define, I think you use the standard HEDIS criteria.

Although one of the challenges traditionally has been that using asthma as a principal diagnosis can oftentimes, you can actually have problems finding people, especially those who are acute because they'll come into the ER and they'll be diagnosed as having bronchitis and asthma. It's always in the second one.

Have you looked at whether or not you could enhance the measure by adding second or third diagnostic?

MR. HAMLIN: We actually do include second or third diagnostic and using the two-year criteria for identification. We tend

that we hope that that's a more specific criteria.

We are also currently looking at additional criteria for future measurement to be more specific with the new clinical data that's coming out through more enhanced systems.

But right now we're sort of sticking with the administrative version that we've got over two years of looking at asthmatics.

It's diagnosed as, but principally of many of the people getting using a combination of diagnosis and anti-asthmatic medications. So it's that combination over the two years that defines the persistent asthmatic population.

DR. STANFORD: So back to what you just mentioned. In your measurement year, are you taking into account secondary diagnosis as well or is it all based on the primary diagnosis of asthma?

MR. HAMLIN: No, it's not on primary diagnosis only. It's on any diagnosis.

DR. STANFORD: Any diagnosis. So you're looking at any diagnosis that occurs and you're attributing cost to asthma, correct?

MR. HAMLIN: That's correct.

DR. STANFORD: Or were you looking at all costs? Just cost of asthma?

MR. HAMLIN: It's all costs for anyone identified with asthma. So any services that are rendered to that person during that year if they've been identified as asthma are attributed, yes.

DR. MOSENIFAR: I had a question. The fact that most of the asthma's delivered in smaller health plans, you know, one- or two- or three-physician groups, have you done any testing or is it possible to test that if in two years' time if the ten smallest groups combined for a variety of reasons and make up

larger groups, will this be applicable in terms of generalizable?

MR. HAMLIN: Well, I'm going to try and answer your question two ways. We have criteria for plans or groups that combine so that are, you know, how you would include that data. Because it's a two-year frame, a lot of things could happen over that timeframe.

So we do have very specific guidelines on how you can lump the populations together and which ones you include, which ones you do not include for the denominator identification.

Again looking at the resource use over that year you have to have the data for that year. Whether it's been aggregated or not is sort of up to the organization itself on whether they can or cannot report that.

So if you have the data over that two years and the complete administrative data for that one year for that population you can report it.

We don't allow any imputation on the resource use so there's no assumptions or imputation allowed for the resource use side.

CO-CHAIR MAURER: Could I go back to the answer you gave just before about picking up asthma as well as chronic bronchitis or bronchitis?

In your identification criteria, you use asthma. You use an ED visit or an acute inpatient discharge, but asthma has to be the principal diagnosis.

It looks to me like the only way you would pick up anything other than asthma would be in the four outpatient visit criteria. Is that correct?

MR. HAMLIN: Yes, I'm sorry. I'm trying to do this from memory. And my computer died and I was trying to be green and not print things out. That is correct.

For ED visits it has to be principal diagnosis. For the outpatient it can be any diagnosis. And the proportion of

people that come in through the ED is actually very, very small in the population.

And then chronic bronchitis also is one of the clinical exclusions, so anyone with an ICD-9 for clinical bronchitis would also be excluded from this population.

CO-CHAIR ELWARD: In the comment also you made about including total cost. I mean if a kid, unrelated to his asthma, falls down and breaks his leg, you know, assuming he wasn't wheezing and stumbling because he's short of breath, why does that make sense to include it in total cost?

MR. HAMLIN: Because we're looking at health plan populations. We're looking at the total cost for managing a person identified with his condition.

And again it's population-based so for the plans themselves to understand what the resource use for this person would be regardless of how you're attributing these different services to the asthma or not.

I mean again it's harder to make the argument that the kid fell and broke his arm because he had an asthmatic attack on the jungle gym.

But again it's really looking at the, you know, what services are used by this population identified with asthma.

It's not really what services are attributable directly to asthma, because we don't want to have to make those distinctions of what can be attributed and what cannot be attributed to asthma itself.

CO-CHAIR ELWARD: Michael?

DR. SCHATZ: Yes, I think the point is really the converse, which is the difficulty you run into when you try to say okay, well, this is asthma, this is not.

So I think the balance is in favor probably of doing it this way because of the problems on the other side.

CO-CHAIR ELWARD: And sort of what you'd have to do is, if you saw an inordinate

amount of cost, you'd have to go back to your own claims data and say okay, were there surgical codes or were there appendicitis codes that would account for that?

MR. HAMLIN: And the service categories that we use are very, very detailed so you could go back and, I mean the results are all presented by that.

So you could look at if you had an inordinately high acute inpatient utilization for this population, you could then go back in your data and look and see okay, well, what is this telling me now?

But again we just want to capture everything. It's a snapshot of everything for managing this population with this identified condition.

CO-CHAIR ELWARD: Thank you.

DR. STANFORD: And I think that probably plays into your sample size. Sample size having to be large, if you account for some of the disparities that you may see

across plans.

CO-CHAIR ELWARD: Thank you. Any other thoughts about that in terms of 2b? I think we had a discussion about exclusions is why I just now, have it, about 2b 4 or 5?

DR. BLAKE: Were you finished Carlos? Did you finish everything?

MR. ALZOLA: Just another comment. They restricted the population to patients age 5 to 50 years old.

MR. HAMLIN: A new update. Actually, that age range has been expanded to 64 now for this next year. We just recently finished some testing and we're now 5 to 64.

MR. ALZOLA: And one other comment is that, when they tested the measure, they tested in both commercial Medicaid and Medicare?

MR. HAMLIN: Commercial Medicaid.

MR. ALZOLA: Medicaid only?

MR. HAMLIN: Right.

MR. ALZOLA: Okay.

DR. STANFORD: So you didn't test it in fee-for-service Medicaid, is that what you're saying? You used commercial Medicaid, right?

MR. HAMLIN: Yes, that's correct.

DR. STANFORD: Have you thought about using fee-for-service? There's a big difference between managed Medicaid and fee-for-service Medicaid.

MR. HAMLIN: There is. And we've been looking at new testing options. Right now, actually, we're looking at some of the larger database aggregators to see if we can try and have some new testing.

Basically the way we've been testing in the past is contracting with plans individually and having them provide data to NCQA for testing which is a very laborious and very expensive process.

And so we're trying to again look at some of the variation between the different plan types to see what their resources might

be, but at the moment we're sort of limited to our current testing strategies. But we do acknowledge there is some differences and there's also been some interest in the Medicare population.

We're looking to test the effect of, and looking at the resource using the Medicare population. But again we have a very high threshold for releasing specifications for a population that we haven't tested very thoroughly.

MS. WILBON: Ben, this is Ashlie here. Just a quick question. Did you say you did not test in Medicare? Is that what you said?

MR. HAMLIN: That's correct, because the age range pretty much precluded the majority of that population. We would have had a very hard time finding Medicare-eligibles with a population of 400 or greater.

MS. WILBON: Just a note. We'll follow up, but on their submission it says

that you tested Medicare. So I think we had a question about that as well.

MR. HAMLIN: Sorry about that, yes.

MS. WILBON: Yes, we'll just want to clean that up before we post it. Thanks.

CO-CHAIR ELWARD: 2b5, data analysis? Have we gone that far?

DR. BLAKE: Did you want me to do, let's see. We didn't do 2b2.

CO-CHAIR ELWARD: Oh, I'm sorry. Go back to 2b2, excuse me.

DR. BLAKE: Okay. So the data in the sample: that was the same as previously, or that I previously discussed.

In terms of the analytic method, the risk categories that they had assigned via the episode risk groups appears appropriate and it made clinical sense to group patients according to their observed mix of episodes.

But their other method, categories based upon age, seemed less appropriate. Maybe I just didn't understand, but it seemed

like, you know, severe expensive asthma events can occur at any age so I wasn't really clear on why they were categorized based on age. Can you comment on that?

MR. HAMLIN: Sure. The age strata for the risk adjustment are effectively designed around, sort of, known utilization patterns. And we use a fairly large Ingenix database of the commercial population to determine those age categories.

The age categories on the RRU side or the Relative Resource Use side are really truly around utilization.

The age categories on the clinical side are around both sort of clinical treatment patterns between the children and adults and also for several reporting strategies.

So our clinical measure is part of the CHIPRA core set, so we have sort of an age strata of the 5 to 18s because that's what's in the Child Health measures.

So there's kind of a difference. In the strata for reporting on the RRU side, it's based on utilization patterns. On the clinical side, it's based on clinical and other reporting needs, if you will.

DR. BLAKE: Okay. Then in terms of testing results that appeared appropriate, in terms of their finding statements, there's a mention of four different methods. And I couldn't figure out what these four methods were whether it was episode of care, disease identification. Were those two of the methods?

I'm on Page 31 of your document. That's where SA 2.4 starts on the bottom of Page 30 and continues to on the top of Page 31.

And there's a bullet on Page 31 and you say under methods, four different approaches were used by the study to measure Relative Resource Use, varying the risk adjustment methodology employed and the focus

on total services versus disease-related costs. So I wanted to know what those four methods were.

MR. HAMLIN: We tested several different risk adjustment methods from the age, sex, comorbid yes-or-no type method up to the current HCC stratified population into 13 different risk cohorts by severity of comorbidity and there was a couple of variations in between.

So we tested several risk adjustment methodologies to determine what was the most appropriate for cost-related and utilization-related factors for this population. And the current method, the HCC method was the one that sort of came out on top, if you will, for the --

DR. BLAKE: The which one?

MR. HAMLIN: That HCC. We use an HCC derived, so it uses the same as HCC approach that we sort of drill down a little bit.

We don't use the entire HCC classifications but we use a good proportion of them for categorizing the population. And during the early testing, I believe there's a 2005 report that was included in the materials that looks at the effect of the testing of the different risk adjustment approaches on the Relative Resource Use populations.

DR. STANFORD: Is cost during the identification, is that one of your risk adjustment measurements?

MR. HAMLIN: During the measurement year, yes, because it looks at the -- the HCC uses any services delivered during that year to appropriately categorize them into one of those 13 cohorts.

So it looks at other diagnoses and, you know, severity and frequency to put them into a severity category.

And it's sort of, you know, 1 is asthma with low severity and then 13 would be multiple comorbidities, high service

utilization and other things that would affect the cost of that population.

DR. STANFORD: Those are cost-driven. Those are --

MR. HAMLIN: They're cost-driven, yes. They're not clinically driven. No.

DR. STANFORD: I'm sorry. So are you using dollar values in your risk adjustment or are you using count values?

MR. HAMLIN: What we use is the associated service codes. So we're looking at ICD-9 and procedural codes to identify them for services rendered to categorize them into that population. And then we go back and look at the number of times those services were offered to that population and correlate that with their HHC cohort.

DR. STANFORD: Have you looked at cost in the identification year as part of your risk adjustment? But I mean have you compared the two whether they --

MR. HAMLIN: Well, we don't

actually get actual costs reported at NCQA, so we don't, the research database that we use to develop our standardized pricing methodology has that information in there and they use, that is how they derive the standardized prices.

So what we're looking at really for what we get is sort of a PMPM for that category of all members who meet that risk criteria.

And the determination, I believe that the steps are laid out for how you would, you know, categorize someone into that HCC cohort. It's primarily driven by, you know, ICD-9 codes that classify the number of comorbidities they have as well as other factors that would push them into --

DR. STANFORD: That's a yes/no count. They have one code, It's counted as, you're not counting multiple þ

MR. HAMLIN: Multiples are counted as part of the process. So if you have

multiple comorbids, that factors into the risk adjustment. Multiple comorbids will put you in a much higher risk adjustment category.

DR. STANFORD: What I meant to say is that, if you have two codes for cardiovascular disease, for instance, or two, those are counted as one or two?

MR. HAMLIN: They're counted as two.

DR. STANFORD: All right, great. Thank you.

DR. BLAKE: Next I reviewed 2b4.

CO-CHAIR ELWARD: Please go ahead. Thank you.

DR. SCHATZ: Well, I think just to say we heard that the sort of the two main exclusions are for high cost outliers, which seems to make sense, and then what is always a concern with asthma is COPD.

So even though the ages, and that's why the age has historically been so low or at least one of the big reasons. But with the

age range rising to 64, but the ability to exclude COPD codes, that seems appropriate as well.

The only thing I wondered about although I think in the high cost arena, by excluding people with acute respiratory failure you are excluding potentially poorly managed patients, which gets at just the whole reason for, at least one reason for doing all this, but probably the cost outlier piece outweighs the situation.

But that was the one thing that occurred to me that, you know, a person could be excluded. It could have acute respiratory failure because they just hadn't been properly cared for, and to exclude them seems questionable.

MR. HAMLIN: So in our recent field test that we did when we were looking at the upper age groups, we did look at the effect of the different clinical exclusions on the populations of interest.

COPD was, by a vast majority, the biggest effect. About 38 percent of the people were eliminated from the population.

The others, acute respiratory failure, were very small percentages and so there's still always some debate on whether, you know, we should be excluding them or not.

But I think at this point, they were less than three percent I think at this point so they don't meet our sort of five-percent threshold of concern.

But we continue to look at the effect of those diagnoses on this population not year over year, but every time we do the test we sort of retest these clinical exclusions on this population to determine, you know, what the effect on the results might be.

CO-CHAIR ELWARD: And again, going back to the issue of ER visits, have you looked at the relative contribution of first and second diagnoses?

I mean, in the health plan I consult with, with Coventry, we actually increased our yield about 30 percent when we went to the second diagnosis code.

MR. HAMLIN: Right. The first and second we have not tested but we have tested the effect of the, you know, who comes in via ED visits versus the other algorithm. And again it was a very small proportion. It was less than five percent of the population came in through ED visits overall.

But we are continuing to look at ways of refining the denominator, because personally I believe, the ED visit is sort of the weakest link in the denominator chain, if you will, because it is possible to get people in there who might have either mild, persistent or intermittent asthma. You know, and they have one one year and one the following year.

But again it's sort of at that low percentage threshold that we're, it's of low

concern but it is still of concern for us.

CO-CHAIR ELWARD: Michael?

DR. SCHATZ: Yes, I mean I think it's a sensitivity/specificity issue and for a measure like this, I think we'd prefer the specificity.

CO-CHAIR ELWARD: A couple people were, in just looking at the overall ratings a couple of people also expressed some concerns about this aspect. Any thought? Gerene?

DR. BAULDOFF: Sorry, I keep forgetting that, that I marked it low. I think what I was doing was that I kind of missed a piece there.

I was very dependent -- thank you so much for the statistician review that really helped a lot, and I really missed the measure score. I was really going by the data elements section whenever I marked that low, so I apologize for that.

CO-CHAIR ELWARD: No, that's very

helpful. Thanks. Other thoughts on that aspect? Go to Kathryn next.

DR. BLAKE: Did you want Carlos to talk about the validity section?

CO-CHAIR ELWARD: If he would like to.

MR. ALZOLA: Yes.

DR. BLAKE: That's what I already talked about, but then I realized he didn't.

MR. ALZOLA: Again, we are looking here at face validity. That's what most measures have been focusing on.

And they, in support of the validity of the measure score, they included a lengthy paper which -- where they compare these four different risk adjustment methodologies.

It wasn't completely clear to me how that related to the methodology they actually selected: the HCCs.

The most single thing I that I found was the methodology where they just made

an adjustment by age and sex groups, and they compared that to the episode treatment groups approach and they found that they had a really high correlation.

So I think in general terms, that's good evidence of face validity, but I wish you had been a little more specific for the method you actually used. Although I'm not going to argue with the face validity of the HCCs, but I just don't think, I thought it wasn't addressed really in a more specific way.

The other evidence that I saw was how the costs were distributed in the different lines of service.

And one thing they found was that the majority of costs were attributed to prescription medications and followed by inpatient costs, which for an asthma population seems reasonable and for clinicians, no. They know better than I whether that's reasonable or not.

DR. SCHATZ: Well, I mean I think

and consistent with some prior independent separately done cost analyses of asthma.

DR. STANFORD: Fee-for-service Medicaid would not fit this category, so it's actually higher resource utilizations in that population.

MR. HAMLIN: And on the results side, one of the things we don't actually do at NCQA is say whether high is bad or low is good and so on and so forth. Again it's a snapshot of utilization.

I mean we have seen some non-statistically significant correlations between high pharmacy and high ED and low inpatient use for some of these populations, but again we're presenting the results as they are. We're not making any value judgments on high bad, low good, kind of thing.

CO-CHAIR ELWARD: Kathryn, go ahead.

DR. BLAKE: Okay, my next section had to do with risk adjustment and

stratification method. This is 2b4.

And the risk adjustment appeared appropriate but I still have to ask, why do you stratify by gender? Your age stratification, 5 to 17, 18 to 44 and 45 to 50 seemed appropriate, but why do you stratify by gender? Is that standard?

MR. HAMLIN: Well, it's part of the fact that we use the same methodology for all of our measures, all of our risk, our RRU measures are all stratified by age and gender categories.

The age categories are slightly different depending on the different measures. But at this time we feel that it's important.

The age and gender are both weighted as part of the HCC approach, so there's a weight assigned to each of those based on utilization from the data that we used to determine that and then are reported back out by age and gender categories that we feel are relevant to utilization patterns.

Whether or not there are differences between the gender patterns, again, we're just looking at that to see if there are any kind of disparities or differences. We don't do any kind of reporting by gender only, if you will.

DR. BLAKE: Okay. Do you want to go on. My next section was 2b5. 2b5 just had to do with the scoring and I felt that was fine. The types of scores were frequency, distribution rates and proportion ratios, weighted scoring, composite scores, and these all seemed very appropriate.

The interpretation was appropriate. The detail score estimation was appropriate and the discriminating results approach seemed appropriate as well, unless Carlos has any other comments on that.

MR. ALZOLA: No. The main thing is whether, I can see, like, initial comparing statistically significant versus clinical significant differences. I didn't see that

addressed.

However, the good thing is that they did provide the actual sample sizes under formulas to calculate confidence interval and that's really just about everything anyone needs to determine whether any differences are significant in which ever way they want to interpret it.

CO-CHAIR ELWARD: Thank you. Any other comments on the section 2 before we go to Michael? Section 2b. Michael, 2b6?

DR. STANFORD: Well, 2b6 and 2c are equal to what we heard before because multiple data resources aren't being used in this and we sort of agreed that data on disparities is inadequate.

But I do think that if it were there from what the guidance before, we would be able to see that relative to what is being captured here. And asthma, like cardiovascular disease, this would be an important issue.

CO-CHAIR ELWARD: Agreed. Thank you. Any other thoughts about Item 2 overall, the overall category? Great, thank you.

MS. WILBON: So what we'll do now, if everyone's okay with the discussion and there aren't any more comments, we'll go through each of the subcriteria on the voting tool and have you guys vote. And if we get to any of the subcriteria and you feel like it hasn't been adequately discussed we can, you know, pick up a discussion there. But we'll go ahead and just run through them and see how everyone feels about them.

CO-CHAIR ELWARD: And if there are broad discrepancies on any one item, we might stop and try to work through that.

MS. FANTA: So 2a1, is the measure precisely specified so it can be implemented consistently.

MS. DORIAN: Can you guys try to keep voting? We're just missing a few votes, and it won't count it twice if you keep going.

MS. FANTA: And the results are nine high.

Moving along to 2a2, does the reliability testing demonstrate that the results are repeatable, producing the same results a high proportion of time when assessing the same population in the same time period and/or that the measure score is precise? And the results are eight high and one moderate.

And now we're going to do a vote on overall reliability, and that encompasses 2a1, precise specifications, and 2a2, reliability testing. Are there any comments that anyone wants to make before we vote? Okay, then we'll go ahead. The results are eight high and one moderate.

Moving along to 2b, validity. Are the measure specifications consistent with the focus of measurement and the measure intent? The results are six high and three moderate.

Moving along to 2b2, does the

validity testing demonstrate that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided adequately distinguishing high and lower costs or resource use? And the results are six high and three moderate. Moving along to 2b3, exclusions. Are exclusions supported by the clinical evidence or analysis of frequency and distribution? Is information about impact of exclusions for patient preference transparent? And the results are six high and three moderate.

Okay, 2b4, risk adjustment. For Resource Use Measures is there an evidence-based risk adjustment strategy or rationale or data to support no risk adjustment or stratification? Okay, and we have seven high and two moderate.

Moving along to 2b5, are performance results reported? Do they identify differences in performance or overall less than optimal performance? And the

results are eight high, one moderate.

Okay, and now this is an overall vote on validity that encompasses that the specifications are consistent with resource use and cost problem validity testing risk adjustment and identification of meaningful differences. The results are five high and four moderate.

MS. WILBON: One last vote on disparities.

MS. FANTA: Okay, so 2c, if disparities in care have been identified, do measure specifications, scoring, and analysis allow for identification of disparities through stratification or results or is there rationale or data justifying why stratification is not necessary or feasible? Okay, so the results are five high, three moderate, and one insufficient.

CO-CHAIR ELWARD: Thanks very much. Usability, Kathryn?

DR. BLAKE: I had usability. This

was fairly straightforward. For current use I found that is acceptable.

For the use in public reporting initiative I found that was accessible. For the use in, what's Q, by quality improvement, in my notes I put non-applicable. Oh, he didn't have anything in there. That's right. There's nothing in there, so not applicable.

And the same for the use in other accountability functions, that's not applicable because they said Relative Resource Units measures are not used for accreditation scoring. So those two didn't apply.

Then in terms of 3, let's see in terms of 3b for understanding or usefulness of the description of the data, method, and results, I have to say NCQA did an excellent job in this respect. I found it very straightforward and easy to interpret.

In terms of the interpretation of the score the observed to expected ratio for their interpretation of the data I found

acceptable. The detail score estimation I also found acceptable and the discriminating research approach I also found acceptable.

MS. DORIAN: Kathryn?

DR. BLAKE: Yes.

MS. DORIAN: Just one question. Would you mind just telling us a little bit around why you found these acceptable, or if anybody else has any comments just for our notes?

MS. WILBON: And summarize what they have, because everyone might not have reviewed the measure. So everyone can vote, we just need to kind of have a summary of what they had in there and then what your judgment was.

DR. BLAKE: I don't remember much. I tried to put down in my notes what I felt like was important to bring up in terms of problem areas. I didn't do as well in putting in what was appropriate comments. So if I think of it as we go along then I'll bring it

back up.

MR. HAMLIN: I can give you an overview of what we do if anyone is interested.

So as I mentioned earlier these results are published in through our Quality Compass module which publishes the individual plan results by detailed service category and also with their quality score.

Now obviously since Relative Resource Units are much more complex measures it requires a little bit more interpretability and so we have subsequently published several resource -- we have a dedicated web page that describes our methodology in great detail. We've conducted a number of webinars that sort of outline Relative Resource Use and how we go about calculating them and how you might go about using them. In addition, we've also created several documents. Some are very brief, what we call four-pager for specific stakeholders like purchasers and employers.

Also there's another more detailed document that explains how you might use the results and how you might interpret the results.

So here you can see a sample report for the different service categories for, I hope this is for asthma. So this is what the results might look like.

And then we have these additional documents and resource guides that will help all stakeholders, targeted to specific audiences that will help them interpret these scores.

And also we offer some opportunities, since I mentioned we only receive aggregate plan level standardized cost data.

We then also have programs where we go back to the plan and say okay, well, if you plug in your actual costs, your real costs using the same methodology here's how you might be able to identify opportunities to

improve some of your utilization patterns. You know, here's where you can look for opportunities to reduce costs if you will. But again we don't have that data, but we offer a lot of education around how you might be able to identify opportunities using your own data. You can plug into this same methodology and therefore try and understand where you can identify those opportunities.

Often it's not the high utilization low cost areas, but sometimes it's the lower utilization higher cost areas that you can, when you go back and plug this in you can actually find some significant opportunities. And so again we work with plans individually.

We also, again, publish these resource documents, make them available broadly to the public through our public website for anyone who's interested in either the methodology or the opportunities. But they are very complex measures, and so that's why we provide this much information.

CO-CHAIR ELWARD: Yes, I noticed a couple places where you reference in the measure that you have to go to a website and get a login and password. That's just a formality, right? I mean is there --

MR. HAMLIN: Actually there is no more login and password anymore. So the standard pricing tables are now moved to our public website so the updated URLs which I'm not sure were in these two measures.

I believe we updated them on the CV and diabetes ones, now are just really available to anyone who wishes to use them, and we encourage their broad use.

CO-CHAIR ELWARD: Yes, the ones in there still want you to log in and all that.

MR. HAMLIN: Okay. Yes, there was no required subscription. We were initially tracking who was logging in and how many people were logging in.

Checking them now, it's just moved up to a more public site where you just go

directly and it takes you right to that site.

DR. STANFORD: Can you -- in treating my question around standard pricing, so when you say standard pricing across health plans, how are they implementing that within the confines of their data? Is that an aggregated or is that an individual?

MR. HAMLIN: They look for individual codes, so they look for ICD-9, they look for CPT, UB revenue for some service categories, and there's actually a standardized price in those tables that is assigned to that specific code. And they do that for each of the service categories so the prices are adjusted for whether they're inpatient or outpatient.

So it depends on where you find it you'll apply a specific code. They then aggregate that and report that to us at the plan aggregate level, but they have to do the individual member level mapping. So it's a fairly significant effort for reporting these

measures.

CO-CHAIR ELWARD: How often do you have a review of the usability, I mean from the feedback you get? How often does that occur?

MR. HAMLIN: Annually. We, actually even more frequently nowadays, because when the measures went public last year, the results went public last year, we had this sort of blitz of information and feedback. And we try and sort of recycle that and get that back out again to the public in a regular fashion.

But we're always taking, you know, right now we're doing sort of an experience analysis if you will. We're contacting clients directly and finding out what their experiences were.

We're contacting some of the employer groups and finding out what their experiences were with this information, how they used it and how we can then tailor that

to try and improve the guidance documents that we provide to people.

So it's a continuous thing, but at a minimum we annually look at the results of the individual plans and how they compared and where they moved and how they did.

You can't trend this information year to year because we calculate the expected on all the plans that submit every year to NCQA so there's some variation there. But we are looking at other ways now with enhanced data to try and perhaps create some trending strategies too for plans.

CO-CHAIR ELWARD: Yes, and it may be my intellectual limitation, but one of the things it seems would be nice is to have some actual examples, you know, with imaginary data that somebody could go through and play with a little bit and say okay, here's how it makes sense.

Just to draw out, you know, if I have a population of 5,000 and I had these

variables, what would it look like? Just to give people an idea of how they would use it.

MR. HAMLIN: We've been doing that annually at our HEDIS update conference. So it's a day and a half conference, and then we save that second afternoon solely dedicated to RRU and that's where we do the opportunities to improve. And so we do exactly that.

We take fake data and we show them the opportunities calculations based on that fake data and the different variables that would affect that. And anyone's welcome to come in and sit in on those. I mean you have to register for the conference then you can go in.

And we actually have people who, experts who will go through and walk you through all those calculations, and we use sort of fake data sets to do that. So that's part of our annual presentations but we also do some webinars of that as well.

CO-CHAIR ELWARD: Okay, thank you.

That's helpful.

DR. STANFORD: I always get hung up on these measures around public accountability and public reporting. Who do you think the primary user, I mean we talk about publicly reporting.

Who's using this the most do you think? I mean is it purchasers or is it the health plans themselves?

MR. HAMLIN: I think actually right now, my impression is equally both. So plans are using it to identify their, you know, how they compare to their peers. Because we only report plans in peer groups, so HMO only, PPO only. They're only compared to other HMOs in their region, HHS region, which is fairly big.

The purchasers are the ones who really are driving the need for this information. So the purchasers and some of the large employers want to see this because it's what they want to use.

You know, previously all they had

was a premium price that they could say well, you know, you're charging me this premium and you're charging me this premium. Here's your HEDIS quality scores.

But this now is one more piece of information that they can then bring into that conversation so they can look at specific plans based on premium, utilization, adjusted for, you know, their peer group and also the quality scores as well.

And so we're finding that plans are interested in their results themselves and that purchasers are interested in the different plans and how they look compared to each other, you know, like year by year.

So we're seeing a lot of, an increase in our Quality Compass registrations on the purchaser/employer side which has been nice for us.

DR. STANFORD: And then the other thing that goes along with that, you had mentioned earlier that you really haven't

found associations with at least most of your quality metrics. And I assume most of the quality metrics are the HEDIS measures and primarily processes of care.

I mean have you looked at outcomes? I mean obviously the goal here I think with a lot of these resource measures is to improve value.

You know, a lot of people talk about improving quality, but frankly I haven't seen many of these measures that I've looked at that actually have evidence that they correlate with quality.

MR. HAMLIN: It's hardest with asthma, I'll be perfectly honest, because we only have one quality measure to associate, more are coming.

But again we only use HEDIS measures, HEDIS quality measures that have fulfilled the full process, which means we cannot have first year measures that are not publicly reported on the HEDIS side. So we

are expanding that.

There are two measures coming, but again they haven't finished that process yet. So I think as we enhance the quality side we will start to see more correlations that will be more statistically significant or valid.

But we again have a very small and very high threshold for what we consider statistically significant correlations versus what we, we sort of see some correlations but we're not going to report those out because we want to really keep the bar very high there.

So I expect as the quality side gets better particularly with enhanced information as we move more into admin plus electronic medical record information on the quality side, I think those correlations will start to become much more apparent and we'll be willing to make those public announcements of those correlations. But right now we see some but we're not going to publish those.

CO-CHAIR ELWARD: That does raise

a question, this is primarily going to be from claims or --

MR. HAMLIN: Yes -- data.

CO-CHAIR ELWARD: Okay.

DR. BLAKE: Going back to the question, I can't remember who asked how I came up with what was acceptable. I went back and I was just relooking at some of the descriptions, and part of the resource use refers you back to items earlier.

And those earlier items really just define how the data is described, which again is in terms of observed and expected ratios which makes very intuitive sense. So I thought that the reason for that was appropriate.

When they looked at levels of analysis it was by group practice, by health plan, by delivery system, by national population, by regional population, and again that to me made appropriate sense.

And then there was further

discussion which we've already talked about of why a sample size of 400 was the minimum cutoff for the estimates.

So looking at all that, and we've talked about it some before, that's why I rated most of these acceptable.

CO-CHAIR ELWARD: Thank you very much. Any other thoughts, comments? I guess we're ready to vote.

MS. FANTA: Okay, so moving along to 2a, usability. Are the measure performance results reported or suitable to report to the public at large in national or community reporting programs? Is there evidence that the measure performance results are available for public reporting? Okay, so we have eight high and one moderate.

3b, did submitted information demonstrate that results produced by the measure are meaningful, understandable and useful for information for quality improvement and public reporting or was a credible

rationale presented? Okay, six high and three moderate.

3c, are the data and result details maintained such that the Resource Use Measure including the clinical and construction logic for defined unit of measurement can be decomposed to facilitate transparency and understanding? And the results are eight high, one moderate.

Moving on to feasibility, feasability discussion.

CO-CHAIR ELWARD: Okay, all right. That's where we'll -- in terms of looking at 4a, for clinical measures required data elements routinely generated, that seems that that's the case.

And 4b, I think it's just very well outlined in there. For the required data elements being available in electronic health records, it doesn't appear that that's the case. It appears that it's administrative data. Is that correct?

MR. HAMLIN: Currently it's only administrative data that's used for that.

CO-CHAIR ELWARD: Are there any plans of being able to draw that out of EMRs?

MR. HAMLIN: There's guidance for plans that use EMRs to map to the claims codes that they would use for then standardized pricing, yes, but that was, I do not believe that part was included because that's a separate set of guidelines. But we have that methodology that's available.

DR. STANFORD: You're linking the two. You're not saying electronic medical records only, you're linking the electronic medical records þ

MR. HAMLIN: So these are officially claims only, but for systems that use electronic medical records for their billing and other, we actually have a methodology that allows them to map those to an appropriate claims code that would then be standardized price that could then be included

in this methodology, so it's an additional step at the current time.

CO-CHAIR ELWARD: On item 4c, susceptibilities between accuracy errors or unintended consequences related to measurement can be judged to be inconsequential.

I think they did a good job recognizing where the challenges are and, you know, that they understand that there's still issues around data collection methods that can vary and errors derived from other sources may affect the results. So I think they've addressed that well, and I thought that was adequate.

4d, the data collection and measurement strategy can be implemented as demonstrated by external programs or testing that is not identified varies to operational use.

My impression in looking at what you've addressed says that the measure's currently in use anyway. The data collection

strategy sounds like it's fairly straightforward. Well, not straightforward, but well established, I'll put it that way.

Can you tell us whether there have been feedback in terms of any difficulties people have had in terms of operationalizing this?

MR. HAMLIN: Well, I mean as you've heard it's a very laborious process. We lay it out in a stepwise fashion for plans and programmers to go through and really understand what has to be programmed, what has to be done in order to report the measures.

All the data has to go through a certified auditor before it's reported to NCQA so auditors can go back and look for errors in the reporting.

And we also can as part of our annual analysis we look at outliers. So the number of people who are identified as outliers so their results are just off the chart if you will, and we go back to those and

we sort of ask why.

But really nowadays the number of outliers is significantly -- is less than half a percent for some measures, less than one percent for others.

So it really, you know, over the four to five years of reporting measures we sort of ironed out all the issues with collecting and reporting the data and most plans now are able to report. And like I said I think we have one or two outliers a year out of a thousand or so plans that report.

So we'll go back and work with those findings to find out, you know, we have validations built into the reporting systems, so validation alarms will come on if something is before this that's flagged as sort of an outlier or flagged as questionable, and the auditors then will go back and also work with the plans to understand, you know, what they're reporting is valid.

CO-CHAIR ELWARD: Are there

different levels of user or support? Are there variable charges for this level of support versus being able to, for people who have questions?

MR. HAMLIN: Nowadays really we sort of have a, you know, if you have trouble and want to contact us you can work through our system and, you know, hire somebody to come over and help you.

We don't really have much of that anymore because now like I said, that happened much in the past when we were still collecting data and there were many more outliers. Now we've essentially published everything on the website. The methodology is transparent.

And this year moving forward, we're actually going to be pushing out XML specifications, so it's even more detailed and more, sort of less room for interpretation, if you will.

So we're going to be using an XSD XML strata for these measures, not just a Word

document that a programmer's going through and trying to interpret.

We're giving them the programming and that they can go through and they have to do certain mapping to do that. So we're really trying to facilitate the complex process through any tools we can.

CO-CHAIR ELWARD: All right, thank you. That was a little bit of addition. I realize that the data collection search, you've been doing this. It's already in use, so that's probably not applicable.

MR. HAMLIN: Yes. And we have noticed a significant decrease in outliers over the years to the point now where we're significantly less than one percent out of all the plans.

CO-CHAIR ELWARD: Thanks. Any other thoughts, questions, or comments about feasibility?

MS. FANTA: All right. So moving along to 4a, are the required data elements

routinely generated and used during care delivery? Example, blood pressure, lab test, diagnosis, medication order. So eight high, one low.

CO-CHAIR ELWARD: Any other thoughts? Any questions about the, I don't want to single people out, but are there anyone that frequently had challenges with that? Okay.

DR. STANFORD: Well, I'm the low.

MS. DORIAN: Can you use your mic?

DR. STANFORD: Oh, I'm sorry. This is strictly administrative claims data according to what NCQA is saying, which is in general just fine.

So I don't -- my interpretation of that is that, do we have these elements in the confines of the data set. And we don't, unless you have the ability to link it to some kind of electronic medical record. So that's why I said low.

DR. SCHATZ: But they're not

required data elements for this either.

DR. STANFORD: No, that's true. But the examples you're saying are blood pressure, lab tests. So my interpretation of it is are these available for the actual assessment, and they aren't.

MS. WILBON: Yes, so just to clarify. The examples in that are actually probably more so for a quality measure. I think we've kind of borrowed some language for our criteria from the quality measure criteria. So we probably shouldn't have had those examples in there. It can be a little bit confusing.

But I think the best way to frame it is based on how they've specified the measure for the data elements that they are asking for. In order for someone to implement the measure would those data elements be available electronically?

MR. HAMLIN: So lab test performed, diagnosis codes and medication dispensed are

all routinely available in the claims data, and we use those for these measures. And those have been tested, and they have been shown to be reliable.

DR. STANFORD: I'm sorry, but for this particular asthma measure you don't use any laboratory values, correct?

MR. HAMLIN: No, but we do report laboratories as a service category, so laboratory use for asthma.

DR. STANFORD: Yes or no? It's yes or no.

MR. HAMLIN: Well, it's standardized laboratory. Like it's a ratio that's reported to observe to a respective laboratory. Diagnostic laboratory use for people with asthma is a service category.

DR. STANFORD: Like spirometry, for instance.

MR. HAMLIN: We use any submitted laboratory claim. Any of those, again, procedure codes that are used for laboratory,

CPT primarily.

DR. STANFORD: Right.

CO-CHAIR MAURER: And do you use the values or just that fact that it's been done?

MR. HAMLIN: Well, it is actually, it's a priced category. So it is something where we look, you look for CPTs with modifiers, and those have specific prices assigned, and the aggregate doesn't report those. It's a price service category for asthmatics, for people with persistent asthmatic.

MR. BRATZLER: Well, one thing I didn't notice is what, you have pharmacy data consistently for all the plans, or I know that was a problem with some of the measures I reviewed. So you consistently --

MR. HAMLIN: There's still, you know, for some PBMs there are still some difficulties.

But we do require that the plan,

the member have pharmacy benefit for these measures, and the plans are responsible for obtaining that information to report the measure.

So, yes, the pharmacy is reported separately from the other medical components, partially because we want to see the correlations but also partially because there is some differences in the data.

CO-CHAIR ELWARD: So just getting clarification with the staff. So in general how we would be looking at 4a is to take off the first clause in the parentheses. Just saying the required data elements were routinely -- okay.

MS. WILBON: That's okay. That's one of the challenges we've had throughout this process, so I understand that. So did you --

DR. STANFORD: Can I go back and revote?

MS. WILBON: Yes, can we revote?

Okay.

DR. STANFORD: It wouldn't be low, it would be high.

MS. WILBON: Okay. All right.

CO-CHAIR ELWARD: Thank you. No, thanks for clarifying. That was just very helpful.

MS. WILBON: It's helpful for us as well because we can kind of clean that up going forward, so thank you.

MS. FANTA: Nine high.

Okay, 4b, are all the required data elements available in electronic health records or other electronic sources? If not, is a credible near term path to electronic collection specified? Eight high, one insufficient.

CO-CHAIR ELWARD: The only reason, I put that down just because I wasn't sure. You said it was in development, and I wasn't sure that there was specific criteria in how to do that yet, am I right?

MR. HAMLIN: Which was in development?

CO-CHAIR ELWARD: As far as mapping to electronic data records.

MR. HAMLIN: Yes, for EHRs on the quality side we're retooling all the asthma measures we have which is about 12 right now, and enhancing those and how you would capture that out. But on the resource use side we're still looking at administrative claims.

CO-CHAIR ELWARD: Okay. Do people feel comfortable with that? Okay.

DR. BLAKE: How would they even get data out of the medical record given there's þ

CO-CHAIR ELWARD: You'd have to create it. If I understand you'd have to, they would know better, but I think you'd have to create a mapping program so that one data element that you have in your EMR directly quote "means" or relates to the billing record. I mean it's challenging.

DR. BLAKE: Yes, okay.

CO-CHAIR ELWARD: Sort of a crosswalk.

DR. BLAKE: Is that what you're saying is being done of some sort?

MR. HAMLIN: Well, each plan is responsible for, you know, again we have guidelines on how to map using EHRs. I think there actually is a separate guidance document that maps to that.

So, you know, capitated plans that use an EMR particularly have to do a pretty significant mapping exercise to track the utilization patterns and map those to the administrative claims codes that are published in the SPTs, in the Standard Pricing Tables.

Again it's an additional layer of complexity, and I think in the future when, you know, the standards in EMRs for maybe a CCD might be more broadly used, we'll probably specify for those as well. However, at this current time it's the plan responsible

for mapping and the auditor responsible for ensuring that that mapping's appropriate.

MR. BRATZLER: I would only point out though for this particular criterion the EHR is not required. It's simply, is it available in an electronic source, which includes administrative claims data. So the answer is yes. The data is all from administrative claims, at least as you read this.

CO-CHAIR ELWARD: Got it. Then I'll change mine to one.

MS. FANTA: Okay, so we're going to revote on that. Nine high.

Okay, 4c, are susceptibilities to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems identified? Seven high, two moderate.

And lastly, 4d, can the data collection strategy be implemented? Is the measure already in operational use or did

testing demonstrate that it is ready to be put into operational use? Eight high, one moderate.

DR. BLAKE: Can I ask just a general question? Do the health plans pay in order to have their -- to be assessed for their efficiency and things like that? Is this something that's required of health plans or do they voluntarily do it and then pay a fee, say, to NCQA to do this?

MR. HAMLIN: So reporting to NCQA is all voluntary. Health plan accreditation has fees associated with it, so it's a base fee plus on a per member charge.

We actually have a number of plans that are not accredited that do report measures to NCQA to be sort of included in the calculation mix, but it's all voluntary.

CO-CHAIR ELWARD: Thank you, good question. Thank you, we're right on time, and I think it was a great discussion, especially starting off for the first one I think we got

a long ways, since we're about ready to take a break and spend about five minutes.

I'm sorry, 15 minutes. Okay, don't want to work you too hard. So 15 minutes and we'll start again just around 11:00. Great, thank you. And we'll be doing 1561.

(WHEREUPON, the meeting in the foregoing matter went off the record at 10:42 a.m. and went back on the record at 11:06 a.m.)

CO-CHAIR MAURER: Okay, we'll get started on the second measure, which is the Relative Resource Use for People with COPD. It's the second and last NCQA measure for today. And I will start with the importance piece of it.

So does it focus on a significant national health priority? It's the first, fourth leading cause of death in this country, and it's heading toward the third leading cause of death. It' the only one of the top four that's actually increasing in its

percentage.

It's the third largest global disease burden, expected to be by 2020. Twelve million people in the U.S. are diagnosed with COPD, and it's thought that there are about 12 million that aren't, with a 125,000-plus deaths per year currently due to COPD.

Now in terms of 1b, which is the demonstration of resource use or cost, the most recent statistics suggest that there's about $18 billion per year spent in direct cost and another $14 billion in indirect cost in this country. And it's a particular burden in Medicare patients because as we all know it's a disease of older people.

In terms of the purpose, objective of the measure and whether or not there are opportunities for improvement, there's obviously increasing morbidity and mortality which suggests that there might be opportunities for improvement.

NCQA in presenting this measure has identified a lack of control of risk factors. They have identified a lack of control of preventive measures such as influenza, a variation influenza vaccination.

They've identified medication adherence as an issue, and the increasing implications of the financial and disease burden in the country, so there are multiple areas in which improvement can be achieved.

In terms of variation in care across different populations, there appear to be higher mortality rates in African Americans, though most of this disease is reported in Caucasians.

The mortality in women is also increasing, and in Caucasian men the mortality appears to have leveled off though it has not in the African American population as I mentioned.

There isn't a lot of referencing of the disparities. There's a lot more out there

in the literature about disparities in care than is actually referenced here.

The purpose of this measure is to measure the total costs in COPD patients. Now Ben is going to mention more about this, but as the measure is described in the document that was submitted, it sounds like it applies only to people who are newly diagnosed requiring 730 days prior without any COPD diagnosis. However, I mean then that is a criteria that is currently applied to the HEDIS spirometry measure.

However, Ben tells me that that was erroneously put into this measure and that this measure is actually supposed to apply to anybody with a single diagnosis claim for COPD in the measurement year.

So it's very different from what it actually says in the measure. So think of it in the way that I guess that it's intended because you'll fix that, right?

MR. HAMLIN: This is one of the

issues where we have, in pairing this measure with quality measures, for COPD we use two quality measures from HEDIS both of which have very different eligible population identification algorithms.

The RRU side is fairly simple in that it uses a COPD chronic bronchitis or one other diagnosis code. I can't remember, emphysema -- thank you, during the measurement year to include.

However, that's paired with the quality measure side and I think in the form we weren't specific about defining the different ways.

So on the quality side we do it this way, on the RRU side we do it this way. I think that could have been much more clear, and I think that may have been where a lot of the confusion was.

So the identification population that's listed in SA-2 defines the SPR and the PCE populations on the quality side, and we

just didn't make that explicit I think in our description of how we define this population.

So we're going to go back in and make some significant corrections to make sure that's explicitly clear that that's how we do it for the quality and then the RRU is done as I mentioned.

CO-CHAIR MAURER: Thanks, Ben. This measure is also designed to measure total costs of that population in a 12-year period. It's a calendar year measurement, and it is designed to be used by health plans primarily. It does not attribute costs specifically to providers or to practices.

Okay, any questions about the importance? Did we get everything that we need?

Okay, so our discussers for this measure are Kurt, Richard, and me, and Kurt is going to start with 2a1.

MS. WILBON: Janet, real quick. We'll just vote on importance.

CO-CHAIR MAURER: Oh, sorry. I forgot about the voting.

MS. WILBON: No, it's okay. Unless there's any discussion on any of the things that Janet brought up, we'll just go ahead and again vote and if you have any discussion to add as we go through we can do that as well.

MS. FANTA: So 1a is does the measure focus address a specific national health goal priority or was data submitted that demonstrated a high impact aspect of health care? And the results are nine high.

1b, was data submitted that demonstrated resource use or cost problems for improvement? For example, variation in the delivery of care across providers and/or population groups, disparities in care.

MR. BRATZLER: Can I ask just a clarification? So one thing I noticed consistently was discussion of variations usually was related to clinical care and not resource use. Now subsequently they show

variations in resource use and in after the development of the model but þ

CO-CHAIR MAURER: It translates, you know.

MR. BRATZLER: Yes, it probably does. I just wanted to highlight that, that in all of the conversations ran variations particularly in these submissions, most of the discussion was in variations in clinical care and clinical quality and things like that rather than, since these are Resource Use Measures. So I'm sure there's resource use variations on those.

MS. FANTA: Okay, so we have seven high and two moderate.

1c, is the purpose objective of the Resource Use Measure including its components and the constructs for resource use cost clearly described? Okay, and we have eight high, one moderate.

1d, are the resource use service categories types of resource use costs that

are included in the Resource Use Measure consistent with and representative of the measure concept? Okay, and we have nine high.

CO-CHAIR MAURER: Now we'll go ahead with Kurt's doing 2a1, and then we'll have Ben give us any more information.

CO-CHAIR ELWARD: This is the item on the definition and precise specifications that can be implemented consistently within and across organizations. They mention EMR, did we leave off EMR category in there?

Based on the quality of the assets so I'm not sure that would qualify, but it appears that they have defined the measure well.

Under S4 the target population is left blank, but I'm understanding that that is a population of COPD diagnosed within the last -- records are available.

And the data dictionary and the code tables are available, so I thought they did a good job in terms of how you count the

service, looking at the E&M codes, pharmacy codes, so I was fine with that.

In terms of -- any questions on, yes?

DR. SANTO TOMAS: Just to clarify that I thought that earlier Dr. Maurer had mentioned that although the initial intent was to limit it to those who had been newly diagnosed, and is that by spirometry? But actually you're going to revise that.

CO-CHAIR ELWARD: I made the comments in reference to that. Before it wasn't clear.

DR. SANTO TOMAS: So we were advised though to include actually anybody who has been diagnosed with chronic bronchitis.

MR. HAMLIN: Right. So, yes, I think the clarification point was in here in the clinical framework we were describing the quality side, and we didn't distinguish that adequately from the resource use side. So the resource use looks at anyone with a diagnosis

during the measurement year of COPD. On the quality side we have a couple different ways of identifying people for the quality measures alone.

So there's a little difference there, and I think we just need to be more explicit about how we describe that in the form.

CO-CHAIR ELWARD: So with the clarification that Janet helped with, I think that was well done.

CO-CHAIR MAURER: Okay, is there anything else that you want to add about that, Ben? Is that þ

MR. HAMLIN: Again, these are admin-based measures only. We're looking for aligning with QDM now, so you need to update your form. It's not QDS anymore.

However, that's right now they're administrative only. And we have found that all these data points are available, the pharmacy, the diagnosis codes. They're all

reliable within the administrative claims that we receive from the plans.

CO-CHAIR MAURER: Okay, thank you. Now for these measures that are, sort of use the same methodology come from the same place. We're going to try to pick up only the things that are really different between the measures because we've already talked about a lot of the general stuff.

So, Carlos, do you want to tell us if there's anything different about this from the asthma measure in the reliability and validity?

MR. ALZOLA: No. The same methodology was used for everything in terms of risk adjustment and how they demonstrated reliability and face validity. The only difference is in the selection of the populations.

CO-CHAIR MAURER: Okay, thank you. So, Richard, then do you want to talk about 2a2?

DR. STANFORD: Yes. I think we can take a lot of what we just talked about from the asthma measure and bring it over to the COPD measure, and I appreciate the clarification around the one-year measurement and the one-year identification period which is similar to what we did for the asthma part of it.

So if you look at 2a2, which is the reliability testing. And testing demonstrates that the results are repeatable, producing the same results a high proportion of the time. I rated this as a moderate.

The issues that I have with reliability testing is this issue around multiple populations being studied. And I don't recall you guys saying this data is in the Medicare population. Maybe I misread that, but I don't think it has been tested in the Medicare fee-for-service population.

MR. HAMLIN: Specifically fee-for-service, we haven't distinguished that from

just the general commercial Medicare provision, but it has been tested.

DR. STANFORD: Which there is a slight difference and I appreciate what you're saying, is basically Medicare-eligible populations in the confines of a commercial dataset, which those folks could be retirees as well but not necessarily. You're one of the new Medicare population.

So but overall, I think it was well done outside of the fact that maybe just some more data in a select population would be very helpful.

MR. AMIN: Can I just add a clarification, Ben? So the testing population listed on the application includes Medicare, but this is the same clarification that you provided us for the other measure that it's really intended for commercial?

MR. HAMLIN: This one is actually commercial Medicare and Medicaid because of the age range. The age range is different.

The asthma we did not test Medicare because of the age range cutoff at 64 precludes the majority of the Medicare population. And those that might qualify earlier would not meet the small sample size requirement.

MR. AMIN: Okay, so it would include Medicare?

MR. HAMLIN: This one does include Medicare, yes. What we don't distinguish is we don't distinguish the fee-for-service versus the general eligible population in Medicare. We might be able to in the future, but at this point in time we include all Medicare together and report it Medicare/HMO.

DR. STANFORD: So were you able to separate out the MAPD folks versus the Medicare-eligible individuals?

MR. HAMLIN: We did not do that, no. In the future we might be able to as we increasingly gain more data, but right now we're sort of, if it's aggregate data. But

it's still a huge volume and so we're trying to break into slowly into what we're expecting from the different plans and who we're comparing.

DR. STANFORD: So to me I think that how I usually state those populations is that they're Medicare-eligible as opposed to when you say Medicare, most people think fee-for-service dataset, which is fine. I think there's a little bit of differences between the populations but I'm okay with that.

CO-CHAIR MAURER: And there is no upper age limit on this?

MR. HAMLIN: Not at the current time, no.

CO-CHAIR MAURER: Okay.

DR. STANFORD: Well, there's no upper age limit, but what is the upper age limit in these datasets? Isn't it around 85 years of age?

MR. HAMLIN: Death.

DR. STANFORD: But technically

there is.

MR. HAMLIN: Yes, technically there is. You know, anyone who dies during the measurement year is excluded from the measures so, you know, it is anyone who's alive.

CO-CHAIR MAURER: Okay. Any other comments?

CO-CHAIR ELWARD: I noticed in, it may be more appropriate in the calculations, but I noticed that the population in each service category by cohort is, I'd like to understand a little bit better. It's 42 to 44 and then it jumps from 45 to 64, you know, clinically speaking for COPD seems a little strange for me.

MR. HAMLIN: Right. That again is because the resource use categories are based on utilization not on the clinical side.

So the clinical strata on the quality measures may be different from the resource use strata that are reported out for the ACC service categories. And there are

some weird ones when the age range is limited from 40 and up.

The resource use categories, because again they apply across a number of different measures, they are sort of restricted to 18 to 44 and if you start at 40 then you kind of go from 40 to, you know.

So there is some weirdness in the strata on the resource use side, but again those are derived from utilization patterns in the datasets that we have. They're not necessarily clinically relevant.

CO-CHAIR MAURER: So that brings to mind another question that I have. Somewhere in here it says that you want to kind of marry the resource use with the quality measures. And if you're going to use different stratifications and so on, how are you going to do that?

MR. HAMLIN: Well, the quality measures aren't risk adjusted right now. So we're basically comparing a risk adjusted

population to a more non-risk adjusted quality measure.

As the quality measures further increase and perhaps in the future become risk adjusted, we might be able to compare specific populations to specific populations.

But because the quality measures are, you know, population-based non-adjusted HEDIS measures that's really all we can do. So you're looking at specific subsets of the population compared to a population that will heap quality measure.

And we only report, the results are sort of plan level, total population, total medical against the quality measure and the total pharmacy against the quality measure. We don't report each male 18 to 44 against a quality measure specifically because of that very reason. Does that make sense?

CO-CHAIR MAURER: Do you have a question, Richard?

DR. SCHATZ: No, I just wanted to

clarify this difference in clinical versus resource. The ages that are listed in this measure are because they represent relatively homogenous resource use categories, correct?

MR. HAMLIN: Correct. That's correct. And it's mostly derived from experiential over the last five years of this and the datasets we've been using to determine these categories. But they don't follow clinical logic. That's where the disconnect is.

DR. SCHATZ: Well, just to clarify one thing. And within the quality measures results are reported with different strata. But are results reported in these different age strata or just adjusted for it?

MR. HAMLIN: They're adjusted for and the plan gets a report that gives them the detailed information both regional and national for each of these individual age and gender cohorts.

How much of that is publicly

reported is still under debate. Most of that right now is just in aggregate, so it's plan level not by individual HCC cohort if you will.

But that information is available so you can, we provide to the plan and researchers are allowed to buy a data download extract which is an enormous file, they can then use for research purposes.

But right now the public reporting is only this sort of high level aggregate plan level population reporting.

DR. SCHATZ: But again, that is where Janet's point would make sense that if the age strata are reported then it would be nice to see some harmonization.

CO-CHAIR MAURER: Well, and they state that they want to kind of bring them together.

MR. HAMLIN: Right.

CO-CHAIR MAURER: Any other comments about this? Are we at 2b1? That

would be Kurt.

CO-CHAIR ELWARD: I think the measure specifications are consistent. As the specification I think is clearly delineated. I think it's on S6-1 and it looks at both demographics, complete data or any clinical diagnoses.

Again one of the challenges with this measure will be, and I think we talked about this before, is whereas asthma has relatively few comorbidities for a lot of people, COPD has a lot of comorbidities going on.

And that's going to be, you know, it may be difficult to know whether you're measuring COPD or their CHF. So it'll be, you know, if you can give me any clarification on how you're going to approach that I think that would be helpful.

MR. HAMLIN: Well, our current approach is that in, you know, by risk adjusting to the specified level using the

HCCs and the 13 different cohorts were comparing relatively similar plan populations to each other.

We're not necessarily able to sort of draw conclusions about resource use with the number of comorbidities, but at least we're drawing what we think are fair comparisons between plans by using that risk adjustment that takes the multiple comorbidities into account over a scale versus just a yes/no which was sort of our previous iteration of how we adjusted.

So we're comparing it Plan A and Plan B relatively equitably by ranking how much of a train wreck these patients are to each other.

CO-CHAIR MAURER: And you're recording total cost, so you don't really care as long as they're risk adjusted.

MR. HAMLIN: Right.

DR. SCHATZ: Just in this issue of the cost quality issue, can you briefly remind

us what the quality measure for COPD?

MR. HAMLIN: The quality index that we use is a combination of two measures so we use the diagnosis of, I'm using spirometry to confirm a new diagnosis of COPD as one HEDIS measure.

The other one is the pharmacotherapy for exacerbations which has actually two rates in it.

So the quality composite that we use is actually a weighted composite using a weighting of the two results in the PC or the exacerbations measure, plus the results from the spirometry measure.

So it's a combination as a weighted average of those. And that's what the quality score is for this measure unlike asthma which just has one which makes it easy but the COPD has use of the weighted composite.

DR. SANTO TOMAS: Since there's administrative data, although at least in those, if you use it for those patients with,

had spirometry you're not actually able to get that particular number, right?

MR. HAMLIN: Well, the spirometry, yes, uses an administrative data measure that uses the procedural codes for spirometry and it's pretty limited.

It's looking at particular visits, you know, with particular providers to ensure we're not getting, we had a lot of noise in the past with spirometry just showing up in records from respiratory therapists and things like that.

But we think we've refined the quality measure side to be fairly specific to a confirmation of a new diagnosis. And those are available in claims code and that is a HEDIS measure that's been around for a number of years and we have found it to be reliable.

DR. SANTO TOMAS: Now when we were talking about the asthma you had mentioned, I don't know how fair it is to kind of ask that about this too but, you know, having linked to

electronic medical records, but is there a thought of that with this as well? Mainly I ask only because one of the main drivers of utilization in COPD is how severe the COPD is.

And of course, yes, these people have a lot of comorbidities. But the attribution to the COPD itself as far as utilization is related to the severity of the COPD.

MR. HAMLIN: Right. So we won't be attributing specific procedures to the COPD itself, but I think future iterations as the version 3.x of the QDM is released we are designing electronic medical records specifications on the quality side that take those different factors into account.

All the steps of the care coordination process and staging of the disease will be included in those specs. But those are specs that are still in development under contract to CMS. Those will not be in this until they've been thoroughly vetted and

validated through our usual process.

CO-CHAIR MAURER: It must be very hard to get disease severity, because I noticed in none of the measures did it really talk about disease severity.

DR. STANFORD: Yes, and asthma and COPD both have issue around, you know, you don't really have the severe event.

What you do have though, for instance, in COPD is probably maybe easier to do than asthma, is that we know that multiple drug therapy is related to disease severity and compliance is actually related to disease severity.

The more compliant a COPD patient is probably the higher severity levels they have, which goes to this issue, it's almost a chicken and egg issue. Like if you have a high cost patient you're going to have a high cost patient.

So that's why I was asking earlier around, you know, what your risk measurements

were in terms of your risk adjustments. Because if you take a COPD patient who's a high utilizer that's probably a much more severe patient than a low utilizing subject which with the goal guidelines they're all, it's some that's based on exacerbations as well.

So when you take that into account, does that help with the risk adjustment or is that something that you don't do?

MR. HAMLIN: It wouldn't help with the risk adjustment. Again, you know, on the quality side we kind of cover either end of the spectrum and not a lot in between because that's the limitation of the administrative claims.

I expect that in the future we'll be able to begin to look at some of those correlations in the middle of the spectrum that we can then relate to the utilization side. Right now we're still just looking at the snapshot because that's all we can do.

I mean again there's lots of interesting work, but we need more information before we can start testing those ideas. And so we're not just there yet but it's definitely on the plate.

I mean the COPD measures are like the asthma measures, very high priority moving forward and using clinically enriched data to start comparing to the utilization side, but right now we're just tracking what we can.

CO-CHAIR MAURER: Other comments on this one? Okay, I think we're at 2b2, and that is Richard?

DR. STANFORD: That's the validity testing and it really is around demonstrating that the measure elements are correct and measure the score correctly.

I guess my questions that I have and probably may fall in line with 2b3, is this issue around outliers. So I mean did you treat outliers in this particular population similar to how you treated in the asthma

population?

MR. HAMLIN: We look for O/E ratios below 0.3 or above 3 and those are identified as outliers.

But again for the results from last year and I'm expecting the results any day for this year that proportion to be very, very low. It's less than one percent. So again the plans are doing a very good job of providing the right data to us and calculating these appropriately.

DR. STANFORD: And how do you handle length of stay? Handle length of stay the same way or is it just when you do standard costing?

MR. HAMLIN: It's part of the standard costing process and it's recorded as a calculated metric in the results.

So you see average length of stay, days in average length of stay in that plan report that is given out. Those are components of that plan report, so it's there.

DR. STANFORD: Okay, great. So I was being, I think I rated it as a high from the standpoint of validity testing.

CO-CHAIR MAURER: Any other comments on validity testing similar to the asthma validity testing? And Kurt, you have 2b3 which we might have covered a bit.

CO-CHAIR ELWARD: Yes, I think the exclusions were well stated. They're very similar to the ones in asthma I think and I didn't see anything that, it all seemed very reasonable and important in that age group.

CO-CHAIR MAURER: Okay. And 2b4, Richard? Risk adjustment.

DR. STANFORD: Yes, and I think you have clarified some of the issues I had and I think it's fine. I think what you've done is in the scope of the data that you have.

You know, you can only do what's there in front of you. So I think risk adjustment in terms of, and I like the way you presented it in terms of it's really around

comparison across populations.

And cardiovascular disease to your point is probably the most to me I would think the biggest driver especially around severity of that disease.

And counting multiple instead of yes or no is I think is a much better, precise measurement in terms of risk adjustment. So I was fine with how they did that.

CO-CHAIR MAURER: Gerene, I think you rated this one low. I'm wondering if you had anything specific you wanted to bring up.

DR. BAULDOFF: No, I'm inclined to change my vote at this time.

CO-CHAIR MAURER: Okay. I just don't want to miss anything that people have identified.

DR. BAULDOFF: No, I don't have anything else to bring up.

CO-CHAIR MAURER: Okay. 2b5, Richard?

DR. STANFORD: Yes, it was the

same. I think what they've done currently to look at the score is fine and there was a lot of detail in how they did it.

I mean I think NCQA did a nice job of presenting their data at least from my standpoint in understanding exactly what was done.

CO-CHAIR ELWARD: One question I had and this may not be what NCQA will provide, but should that be a health plan or a physician group that says okay, I've got COPD patients that are causing, that really have a lot of utilization.

Can I use that data to break down how much would be used for COPD medication as well as cardiovascular medication? Can I break that out using your datasets?

MR. HAMLIN: The data that I have access to, no. But the plan could, in fact, go in and look at and categorize their pharmacy by category if they wanted to and apply the same methodology. What it wouldn't

give you is, you know, the expecteds are calculated for each plan individually.

So on the pharmacy side you wouldn't be able to look at CV medications expected versus -- COPD medications expected because, you know, we calculate that using all sort of plan data so we couldn't be that specific.

But you could, if you have particularly high utilization on the pharmacy side and, you know, perhaps on the inpatient side you might be able to break that pharmacy down individually and see where you're -- if it's skewed in one direction or the other. But you couldn't relate it back to the expected calculation because that's a group of all pharmacy.

CO-CHAIR ELWARD: Oh, yes. It would not be expected. I mean one of the challenges still with COPD is there's less but still a significant amount of clinical sort of nihilism among, but what can you do about

COPD. So that getting the right medications to the people has been a particular challenge over the last few years.

So it could be the high utilizers have real high pharmacy costs but they're for all the complications not the medications that they need.

MR. HAMLIN: Okay, thanks.

CO-CHAIR MAURER: Other comments about that? Are we doing 2b6? Skip the b6, and 2c, disparity?

MS. WILBON: 2c we can briefly talk about.

CO-CHAIR MAURER: So that's yours too, Kurt, the disparities.

CO-CHAIR ELWARD: Again it appears that there's attention to that. I think the same issues apply as applied for asthma.

I think it's even more important to have the differences in racial disparities able to be identified, and I think they do that.

MR. HAMLIN: It's currently information not available in admin claims reliably. The recent test showed all the way from zero to 98 percent availability of race/ethnicity data in the administrative claims.

So we're not there yet where we can, that we know the standard that we can apply. So we do gender. That's all we have right now.

CO-CHAIR ELWARD: Oh, I'm sorry. So you don't do race?

MR. HAMLIN: Well, we would measure it if we could, but every time we test it, there are plans that are actively not collecting race/ethnicity data for a whole host of reasons and we keep testing to see what the availability of the data is in these datasets, and again we see the range from zero to about 98 percent.

So we just can't include that as a factor because it's not there in the claims.

And you can't push that.

DR. BLAKE: Why would the plans not collect this data? You would think that would be a driver of costs.

CO-CHAIR ELWARD: I can give you a cynical answer, but the one that I've heard is that if you don't record racial disparities you can't be sued for it.

CO-CHAIR MAURER: There's some issue about giving race too. We can't require that people give their race.

DR. STANFORD: It's not a required field for health care in general. A lot of the racial data within these datasets is actually survey data.

Yes, I mean they don't have, some are trying to actually take Census data and plop it on top of there, but that's not a very good way to do it.

You can get race in the Medicaid data, but you have to go to each individual state for that. There's not aggregated in a

large Medicaid dataset.

MR. HAMLIN: There's also some issues of provider recorded versus patient reported. So there's some consistency issues as well, but I've also heard many cynical answers of, you know, protectionism and fear of protests and lawsuits and other things.

CO-CHAIR ELWARD: Although on the other hand I've heard a couple of medical directors simply say, actually one person walked in the office the other day and said, you know, we really need to be doing better about racial disparities.

And I looked around and said, where did that get in the business plan? But I think there are some people who really are looking at that more carefully. But it's unfortunate.

CO-CHAIR MAURER: Yes, it's hard to make interventions if you don't know who they are, you know?

Okay, any other comments about

disparities with respect to COPD? So that brings us to usability, and that's mine.

Are we going to vote?

MS. WILBON: That's okay. She's trying to get us done early here which is we appreciate that. We're going to go ahead and vote on the subcriteria for scientific acceptability.

MS. FANTA: Okay, so I'll start with 2a1. Is the measure precisely specified so it can be implemented consistently? We have nine high.

2a2, does the reliability testing demonstrate that the results are repeatable producing the same results a high proportion of time when assessed in the same population in the same time period and/or that the measure score is precise? Eight high, one moderate.

Okay, now we're going to vote on overall reliability testing which includes precise specifications and the reliability

testing. And the results are seven high, two moderate.

Okay, 2b1, validity. Are the measure specifications consistent with the focus of measurement and the measure intent? Eight high, one moderate.

2b2, does the validity testing demonstrate that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided, adequately distinguishing high and lower costs or resource use? Six high, three moderate.

Okay, 2b3, exclusions. Are exclusions supported by the clinical evidence or analysis of frequency and distribution? Is information about impacted exclusions for patient preference transparent? Four high, five moderate.

Okay, 2b4, for Resource Use Measures is there an evidence-based risk adjustment strategy or rationale or data which

supports no risk adjustment or stratification? Six high, three moderate.

2b5, are performance results reported? Do they identify differences in performance or overall less than optimal performance? And we're just missing one vote.

If everyone could vote one more time, please. It won't count your vote twice so if we could just keep voting. Got it, okay. So we have five high and four moderate.

And now is the vote on overall validity testing, which includes specifications which are consistent with the resource use or cost problem validity testing risk adjustment or identification of meaningful differences. Four high, five moderate.

And then 2c, if disparities in care have been identified do measure specification score in data and analysis allow for identification of disparities through stratification or results or is there a

rationale or date justifying why stratification is not necessary or feasible? Five high, four moderate. On to usability.

CO-CHAIR MAURER: So we have kind of a Supreme Court on the validity and reliability.

And I'm just wondering, before we move on to usability does anyone feel that we need more discussion around this or are we really reflecting sort of the impreciseness of the data that we can gather and the inability to make it more maybe clinically relevant?

DR. MOSENIFAR: COPD is just a vast area with a lot of comorbidity. So I think what you're seeing is really a true reflection of the mixed feelings about it that it really encompasses a lot of comorbid factors. Asthma is a much tighter disease.

DR. SANTO TOMAS: Yes, I mean I would in a sense echo that concern in a sense that if this is to be used for quality improvement, benchmarking and other things

that may not have been even originally intended for use, it then becomes at some point it becomes a disincentive to clinicians to look at this and all and say well, why am I being dinged for this?

When again, as he said, you know, COPD unlike asthma is a little harder to put into a good niche just because, although the comorbidities definitely, you know, heart failure and other cardiac problems definitely affect its course or at least hospitalizations and other resource use.

But a lot of it is driven by the severity of the disease itself. I mean and there are different, you know, from a simple as their, you know, patient's weight and nutritional status and those kind of things which I don't know how much of that is taken into account.

I mean some of that could probably be looked at in the administrative data as well. I mean, you know, malnutrition, for

example, as in a sense reflective of maybe how severe the disease is.

But there's just so much there that is really when we talk about risk or severity it doesn't really reflect the severity of the COPD but more of the comorbidity.

CO-CHAIR MAURER: So what I'm hearing is that you're saying that this is a much more heterogeneous disease than we're talking about with asthma, and the inability to actually record facts about the disease itself in that person you think might impair our ability to really look, you know, really accurately at the resource use?

DR. SANTO TOMAS: Yes, I think I'm hoping, I don't know how much of this is being done already in, for example, acute renal or kidney disease, but I believe there are either modifiers or codes.

For example, if you have somebody with chronic renal disease then you have, you know, then you have the stages, which maybe

we're not there yet with COPD.

And like as I said I think in renal disease they actually could classify, you know, this is stage 1, stage 2.

But I think maybe instead of pushing something, which I recognize is very important, it is a major driver of resource use in health care in general, but maybe it's premature until we fix this prerequisites. I know it's --

CO-CHAIR MAURER: Well, maybe also something that's bothering you and me too, is that we don't have a comparison between the use of these nondisease specific risk factors and actually disease specific risk factors. So we don't know if they really reflect the same thing exactly.

DR. SANTO TOMAS: Yes.

DR. STANFORD: Yes, and I think the other thing is how I see this from a data administration and risk factors is the heart failure, cardiovascular disease component.

And is there a way to marry the two as one measure? Because to me a lot of the risk factors for COPD costs are really related to how these patients, if they have comorbid cardiovascular disease, for instance, may be related to, equally related to not only their COPD treatment but also their cardiovascular disease treatment.

So those costs that are driven by total cost may actually be a factor of both of those equally. So I mean have you thought about that as well?

MR. HAMLIN: We are actually looking at different composites of the Relative Resource Use, so we're looking at utilization of COPD and CV if you will.

But for now the best we can do is a population to population balance and comparison between one population of a plan and another on the data that's available.

So we feel that these are the best measures of utilization for comparison of one

plan population to another as long as they're within a specific peer group that we can do with the data that's available.

We are looking at additional ways of incorporating total cost information and various composites of different disease states as a utilization measure.

But again there's, that it sort of falls into the same issue of attribution on the episode-based measurement approach, you know, which ones that are actually more relevant than others and how do you adjust for that and so on and so forth. So, you know, this is the best in class at the moment until, you know, further clinical information is available for us to use through other means.

But again that's why we've pretty much limited it to the population plan level comparison at this point because we feel that there are a whole bunch of host of factors, and we've had this problem with the quality side as well.

For our COPD measures we're really missing a great component in the coordination of care and staging of the disease and so on and so forth, but that's just not in admin claims and we're reluctant to, and it's not even really in a lot of medical records as well.

A lot of that's done through disease management companies or other, you know, employer programs or things like that that just are not part of that sort of record if you will that we can ask plans to use to report measures to us.

So we're always thinking about innovation there, but like I said this is kind of the best we can do at this moment in time given the information that's available. And that's why we do limited to the plan to plan, large population to large population.

We were hoping that, or the assumption is that most of those variations will balance out, because one plan population

when it's risk adjusted for severity of other disease states will look relatively similar to another plan population in a peer group that has the same risk adjustment approach for factors of cost utilization.

CO-CHAIR MAURER: So let me ask the group. Does this discussion accurately reflect the concerns that we saw up on the voting? Does anybody have anything else to add? Okay.

So we'll move to usability. This one is mine and this is, are the measure results reported to the public at large? Yes, they are. They are, we heard about the sort of standard reporting that NCQA does and I ranked 3a as high.

The measure performance results are considered meaningful, understandable and useful to the intended audience for both public reporting and informing quality improvement.

I've used the quality results which

are similar to, which are reported in a very usable way and I believe that the relative resource results are also reported in a very usable way.

And based on the examples that were given with the submission, I believe that these are very usable and understandable.

3c is, data result and detail are maintained such that the resource use including the clinical and construction logic can be decomposed to facilitate transparency and understanding.

And I believe that NCQA does that for the health plans that report to you. And I also, one thing I also want to mention about NCQA is they do do very extensive audits of their material and they do this over several domains on a regular basis.

And I think that's a really good, the quality control that they do that way by auditing. So I ranked all three of these high. Now we can vote.

MS. WILBON: Yes. Just what we've been doing kind of as staff, particularly across measures from the same developer, is to kind of share the voting results that you had for the other measure particularly around usability and feasibility, which I would say can be generalized to a developer.

Usually the underlying kind of construction and methodology are the same across all the measures. So what we'll do is I'll have Lauralei kind of read aloud your ratings for the other NCQA measures just so you kind of have an idea, not that you have to duplicate it but so we're kind of consistent.

And if there's anything different about this particular measure that you think needs to be identified before we vote so that can be kind of reflected in the votes we'll do that, okay?

MS. DORIAN: I guess I'll go subcriteria by subcriteria. So for 3a everybody said eight high and one moderate.

(Off microphone comments)

MS. DORIAN: Keep going? Okay, 3b was six high and three moderate. 3c, which is transparency was eight high and one moderate, and that's all.

MS. WILBON: So generally highs and a few moderates in there, so just so you have a little bit of context as you're voting. That'd be great, thanks. You can go ahead and vote.

CO-CHAIR MAURER: And feasibility?

MS. WILBON: We'll vote on usability and then we'll come back.

DR. SANTO TOMAS: That was from the previous one.

MS. WILBON: For the previous NCQA asthma measure, yes.

MS. FANTA: So 3a, are the measure performance results reported or suitable to report to the public at large in national or community reporting programs? Is there evidence that the measure performance results

are available for public reporting? Nine high.

3b, did submitted information demonstrate that results produced by the measure are meaningful, understandable and useful for information for quality improvement and public reporting or was a credible rationale presented? Five high, four moderate.

3c, are the data and result details maintained such that the Resource Use Measure including the clinical and construction logic for a defined unit of measurement can be decomposed to facilitate transparency and understanding? Six high, three moderate.

Okay, you can move on to the feasibility now, discussion.

CO-CHAIR MAURER: Okay. So the first one is, are feasibility, are the required data elements routinely generated and used, are generated in the same way that they're generated for asthma through

administrative claims? So yes.

And the next one is, the required data elements are available in electronic health records or other electronic sources. They're all transmitted electronically through the IDSS system I think?

MR. HAMLIN: Yes, currently.

CO-CHAIR MAURER: And 4c is susceptibility to inaccuracy, errors or unintended consequences.

Again, this extensive auditing process that they have helps to, you know, mitigate this and they chop off either end at 0.33 and 3.0, so they do mitigation around inaccuracies and errors, which is I think as much as probably can be done with the data. So I ranked these all high.

MS. WILBON: If there's anything different again for feasibility from this measure or from the asthma measure I would just kind of encourage you to voice that now before we go ahead and vote.

Otherwise, again, we can have Lauralei read the ratings, but I think if we can be consistent that would be great.

CO-CHAIR MAURER: Okay, I'll go ahead and read the ratings then. We had nine high for 4a, the byproduct of care. Nine high, available electronically.

Seven high and two moderate for susceptibility to inaccuracies. And eight high and one moderate for barriers to use. So are you all ready to go ahead and vote?

MS. FANTA: So 4a, are the required data elements routinely generated and used during care delivery? Nine high.

4b, are all the required data elements available in electronic health records or other electronic sources. If not, is a credible near term path to electronic collections specified? Nine high.

4c, are susceptibilities to inaccuracies, errors or unintended consequences and the ability to audit the data

items to detect such problems identified? Six high, three moderate.

And 4d, can the data collection strategy be implemented? Is the measure already in operational use or did testing demonstrate that it is ready to be put into operational use? Okay, eight high, one moderate.

CO-CHAIR MAURER: Okay. So we're a little bit early but what we're thinking we might want to do is break early, come back early and that'll help with the people who have transportation issues later on in the afternoon. Is that okay with everybody?

So we're ten minutes early so we should come back at 12:30.

MS. DORIAN: Katie, are you there on the line?

OPERATOR: I am.

MS. DORIAN: Can we open it to public comment at this time if anybody's there?

OPERATOR: Sure. And if you'd like to make a comment, please press star 1 on your telephone keypad at this time. And we have no comments at this time.

MS. DORIAN: Great, thank you.

CO-CHAIR MAURER: Timing still stands. Come back at 12:30.

(Whereupon, the above-entitled matter went off the record at 12:03 p.m. and resumed at 12:38 p.m.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

12:38 p.m.

CO-CHAIR ELWARD: Tom, are you here from Ingenix?

DR. LYNN: Yes, I'm on the line.

CO-CHAIR ELWARD: Thank you very, very much for being here.

DR. LYNN: No problem.

CO-CHAIR ELWARD: Let me explain how we've done this today. We've been taking the measures and then addressing the overall importance as the first issue, and that's what Janet and I have been doing.

And then initially going over some of the -- well, actually what we'll do is we'll turn it over to the rest of the panel to ask questions about the overall importance issue, and then I'll turn it over to you to see if you can explain the measure and talk a little bit about it and have some other questions answered if needed.

And then we'll be going through

each item and having individuals report on that and outline what their response is to various criteria that we have. And we'll be touching base with you intermittently for questions that they'll raise.

DR. LYNN: Sounds good.

CO-CHAIR ELWARD: Okay, great. Thanks. The next on our agenda is 1605, which is the ETG-based Asthma Resource Use Measure. And we're relying on Michael and Kathryn again to help us out.

The overall importance of the measure I think is very similar to what we had this morning. It focuses on clearly a national priority and one that's high impact.

This is the measure that describes the overall use of resources for asthma care, and specifically it says that it focuses on resources used to deliver episodes of care for patients with asthma.

And it will be defined as episode treatment groups using that methodology and it

describes the unique presence of condition for patients and the services involved in diagnosing and managing and treating asthma.

The measure that's proposed does demonstrate cost problems and the opportunity for improvement, and also I think outlines some of the experience that the developers have had. So I think that's acceptable.

I think the purpose and objective of the Resource Use Measure has been defined and the service categories at least in my reading seem to be very consistent with what they're describing what they're trying to measure and of what's important. So I would say it meets criteria in all of those.

Any thoughts or questions? Do we need to vote on that? Okay.

MS. FANTA: Okay, so for 1a, does the measure focus address a specific national health goal priority or was data submitted that demonstrated a high impact aspect of health care? We're just missing one vote. So

we have nine high.

Moving along to 1b, was data submitted that demonstrated resource use or cost problems for improvement that is variation and the delivery of care across providers and/or population groups? Okay, and the results are eight high, one moderate.

1c, is the purpose objective of the Resource Use Measure including its components and the construct for resource use costs clearly described? Seven high, two moderate.

And 1d, are the resource use service categories that are included in the Resource Use Measure consistent with and representative of the measure concept? And the results are seven high, two moderate.

CO-CHAIR ELWARD: Thank you. Tom, we'd like to hear from you about the measure 1605 and how, perhaps give us some background that might help us as we start consideration of the measure.

DR. LYNN: Sure. I think this is

a measure that's part of our sweep of creating episodes around diseases and conditions called the episode treatment grouper.

And the purpose of this rule and that product is to identify claims that should be part of an episode of asthma and then it divides those episodes into year-long segments of, treat asthma as a chronic disease.

This particular rule then goes on to identify how you would aggregate episodes of asthma across entities and measure how cost effective the treatment was with statistical methodologies.

In addition to that, these episodes are severity adjusted using clinical markers that come from within the episode that we call condition status factors, and clinical markers that come without the episode of possible morbidity.

And we are able to create a severity model for asthma and use these clinical-only markers not utilization markers,

just clinical diagnostic markers to identify higher cost episodes of asthma differently from lower cost episodes of asthma.

I want to say that the story for COPD is pretty similar. And there is one sort of special feature for asthma and COPD that has not applied to rules prior, and that is that grouper assumes that you can't have asthma and COPD at the same time.

So the grouper takes a special step, and if you do have both of those episodes at the same time due to some sort of coding error that the grouper will make a determination about which one is correct by counting the number of face-to-face connections between a provider and a patient.

And whichever episode has the most they're merged together and become either asthma or COPD.

So that's one special consideration for these two episodes that has not come up before in discussion of our measures.

DR. SANTO TOMAS: I'm sorry. Can you repeat that? How did you try to distinguish the two or þ-

DR. LYNN: Yes, so if you have two episodes running concurrently, one asthma and one COPD, then the grouper looks at, when you read the description about how the grouper works, it looks at the anchors which are, anchor records are records where the claim represents a face-to-face encounter between a clinician and a patient, and it counts those encounters. And whichever episode has the most that's how you label the episode and it lets those merge together.

So for example, if you had an episode of asthma that had four or five office visits and then at one point there was an office visit that was coded as COPD instead of asthma, then instead of having one asthma episode that's sort of missing a claim and one COPD episode that's very small, we merge them together and we say hey, this asthma had five

-- and the COPD only had one. So we're going to merge these episodes together and make them one episode and we're going to call it asthma.

DR. STANFORD: This is Richard Stanford. Can you give me an example of what would be a minimum ETG? Like what would go into a score to base somebody as an asthma episode minimally, like what would be the minimum criteria for that?

DR. LYNN: Minimum criteria would be that there would be an office visit or some sort of encounter between a clinician and a patient.

In the case of asthma it would mostly likely be an office visit but it could be an emergency room visit or it could be an admission. But the minimal ones probably would be the office visit and that claim can start an episode of asthma.

DR. STANFORD: And that's based off of an ICD-9 code, is that correct? Or is that based off of other codes as well?

DR. LYNN: That's a great question. So it's mostly based on the ICD-9 code for asthma, but it's also based on in order to, we only let certain claims start episodes.

Those are claims that we call anchors, and that requires that the provider be a clinician or that has a certain procedure code that shows that the clinician and the provider, you know, were face-to-face.

So it mostly uses the diagnosis code, but it uses the provider's specialty and the procedure code to make sure that it's a claim that we give the power to start an episode to.

CO-CHAIR MAURER: I have a question as well. I didn't quite understand. Let's say you have a patient who has an anchor record that starts an asthma episode grouping, and a month later they break their leg and they have that as a primary diagnosis and asthma as a secondary diagnosis.

Is that a record that starts

another anchor for broken leg or does that go into the asthma treatment group?

DR. LYNN: That's a great question. So the second, let's assume the second one is an anchor record. I think you said that and I appreciate that. The second one is an anchor record.

Then what would happen is that there would be, there's a set of tie-breaking logic that we go through to see whether that claim should group to broken leg or should group to asthma. And also there's a check to make sure that the procedure code makes sense for asthma.

So let's take an example. If the procedure was say we put a cast on your foot so you broke your foot. And the person dies, you put a diagnosis code of fractured foot on that claim and a diagnosis of asthma because they had, you know, they knew they had that sort of comorbidity.

Then the grouper says look, you

don't put casts on people's feet for asthma. So it doesn't do anything. It doesn't group the asthma episode at all. It has no effect with the asthma episode.

So now let's take an example where it was an office visit. The office visit and then the first diagnosis code is fracture of the foot and the second diagnosis code is asthma.

Then what the grouper would do, it goes through a bunch of tie-breaking logic, but one of the last tie-breaking logics is that the foot fracture was the first diagnosis code on the claim and asthma was the second one.

So an office visit could have had an effect on asthma or it could have had an effect on fracture of the foot. So what the grouper will do is it'll start an episode of fracture of the foot and the claim will actually group to that episode.

But the claim will have some effect

on the asthma episode in that it will, it can gather other claims to it and it has a what we call a phantom relationship with asthma. So it restarts the claim period for asthma.

It allows other claims related to asthma to group to the asthma episode through this anchor, but the claim itself would not group to that episode.

CO-CHAIR ELWARD: Michael?

DR. SCHATZ: On I think that related question, so certain costs that are not felt to be attributable to asthma during the year-long episode are not attributed to asthma. They don't end up in the quote, total cost, for asthma? Am I correct or not?

DR. LYNN: That's correct.

DR. SCHATZ: So there is a determination by the grouper or by something as to what would be really related to asthma versus what wouldn't?

DR. LYNN: That's correct.

CO-CHAIR ELWARD: It sort of holds

it, in my understanding in reading one of the documents that it sort of holds it as a what you call a phantom episode and then at some point the logic says okay, yes, this really was part of the asthma code, or there's a decision point where it drops off. Is that right?

DR. LYNN: Yes. So and we have a phantom relationship, so we have, let's say we'd take that example a little further and, you know, the office visit was for foot fracture and asthma.

And while they were fixing their foot they realized that they were having a little trouble with the asthma so maybe they did a peak flow test and they charged for it or something like that.

Then the peak flow test which can't sort of continue the asthma episode itself because that does not have the power of being an anchor, but it will group through the asthma episode via that office visit that had

the asthma ICD-9 code attached to it. So that office visit, although the dollars don't go to asthma itself, that peak flow test gets to the asthma.

CO-CHAIR ELWARD: Okay.

DR. STANFORD: So maybe I didn't read the document well enough, but so this level of analysis, it's at the ETG level or is that patient level? I mean that's where I was a little bit confused.

For instance, can a single patient have multiple ETGs and be kind of multiple times or is once they've been categorized as an asthma episode you don't include any other episodes within the confines of that patient only for another ETG? I guess is what I'm asking.

DR. LYNN: Yes, so in the case we're describing where someone during a year or during sometime has claims related to asthma and claims related to the foot fracture, what the grouper does is it says,

and this stuff that is related to the foot fracture is in a separate episode.

Remember, this is part of a sweep, well, it's part of a product that groups all episodes. So we put that into a foot fracture episode and then we only put the claims related to asthma into the asthma episode.

DR. STANFORD: Right. So I guess my question is, you're saying that the episodes are 365 days, is that correct?

DR. LYNN: That's correct.

DR. STANFORD: All right, so each --

DR. LYNN: Like a foot fracture wouldn't be because that's an acute illness, but yes, asthma --

DR. STANFORD: An asthma episode is 365, so in essence a patient is only counted once and an ETG is related to one patient.

DR. LYNN: That's correct. Now you could group two or three years worth of data, but a single patient is only going to have one

asthma episode every year, but you could group three years of data and have three episodes of asthma.

DR. STANFORD: May I assume we can ask any questions about the --

CO-CHAIR ELWARD: Please, please.

DR. STANFORD: Yes. Am I correct that I believe I read that you allow missing pharmacy data?

DR. LYNN: Okay, so here's what we do about pharmacy data. If the member has eligibility for pharmacy data, for pharmacy, in other words we have their pharmacy data, then we obviously group it to the asthma episode.

If a member does not have pharmacy data, then we don't group pharmacy data to the episode. And then in the analysis of the episode one of the adjusters for the episode is, did you have pharmacy data during this episode?

And it has to be that you've had

pharmacy eligibility during the entire episode. If you only had it for part of the episode then we don't group pharmacy, or we don't take into account the pharmacy dollars for that episode.

And sort of once you've created the episode there's a subsequent analysis, and during that analysis if there is an episode of asthma, there's an expected value for an episode of asthma that has pharmacy data and a different expected value for an episode of asthma that does not have pharmacy data.

And there's one of those for each of the severity adjusters for asthma, for each of the severity levels for asthma.

DR. STANFORD: So do I understand you correctly then, in the presentation of the information in each severity stratification there's a mean cost per episode with pharmacy data and a mean cost per episode without pharmacy data?

DR. LYNN: That's correct.

DR. STANFORD: Well, while I'm talking then in terms of what's presented, I gather that what's presented are severity stratified results as was mentioned.

You mentioned that a composite is constructed. Is that composite presented as well? And how exactly is that weighted or semi-exactly, how does one make a composite out of these different strata?

DR. LYNN: Yes, sure. So the, I think asthma has three levels. I'm not positive about that. If asthma has three levels and there's with and without pharmacy data, then you basically have six buckets, right?

So you take the data across all of the data that you have to analyze. And we also, you know, we do the analysis differently for different peer groups.

So if you were doing an analysis of pulmonary doctors taking care of asthma, you'd only look at episodes of asthma taken care of

by pulmonary doctors, and you'd look at the six different strata and you'd calculate across all of the data that you had for this analysis. What's the average cost in each one of those strata? So you have six expected values.

And then if a doctor, just take a simple case. We'd never do it for a doctor that had three cases, but just to sort of keep it from getting, you know, us having to be computers and we just be human beings.

But you could have a doctor that had two cases where they were level 2 and they had the pharmacy data, and maybe one case where it was level 1 and they didn't pharmacy data.

So you take the actual dollars spent on those three cases that the observed cost, and you divide that by the mean of the severity level 1 without pharmacy data plus the mean of the severity level 2 with pharmacy data plus the mean of severity level 2 with

pharmacy data.

So it's the observed cost divided by the expected cost and that gives you a ratio. And of course if you sort of did that for the whole set of data you'd get 1. So the expected value for that measurement is 1.

And of course, obviously you can expand that idea to many, many cases of asthma and, you know, you could expand that idea to the asthma and COPD or asthma and COPD and pneumonia as long as they were diseases that you would expect that specialty is there for.

MR. BRATZLER: Yes, I don't remember for the asthma measure. I know in the pneumonia measure it states that, I guess at the plan level I'm assuming you can either use actual payments or you can use the standardized resource costs for the episode.

DR. LYNN: That's correct. And it depends on what your interest in it. It also depends on the data you have available. Sometimes you may have data that doesn't have

price. It has all the utilization but it doesn't have the exact price, so you could standard price the data instead of that.

The other thing is you maybe you're not interested in the overall cost, you're really interested in trying to get at utilization. And then in that case you can use the standard price to get actual utilization.

Other times, you know, you want to look at utilization in the setting of the contracted rate and you want that to be a part of the analysis then you would use real dollars.

MR. BRATZLER: Boy, I think that'd be obviously particularly relevant if you were trying to compare across plans or across provider groups to use some, so the actual use of the actual cost would be misleading.

DR. LYNN: Again it depends on what you want to know. But if you wanted to know utilization, the actual cost would be included

and they should be using standard pricing.

CO-CHAIR ELWARD: Other questions right now? Great. Thanks, Tom.

Well, then let's start off with 2a1. Michael? We may have discussed some of this already, but þ

DR. SCHATZ: Right. Well, two things I'd say. Number one, I think that one concern with 2a1 is given that it's not necessarily standardized costs, the playing field I'm concerned is not level.

So that I don't think it can be implemented consistently across organizations if one organization is using standard cost and another is using actual payments. So I'm concerned at least as I see the definition of this measure, I mean the criterion.

And then, and I really defer to Carlos then on the rest of this, but I see that risk adjustment methodology is part of 2a1, at least a part of Carlos's report, and he does not feel that there is sufficient

detail according to the report. And so I defer to you, Carlos, for any other comments on 2a1.

MR. ALZOLA: Yes, the risk adjustment methodology is presented in a way that's a little bit mechanical in the sense that they present coefficients, they divide the range of risks into four groups, and without giving much explanation about what are the goodness of fits of the measures and the calibration of the risk adjustment model. So when you're looking at the ratios of service that's expected, it's important to see how well we're predicting the low end and the high end to really be able to assess whether those people are really high, are being efficient or not.

So basically what I would like to see in this submission is more information on the R-squares and the calibration of the models and also how they chose the cutoffs to the right, the risk score into four groups.

CO-CHAIR ELWARD: Tom, can you address that? I mean on the one hand I also understand that you're trying to get to a level that's quite a bit beyond what normally we see presented.

So we really appreciate that, but can you give us an idea of what Carlos, or address what Carlos has raised?

DR. LYNN: The best way to address it is to provide that information on which I can't do off the top, but we could definitely provide on the information that Carlos has requested.

CO-CHAIR ELWARD: So you have done the analysis and have the, you know, of the R values and things that Carlos has mentioned?

DR. LYNN: I can't say that we've done specifically for asthma, we actually did it for diabetes, but it's doable with the data that we have.

MR. ALZOLA: Yes, we saw it for diabetes and the R squares were good. I like

to see that. The one thing that wasn't there with the diabetes information was the calibration.

So how do the observed relate to the predicted? Some kind of graph of care would be very useful there.

DR. LYNN: And Carlos, I apologize. I'm not a statistician. I understand enough of this stuff to be, well, some of the stuff. So if my folks have any questions about what calibrations that we can get --

MR. ALZOLA: That's fine. But they would know what I'm talking about.

DR. LYNN: Yes, that's great.

CO-CHAIR ELWARD: Well, Tom, can you give us an idea just to make sure everyone knows, how do you develop the coefficients?

DR. LYNN: Sure.

CO-CHAIR ELWARD: I mean you explained some of that, but just to help us understand, you know, where they come from in the, based on your dataset. And do you go

backward from total utilization and then derive those or how do you do that?

DR. LYNN: Right. So first of all, our coefficients are calculated using a large dataset. We do have I believe it's either 25 or 30 million members worth of data. It is all standard price, so the coefficients are based on a standard priced claim.

So basically what is done is that you create a bunch of episodes of asthma through the standard grouping process that we've been talking about. And then we identify clinically with input from a pulmonologist, what's sort of the wide net.

If you were going to cast a wide net study, what clinical concepts affect the resource utilization on asthma what would they be? And we'd get a set of concepts and we develop markers that are only based of course on diagnostic information.

And then we run a model of, it's basically a regression model, with all these

clinical markers. Some of them are internal for the episode which we call condition status markers.

And then there are markers that are outside of the asthma episode that basically said hey, this person has congestive heart failure at the same time and this person had renal failure at the same time that had an indirect effect on the cost of asthma, because the actual cost is actually their episode.

So we do a model and then we present results to the expert, clinical expert, and then we know what's statistically significant and what wasn't and what's clinically significant and what wasn't. And then we come up with a finalized set and then we run it through the progression models and come up with those coefficients.

CO-CHAIR ELWARD: Okay, just so I'm clear, when the, say a pulmonologist gives you weights.

DR. LYNN: No, no, no, no. No, the

pulmonologist doesn't give us weight. What he gives us is what the markers.

So what we do with the pulmonologist, if he says hey, you know, if you have status asthmaticus, you know, in effect the resource utilization for asthma, and I think that if you have congestive heart failure it would indirectly affect the resource utilization for, of asthma. And they tell us what are the clinical concepts. And we ask them to cast a wide net because, you know, if they pick something that doesn't have a relationship it'll show up a little bit. So it's more important in that first step to include this.

But there's no way anybody no matter how brilliant they were, a clinician, could figure out what the coefficients are. So we let the model tell us what the coefficient is and then we review the results with the expert to help us, you know, come up with a final set.

MR. AMIN: Tom, we're having a little bit of trouble hearing you. You're going in and out a little bit. If you can either talk into the phone a little bit more, I don't know if you're on speaker phone, but that would help just because it's a very important part of the conversation. We don't want to miss anything.

DR. LYNN: Absolutely. Is that better? I apologize.

MR. AMIN: It seems to be, yes. Thank you.

CO-CHAIR ELWARD: Great, thanks. Other things before we go onto the specific elements? All right. Then let's start with 2a1, which I think you've addressed.

And Kathryn, do you have 2b2?

DR. BLAKE: 2a2? Yes, I had 2a2. Do you want to scroll it for me so everybody can see? So this is Tom? Is that who's on there? So Tom, I'm reviewing the data and sample aspect of this. And I had a question.

You have 25 million member sample to pick from. Why did you, or what was the reason for picking four million of those for the face validity testing, seven million for the reliability evaluation and 75,000 for the content validation testing? Why did you pick those subgroups and how were they picked?

DR. LYNN: Oh my gosh. I don't know.

DR. BLAKE: Sorry.

DR. LYNN: Sorry.

DR. SANTO TOMAS: Was that just a technique as far as sampling and validation? I don't know, maybe Carlos could even --

DR. LYNN: I'm sorry, I don't know the answer to that question. Is Dan Dunn on the line by any chance? I'm sure he's not.

MR. ALZOLA: Can you repeat the question?

DR. SANTO TOMAS: I'm just guessing it's a technique for validation like meaning, you know, certain proportion of your sample

you pick for well, testing, and then later on you kind of test your, like coefficients and see if that actually still applies, right? Is that --

DR. LYNN: Right. I know we looked at two different payers basically to do some of that validation work. But I don't know whether þ-

DR. BLAKE: I guess I was just kind of wondering, I mean why wouldn't you just pick, you know, seven million for all of them or something like that since you have a 25 million person database?

I mean maybe it doesn't affect the results at all. I was just curious. And when I was looking at it and wondering if there were differences between those three populations that could have impacted the results, that's where I'm coming from. But if we don't have the answer now that's okay.

And I guess the other thing that I didn't see and I couldn't find anywhere was

just some more characteristics of the population such as ages and things like that I couldn't find anywhere. So that would have helped a little bit when I looked at some of the other information.

In terms of the analytic methods, they seemed appropriate and your expected results were compared with benchmark database using regressions, which I thought was appropriate.

But tell me, why did you compare two different softwares? I couldn't really get why you were comparing SAS with the Resource Utilization Measures software.

DR. LYNN: What we were doing there is just trying to show that our SAS prototype comes up the same answer as our actual application.

So there was a question about repeatability of results and that's one of the ways we tried to show that we always get the same answer.

DR. BLAKE: Did you come up with some sort of I guess component of SAS software that you used and was the Resource Utilization Measure, is that something that's proprietary that is yours?

DR. LYNN: Right, so basically what this is is that we used SAS to write the prototype of the methodology and then we have a I think it's a C++ version that's actually packaged and sent out to people.

So it's a comparison between running the data through those two, but they're coded to do the same thing.

DR. BLAKE: But again going back, why do you need two different softwares?

DR. LYNN: Well, again we're trying to answer the question is if you use our method and you run it through two different creations of that method you come out with the same.

DR. BLAKE: Okay.

DR. LYNN: It's sort of an obvious

thing but we felt like the NQF asks that question. If you run it two different times do you get the same answer?

DR. BLAKE: Okay. So that's clear. And then the results, the consistency between the two softwares was obviously very good in your testing results.

DR. LYNN: Right.

DR. BLAKE: In that section, testing results, you said that there was a lot of consistency among the groups, the peer groups for instance.

What were you basing consistency on, because like in the pediatric peer group, and this might be getting into too much detail that you might not have on hand, but like there was a more than say two-fold difference in maybe the cost per episode.

I mean that doesn't seem consistent to me, but am I missing what you're trying to measure as consistent? Because it said that --

DR. LYNN: I'm sorry. I can't see exactly what you're looking at.

DR. BLAKE: You provided a table and this is the table for reliability and validity testing.

And in these tables you have your pulmonary peer group, pediatric peer group, family practice and internal medicine peer groups.

And if I understood what you were doing correctly, you were comparing to look at consistency across these nine different organizations, health care organizations.

DR. LYNN: Right.

DR. BLAKE: And the comment in what you provided was that they were consistent.

So I was wondering what you were using to call something as being consistent, because when I looked at it, just my eyeballing it like under the pediatric peer, there was a lot of variability, a two-fold difference between the health care

organizations.

So are you talking about something different than what I was interpreting it as? Am I making sense?

DR. LYNN: I don't think we were. I'm sorry, I can't see that. I can't find the exact table that you're looking at. Can you tell me what section it's in like SA 3?

DR. BLAKE: It was on Page 41 and the top of Page -- no, excuse me. The bottom of Page 40 and the top of Page 41. You reference the table SA Reliability Validity Testing.

And when I went to that table, this was your table that has different peer groups, pulmonology, pediatrics, family practice and internal medicine with the nine health care organizations.

MS. WILBON: Tom, this is Ashlie from NQF. So the committee is looking at the PDFs that are posted online. You have access to those, correct?

DR. LYNN: Well, I have access to what we used to build those.

MS. WILBON: Okay, yes. It probably would help to, just because they're referring to page numbers that are consistent with the documents we have posted online.

But if you refer to your attachment you have for SA, Item SA Reliability Validity that was attached to the document, they're looking at the results across peer groups tab. It's an Excel spreadsheet.

DR. LYNN: Yes, I've got it now, sorry.

MS. WILBON: Okay.

DR. LYNN: Results across peer groups.

MS. WILBON: Comma utils.

DR. LYNN: Right, comma utilization, okay. And then we have, so specifically we were looking at peer, pediatric peer definitions and we were looking at, because these aren't, is what I'm looking

at is different severity so you'd expect there to be differences.

DR. BLAKE: I'm sorry, I couldn't hear that.

MR. ALZOLA: I think you're looking at the next aisle on the spreadsheet which is --

DR. LYNN: Yes, it's reliability --

MR. ALZOLA: -- results across peer groups. And I think we're talking about the reliability across HCOs.

DR. LYNN: Right. Yes, so the episode quantity, so the two-fold difference if we're looking at pediatric peer definition.

DR. BLAKE: Right. Like cost per episode for instance.

DR. LYNN: Yes, from a low is like 709 to a high of 1,057?

DR. BLAKE: That, yes, that and there's another one under the family practice group, a low of 70 and a high of, what was the high? A high of 153, for instance.

I mean is that what you're, is that consistent? Is that what you're saying is consistent?

DR. LYNN: Yes, that's what we're saying is consistent. There are no other differences here around, you know, there's differences around what the geographies -- that so we're arguing that that's consistent.

DR. BLAKE: Say that last thing again. I couldn't hear it. Your voice dropped off.

DR. LYNN: Oh, I'm really sorry. Is that better?

DR. BLAKE: Yes.

DR. LYNN: Oh, we're arguing that's consistent because there are, you know, differences in what geographies these health plans are pulling from and other things like that that could make these different, so we felt like that was pretty consistent. I mean family medicine goes from 854 to 1,074, 1,090?

DR. BLAKE: Yes.

DR. LYNN: I mean, you know, these are completely different populations except that they're all commercial so we didn't do any statistical measure to say that that was --

DR. BLAKE: I mean it might have helped if you had said consistency in everything is within all the NCOs, or health care organizations are within 30 percent of each other or something like that just to give us some sense of what you were describing as consistency.

And on this same one, how come for the pulmonology group you don't provide the data across the nine different health care organizations? You have it for the pediatrics, the family medicine and the internal medicine. But why wasn't it there for the pulmonology groups?

DR. LYNN: I think sometimes we have challenges in our data figuring out what the provider specialty is and it's easier in

the primary care.

DR. BLAKE: Okay. All right, well, thank you.

CO-CHAIR ELWARD: This is Kurt. I'm just wondering if I, while Kathryn's getting prepared for her next comment, under let's say the family practice peer definition, this second version where we have, start with episode quantities, specialist visits per thousand.

Why do those end up being like a thousand, like 16 or 28 specialist visits per thousand? Is that, seems like that might be a little bit, that's a lot of specialty visits. Do you know how that's, it just seems like that's a lot of, pretty high rate.

I'm sorry. I'm on the tab, reliability across HCOs.

DR. LYNN: Yes, I got it.

CO-CHAIR ELWARD: And then the one that's, just two peer definitions per group. There's family practice peer definition and

then the table just below that which is the same thing. It says episode quantities, specialist visits per thousand.

Is that -- you calculate just how many visits that are coded as specialist visits per thousand episodes that are -- pardon?

DR. LYNN: Per thousand episodes. That's correct.

CO-CHAIR ELWARD: Okay.

DR. LYNN: It does seem a little high.

CO-CHAIR ELWARD: I guess, Michael?

DR. SCHATZ: I'm getting confused now. The issue of reliability of course sort of implies that you've got, you're doing the same. If you had, if it should be the same it is the same.

But when you're comparing across groups there's all, I mean if it were the same we wouldn't be doing it. So I guess I'd go back to Carlos.

The analyses that really reflect reliability, do they reflect reliability and the things that don't appear to be repeatable, should they vary? I guess I now don't know how to interpret these data when it comes to reliability.

MR. ALZOLA: Okay. When I look at reliability, at first I look at that as the ability to replicate the resource. So and there are a couple ways of looking at that.

One thing they did was to do the two software, the independent software development approach, so that's a way to validate their software. So if you, it's very, it's common to use from the surgical industry.

You have two independent programmers do the same process and they would all of arriving at the same results. So if they start from the same data, and using two independent software approaches you arrive at the same results, then you can say that the

results are reproduced.

On the other hand, when I look at these tables I look at them more from the point of view of validity. And honestly they talk about reliability across HCOs, I didn't see much use for it in terms of what we're trying to assess here.

Yes, there are going to be differences across HCOs, especially since we are not using standardized prices. Whether the differences are too big or too low, I really can't say.

DR. BLAKE: That's a good point. I think it was difficult. I wasn't sure if they were or were not using standardized pricing, and I think that would've made a more clear, you know, whether or not we were looking at changes across the, or differences across the HCOs, which would've been relevant if pricing were standardized.

Or like Mike, like you were saying, if from their 25 million population, they had

taken a subset of seven million and run the testing, and another seven million and run the testing, and another seven million and run the testing, and then if you got consistency among those three groups of seven million, then you would say that there was reliability there.

And that's how I would come from it from, you know, my background. And that's why I'm having a harder time I think, understanding it from this perspective.

CO-CHAIR MAURER: That would still be pieces of the same, that population would've been gathered in the same way so that would be okay. But it would be even better, wouldn't it, Carlos, to have a different population that comes from a different source to run it against?

DR. ALZOLA: It would be about the same as comparing two different HCOs. There will be differences then. Why? Why is because they are different, and how different they should be, I really can't say.

CO-CHAIR MAURER: Well, comparing them in two different populations is one of the things that's mentioned in the measures that are sent, you know, as part of that, as part of measuring reliability. So is that not so important then?

DR. STANFORD: But these are such a vast number of populations.

CO-CHAIR MAURER: Yes, I know.

DR. STANFORD: That even minor differences really go away. I don't see much of a statistical issue with this.

This is, you're talking seven million people. People predict elections of, you know, a country that 140 million people vote by a thousand people, this is a very robust statistical system. I think we are nickel-and-diming this.

MR. AMIN: Can I just add one piece of guidance here? In the reliability criteria, it asks whether it produces the same results a high proportion of the time when

assessed in the same population, in the same time period.

So this threshold of split sample validation would be sufficient, as Carlos is describing. So it doesn't necessarily have to be, while that's clearly a more robust method to evaluate across different populations, it doesn't, a split sample validation approach is sufficient.

MR. BRATZLER: But you're confusing the term validation, or you just used the word validation. So I completely agree with you on reliability.

MR. AMIN: Right.

MR. BRATZLER: They demonstrated a great reliability. I have the bigger problem with validity because it's been tested in their dataset, and that's where I struggle a bit with the, you know, that with these measures that haven't been tested across different sources of data for checking validity.

MR. AMIN: That's fair.

MR. BRATZLER: But reliability, I completely agree with you. They meet the definition.

DR. BLAKE: Okay, and then their finding statement I thought was appropriate and fine, so that's the end for 2a1. Is 2b2 next? Let's see.

CO-CHAIR ELWARD: 2a?

DR. BLAKE: 2b1, okay.

DR. SCHATZ: And you'll get my interpretation of 2b1. My concerns would be that number one, the asthma costs are not transparent. There is some determination as to what's an asthma cost and what's not an asthma cost, and that's not transparent. So I interpret that as being a validity issue.

And I personally think that since pharmacy costs are greater than 50 percent, to have any information presented when pharmacy costs aren't available is not valid.

So I really am concerned that, you

know, that any results are going to be presented, when greater than 50 percent of the costs are not represented. So those to me are two validity issues as I interpret 2b1.

CO-CHAIR ELWARD: Although one thing, if they make it clear, I mean on the, you know, to Ingenix's credit, if the data aren't there they aren't there, but they can at least say hey, it's not, and they can say it is.

DR. SCHATZ: Well, and people can look at it differently. I just think that when what isn't there is greater than 50, is likely to be greater than 50 percent, to see the less, I mean to see the other end, I guess what I'm saying, I'd rather have it not there than even try to interpret information that doesn't have the majority of the cost.

Again with the concept that is this measuring what you're trying to do, the cost of asthma care, and you've got a component that doesn't include more than 50 percent

that, I mean again we don't have to argue about that. That's a concern --

DR. LYNN: No, right.

DR. SCHATZ: -- and I mean, no question, I'd rather have it stratified that way than not know, but I just question whether that's useful information when that's such a high proportion of the cost of asthma.

DR. LYNN: Right, and I just want to point out that, note that I think, you know, I think that's a good point. I'm not going to argue with that point.

But just to say that that would be a pretty easy change to the methodology that, you know, we'd rather that you guys -- well, we can say that you look at the pharmacy information to do this instead of trying to do that.

CO-CHAIR ELWARD: I'm just wondering how, well, NCQA has the same problem though, right, if they don't have access to pharmacy data.

They just don't look at it.

(Off microphone comments.)

CO-CHAIR ELWARD: I'm just wondering if they can yes or no though.

(Off microphone comments.)

CO-CHAIR ELWARD: Pardon, go ahead.

MS. TURBYVILLE: So that's right, a measure can, in its specifications, state you must have this medical benefit, you should be in the minimum, and you'll see others that say you must have pharmacy to report it, so that's right.

And what Ingenix has done is they say if you have pharmacy data, you know, that's I'm guessing ideal. If you don't, they're giving you an alternative. They just say don't combine them.

But I think, Tom, you broke up a little bit on what you were responding to the committee that Ingenix would or would not, I didn't hear what you said in response to the pharmacy benefit being a requirement or not?

DR. LYNN: Can you hear me better now. I'm really sorry I've been having trouble with the, is that better? I just took my headphones off.

MS. TURBYVILLE: For now it is, so we'll let you know if it breaks up again.

DR. LYNN: Yes, so what I'm saying is, you know, we could certainly entertain, more than entertain, you know, that if we felt like this should only be done with folks that have pharmacy information, you know, we could make that change.

CO-CHAIR ELWARD: Okay, so you could do that then?

DR. LYNN: Yes, that would be pretty straightforward.

MS. ZIELINSKI: Well, this is Cheri with Ingenix. I think it's also important to know that pharmacy is not a requirement --

DR. STANFORD: She just said that pharmacy is not a requirement to get into the episode of asthma care.

DR. SCHATZ: To get into the episode, right.

DR. STANFORD: Correct. I guess that has to be.

DR. SCHATZ: Right, right. I mean that, I'm glad of that. Again, my concern is still there.

CO-CHAIR ELWARD: Let's go to 2b2, the validity testing.

DR. BLAKE: In terms of 2b2, some of my comments were the same as under the reliability, so I won't repeat those.

But I thought that the face validity was appropriate under the analytic method. And my comments under the testing results and the finding statement are similar.

CO-CHAIR ELWARD: Carlos, can you mention something about validity? In your comments you raised some concerns.

MR. ALZOLA: No, not really. I think that the only concern I may have raised is whether the database would be sensitive of

the over 65 population.

But the measure is, even though it includes people, I mean it's designed to work with people over 65 and is not being considered for that group. So that was my only concern.

CO-CHAIR ELWARD: But you had mentioned the data integrity checking wasn't found?

MR. ALZOLA: It was not found, but this is a very large database vendor so they really, and I'd be shocked if they didn't do any, a very thorough checking of their data.

DR. STANFORD: This is the impact data?

MR. ALZOLA: Don't know.

DR. STANFORD: Is it the impact data?

DR. LYNN: Yes. It's the --

MR. ALZOLA: Impact, yes, I know.

DR. LYNN: Yes, the impact dataset. It's actually, that's what it is. I don't

know why somebody knows the impact database.

DR. STANFORD: Because I use impact a lot.

DR. LYNN: Okay.

DR. STANFORD: The thing about the impact data which is to issue around costing, is that all of the costing data is pretty standardized across multiple health plans.

So if you're going to do a validation test this is what you do it in because you don't have issues around paid amount versus charged amount versus no amount. So it's appropriate into your point, a lot of that's already been done ahead of time to make sure that it does meet those standards.

CO-CHAIR ELWARD: So it should have a very high level of data integrity?

DR. STANFORD: Yes.

CO-CHAIR ELWARD: Okay, perfect. Thanks.

MR. ALZOLA: And just being, go where what I seen in the submission that that

was the only pondering. I had to complete that section and I didn't see it, but I really didn't have any reason to suspect it.

CO-CHAIR ELWARD: All right. So that should be a strong point then. Okay, thank you.

Then exclusions, Michael?

DR. BLAKE: Me? Is that me?

DR. SCHATZ: I think it might be me.

DR. BLAKE: Oh no, sorry.

CO-CHAIR ELWARD: 2b3.

DR. SCHATZ: Well again, Carlos, maybe I'll defer to you because you did seem to have concerns about the exclusions. That there wasn't sensitivity analysis, reasons aren't really addressed.

Again, just I'm not sure this fits into exclusions, but I am again concerned as to there are obviously certain costs excluded, but it's not transparent what they are relative to what are considered by the group

are non-asthma costs.

MR. ALZOLA: I have the benefit of having reviewed many of their measures so they know how they work. So a lot of the things that I list in here is because they do not show in the submission.

But I have learned through other reviews of how they choose their outliers. Why do they exclude the low outliers and how they define the high outliers.

And I think and, Tom, you can correct me if I'm wrong, is that the reason they exclude the low outliers is because they tend to be incomplete episodes. So it's not, all the data is not there and the cost would be artificially low.

And as for the high outlier, they Winsorize so they just exclude like the top two percent or something.

Again, that's not explicitly mentioned, but I think that's how they've done it in other measures and I'm guessing that's

how they did it here.

DR. LYNN: Yes, it's done the same way, where we do exclude the costs if it's the fifth percentile. We exclude all episodes that are below that cost. That our thinking is that many of those are incomplete for some reason, or mistakes, or rule out diagnoses, things like that.

And then we cap the high ones because they're real, they really happened. We just don't want to necessarily, you know, include all of those dollars in the cost of that episode. So that's why we do what we do.

DR. SANTO TOMAS: And then maybe just to clarify or follow-up on awhile ago, we mentioned those people who have both the diagnoses of asthma and COPD. So I take it then the reverse of what you said a while ago, then if somebody has both diagnosis, if somebody has more COPD, then do you exclude those patients then?

DR. LYNN: So what we do is we

would, right, we would include those cases.

DR. SANTO TOMAS: So you would exclude actually those?

DR. LYNN: We would exclude them from the asthma role, they'd be included in the COPD role. But yes, we exclude them from this role.

CO-CHAIR MAURER: In the previous measures we were looking at before lunch, very high costs, people like cancer patients and transplants and so on were excluded.

Does it turn out that when you Winsorize the top two percent, you're effectively excluding those kind of patients, or how does that work?

Because you certainly could have dual diagnoses of say lymphoma and asthma, and asthma could be the first diagnosis on the claim in some cases.

DR. LYNN: Yes, so I think that, you know, that the rules that you were looking at before, I wasn't in this morning but I know

that there are an overall, in the past they've been in overall patient cost. Whereas, we're trying to divide out the costs directly associated with asthma.

So, you know, that's less of an issue because you're right, I mean if sometimes the episode, if sometimes the claim has the asthma first and lymphoma second, it may end up in asthma. But only under narrow circumstances where the claim could be for either one of those things.

In other words, even if the first code on a claim for say a bone marrow, well, not for lymphoma but some sort of, you know, some sort of lymph node procedure, even if the first code's asthma and the second code's lymphoma it's still going to group with lymphoma. But there are cases where that could happen.

But it's basically, you know, trying to not include in the denominator things that, of the cases that may have been

extremely complicated or extremely costly.

And you can include them but you only include them at the dollar amount, at the threshold, the 95th percentile threshold. And there's many reasons why, you know, something may be very high cost.

CO-CHAIR ELWARD: Other questions? Let's go to 2b3, the exclusions.

DR. BLAKE: That's mine, 2b3, no.

CO-CHAIR ELWARD: I'm sorry, yes. But 2b --

DR. BLAKE: 2b4?

CO-CHAIR ELWARD: Sorry, so any other thoughts on 2b3? I'm sorry. All right, 2b4?

DR. BLAKE: This was the risk adjustment information. So this is a question for you, Tom. When you determined your severity levels, this is based not on what I guess we as clinicians call different severities of asthma, but this is based upon the comorbidities that were assigned to an

asthma episode?

DR. LYNN: Right. So we take those, we have coefficients for each of the markers whether they're condition status which are inside the episode, or comorbidities which are outside the episode.

Can everybody hear me okay?

MS. WILBON: Yes, we can hear you.

DR. LYNN: Okay. And we build a score which is a real number for how much, you know, based on the clinical markers with the expected resource utilization for that episode of asthma, compare it to the average cost of all episodes of asthma.

And then what we do to actually pick the level is we take a distribution of that real number and we look at places where the number is relatively, the severity score is relatively flat for a period of time and then they jump up in certain cases.

And that's how we choose where we have our different severity levels, trying to

maximize the homogeneity of the bucket that's created by that threshold.

We also want to choose, create buckets that have a number of cases of, you know, a high enough percentage of cases that our users won't have trouble creating expected values for those buckets.

So those are the two criteria we use when we create the severity level from the severity score.

DR. BLAKE: Okay. All right, thank you. So that answers the question. I feel like that was appropriate.

And the same goes for the stratification method, with what he just described that appeared to be appropriate as well.

DR. BAULDOFF: I have a quick question. I had some concerns about the risk adjustment just because I didn't see the specific detail related to the asthma that we had discussed earlier.

I guess my, or just for clarification, if because they were able to do it for diabetes and that it can be done for asthma, do we base our vote by what we've seen here? Or do we base our vote on what we know that they're able to fill in for us?

MS. WILBON: We'll be asking you to base your vote today on what you have in front of you. We'll get the additional, and we'll be documenting your rationale for those ratings.

That, your ratings along with your rationale for that, will be passed on, and any additional information that they give us between now and the Steering Committee meeting, that will go to the Steering Committee. But that's kind of in an effort to kind of keep the process moving.

But obviously, the object of using you guys is to identify issues like that so that the Steering Committee can evaluate them. So we'll just be asking you to evaluate what

you have and then we'll move that forward.

CO-CHAIR ELWARD: Yes. Tom, there's an item on 17 of your ETG construction logic, Asthma 2, the Microsoft Word version.

I think my impression in looking at that is that you already have done analyses of how comorbidities affect this and the risk adjustment. Is that correct?

DR. LYNN: Yes. So that's part of creating the models, right, is looking at how the comorbidities affect the cost of the episode.

CO-CHAIR ELWARD: Okay. Can you describe for us how you've done that risk adjustment, how you've tested that? Seems to be some, you know, uncertainty about how refined that is.

DR. LYNN: Right. So what we can do to look at how refined it is, is to provide the R-squared for the different severity levels, and how that predicts resource utilization.

And that's something we have not done but can provide to you. It's something that we did for diabetes because when it was evaluated they asked for us to do that.

CO-CHAIR ELWARD: Okay. Michael?

DR. SCHATZ: Well again, as I read, Carlos, your report, it looks like a lot of the information you would've liked to have seen wasn't there. So based on what is here, do you think we have enough information to answer that question?

MR. ALZOLA: Personally, I need to see that information about R-squares and calibration to answer.

DR. BLAKE: At this point you would consider it insufficient to be able to make an assessment?

MR. ALZOLA: Yes.

CO-CHAIR ELWARD: And number 5?

DR. BLAKE: 2b5?

CO-CHAIR ELWARD: Yes.

DR. BLAKE: This is type of score

and the four types for a continuous variable count, rate or proportions, and ratio.

And I could not find this S12 Sample Score Report to see exactly what they were doing. I don't know if you all are able to pull it up.

So while you're doing that, my interpretation though that they, of the description of what they provided seemed appropriate in terms of the interpretation.

And then in the detail score estimation, they again provided observed versus expected ratios, which was appropriate.

But they did have in there, they referenced Section S9.5, which doesn't seem to have anything to do with continuous cost measures. It had to do with complementary services. So I think that was just a wrong reference to a section in there.

Is that what you thought? Were you able to pull up that?

MS. DORIAN: Do you know what page

that's on, Kathryn?

DR. BLAKE: That's the S12 Sample Score Report? Okay. So those are ratios? Number of -- well, quite honestly I'm not sure how that report helps understand anything.

I mean they said, the types of scores are continuous, variable, count, rates and proportions, and ratios. Is that what that is there?

MS. WILBON: So my interpretation is, in the form that we give them to fill out online we instruct them to select the different types of scores they report out using so, and various parts of the report they probably utilized those different types of scores.

DR. BLAKE: Okay.

MS. WILBON: It might help perhaps if Tom could kind of walk through a sample report and explain how the scores are used and what information that provides by the different types of scores, which was the

purpose of that section.

DR. BLAKE: Okay.

CO-CHAIR ELWARD: Yes, Tom, can you go through if you have a S12 sample score? There aren't any asthma measures on that, but again, as Ashlie mentioned, maybe you could walk us through how this would work.

DR. LYNN: Right. So what, basically what is done, I mean the main measurement is the O/E ratio.

And what the process, it's an O/E ratio metric, is to look at the different buckets of the asthma based on the severity score, and whether or not they had a pharmacy benefit or not.

And then assign a cost to those, an expected cost to those buckets, which is the average of all of the episodes across the peer group based pediatrician. And that gives you the expected value for each of a physician's episodes.

So the numerator of the O/E ratio

is the cost of all the episodes of asthma. The denominator of that ratio is the expected costs for each one of those episodes come together that we calculated across the peer group for each of those severity buckets. And that's how you get the observe to expected ratio.

And then there's a technique that's referenced in lots of RAND work, some on these sorts of measurements that creates a standard error for around that O/E ratio, and we use that methodology to create a confidence interval around that measurement.

DR. BLAKE: Thank you, that makes it clear.

CO-CHAIR ELWARD: And then 2b6?

DR. SCHATZ: Well, that's easy because they aren't.

CO-CHAIR ELWARD: And 2c.

DR. SCHATZ: And again, well, I mean I think if data were available I do think it'd be there.

CO-CHAIR ELWARD: Okay. Let's go to use, do you want to go to that? Well, just --

DR. STANFORD: Well, can I ask, I'm sorry. And I know about if the data were available it would be there.

And maybe it goes back to what Carlos was talking about not having the data. And if you're generating these episode groups, how would that factor into your, how would these disparity measures factor into that episode group? Or is that just an outside variable that patients are stratified by that outside variable as opposed to the episode treatment group, right? Is that what --

DR. SCHATZ: Yes, well, I mean you're right. I mean that's a good question. If the information were there, one could theoretically use it the way people would want. But I guess that's a good question to ask.

Are there any of the measures you

use that, Tom, that you adjust for gender, race, ethnicity, socioeconomic, something? How would that figure in, especially where if you had information on socioeconomic status or race/ethnicity, for example, how would that be figured in such a measure?

CO-CHAIR ELWARD: And Tom, we also realize that that's hard to get.

DR. LYNN: No, no, right, right. So we do have, we do that for of course gender and age because we have that. We don't do it for these other markers that, you know, we all think would be very interesting, including me.

You know, I think what you would do is, potentially it would be another marker in the model, right. So the race would be a marker in the model, or the socioeconomic status would be a marker in the table, and you could see how that affected the cost.

You know, one of the potential concerns is that just some of these markers have a, for example, I've seen some evidence

that, you know, a socioeconomic marker that marks someone as low income is sometimes, those folks are actually less expensive, because they live in places where they don't have access to care and it may be harder for them to go to the doctor and those sorts of things, when they're in the commercial population particularly.

So, you know, I think you'd have to look at, you don't want to create a methodology that says that you should spend less money taking care of folks without, with fewer resources.

So although I think we would all like to have those markers and we'd all like to use them, I think you'd have to be a little careful with them.

DR. SCHATZ: But I guess I'd make the point that when you've got potentially ten different results, I'm thinking four severity levels, a composite with and without pharmacy data, you'd almost have to use it in the model

as opposed to stratification by those factors.

And I would submit that maybe stratification would be more useful. So I think it's a bit of a disadvantage to have so many outputs. It sort of precludes stratification for this situation, which I think might be better. Anyway, it's just an observation.

DR. STANFORD: It does say stratification and not including it in a risk adjusted model.

MS. WILBON: Right. And I'll just clarify as well that the criteria for this does indicate that if there is a valid rationale for why it's not addressed in the measure that that is sufficient. Why it's not feasible is part of the criteria. So again, just kind of context for why you're voting.

CO-CHAIR ELWARD: Any other items before we vote on -- my own, I was just talking with Ashlie.

It seems like there are a number

of, you know, the ETG methodology is action in many ways very exciting because it has, you know, the potential to get past just overall utilization.

And so you just dump everything into the box and see what people cost, as opposed to, you know, looking at a specific episode of care. At the same time there are a lot of questions about how you can do it right.

So my understanding is we can submit what we, we should vote on these today and, you know, with what we have.

But there would be an opportunity for us to feed specific questions back to Ingenix and say, can we receive these answers again in clarification, the R-squared values, things like that, so we could reassess that at some point?

MS. WILBON: So I would just say, particularly since this discussion, there's been a lot of I think just trying to

understand the underlying methodology for this.

As we go through each of the subcriteria, staff can try to summarize kind of what we've got.

And I think it's going to be really important for us to make sure that we've actually captured your sentiment about each of those, and make sure that we've captured any follow-up items for that particular criteria.

So if, co-chairs, it'd be great if you could kind of help summarize maybe what you heard. We can make sure that we've reflected that in our notes.

So that as we move forward, and I think a lot of this discussion, even though it's a little arduous for this measure, I think a lot of the things you'll find as we move forward to the other Ingenix measures, again just like NCQA, a lot of the stuff will carry forward and then we can pull out some of the condition specific stuff.

So we appreciate the deep dive because this is the stuff that we really need to address and move forward.

DR. SCHATZ: Well, then I think in that regard I guess I'd make one other point.

MS. WILBON: Sure.

DR. SCHATZ: One of the things that stratifies this is disease specific, and that seems to be heavily related to exacerbations.

And what concerns me a little bit is, again if we're trying to look at overall management practices that are positive, people end up with exacerbations can be thought of as a failure.

And yet by adjusting for the exacerbations you're sort of eliminating that. Now I know you can look and see what proportion of patients fit into various things.

But I mean I agree with you, Kurt, that it is nice to try to dive down and get at different segments, but again I would just add

to the mix here, this concern that the disease specific exacerbation issue may hide some poor care issues, as a concern.

CO-CHAIR ELWARD: And Tom, am I understanding this correctly that what would also trigger an episode of care would be like a planned visit for asthma? Like if I decide to have one of my asthma patients in and treat them but code it as asthma, that would start an episode of care, is that correct also?

DR. LYNN: That's correct.

CO-CHAIR ELWARD: All right.

DR. SCHATZ: Yes, unless it's, I mean starting an episode that's good. I'm concerned about the severity stratification.

CO-CHAIR ELWARD: Well, with that one, let's go --

DR. STANFORD: I'm sorry.

CO-CHAIR ELWARD: Yes.

DR. STANFORD: Can I ask one question before I answer this question?

CO-CHAIR ELWARD: Okay.

DR. STANFORD: Because it's going to help me answer this question.

CO-CHAIR ELWARD: Yes, please.

DR. STANFORD: So this, Tom keeps talking about the logic of the program and how it's going to relate to getting a patient into a particular episode group.

From a plan level perspective, and this comes out with the issue about implementation consistently, how does a plan implement this particular program in their database? I mean is it, that's my question to Tom, I guess.

CO-CHAIR ELWARD: Yes, Tom, do you understand the question?

DR. LYNN: I think so. I mean, you know, we have a lot of different customers who are doing, were using this methodology and they may use them for, you know, we have employers that use them for measuring their employees.

And we have, you know, they're done

for certain sort of financial analysis. But they're also used for measurement and they have been used for pay-for-performance programs along with quality metrics. They have been used for public reporting of scores.

Now having said that, we don't necessarily use asthma alone in that particular case. We may use asthma along with other episodes that are treated probably by a different speciality which helps define a peer group. So those are the ways that some of our customers use this methodology.

CO-CHAIR MAURER: I can see how plans would use this. I mean if they have a group, if they can look across a set of pediatricians and look at their different costs of care, they can use that for pay-for-performance.

It means one thing about this methodology unlike NCQA, is it attributes to an individual provider, to a practice, to a much more granular level that can be rolled up

to, you know, plan level actually.

DR. STANFORD: And what I'm really looking at is the, what is the programming burden on a health plan? For instance, if I'm a small health plan with very limited resources in even being able to do this type of programming, will I be able to do it? As opposed to, you know, if I'm United Healthcare or Aetna, easily I can find somebody to do it.

That's what I'm talking, that's what my question really is around. How simple is it for somebody to implement?

DR. MOSENIFAR: I guess I can think of that, they're such a massive silo data with a bunch of little silos.

So if they have your company that your particular interest in the silo sub A, you could go to them and say that, you know, what is your database for silos subtype A, and they can really extract it and give it to you.

Although, you're right though, your massive silo may not be applicable to your

company, but those subtypes will be applicable. So I think that's really the issue.

MS. WILBON: Tom, could you kind of expand on that a little bit for the committee, and this is actually kind of going to come in again in the usability and feasibility discussion.

I think it's a very important question. I wonder if we should maybe table it until we come back and then we'll queue Tom to kind of bring that back, if that's okay with everyone.

We can get through the scientific acceptability. Again, I think all this discussion is going to help us for other measures, and even though we're moving slow, it'll help us later.

DR. LYNN: Do you want me to talk about that when we talk about feasibility then?

MS. WILBON: Yes, I think that

would be better.

DR. LYNN: Okay. Great.

MS. WILBON: Thank you.

CO-CHAIR ELWARD: And more or less not to delay for the -- going back to what Michael had said earlier, whether you use charges and how you create the user inputs. I mean you mentioned, I'm trying to recall your point again, because I thought it was a really important one.

DR. SCHATZ: Well, I was concerned that different plans can do it differently. They can use what was actually paid or they can use standardized costs. And so I think comparing across plans, I mean I can see it within a plan, you know, Janet's point is good.

But across plans, where one plan uses standard costs and one plan uses payment, and now you want to compare. I think that's difficult.

CO-CHAIR MAURER: But isn't what

Ingenix does, just produce the groups, and then you who have purchased it from them do what you want to do with it? Isn't that, if you want to use standardized costs, you do that or whatever, no?

CO-CHAIR ELWARD: Well, Tom, yes, can you answer that for us? I mean do you have a preferred way of doing it? Or how do you handle that because you can get a bunch of different versions of the same thing.

DR. LYNN: Right. And that's of course what, you know, you guys are helping to start this ball. But, you know, we have clients, and if I say anything wrong Cheri will jump in if she's on the line.

But we have clients that sort of use this product at different levels. So, you know, we have large clients that they just purchase the episode creation part and then they have their own proprietary methodology for, you know, creating the O/E ratio or doing some other measure of costs.

And then we have smaller clients, not always smaller clients but clients that are maybe less technically sophisticated, that would buy the product that not only has the ETG in it, but also does these calculations to create an O/E ratio and implement, you know, most of that net level of methodology as well.

So it depends on the client's IT sophistication, and some are more sophisticated and they build that second part themselves.

And they're less sophisticated than, or they don't want spend the money I should say on that second part, then they buy a product that has that built in as well.

MS. ZIELINSKI: Yes, I would just add to that. This is Cheri. You know, we have never really came out and said this is the right way to attribute providers to episodes. This is the right way to make peer groups, because each, you know, each user, each application of the product really calls

for different ways and different methods to do those kinds of post-grouping activities.

So that is why like, you know, sometimes a provider can get a report card using ETGs from Aetna, you know, and their methods are different than what CIGNA's report cards would be using ETGs as well.

So there are post-processing decisions. We give guidance. You know, we've told people how different options to do things and why they would pick one option over others. But there has not been any, you know, this is the way to do X.

DR. LYNN: And we have white papers on those subjects.

MS. ZIELINSKI: Correct.

CO-CHAIR ELWARD: Thank you. Why don't we go ahead and vote then, unless there's other questions. Go ahead, for reliability.

MS. FANTA: So I'll start out with 2a, reliability. Is the measure precisely

specified so it can be implemented consistently? Okay, so two high, six moderate and one low.

MR. AMIN: I'm going to attempt to summarize a little bit of the discussion to make sure that we have it all captured.

It's not meant to be exhaustive, but if there are other inputs, please, because the conversation sort of went along multiple of the criteria at the same time.

So there was a bit of discussion around what the comparison of without pharmacy claims would mean, the interpretability of that would be complex to say the least.

The comparability between actual prices and standardized prices, more detailed on the R-squared which actually goes a little bit more in the risk adjustment but was discussed at this point too.

Where the cutoffs are determined for the severity levels, and I think that was the majority of what we had discussed.

CO-CHAIR ELWARD: And I guess one way of my looking at it is, is while the flexibility of their, you know, input and analysis is a strength for their clients, it may be a challenge in terms of creating a generalizable measure that could be used, that people would know is consistent across the plans.

MR. AMIN: Okay, thank you.

MS. FANTA: Okay, and moving onto 2a2. Does the reliability testing demonstrate that the results are repeatable, producing the same results a high proportion of time when assessed in the same population, in the same time period and/or that the measure score is precise? Okay, and the results are three high, five moderate and one low.

MS. WILBON: And so for this one and I'll open it up to my colleagues to piggy back on anything that I have. Again, I'm going to try to attempt to summarize with some of my notes here.

But that they use some split sample testing in terms of trying to determine reliability, but then later it was thought that maybe this method is more applicable to validity testing.

They were missing a few details about how the testing population was identified and some of the characteristics of that population.

But that repeatability was also demonstrated by the programming of the measure and the two different, not different databases but in the two different software between SAS and their Resource Use Measure software.

There was some discussion in reference to the scientific acceptability, reliability and validity attachment and how consistency was determined.

But ultimately with input from Carlos, that the TAP felt that there was reliability demonstrated based on what they have submitted at the threshold level of

repeatability and reproducibility.

DR. BLAKE: Ashlie, I think, yes. Early on I think you didn't have it quite right.

MS. WILBON: Okay.

DR. BLAKE: At least how I recalled it. When we talked about the split sample, that had to do more with reliability not validity.

MS. TURBYVILLE: Not to call anyone out, but given Carlos' input and the split sample demonstrating the repeatability, which is what this criteria focuses on, it doesn't touch into the validity part, if someone could speak to maybe why they were a low moderate or the low voter on this particular criteria, it would help us understand how to communicate that.

Yes, just so that the data are repeatable, producing the same results in the same population at the same time period.

DR. BLAKE: I mean I'll speak to

that because I brought it up. Is I didn't really understand why they had those three separate populations of different sampling sizes, and each use for a different measurement of types of repeatability or validity.

Because I would think that you would take a portion of the larger population and test it multiple times in order to look for reliability. So that's why I squirted as a moderate.

MS. WILBON: Anyone else who is on that threshold want to add to Kathryn's?

CO-CHAIR ELWARD: I guess I was on the moderate range. It's just that I understand that the data may be there, I would just like to see them. And so I didn't feel comfortable giving a high measure until I actually saw that there was some, they could provide some of those data.

MS. WILBON: So better communication on the approach as well as more

presentation of the data?

CO-CHAIR ELWARD: Yes.

MS. WILBON: Yes.

DR. STANFORD: Yes, I was the low person. And that was why I graded it low, was the issue around just not being able to see the, I mean based on what we had discussions around the reliability, it's a pretty low hanging fruit to be able to get just, you know, repeating it one or two of multiple times.

So it's not something that they shouldn't be able to provide us in terms of detail information. So I'm sure I'll be rating them moderate or high. Just based on what I have in front of me, I can't go above that.

CO-CHAIR MAURER: I think that's going to be a --

DR. MOSENIFAR: I mean all metrics. Sorry. I mean all metrics show that the data is very detailed, but it's just, it's not

there yet in terms of analysis and availability, so that's the point, sorry.

CO-CHAIR MAURER: And I think that's a problem that we're going to see as we go along here, is that we are asked to take a lot on faith. You know, it may've been shown in diabetes, but diabetes isn't asthma, you know.

MS. TURBYVILLE: That's very helpful, thank you.

MS. FANTA: And now we will vote on overall reliability testing which encompasses precise specifications and reliability testing. Okay, so we have eight moderate and one low.

DR. SCHATZ: Well, and again I think by the words here something has to be implemented consistently across organizations. And what are the values of being able to customize, it I believe, fatally interferes with consistently interpreting that across organizations.

MR. AMIN: And I would add, there was a little bit of discussion, I'm not sure that it totally fits in this category but it was a discussion around whether the database was representative of the over 65 age group, was another concern that we'd have.

CO-CHAIR ELWARD: Although I wouldn't expect it to be for asthma. Oh, I'm sorry, COPD.

I would say one of the things that would be encouraging is if Ingenix could provide us with, you know, a plan for how they would handle the balance of, you know, customization versus if this were to be a measure.

I mean people could potentially do anything they wanted with the database on their own, but for cross-plan compares that this is going to be a national measure, could there be some criteria that they would say, this is the way it should be reported if you're going to report it and expect it to be

compared nationally. So it's just a thought.

MS. FANTA: Okay, moving along to 2b1. Are the measure specifications consistent with the focus of measurement and the measure intent? Okay, so we have two high, five moderate, one low and one insufficient.

CO-CHAIR ELWARD: Anyone want to comment on --

DR. BLAKE: I said insufficient based upon what we had talked about when Carlos gave his opinion, and I asked him if this would be insufficient.

DR. SCHATZ: And I continue to be concerned that if asthma costs are the focus, and not being able to be transparent in terms of what's an asthma cost as well as, you know, my belief that anything presented without pharmacy data for asthma is not meaningful.

CO-CHAIR ELWARD: Good point.

MS. FANTA: 2b2. Does the validity testing demonstrate that the measure data

elements are correct and/or the measure score correctly reflects the cost of care or resources provided, adequately distinguishing a high and lower cost or resource use? One high, four moderate, two low, two insufficient.

CO-CHAIR ELWARD: We could ask who the optimist is. But I mean I join the insufficient category just because of the same, you know, that would like to see this, you know, I have some sense that that'll be deliverable, but I share some of Michael's concerns also.

CO-CHAIR MAURER: I also think based on Carlos' comments and others, that we don't have the data in front of us, you know.

MS. FANTA: Moving along to exclusions, 2b3. Are exclusions supported by the clinical evidence or analysis of frequency and distribution? Is information about impact of exclusions for patient preference transparent? One high, seven moderate, one

low.

MS. WILBON: So I can try to recap a little bit of what I have in my notes here. Based on some of the analysis that Carlos did, he mentioned that he would've liked to see a sensitivity analysis to the exclusions. That was missing.

It's not totally transparent how the grouper is identifying outliers in exclusions. I guess, sorry I don't have the rest of that. It was just a brief cutoff sentence there.

That they do use Winsorization for the top two percent, and identification of outliers for incomplete episodes are excluded. That they cap the high costs and so not all the high cost episodes are included.

MR. AMIN: I think that generally, I think the conversation was around who was in that excluded population, and some statistics in some way would've been helpful I think for the TAP in their review, especially in the

high cost population that was Winsorized, who they represent.

MS. FANTA: All right, and then 2b4. For Resource Use Measures, is there an evidence-based risk adjustment strategy or rationale data to support no risk adjustment or stratification? One high, four moderate, two low, two insufficient.

MR. AMIN: Seems like the majority of the conversation here was around the R-squared values providing that for the TAP. Is there anything else in particular?

MS. WILBON: I'll just add there was some discussion around how comorbidities impact the severity ranking, and how the severity ranking, how exacerbations for asthma are, I guess computed in the severity ranking and addressed in the total cost for the episode. Is that right?

DR. BLAKE: I was the high person, and I was thinking I was focusing what on, the last part that you just said, Ashlie, I'd kind

of forgotten about the R-squared stuff.

About when he said that it was based on both exacerbations or like an office visit. So that's why I thought that was appropriate and I rated it high.

MS. FANTA: 2b5. Are performance results reported? Do they identify differences in performance or overall less than optimal performance? Eight moderate, one insufficient.

MR. AMIN: I think the only thing I have here as far as the discussion goes, was the S12 report that Kathryn brought up during the discussion around the confidence interval of those point estimates, what they actually mean, the interpretability of those measures. And yes, just I'll leave it at that.

DR. BLAKE: And then there was the table S9.5 that didn't have anything to do with detail score estimation, so I think they just had the wrong, they were referencing the wrong section. It had to do with

complementary services. It didn't have anything to do with detail score estimation.

MS. FANTA: Okay, and then we can move along to overall validity testing, which encompasses specifications consistent with resource use, cost problem, validity testing, risk adjustment and identification of statistically significant or meaningful differences. Six moderate, one low, two insufficient.

2c. If disparities and care have been identified, do measure specification, scoring, and analysis allow for identification of disparities through stratification or results, or is there a rationale or data that justifies why stratification is not necessary or feasible? Two high, six moderate, one insufficient.

MR. AMIN: Just to summarize some of the concerns in this area, which I believe they're reflected in these scores, were a lot of the points that Michael brought up around

the number of outputs across the various severity levels.

Stratified by race would make interpretability also relatively complex, and the current stratification factors on severity, which actually may be a result of poor care.

MS. WILBON: Did we capture I think everything? At least we'll have the ability to go back and listen, and look at transcripts, but do we feel like we've captured everything? Okay.

DR. SCHATZ: And again, I think this is where we wanted some additional information.

But just one general piece of information, Tom. You do list in the application that a lot of people use these measures but as you had mentioned, not clear how much is asthma alone versus the whole suite. Is there any information about how the asthma measure specifically has been used?

DR. LYNN: There isn't, because we really don't have, if we have customers that are using asthma specifically, I don't know about it.

Almost all of our uses are in the area of measurement, or combining asthma measure with the COPD measure with the pneumonia measure, and potentially even wider than that.

DR. SCHATZ: So again my concern would be, I just, in trying to answer these questions of usability, you know, in theory the concerns that have been raised up to now would put some questions. And then in reality, how it's been used or how useful it is, we don't know. And so I guess those are my concerns.

CO-CHAIR ELWARD: Other thoughts, I mean any other comments on 3a? Okay. Oh, 3b?

DR. SCHATZ: Yes, I was sort of lumping all of the usability to some extent,

which I maybe shouldn't have.

CO-CHAIR ELWARD: It seems like they can be reported. It's evident strictly looking at 3b's criteria, that they need to be, there needs to be a careful explanation. It'd be easy for people to misunderstand what the measures mean in terms of the, it being understandable and useful to the intended audiences.

And not to take away the importance of what ETGs represent because episodes of care are huge, you know, hugely important in terms of thinking about what, you know, what actually happens to a patient with asthma.

And specifically what kind of specific care they get. But it seemed like there were a lot of questions that we had about, when you report these what are they going to mean to people? It's just public consumption.

DR. SCHATZ: Yes, I mean I think 3a definitely could be publicly reported. It's

more 3b, you know, is it meaningful and useful, where I guess I have more questions.

I think 3c, other than the asthma cost which maybe that too could be decomposed, I think that 3c probably does if you look hard enough, allow you to really get into it. But I don't know if that makes it usable.

CO-CHAIR ELWARD: Yes, it seems like we can break it down even more.

DR. SCHATZ: Yes, I mean I guess my biggest concern would be with 3b, and partly that I just don't know.

CO-CHAIR ELWARD: Other comments from others on -- Tom, do you have any responses to Michael's comments about, you know, how to handle the reporting issues, or how your users are addressing the differences in the different reports that can be produced?

DR. LYNN: Well, I think this is a different reporting system that can be produced with the different methods that can be used. You know, we don't have, obviously,

complete control over how our users use the product.

We do have some ideas about how they use it. The question is really around some of the choices that we allow. Like whether they use standard price or whether they use their own price. I think that's one of the big ones.

And I think we probably, that might be close to 50/50. And it depends on whether the focus is on the dollar amount or is the focus on actual utilization. And so I think it's used both ways in that regard. I don't know if that answered the question.

CO-CHAIR ELWARD: Okay. That's helpful. Can you give us some other thoughts about, or information about how when we decompose the results, maybe that's not the, sometimes I'm not sure that's the best term, unless we're talking pathology.

DR. LYNN: Right.

CO-CHAIR ELWARD: But deconstruct

the measures for, or how some of your customers have deconstructed the measures to answer specific questions?

DR. LYNN: You mean how they might have used asthma? Oh, oh, I'm sorry. Well, I guess the question is, you mean how they would've used asthma specifically out of a list of diseases, or how they would use the different metrics, like number of ER visits and in dollars for laboratory for a measurement?

CO-CHAIR ELWARD: I'm thinking about when they had questions about their asthma care, saying how would they break down different aspects.

For example, being able to say okay, for a given number of patients who have asthma, how can we break down what impacts their ER use?

Or can we even get to some of Michael's concerns about, you know, under-prescription of care? Or just how they've

addressed the asthma, understand their asthma measures better.

DR. LYNN: Well, I think, you know, if you look at what some of the metrics were that were included with this, you know, there are ways to sort of drill into different aspects of care to see how they may be driving the score. And, you know, that too, and that's why we included those here. So how we recognize that the measurement is more effective if there is ways to get it where the drivers are.

As far as underutilization, we certainly feel that, you know, these sort of cost measurements should be used alongside some quality measures that may help evaluate that.

CO-CHAIR ELWARD: Yes, I think when I look at the best, some of the best material you presented for that I think is that SA reliability validity testing.

Where you really do have a lot of,

seems like there are a lot of data on how you can break it down by lab encounters, ER visits and patient days, things like that.

DR. LYNN: Right, try to get at drivers, right.

CO-CHAIR ELWARD: Okay. All right. Any other, before we go to any other questions, comments? Okay. Go ahead.

MS. FANTA: 3a. Are the measure performance results reported or suitable to report to the public at large in national or community reporting programs? Is there evidence that the measure performance results are available for public reporting?

Oh, did someone leave the room?

(Off microphone comments)

MS. FANTA: Oh. So at least we're just missing one then. Got it. Okay, so two high, four moderate, two low and one insufficient.

MR. AMIN: I just want to confirm that these scores reflect the major concern

that Michael brought up around how the asthma measure itself would be used as opposed to the suite of measures that would come with the Ingenix product, is that correct, or is there others?

(Off microphone comments)

CO-CHAIR ELWARD: And also, how the public would interpret a given report given the variation that, you know, flexibility that they have.

MR. AMIN: Okay, so that would likely be in 2b. They're in 3b also I would bet.

CO-CHAIR ELWARD: Yes.

MS. FANTA: Okay, 3b. Did submitted information demonstrate that results produced by the measure are meaningful, understandable and useful for information for quality improvement and public reporting, or was a credible rationale presented? Six moderate, two low, one insufficient.

CO-CHAIR ELWARD: And I think the

rationale is, I guess you've mentioned it.

MS. FANTA: 3c. Are the data and result details maintained such that the Resource Use Measure, including the clinical and construction logic for defined unit of measurement, can be decomposed to facilitate transparency and understanding? Three high, five moderate and one low.

CO-CHAIR ELWARD: Thank you very much. Let me go over the measures for feasibility.

You know, in looking at the criteria in terms of, are that the required data elements do seem to be routinely generated and used during care delivery to the extent that they can record lab tests, things like that. I think it seems like the feasibility of getting those reports is very high. The required data elements are certainly available. That's not a problem.

Susceptibility to inaccuracy, errors, or unintended consequences. It seems

like they paid a lot of attention to that.

I'm not sure that there were sufficient evidence, you know, with what we saw today, to answer all those questions, and it would be helpful to get more information from you all, Tom.

And maybe you can answer this before we vote, about how you've dealt with or identified inaccuracies and adjusted for possible errors, as with any measure that we review.

Actually it seems like the data collection measurement strategy has been implemented by a lot of your customers, and has been used fairly extensively.

The major question is whether it's been used in external reporting programs. Can you answer the, probably two issues?

One of them is how you address or identify inaccuracies or errors. And second, how has it been used by your customers in terms of external reporting programs?

DR. LYNN: Right. So the first question is the product has built into it a number of error checks, so we can identify if there is a large number of claims that are outside of the reporting period, or have invalid procedure codes, or those sorts of things.

And we have benchmarks for what we expect those values to be. So our customers know that when they run their data that they have an expected number of say invalid procedure codes that's not reported to be effective. So we do have error checking inside the grouper to do that sort of work.

The other answer to the question is, we do have clients that are publicly reporting data based on, you know, the episode grouper methodology, but they usually do it aggregated across multiple episodes.

But, you know, asthma and pneumonia and COPD, for example, like we've been talking about, I don't know of clients that are

breaking it down to the disease level.

CO-CHAIR ELWARD: Yes. You mean they use the whole package of services. It's not like they lump all the costs of asthma and CHF together. I mean they obviously have different reports that they look at.

DR. LYNN: Well, you can drill down to different reports. Using that O/E ratio you can come up with an aggregated score across the data, and we have clients who do that. It just changes the expected value in the denominator, right.

CO-CHAIR ELWARD: Other items on 4? Questions about that? Okay, go ahead.

MS. FANTA: 4a. Are the required data elements --

MS. WILBON: So quickly I just wanted to, Richard, you had a question earlier that we tabled for feasibility discussion. I just want to make sure we get back to that.

For Tom, there was a question about how it's implemented like at the health plan

or something like, I don't know if you remember, sorry. Hopefully, okay, you could recap that, that would be great.

DR. STANFORD: Yes, evidently Tom has mentioned that the health plans that he has in his examples here all use this outside of Ingenix, is my assumption.

But my overreaching question is around, you know, how much burden is this on a programmer to implement into maybe a smaller health plan or even a larger health plan for that matter, sort of a, and I think Tom mentioned the sort of the less sophisticated customer. Just what is the overall programming burden?

MS ZIELINSKI: I can take that, Tom. Actually, you know, I'm going to give you the dreaded, "it depends" answer.

I mean, you know, if somebody comes and it's a small shop and they're familiar, you know, with ETGs, and understands the product and the processes that are needed to

be required to get the data ready to be grouped, and then to perform the post-grouping steps needed to get from output to report, you know, it's a pretty seamless and straightforward process.

For those that are not as experienced or, you know, just starting from scratch with the fee of the, you know, there's unlimited training involved. Help desk support, those things, you know, that can assist somebody in from a base level knowledge, to being able to process groupers and program the groupers to run correctly.

You know, it takes some input from the analyst and the business part of the group to instruct the programmer how to configure the product. You know, which things they want to turn on and which things they want to turn off in the configuration.

That might be some testing that is involved, you know, looking at outputs if you put something on, and then looking at outputs

if you put something off.

You know, and then the third answer I would provide here is, you know, there's also options that we have here at Ingenix where we take the data, run it for them, or we have a product that's called PCQ Connect that kind of takes their outputs and gets it into report-ready formats for them, so they don't have to do the manual processes after the data has been grouped.

So, you know, there is some, while it is a complicated process, there is a great level of support that's provided for them to be able to be successful in the shortest amount of time.

CO-CHAIR ELWARD: Thank you. Thank you.

MS. FANTA: So we'll go ahead and vote on 4a. Are the required data elements routinely generated and used during care delivery? Just waiting on one vote. So we have seven high, and two moderate.

Moving along to 4b. Are all the required data elements available in electronic health records or other electronic sources? If not, is a credible near term path to electronic collection specified? Seven high, two moderate.

4c, are susceptibilities to inaccuracies, errors, or unintended consequences, and the ability to audit the data items to detect such problems identified? Okay, one high, and eight moderate.

And lastly, 4d. Can the data collection strategy be implemented? Is the measure already in operational use, or did testing demonstrate that it is ready to be put into operational use? Okay, four high, four moderate and one insufficient.

(Off microphone comments)

MS. TURBYVILLE: Just a quick question on a and b which was about the electronic sources.

Is it the moderate votings

primarily about the pharmacy data not always be consistently available? Or is there something that I missed in the conversation that would, not trying to persuade vote changes, just want to make sure we understand the moderate votings on the first two feasibilities, which are about the electronic data being routinely available, and then --

DR. MOSENIFAR: I think they had the data. Actually I voted high on the second one, but they have the data. It's just a question of how to get the data in presentable shape.

MS. TURBYVILLE: That there would be a lot of work for the users to do that to map it to the --

DR. MOSENIFAR: No doubt they have tremendous amount of detailed data which would be applicable to pneumonia as well, but the question of, you know, is a little shop is going to be able to use it or not.

CO-CHAIR ELWARD: I think if they

use Ingenix, you know, as their primary source they will have very little trouble.

Well, thank you very much. I think given the amount of time, you know, we've spent a lot of good time on this and I think it will make the other two go easier.

But let's go ahead and take a ten-minute break and then we will reconvene and start working on the COPD measures.

CO-CHAIR MAURER: Actually we're going to do pneumonia next.

CO-CHAIR ELWARD: I'm sorry, pneumonia, yes. We have --

CO-CHAIR MAURER: Zab has to leave so we're going to run through his stuff as soon as we get back.

CO-CHAIR ELWARD: So pneumonia's next. Thank you for correcting me on that. And Tom, you can put, you know, that on mute.

(Whereupon, the above-entitled matter went off the record at 2:54 p.m. and resumed at 3:05 p.m.)

CO-CHAIR MAURER: I want to get started because we're going to lose Zab here pretty soon.

Okay, so we're going to do the ETG Based Pneumonia, the Ingenix measure. And we'll start off with the importance.

This is a measure that's to evaluate episodes of pneumonia. This is an acute ETG, meaning that it has, I believe it has a defined clean period after it. It isn't classified as a chronic condition, so there will be kind of a period at which it kind of washes out.

The proposers say that this is an important measure because there were like 589,000 hospital discharges in males, and 643,000 hospital discharges in females per year. And there are about 175,000 of these are pneumococcal pneumonia.

They've looked at their own database and found out that 0.4 percent of their population gets pneumonia.

The U.S. records about 55,000-plus deaths due to pneumonia or influenza every year, most of which are from pneumonia. And it is, the combination of pneumonia and influenza are the eighth leading cause of death in the country. This is a cost to the health care system of $34.2 billion direct, and $6 billion indirect.

And the proposers suggest that there is an ability to achieve improvement through reductions and variation in care based on their assessment that using one as a sort of an average care that the variation ranges from 0.48 to 1.32, and even higher variation among specialty groups of scores of 0.25 to 1.21. And pharmacy varies from 0.8 to 1.15, so they say that there is a significant variation in care.

And also preventive approaches such as pneumococcal vaccine are not used as much as they could be to help reduce the rate of that type of pneumonia.

So I think it is a important measure. It has significant resource use. There are ways in which improvement could be addressed.

And in setting up their episode-based groupings, they have described in some detail the service categories that are going to use, and it maps to providers and practices, and in an aggregate level to the health plans if you want it to.

I think that's about it for importance. Okay. Now we vote, unless somebody would like to --

MS. FANTA: Oh, perfect, okay. So 1a, high impacts as the measure focus addresses specific national health goal priority or was the data submitted that demonstrated high impact aspect of health care? So we have eight high.

And moving along to 1b. Was data submitted that demonstrated resource use or cost problems for improvement that is

variation in the delivery of care across providers, and/or population groups? So again, we have eight high.

1c, is the purpose, objective of the Resource Use Measure including its components and the construct for resource use or cost clearly described? Okay, and eight high.

And lastly, 1d. Are the resource use service categories that are included in the Resource Use Measure, consistent with and representative of the measure concept? Seven high, one moderate.

CO-CHAIR MAURER: Okay, I'm going to, Dale, you're the first one with 2a1, and then we're going to go to Zab and have him do several of his so he can get through them, okay.

MR. BRATZLER: So much the same issues that we just discussed with asthma. So essentially the methodology is the same. You identify an anchor record.

You first start with all your claims, determine what category they fit in, and then identify an anchor record. And then tie all the additional records subsequently to that anchor record that represent, or could represent pneumonia care, including pharmacy, clinical services, inpatient, other types of measures.

So basically the construct is the same. I think it's important to note that this is pneumonia. It's not community-acquired pneumonia. It's not health care. It includes all forms of pneumonia.

Patients with HIV, CAP, health care associated, immunocompromised, the whole spectrum of pneumonia care.

The risk adjustment works essentially the same way. They've developed, identified comorbid conditions that are each assigned severity weights.

And I think my biggest concern was how those severity weights were actually

determined, because I've spent a lot of time doing severity on pneumonia. I just don't know what the methodology was for assigning different weights.

And they have different weights for elderly. And again, since we don't have much information about the measure population that they actually used to develop the measure, and what proportion are actually in that elderly age group, I just don't know how they came up with different weights for elderly versus the younger population.

I'd raised several issues before during our discussion with asthma, the costing method, the actual payments versus standard payments, standard adjusted payments.

So the clean period is 180 days. So the patient basically, once they have that anchor record diagnosed with pneumonia, all records that are tied to pneumonia care until they have no episodes of care tied to pneumonia for 180 days, then that patient is

then out of that.

Just as we saw with the other measures, a patient could be potentially in several anchored records. So they could be in the pneumonia group, and they could be in a diabetes group if they have an anchor record there also, it works the same way.

When I did the initial evaluation I rated this high because very clearly, highly detailed description of how to go through and develop the measure, very well spelled out.

My biggest concerns at this point are around, how you do the risk adjustment?

And the other thing is I've spent 12 years arguing with clinicians, particularly the pulmonologists in the room, that we need to make sure that we've separated community-acquired patients from health care associated patients, patients with immunocompromised conditions, and they're all rolled up into one.

And so I think that makes it

tougher, from the face validity for the clinician at the bedside, to figure out who you're actually addressing with this measure, which is just generically pneumonia which includes a really a broad array of patients here over time.

CO-CHAIR MAURER: Other comments? Can you tell me where you found the clean period, because I looked for it and couldn't find it.

MR. BRATZLER: Well, I think it's, yes, I actually just went through and did a search.

In the example they talk about viral pneumonia, which I think is one of their examples that falls into this category, because viral pneumonia does fit in this pneumonia measure, and that example used 180 days, so I made that assumption.

CO-CHAIR MAURER: Oh, okay. Because I looked for that.

MR. BRATZLER: Yes, it's in those

last documents.

CO-CHAIR ELWARD: I think it's in the logic model.

MR. BRATZLER: Maybe.

CO-CHAIR MAURER: Maybe, you know, they give other clean periods, but I didn't see the pneumonia one.

CO-CHAIR ELWARD: Tom, are you there?

DR. LYNN: Yes, I am.

CO-CHAIR ELWARD: Yes, were we correct in saying that the clean period for pneumonia is 180 days?

DR. LYNN: Yes, that's correct.

CO-CHAIR MAURER: Okay. Any other comments? No. Okay, so Zab, we're going to move onto 2a2.

DR. MOSENIFAR: Sure. Can you move the 2a2 up there? Anyways, I echo Dale's comments. This is a tremendous, vast amount of data of pneumonia.

Every pneumonia of the possible

causes that you can think of, it's in this data pool. And it's going to require some sophisticated user to have access.

And although that same, the comments about statistics apply to this as well so my comments and vote, actually it's going to be across the board at the moderate level. Because the data is there and the question is that you have to be very sophisticated to get to this data.

I don't think an average user without having a robust statistician will be able to, either they should help or someone else, they should have someone like Carlos next to them, and I see Carlos is missing here so I'm going to be handicapped.

But I think using his comments from the prior discussions will be the same applicable. The data is there. The statistics could've been there. It's not quite there, the details.

CO-CHAIR MAURER: So we need more

information, more detail from Ingenix?

DR. MOSENIFAR: I think the information is there, it's just the way it's going to be user friendly.

CO-CHAIR MAURER: Oh, okay.

DR. MOSENIFAR: That's really the question.

MR. BRATZLER: Well, I mean, and maybe Tom can help answer this, but my issue is, is specific weights are assigned to many comorbidities and they may be appropriate, I just don't know how they came up with the weights.

CO-CHAIR MAURER: How they got that.

MR. BRATZLER: For all these various comorbidities and they have a separate set for the elderly population, how that was generated. You know, what proportion of the population that they've tested these measures on fit the definition of elderly? That's also not clear.

DR. LYNN: I can answer that question, particularly about that one. We had, you know -- this is commercial data, right. So we have over 65 folks in our commercial data, either because the primary insurance is, you know, the insurance health plan acts, or because we have some data where health plan acts, it has some sort of, you know, has taken on Medicare responsibility for a member.

And that's why we have some members over 65. It's not a high number. I think it's probably like eight percent of the population that we used to create focus.

MR. BRATZLER: And so that's important because at least, and my knowledge is much greater on inpatient pneumonia where two-thirds of the patients are probably of Medicare age group versus the younger age group.

My other question is, when these risk adjustment methodologies were developed,

were they ever tested for different populations? So for instance, well, if you don't have many Medicare patients you can't do it.

But one example would be the patient that comes in from an ambulatory setting versus the patient that comes in from the nursing home setting.

And without a lot of Medicare data you probably can't identify, but kind of lumping all of the community-acquired patients with the health care associated patients.

It may be appropriate and maybe the risk adjustment addresses it, I don't know. But I just wonder whether it's been tested, kind of, for that validity at the clinician level, about whether these four strata of severity levels adequately really divide up the patients into severity.

I mean what can you actually do with data at the end of the day? Can you change quality, other than just looking at

resource use?

DR. LYNN: Well this is a Resource Use Measure. But to answer that question, we have not looked at it like the Medicare population or Medicaid population.

Matter of fact, you know, if we were going to use it in a purely Medicare population then, and as a matter of fact we're working on that right now, we would create new weight.

If we used it for Medicaid we'd create new weight. So we don't really, you know, this is meant to be used in a commercial population.

CO-CHAIR MAURER: Any other comments?

CO-CHAIR ELWARD: In answer, Tom, maybe you can follow along with what Dale was mentioning. Can you break down the measure by different types of pneumonia?

For example, if you wanted to say, how are we doing with, you know, pneumococcal

network, bacterial pneumonia versus viral pneumonia. Can you create separate reports on that?

DR. LYNN: You can create separate reports on viral and bacterial pneumonia. I don't know. I can't remember whether we have a pneumococcal condition status factor, off the top of my head.

CO-CHAIR MAURER: Well, your condition status actually lists a whole set of bacteria separate condition statuses that you add.

DR. SANTO TOMAS: There's no ICD-9 code for community-acquired pneumonia, but there's an ICD-9 code, for example, I mean, you know, there's no health care associated pneumonia, but there's a ventilator associated pneumonia, I think it's a modifier as a complication to, you know, there's a pneumonia and it's a modifier.

So I think at least some of those you could maybe tease out. I'm not sure if

that's included in the, I tried to search the methodology. I don't think I saw that in particular.

But those are things that I think, you know, clinically that is important. I mean those are usually of course, high resource utilizers, those kind of complications.

CO-CHAIR ELWARD: This is Kurt again. I'm just also wondering whether, you know, in general when NQF is looking at measures, I think you tend to be more global in your measures, is that correct?

I mean it's not like you look at diastolic versus systolic CHF in your congestive heart failure measures. Or, how do you separate diabetes? For example, do you go into gestational diabetes, Type I, Type II, you don't do that?

MS. WILBON: Yes, so I'm going to take a stab at this and I'm going to open it up to Heidi and Sally to piggyback on.

We do encourage more global measures, but I think where there are clinical differences in populations, we would probably encourage like stratification of those different types, of the differences between those clinical groups can be identified.

I think that would probably be our recommendation, how we've done in other measures, where they would stratify for like Type I and Type II diabetes or something like that. And Heidi or someone can piggyback or --

MR. BRATZLER: So the clinical quality measures for pneumonia, we have certain ones that are kind of the broad population of hospital-based pneumonia patients.

And then we have ones that are specifically community-acquired based on the fact that there are only guidelines for that population for certain things.

MS. TURBYVILLE: Just a little, in

addition to other measurement at first. The steering committee did make a statement and a guiding principle at least for this first effort that they were interested in seeing broad resource use measures.

But again Ashlie's absolutely 100 percent correct. That shouldn't be to the detriment of whether or not clinically it makes sense to combine them. But they were looking for across settings, across services, and so whether this works the way it is, is clinically is really what we're hoping to hear from you all.

DR. MOSENIFAR: Nosocomial pneumonia is really a different kind of a pneumonia. Ventilator associated pneumonia is a different kettle of fish, so it will be disservice actually to combine CAP with HAP.

MS. TURBYVILLE: And those can be differentiated through ICD-9 codes or otherwise?

DR. SANTO TOMAS: Yes, when I tried

to look, the only one that had, again was ventilator, in fact, not hospital-acquired but just ventilator associated pneumonia does have a --

CO-CHAIR ELWARD: VAP has a separate.

DR. SANTO TOMAS: Yes, VAP has a separate, V-A-P.

MR. BRATZLER: Yes, so now I would say that you can't, just from ICD-9 codes alone, you can't differentiate health care associated from, you can VAP, but not health care associated pneumonia, from community-acquired pneumonia, you can't do it.

CO-CHAIR MAURER: Well, health care associated is really a new designation in the last five years, right.

MR. BRATZLER: Right, and there are ways to do it. I mean if you looked at the ABMS, they actually tried really hard to separate the two groups. And I'm not saying that that's a flaw.

A broader measure may be okay. What I can't tell though is if you happen to be a clinician that had a relatively large HIV population, does the risk adjustment adequately address that immunocompromised population of patients who may have much tougher pneumonias to treat?

Or you have a much sicker population that gets gram-negative pneumonias because they're in nursing homes or something. So that's what I can't tell, is how those weights for the risk adjustment were developed.

DR. LYNN: I just wanted to, yes, point out and you said it, and I appreciate it. But the markers are there for risk adjustment, not for hospital-acquired or community-acquired, but for, you know, the bug identified.

CO-CHAIR MAURER: I was just going to say before Zab walks out the door, you have validity and outcome measures. Do you have

any specific comments you want to give us?

DR. MOSENIFAR: Specifically again, the data is there. All the data is there, it's just the same comments that in terms of accessing the data is the difficulty. A good statistician will have no difficulty, but I had difficulty really getting into information that I wanted to know.

So I have no doubt that the data is there because the number of patients that they have really studied is just vast. I'm sure if it's analyzed carefully and with a robust statistician, I have no doubt that every bit is there, but it was not available to me.

(Off microphone comments)

CO-CHAIR MAURER: He said he would vote moderate across the board I think, don't you?

Okay. All right, so 2b1, Dale?

MR. BRATZLER: So 2b1, I also rated moderate. And here again, I think this largely got to, so here part of it is the

inclusive target population, and again my biggest concern was kind of this broad category of pneumonia across the board.

And secondly, the exclusions. The only exclusions again are the low outliers. The high outliers are Winsorized, and there really are no other exclusions as we've seen with some other measures, so that's around high cost outlier patients who are anything else.

So this is broad, broad measure looking at pneumonia with any potential comorbidities or other things hopefully adjusted for in the risk adjustment.

But I think my other concern was partly here around the validity of whether this measure would reflect appropriately other populations such as the elderly, like Medicare because of the age limitations of the derivation cohort, I'll call it.

CO-CHAIR MAURER: Well, so can you use it at all do you think in the elderly

population? I mean if there's a really small number of people that it's been --

MR. BRATZLER: Well, I think we don't have enough information to know that to be honest. I can't tell, you know, how they built that risk adjustment, and they have different scores, different weighted scores for elderly. I just don't know how they came up with them, if they have a small population that they were able to test.

DR. LYNN: Yes, what we did is, we only used the different one, if it was statistically significant to be different. So if we couldn't tell if it was the smaller sample size, then we used the overall one. And it came of course, we submitted this to be used in a commercial population.

MR. BRATZLER: Yes.

DR. LYNN: Not necessarily to be used for Medicare.

CO-CHAIR MAURER: So I think what I'm hearing from you, Dale, though is that we

really need to see how this risk adjustment was done more specifically.

MR. BRATZLER: So I think that's pretty much the same discussion we had for the asthma measure.

CO-CHAIR MAURER: This one, yes, yes.

MR. BRATZLER: That part isn't very transparent.

CO-CHAIR MAURER: Yes. Okay 2b2, this was Zab's, but do you have any comments about the same comments you just made, you mentioned.

DR. BRATZLER: I've already covered it.

CO-CHAIR MAURER: Okay, exclusions.

MR. BRATZLER: Well, again there are none. I mean except for the low outliers.

CO-CHAIR MAURER: Well, there's the upper end.

MR. BRATZLER: Well, they're Winsorized, and the lower outliers are

excluded. But those are the only exclusions, so there are no other exclusions on the measure.

CO-CHAIR MAURER: Any comments about that? Any other comments? Okay, Zab again had 2b4, but do you have any comments? We talked about risk adjustment already, you did. So nothing more on that?

MR. BRATZLER: No.

CO-CHAIR MAURER: Okay. Linus, you read this one too? No?

DR. SANTO TOMAS: Yes.

CO-CHAIR MAURER: Yes, okay. Okay, 2b5, this one was Zab's. Did you have any comments, Dale or Linus? This is, data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and meaningful data.

MR. BRATZLER: Yes, so I was impressed with the data that they have, but there are substantial variations in resource

use between different providers within specialties and others.

So I have no doubt that they identified differences in resource use. How that relates to quality or outcomes or anything else, I don't know. But I think their analytic methods are detailed and they demonstrated variation which doesn't surprise me.

CO-CHAIR MAURER: But do we have the populations appropriately separated or risk adjusted to really assess meaningful?

DR. SANTO TOMAS: Yes, no, I think it goes back to, as far as being meaningful, not knowing exactly what those pneumonias are, you know, how much of that variation is because of what the nature of that pneumonia is, whether it's community-acquired, health care associated, or ventilator associated pneumonia.

CO-CHAIR MAURER: Any other comments about that? 2b6, we're not doing.

And that leaves us --

MR. BRATZLER: Then for 2c.

CO-CHAIR MAURER: 2c, that's --

MR. BRATZLER: 2c, I would've said if the data were available it would probably be feasible, just likely not available.

CO-CHAIR MAURER: Okay. All right. Now we need to vote. Well, this is quick.

MS. FANTA: Okay, 2a1. Is measure precisely specified so that it can be implemented consistently? Three high and four moderate.

MR. AMIN: Okay, I'm going to just try to do a little bit of recapping because we likely will have the time, but I just want to make sure that we're capturing it even though a lot of the issues will be similar.

Here there was a discussion around two major issues which will likely play out across the different criteria, but the separation between community-acquired and health care-acquired, or ventilator-acquired

pneumonia, it really is a grouping of all three of them.

And concern about the severity level across the strata and whether that is appropriate, a lot of information wasn't, more information would've been provided there it would be helpful. And also the severity weights for the risk adjustment model.

MR. BRATZLER: The only other thing I would add is, again we discussed with the asthma measure that if you're using the measure to compare across plans then you need to have consistent cost data across plans.

MR. AMIN: Right, correct. Thank you for that.

MS. FANTA: Okay, 2a2. Does the reliability testing demonstrate that the results are repeatable, producing the same results a high proportion of time when assessed in the same population in the same time period, and/or that the measure score is precise?

MR. AMIN: I would, before we actually start voting on this one, real quick. Sorry, Sarah.

MS. FANTA: Yes.

MR. AMIN: Even though we got that one high vote, this is, just want to remind what this criteria is really aiming at, which is whether the data is reproducible in the same population.

So although there is some concern that the TAP noted around the specificity of the measure score, let's try to make sure it's specific to this criteria. Thank you.

MS. FANTA: Six high and one moderate. Okay, now we're going to vote on overall reliability testing including precise specifications and the actual reliability testing. Three high, three moderate and one insufficient.

MR. AMIN: Not to put anybody on the spot, but maybe we can have a little bit of discussion around with the moderates and

insufficients are here specifically.

CO-CHAIR MAURER: Does anyone want to volunteer? I'm not getting --

MR. BRATZLER: 2b3. 28, I'm sorry, I'm on the wrong page, I'm sorry.

(Off microphone comments)

MR. BRATZLER: Oh, it's overall, yes.

CO-CHAIR MAURER: Well, it's a combination of 1 and 2, I mean basically, and so you had some moderates on one of those.

MR. AMIN: So it was the two basic issues that we had discussed prior? I just want to make sure there's nothing else. If there is nothing else, that's okay, I don't want to drag it out.

(Off microphone comments)

MS. FANTA: All right, 2b1. Are the measure specifications consistent with the focus of measurement and measure intent? Four high and three moderate.

2b2, does the validity testing

demonstrate that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided, adequately distinguishing high and lower costs for resource use? Five moderate and two low.

CO-CHAIR MAURER: Okay, and I think this reflects a concern about whether this is one population, you know, or whether it's multiple populations. And we don't know that based on the data we've been given. That would be my interpretation. Others?

CO-CHAIR ELWARD: I think the only thing I just want to make clear with Tom is that given this measure, you can separate out just the types of bacterial pneumonia or can you separate out, I mean for me one of the issues is, can you separate out the various types of pneumonia either by ICD-9 code or by site of admission?

DR. LYNN: Yes, we're separating them out by the type of pneumonia, the

bacteria, in the case of bacterial types. And we're not stratifying them of course, we're just using them as markers for severity adjustment, risk adjustment.

And we do not use, where they, you know, whether this was diagnosed in a hospital or diagnosed in the office, and part of that is related to not wanting utilization to drive risk or severity.

Maybe we need to do that in this particular case or, but that's why we did it, because we're sensitive to not having markers of utilization drive costs, drive our severity scores for cost.

DR. SANTO TOMAS: Maybe just one thing to think about is as you try to sharpen this particular measure further is, you know, one, most people who are diagnosed as an outpatient community, we've never really checked for what bacteria they have, we just treat them.

Second, even for those who are

admitted to the hospital, in the best perspective and in fact, you know, kind of design study, so this is not even how practice is done so you would think they would have a better yield.

But even in those you only get like 50 percent is the best reports of how much we are able to identify an organism.

So that means a lot of this pneumonias that you would see coded as pneumonia would really have no organism specified. Although, I mean you could list all the possible organisms, but most of them would actually have no organism specified at discharge.

MS. TURBYVILLE: Does that mean you don't think it's possible?

CO-CHAIR ELWARD: No, just practically speaking either the cultures don't show up or they get treated quickly enough that, and people forget their cultures or they just or, you know, I haven't see the last time

somebody, you know, even one of my pulmonology colleagues actually ask for a sputum culture. I just can't remember when that happened last.

CO-CHAIR MAURER: So they have condition status they assign to, I mean they assign different, as a condition status they assign different organisms, bacterial organisms in the methodology they gave us. So they would only be able to do that less than half the time, probably.

MR. BRATZLER: That's true, but that's the reality of pneumonia. Most of the time you treat empirically, most of the time. So that bothers me less than all the other comorbid conditions, or underlying conditions that might effect their severity of illness and their resource use.

CO-CHAIR ELWARD: Maybe differentiating between the different types of pneumonia doesn't make that much difference.

CO-CHAIR MAURER: Well, but that

and, in fact, that's maybe the case that you shouldn't use these bacteria's condition status or something like that, you know.

MR. BRATZLER: I guess if you've got the information, it's useful. So if you know if somebody's got gram-negative pneumonia you can probably predict higher resource use.

So I think if you've got it, it's just we just, and I think they recognize that most of the time you don't have that information.

So again, I think a lot of this could be potentially cleared up if we had just a better understanding of the risk adjustment methodology, kind of the face validity of these four strata.

They divide them into four categories of risk based on this weighted score, and that's what I just don't quite understand.

I've used a lot of different risk stratification protocols for pneumonia

patients, but I just don't have enough information in the materials to know how they assigned risk scores and whether those four categories, you know, to a clinician really make sense. And I'm sure they've had somebody look at it. I just don't know what it is.

DR. STANFORD: Tom, you mentioned, you piqued my interest around this issue around not using resource utilization in your, can you go over that again and explain your rationale for not using resource use as part of your risk adjustment?

CO-CHAIR MAURER: You mean, utilization?

DR. LYNN: Yes, I mean if you're trying to measure whether someone's, you know, managed resources well, which is what we're looking at here, finding that.

Then even though it would improve your R score and you can improve your predictability, you can't really, it's circular to say, well, you know, you waited in

the hospital so you spent more money.

So that's one of the reasons that we try to limit the impact utilization on the severity score. What we're trying to do is look at, what's the diagnostic description and how does that lead to increased utilization?

Because if you say or you do something that is parallel to, if you're in the hospital then you spend more money, then are you really looking at how well they managed the case?

Now, you know, again, that's why we did what we did. We think it's the right thing to do but, you know, I understand there would be some, you know, discussion about that.

DR. STANFORD: Yes, and I understand that. I think that's why you probably need to stratify by that resource use. I understand what you mean. I was a little bit confused around what the identification time period was, what the

measurement time period was.

So in that particular case, you'd want to stratify by that inpatient use or outpatient use, whatever that may be.

DR. LYNN: Right. And we, you know, we get the opposite message sometimes, right, that we should because, you know, we talk about unintended consequences and if, would that clause be open to put their patients in the hospital. You know, then their expected cost for that episode would go up.

CO-CHAIR MAURER: Well, and then by definition --

DR. LYNN: That's another argument. I'm not necessarily, I'm just representing that argument.

CO-CHAIR MAURER: Yes, and by definition ventilator associated pneumonia starts out in the hospital, you know.

DR. LYNN: Right. And that's a good point.

CO-CHAIR MAURER: Other comments about --

MR. BRATZLER: Well, like I said, I would just say again I actually don't have big arguments with the methodology, and I think your assessment of not using resources in the risk adjustment is appropriate.

And I just, you know, what I think about is, if I had a patient that was, if my anchor record was a patient that came into the hospital from a nursing home, and versus the patient who is a walking, you know, 30 year old that had mycoplasma pneumonia where I think the resources might be very different in terms of use, does the risk adjustment adequately differentiate those two patients? And that's what I couldn't tell from the information.

For me the big deal is, does the risk adjustment methodology adequately separate patients that should be separated well?

DR. LYNN: And it's a tough answer. It's a tough question to answer, and not just in providing information, which maybe if we give the, you know, the R-squared information about pneumonia may help.

But, you know, one of the challenges we all have with this is that there really isn't a gold standard. I mean how much are you supposed to spend on someone who, you know, has a staphylococcal pneumonia versus a, you know, gram-negative pneumonia?

You know, part of the challenge is no one knows what that is and so you can't really compare to some sort of gold standard. That's one of the challenges. It's a challenge in evaluating as well.

MS. TURBYVILLE: And one thing to add, Dale, you had also mentioned about the transfer from the nursing home into a hospital, with pneumonia, and had also hinted that it was clear that the data are not also available.

And just as a reminder, this measure like some others, because it was tested on the commercial populations, means it would be endorsed only for use in commercial populations.

So hopefully that at least addresses some, though clearly not all yet. Hopefully we'll see future measures that can capture that information.

MR. BRATZLER: Right. I mean when I read any of these measures, it's all about, is it useful for public accountability across different populations, in different, you know, and so that's where I struggle a bit. But I do recognize this was tested with the commercial population. I do understand that.

MS. TURBYVILLE: Hopefully one day we'll have those data soft forms, right, you're absolutely right.

MS. FANTA: Moving along to 2b3. Are exclusions supported by the clinical evidence or analysis of frequency in

distribution? Is information about the impact of exclusions for patient preference transparent? Missing two votes.

Just missing a, maybe these are giving out.

CO-CHAIR MAURER: Where is it, I wonder if somebody's battery's dead.

Yes, you got them.

DR. LYNN: You mean that literally or figuratively?

MS. FANTA: Okay. So we have two high, four moderate and one low.

MR. AMIN: Since the exclusions were really cost exclusions, I think the concerns of the TAP were the effect of those exclusions and what they actually, you know, some more data around what the effect of the exclusions were, would be, you know, would add value.

MS. FANTA: Okay, 2b4. For Resource Use Measures, is there an evidence-based risk adjustment strategy or rationale or

data to support no risk adjustment or stratification? One high, three moderate, two low and one insufficient.

MR. AMIN: Sounds like additional detail on the risk adjustment methodology, and potentially stratification on different types of diseases that's in this cohort would be needed.

CO-CHAIR MAURER: Yes.

MS. FANTA: All right, 2b5. Are performance results reported? Do they identify differences in performance, or overall less than optimal performance? Two high, four moderate and one insufficient.

Okay, now here's the vote on overall validity testing which includes specifications which are consistent with the resource use or cost problem, validity testing, risk adjustment or the identification of meaningful differences.

MR. AMIN: Sarah?

MS. FANTA: What?

MR. AMIN: Quick, I'm just sorry. It's taking me a little bit to process the scores. Can we go back one?

MS. FANTA: Yes, yes.

MR. AMIN: Can we go back one? I'm sorry.

MS. FANTA: That's okay.

MR. AMIN: We did have quite a bit of discussion around the score for the asthma measure and also the pneumonia, about what the point estimates and the interpretability of the point estimate, and the confidence intervals.

But it would also be interesting, I'm interested to know the insufficient that seems to be a little bit stronger than how we were voting before. If there is anything else that I missed on?

DR. STANFORD: Well, I think I missed this part because I was looking at Tiffany and I was looking at my notes þ

MS. FANTA: Use the, your

microphone, sorry.

DR. STANFORD: I was confused about the question because I was trying to think, did we go over this somewhere? Did we have a discussion about the þ

MR. AMIN: Yes, we went, Kathryn went over the report for the asthma measure.

DR. STANFORD: Right, but in terms of this particular one?

MR. BRATZLER: It's the same.

DR. STANFORD: Is it? Okay.

MR. AMIN: I think that's how it was.

DR. STANFORD: And that was my --

MR. AMIN: Okay.

DR. STANFORD: That was, so I would change it to that one.

MR. AMIN: I didn't meant to influence the --

DR. STANFORD: My former reference is basically around this measure, not necessarily across all of them.

MR. AMIN: Oh, okay.

MS. FANTA: We'll revote on that one then. Okay.

MS. WILBON: So just to clarify, this is a revote on whether or not the measure demonstrated differences in performances, statistically meaningful differences in performance. Sorry, you know what I mean.

MS. FANTA: Okay, so for 2b5, seven moderate.

DR. SANTO TOMAS: That's what happens when you take a vote, regression to the mean.

CO-CHAIR MAURER: You're doing a delphi here, I should keep until it goes toward the mean.

MS. FANTA: Okay, now we'll vote on the overall validity testing which I already mentioned before, so unless somebody needs me to repeat it, we're good.

CO-CHAIR MAURER: It is a long day, isn't it, Sarah?

MS. FANTA: And again, seven moderate.

CO-CHAIR MAURER: We've regressed to the mean.

MS. FANTA: All right, for 2c. If disparities in care have been identified, do measure specification scoring analysis allow for identification of disparities through stratification of results, or is there a rationale or data that justifies why stratification is not necessary or feasible? Okay, so we have two high and five moderate.

CO-CHAIR MAURER: Usability?

DR. SANTO TOMAS: Yes.

CO-CHAIR MAURER: And this is Linus?

DR. SANTO TOMAS: Well, obviously there's a lot of examples given that they're saying, you know, I think like nine health care organizations are using it.

My main concern, which has been I think kind of mentioned previously is, if

we're looking for something more global could be seen across the board, if this is going to be something public then in that particular sense, I guess I don't know how usable it would be.

I could see how, even not distinguishing, for example, you know, community-acquired pneumonia versus health care associated pneumonia.

But for a particular organization, and if they're do this in this in a segmental way, you know, just all their primary care physicians, all the encounters in the outpatient, then they could do it in that way I guess.

But if it's something more to be used in a global sense, then I think then it's usability may be less because then I don't know how to interpret, I guess the results that may come out from that.

Again, just the not being able to distinguish what is community, what is in the

hospital.

CO-CHAIR ELWARD: That's a great point, Linus. I'm just wondering, from NQF's standpoint, are you interested in measures that are more helpful across, you know, across plans and across different strata, or can be used by say, large health systems to know where they stand?

I mean I'm sure both are important. But of the, you're looking for the more generalizable measures? I mean public reporting, I would think so, but --

MS. WILBON: Yes, again I'll take a stab at this, and then Heidi and Sally, or Taroon can piggyback.

So in terms of level of analysis, which is generally the area in the submission form where we ask the developer to indicate at what level this measure would be able to apply, like what level of analysis the measure would be able to applied.

They would check different boxes.

The boxes that they check, we expect them to also demonstrate in their testing data, that it is actually like the data that you would get from applying the measure at that level, is actually meaningful and useful.

So it's more of a, I guess a user, I mean a developer defined item in this submission.

However, even though we want broad measures across all different types of level of analysis, we don't want a measure to be submitted at the group level and then using it at the, and then it gets publicly reported at the physician level. That's not the intention.

We want to make sure that how ever it's submitted and the intended use, that it's been tested at that level, that the testing is sufficient and that it's used at that level. So am I making sense or no?

MR. AMIN: And I would just add, the only thing I would add to that, Ashlie, is

based on this discussion it also should be clinically relevant. I mean so that's an interpretation question for the TAP to decide.

So and that would really fit sort of into 3b, where we're talking about whether it's meaningful for it's intended purpose.

The intended purpose being, as Ashlie described, how the measure is specified and then making sure that it's clinically meaningful. But that would be up to you guys as experts to make that decision.

DR. SANTO TOMAS: So I think I may have made this a medium when I first looked at it, but the more I look at it now I'm actually veering more towards really the low as far as usability.

If, just looking, or taking into account what you've just said as far as, how it's going to be interpreted and used in a more general term.

CO-CHAIR MAURER: Without knowing more about how the risk adjustment affects the

--

DR. SANTO TOMAS: Risk adjustment and specification of, you know, classifying a particular kind of pneumonia.

MR. AMIN: Could I, I just don't, look, could we just get a little bit more clarity for me now, because this question of generalizability, I don't know that we're necessarily giving guidance on the question of generalizability.

The measure has it's intended purpose which has been specified at which level of analysis, and it needs to be clinically relevant for that level of analysis.

But the question of generalizability I don't think that we're necessarily giving any guidance on. So I wouldn't want anybody to make a decision based on NQF guidance on generalizability, so just that being clarified.

DR. SANTO TOMAS: Yes, about

pneumonia, right, but without subdividing that further how useful is this going to be?

MR. AMIN: Right. And that's, so I would classify that as sort of the clinical logic behind it. If it doesn't clinically make sense to you as a clinical expert that the measure of the population, how it's defined isn't clinically coherent, that's totally understandable.

So that, but I also would reference that a lot of that is, it does come in 3b in that it's meaningful. But it's also, a lot of it was covered in scientific acceptability. So that's just how þ

CO-CHAIR ELWARD: If we're thinking, again compartment of quality, but mostly just, how much does it cost to take care of your pneumonia patients?

It seems like the measure would have pretty good reliability on how much it's going to cost us.

What the health plan would need to

do or would then say okay, it costs X amount of dollars. We need to parse it out by the type of pneumonia we had. If we're doing really well, we don't want to, you might not want to bother with it anyway.

But if we want to rank ourselves, we would have the, and maybe Tom, you could answer this for me. Would we then be able to break it down by, I may keep asking the same question a different way, but we would only be able to break it down by bacterial versus non-bacterial, right?

DR. LYNN: Yes, the rule is stratified so that viral is looked at and bacterial is looked at in different episodes. But then the severity adjustment is based on more detail. But in most cases it's just whether or not, or what the organism is like.

CO-CHAIR ELWARD: Okay.

DR. LYNN: And I think there was two questions embedded in there, but I answered at least one of them I think.

CO-CHAIR ELWARD: Okay. No, that's okay.

CO-CHAIR MAURER: So I'm not really sure that will tell you what kind of pneumonia you, I don't think it's generalizable. I mean I don't hear that it's generalizable, necessarily from what we know, do you?

CO-CHAIR ELWARD: Well, if we're looking at just how much does it cost to take care of pneumonia. You know, all it covers, at least it doesn't, seems like it allows, if they're all measured the same way.

I mean again going back to the issue of, whether you do different types of charges, but in terms of population it seems like it's fair across the board. I'm not sure, does that make sense?

I mean everybody will have to split it out, but it does give you an idea of, at least it's measured the same way across populations in terms of what type of pneumonia you're looking at, essentially just lumps it

all together.

CO-CHAIR MAURER: Exactly.

MS. TURBYVILLE: So maybe one way to think about this is to, you know, different criteria and usability get to different issues.

So like 3a talks about, is it being used right now? 3b is, would it as it's being used and as it's specified, would it be actionable and meaningful for examining resource use in pneumonia?

And then 3c has another aspect. So, you know, is the score interpretable, so does it tell you if it's high or low resource use for the population?

CO-CHAIR MAURER: Yes.

DR. SANTO TOMAS: I guess just some of the, going back to, once you have that particular measure out there and you could have unintended consequences, right, as far as how it's used.

So that's I guess my main concern.

Without being able to divide into segments that particular, this broad pneumonia, that even if you tried to say oh, we're just looking at whether it's bacterial or viral.

Because I'm guessing that if you actually looked at all the submitted claims, they're probably going to come out as the pneumonia non-specific as the majority.

I don't know, that would be my guess. But meaning the minority would actually go into bacterial or viral.

MS. TURBYVILLE: So question to help us, it sounds like it's a real issue with the administrative claims data and it's ability to support what would be useful differentiation among the different types of pneumonia, for actionability.

And even if, it would even hinder a health plan, for example's, ability to drill down into their results and really see.

Am I understanding correctly?

DR. SANTO TOMAS: Well, you know,

again you can tell me whether it's possible or not, but if the administrative data claim, I think I heard a while ago that it can be done.

But can it really be done, for example, to look at whether that patient has been in the hospital or some long-term acute facility in the past, you know, 60 days, those kind of stuff.

MR. BRATZLER: So I think we're getting bogged down here.

CO-CHAIR MAURER: Yes, I do too.

MR. BRATZLER: And so I think we all agree that we need more information about the risk adjustment methodology to see whether or not it adequately helps us feel comfortable that you can broadly define these.

I can tell you, NQF has already endorsed, it's already publicly reported the hospital related pneumonia mortality, which does not separate health care associated from community-acquired pneumonia. It's pneumonia mortality.

It's got a detailed risk adjustment methodology, but it is a pneumonia mortality measure. Not a community-acquired pneumonia, a pneumonia mortality.

So there are examples out there already that NQF has endorsed, and they're already publicly reported measures that rolled up pneumonia as a category.

So I'm not going to argue against doing that. I just want to know more about whether it clinically, you know, whether a clinician can do something with the results.

If you're in a health plan, this is a health plan measure, a commercial plan, whether you can do something.

And whether, you know, the doc that takes care of mostly really sick patients, immuno-compromised, others that get pneumonia, whether their data is adequately risk adjusted to separate them from Kurt's primary care office, that he is primarily treating, you know, walking patients with pneumonia.

MR. AMIN: I would just add, point of information just to the end of that, was that there was a question while we were giving guidance around the level of analysis.

I just want to make sure that the way this measure is specified on page 38, the level of analysis is at the clinician individual and at the clinician team level, among others which are at a higher level.

So just keep that in mind as part of this discussion that Dale just pointed out.

MR. BRATZLER: And that's important if you're profiling at the physician level, unless you're comparing them across groups.

MS. FANTA: Yes, let's vote. So 3a, usability. Are the measure performance results reported or suitable to report to the public at large in national or community reporting programs? Is there evidence that the measure performance results are available for public reporting? So we have six moderate and one low.

MS. WILBON: So we're whispering because we're trying to again monitor the whole internal consistency thing. And 3a is one of those things that is, one of those subcriteria that's probably the same for all Ingenix measures.

So we just wanted to, actually wanted to remind you of this before the vote, but we didn't get to it in time, but so you kind of hear how you voted on this for the other Ingenix measure.

MS. DORIAN: Which was 1605, and you voted two high, four moderate, two low and one insufficient.

MS. WILBON: So if you're okay with that, that's fine. But I just wanted to kind of call that to everyone's attention, so we're being internally consistent as much as possible.

CO-CHAIR MAURER: The only thing to me that's not really consistent there is the insufficient. You know, I think the moderates

and the lows are close enough. And I don't remember why that was insufficient before.

MS. FANTA: So moving onto 3b. Did submitted information demonstrate that results produced by the measure are meaningful, understandable and useful for information, for quality improvement and public reporting, or was a credible rationale presented? So one high, five moderate and one low.

Moving onto 3c. Are the data and result details maintained such that the Resource Use Measure, including the clinical and construction logic for a defined unit of measurement, can be decomposed to facilitate transparency and understanding? Do you want to share this first?

MS. WILBON: So this again, before you guys vote is another one where we figure would be pretty similar to the other Ingenix measures, just based on kind of common construction logic.

MS. FANTA: And your ratings on the

previous one were three high, five moderate and one low.

One high, five moderate, and one low.

CO-CHAIR MAURER: Ashlie, just to comment on your statement that these are pretty consistent, there is actually quite a bit of difference between pneumonia and asthma.

Asthma and COPD should look pretty much alike, but this is a much more heterogeneous disease, and that really the discussion has been about how heterogeneous --

MS. WILBON: Yes, okay, that's fair.

CO-CHAIR MAURER: -- this hospitalization --

MS. WILBON: That's fair, that's fair. And as long as you feel like that's reflected in the ratings, then we're fine with that.

CO-CHAIR MAURER: Yes.

MS. WILBON: Okay, thank you.

CO-CHAIR ELWARD: And that's just some of my votes were based on what was just said about the fact that NQF already has an overall global pneumonia measure.

Because I thought, well if we're already doing, you know, if that's been good enough then maybe I need to kind of cut them a little more slack. Not taking away anything that Dale's been saying, which is exactly on target.

CO-CHAIR MAURER: To me, underlying a lot of this discussion is, we need the information. We haven't been given the information, you know.

Okay, feasibility. The required elements are routinely generated and used during care delivery? Yes, that's true. They are electronic. That is also true. They use entirely electronic data elements.

Susceptibility to inaccuracy or errors on unintended consequences and

measurement? You know, I guess that this is reflected somewhat in the discussion that we've had. We don't really know the answer to that I would say.

Although, you know, as we heard before this is a huge database. They maintain it very well. It's productized so, you know, they have to maintain it for their clients. So they obviously do a lot of work to keep a high level of data integrity.

So any other comments about feasibility? We just don't have a lot of the information because they haven't provided it to us.

MS. TURBYVILLE: Just to jump in really quickly, because I know the day is getting very long and this could be me misinterpreting as well.

So one of the things to think about in feasibility, the testing does help to demonstrate if it can done, but also if they've adequately demonstrate that it is

being done. Keeping in mind that a lot of the users won't be using their database, they'll be using their own database --

CO-CHAIR MAURER: Their own database, yes.

MS. TURBYVILLE: -- to measure their own physicians.

CO-CHAIR MAURER: Yes.

MS. TURBYVILLE: Or their own compare organization to organization.

CO-CHAIR MAURER: Well, they've definitely demonstrated it can be done and they use, a lot of health plans use it.

MS. FANTA: Okay, it looks like we're ready to vote on 4a. Are the required data elements routinely generated and used during care delivery? Seven high.

4b, are all the required data elements available in electronic health records or other electronic sources? If not, is a credible near term path to electronic collections specified? Again, seven high.

4c, are susceptibilities to inaccuracies, errors or unintended consequences, and the ability to audit the data items to detect such problems identified? One high, five moderate and one insufficient.

MR. AMIN: I mean the question I would have is that from the discussion it seemed like the majority of these data elements would be available.

The way the measure is currently specified, it seems the majority of these data would be available from electronic sources. Is there any other discussion that we want?

MS. TURBYVILLE: This is 4c.

CO-CHAIR MAURER: It's a long day.

MR. AMIN: I will stop.

MR. BRATZLER: So 4c, for 4c, this one I'll just say, that I think one of the issues that Carlos brought up is just the issue about how the data is cleaned, or what do you do with missing data, or how much is there.

I mean, then did make general comments about they encouraged the plans to have, you know, rigorous data, but so just there was a little, in my opinion just a little insufficiency of information about that. And what do you do if the plan doesn't have pharmacy data, and those types of things, or if it's missing.

MS. FANTA: And lastly, 4d. Can the data collection strategy be implemented? Is the measure already in operational use, or did testing demonstrate that it is ready to be put into operational use? Five high and two moderate.

MS. TURBYVILLE: So now it's big decision time.

CO-CHAIR ELWARD: Yes, I think the measure is enough like asthma that we can go through it. What I'd like to do is as we go through each measure, if the issues are similar, you can say so and we'll move on.

If you have noted something that's

specifically different as it pertains to COPD, let's focus on that, okay?

MS. TURBYVILLE: And especially any kind of clinical inadequacies of the construction of this measure.

CO-CHAIR ELWARD: Exactly.

MS. TURBYVILLE: Keeping in mind that you guys really represent a large amount of clinical expertise that will help the Steering Committee. So in particular, anything that's not quite right for the COPD measure would be helpful.

MR. AMIN: One other process suggestion that I might have, and you can disregard as the chairs, but it might be helpful to go through criterias, subcriteria, by subcriteria and just vote as we go.

CO-CHAIR ELWARD: Oh.

MR. AMIN: And that way Sarah can just read it, we have the discussion. It's when you feel like it's sufficient, Sarah can just vote.

CO-CHAIR ELWARD: Yes, I'll keep it short.

MR. AMIN: So it's not, you know --

CO-CHAIR ELWARD: You will be introducing some variation in that process.

MR. AMIN: I recognize that.

CO-CHAIR ELWARD: Inconsistency, it is.

DR. BLAKE: I have a quick question.

CO-CHAIR ELWARD: Yes.

DR. BLAKE: We may have talked about this before, but based on what Mike kept bringing up about the percent of health care costs related to the pharmacy charges being about 50 or more percent, is that similar for COPD?

CO-CHAIR ELWARD: No.

DR. BLAKE: What is it for COPD?

MR. AMIN: No, it's about 20 percent.

DR. BLAKE: Okay. Thank you.

CO-CHAIR ELWARD: Oh good, okay. That's very helpful. Thanks, that's a very good point. Why don't we go with, let's see we'll start of with Linus, 2a1? Oh, I'm sorry, oh 1a, okay. Well, I'll -- thank you. I'm trying to speed up too much.

I think definitely I would say the, it's the same issues as before with COPD for the NCQA measure, and I think they've done a very nice job of just iterating the high impact and that it addresses it appropriately. Okay.

MS. FANTA: Okay, 1a. Does the measure focus, address a specific national health goal, priority, or was data submitted that demonstrated a high impact aspect of health care? We're just waiting on three votes.

MS. FANTA: Seven high. 2b, or 1b. Was data submitted that demonstrated resource use or cost problems for improvement?

CO-CHAIR ELWARD: Yes, and I think

they did a good job of presenting that also.

MS. FANTA: Waiting on one vote. Seven high. 1c, was the purpose and objective of the Resource Use Measure clearly described?

Seven high. And 1d, are the resource use service categories consistent of the measure concept? Six high and one moderate.

CO-CHAIR ELWARD: Okay, thank you.

MS. FANTA: Yes.

CO-CHAIR ELWARD: Thank you. Linus can you start us off with 2?

DR. SANTO TOMAS: Yes, so measure is well defined, oh sorry, so as far as the measure being well defined, within it's own context I think it is. It could be consistently implemented across organization.

Yes, I think that this may straddle some of the other subsequent numbers. My main concern with this really is, just the risk stratification as well as what it calls COPD severity, is not really a COPD severity, but

actually more of comorbidity severity.

DR. STANFORD: I have a question around the, how they determined their ETG. The choice of 180 days, I'm a little bit, I don't quite understand why they chose that. For asthma they used 365, so why would they not use the same time frame for COPD, since they're both chronic diseases?

CO-CHAIR ELWARD: Yes, Tom, it's a good question, do you know?

DR. LYNN: I don't have a good answer for that, why we chose that for COPD. And kind of a decision made a long time ago, but, you know, if these kinds of episodes get chopped up into year-long episodes, it doesn't make that much difference.

CO-CHAIR MAURER: Are you sure it's 180 days though, because I thought that they said that any product þ

DR. SANTO TOMAS: I thought they did say 180 days also.

DR. LYNN: Let me see.

DR. SANTO TOMAS: I didn't notice though that the asthma was, yes, until you pointed it out.

DR. STANFORD: You were talking, on the ETG it says --

Well, it says here --

DR. SANTO TOMAS: Oh, I think this is the one, isn't this the one that you were pointing out a while ago, which is --

DR. STANFORD: It says for chronic bronchitis, the clean period is 180 days consistent with most, with a more chronic illness. While asthma is, asthma basically is 365.

CO-CHAIR MAURER: I'm sure in here some place, they say it's a year too.

DR. STANFORD: I could not find that, and if it is in here I apologize.

MS. TURBYVILLE: It's on Page 26.

CO-CHAIR ELWARD: On Page 26?

MS. TURBYVILLE: It's page 26, S9.3.

DR. LYNN: I'm referring to the original to see what it really is.

MS. TURBYVILLE: So we're going to be looking at S9.3, Page 26, Response by Ingenix.

It says once an episode is triggered, a year-long episode is created. Maybe there's contradiction in this submissions, Richard, so just --

DR. STANFORD: Okay, I'm sorry.

DR. LYNN: Yes, so it's going to be created. It's going to be made into a year-long episode and the clean period has a much more profound effect on the acute diseases. It has either essentially no affect on the comorbidity, I mean on the year-long chronic disease types of episodes.

Because if you started a new episode in the same year it would be combined based on the year, not the fact that there was a period of inactivity.

We still need it for various, you

know, we use this internally. And it should be, it probably should be different, but it actually doesn't have an impact on the grouping.

DR. STANFORD: As a guideline, the episode included resource focus 12, okay, so it is a 12-month period?

DR. LYNN: Yes, it is definitely a 12-month period.

DR. STANFORD: Okay.

MS. TURBYVILLE: So Tom, we'll want to take a look at the entire submission and make there isn't, you know, some contradictions somewhere in there as well. And we'll get back to you on if it needs to be updated.

DR. LYNN: Yes, absolutely. Thank you. And I apologize for any confusion I caused.

CO-CHAIR MAURER: For 2a2. Oh, going to vote.

CO-CHAIR ELWARD: I'm sorry, we

want to go ahead and vote on 2a1, right? Okay.

MS. FANTA: 2a1, is the measure precisely specified? Four high, three moderate.

CO-CHAIR ELWARD: And I think that reflects in part, as far as I'm concerned that it's more consistency than the pharmacy database.

CO-CHAIR MAURER: There we go. For 2a2, actually this is very similar to the asthma ETG, and so therefore they have the same strengths and weaknesses. I really didn't have anything else additional. I don't know if anyone else has anything else additional.

CO-CHAIR ELWARD: Any other questions? Let's go to 2a2.

MS. FANTA: Okay. Does the reliability testing demonstrate that results are repeatable? Five high, two moderate.

And then moving onto overall

reliability. Four high, three moderate.

CO-CHAIR ELWARD: Thank you.

DR. SANTO TOMAS: Yes, so is there evidence presented in the measures specifications allowed to demonstrate variations and resource use across providers and population groups? And then does the measure and risk adjustment methodology address this variability allowing for fair comparisons?

Again just as far as that particular part of the methodology, it's more of the risk adjustment that I have concern with, which has been voiced earlier.

CO-CHAIR ELWARD: Other comments? All right.

MS. DORIAN: Just to confirm, we are on 2b1.

MS. WILBON: Yes, I was --

MS. DORIAN: Is that what you were --

MS. WILBON: I'm trying to make

sure the one you read didn't sound like 2b1.

CO-CHAIR ELWARD: Well, I think you were on 2.

MS. WILBON: Oh, okay, you read that 2b1, the measure specifications are consistent with the evidence presented.

CO-CHAIR MAURER: You sure you want the Carlos' sheets?

MS. WILBON: Oh, okay.

CO-CHAIR MAURER: Carlos is kind of principle to replace it.

MS. WILBON: Okay.

CO-CHAIR ELWARD: Oh. Okay. So the question is, we want to know whether the specifications are consistent with the evidence.

MS. WILBON: Yes, that would be the intent of the measure.

MS. FANTA: Two high, five moderate.

CO-CHAIR MAURER: For 2b2, again it's similar for data elements. It's

absolutely pretty much right on track with the asthma. For a measure score it is absolutely similar. I think that's it for my information so that should wrap it up from 2b2.

CO-CHAIR ELWARD: Go ahead.

MS. DORIAN: For both Ingenix? Just the asthma, okay. Well, then that was one high, four moderate, two low and two insufficient.

MR. AMIN: And I believe the concern there for asthma was, while the face validity was appropriate the testing method with the customization was a question around the, right, and then comparing the standardized prices versus the real prices.

MS. WILBON: And I think Dr. Schatz, or Mike's issue with the pharmacy costs at 50 percent being half of the claims, was an issue for him and he wasn't quite sure that it was actually measuring cost of asthma care being that some of that would be missing.

So I think the question that

Kathryn had before about, if that same issue applied to COPD, may or may not weigh, and your ratings may be different based on how high or how much you weighted that issue before.

CO-CHAIR ELWARD: Yes. Well, the issue around, which charges do you take into account for would still pertain.

MS. FANTA: So we're voting on whether or not the measure score reflects the cost of care resources provided. Seven moderate.

MR. AMIN: So 2b3?

CO-CHAIR ELWARD: Yes, please.

MR. AMIN: So exclusions are supported by, so my main thing with this is aside from the exclusions or beyond administrative reasons for exclusion it actually doesn't specify any clinical reasons for it.

I mean just an example, age, for example should really be something, you know,

if somebody is labeled COPD, younger than 40, then that should raise a flag. Or it doesn't take into account, you know, or rather competing diagnosis, asthma in particular.

Then I guess how you handled that with the asthma previously is again, you count how many episodes are asthma, and somebody diagnosed them with COPD, is that, I guess that's the same way, right.

But then that should be specified as far as an exclusion. If somebody has more, a particular patient having a diagnosis of both, you know, just make that clear then, that those patients who have more asthma episodes should be excluded I guess in this case.

DR. STANFORD: But they're not excluded. I don't think they're excluded.

DR. SANTO TOMAS: I thought a while ago that if somebody was diagnosed with asthma --

DR. STANFORD: Oh, like they're

more asthma.

DR. SANTO TOMAS: -- yes, so, popularity vote.

DR. STANFORD: They're more asthma.

DR. LYNN: Yes, their results is they're excluded. And we can explain, I think what we were doing there, but we didn't put it in the exclusion area. I understand why maybe we should have.

DR. SANTO TOMAS: And some people actually have both, so that's the other thing.

CO-CHAIR ELWARD: Asthmatic bronchitis, that's a nice try. But just to clarify, were there any exclusions like end stage renal failure, things like that? None, okay.

DR. SANTO TOMAS: So really it's just administrative that they mentioned, administrative --

CO-CHAIR MAURER: They may come out at the upper end when they Winsorize it.

CO-CHAIR ELWARD: When they

Winsorize it, that's a way they did that, yes. Okay.

MS. FANTA: So we're voting on whether or not exclusions are supported by evidence. One high, six moderate.

CO-CHAIR ELWARD: 2b4?

CO-CHAIR MAURER: 2b4, again the risk adjustment methodology was similar again, back to the same issues that we had. While they had a nice description of how they actually developed their approach, there wasn't a lot of modeling that was presented, and that there wasn't the detail that we had discussed as a group that we would've liked to see in asthma. The same thing is happening here.

CO-CHAIR ELWARD: And I think our votes were three high, four moderate and one low. What's that, b4?

MS. DORIAN: I actually have one high, four moderate, two low and two insufficient.

MS. FANTA: So we're voting on the risk adjustment strategy. Four moderate, three low.

CO-CHAIR ELWARD: And that's based on the problem with the R-squared is missing. So that could be corrected if they, yes, but that's going to be important. Right, go ahead. 2b5?

DR. BAULDOFF: 2b5 is having to do with risk factors identified. I'm so sorry. Risk factors identified are associated with statistically significant and clinically meaningful differences.

Carlos found that there was nothing. The one thing I did make a note of was, does the practical significance, is it indicated by the relative cost ratio, which was reflected on pages 32 and 33, was the only other thing that I brought out there. Otherwise, it is extremely similar to what we found in asthma.

MR. AMIN: And the issues that were

discussed in asthma, I would just add were the confidence intervals, the point estimate and the report that Kathryn brought up. The S12 report and the S9, detailed list and nation score report.

MS. FANTA: So we're all set to vote on 2b5. If they identify differences in performance or less than optimal performance? Seven moderate.

And now it's overall level of validity testing. Seven moderate.

MR. AMIN: Yes, so the last one, disparities. If they have been identified, I mean really as far as this, it looks like there should be only two that I found that they're looking at as far as demographic features, age and gender, which is in a sense similar to if you just --

CO-CHAIR ELWARD: That seems to be a problem for all the measures throughout.

MS. FANTA: Okay. Go ahead and vote on whether or not disparities are

identified or if there's justification for why they're not. Two high and five moderate.

DR. STANFORD: I think it's the same issues around, I apologize. It's the same issues that we stated around the asthma measure as well.

I like the fact that they have used them in other managerial organizations, but it's difficult to understand exactly across what types and sizes and how these are being used. So I think I rated them as a moderate for usability.

It would be nice, and this is true for even the NCQA stuff, it would be nice to see these testing in much more broader populations, much more broader data sets.

It would've been nice to have everybody use the same database basically. It would've been nice to see that.

MS. FANTA: All right, so 3a is, if the measure results are publicly reported. Oh, sorry.

MS. DORIAN: Shall I, I'll remind everybody quickly of your votes on the previous one which were two high, four moderate, two low and one insufficient.

MS. FANTA: Seven moderate.

DR. STANFORD: And this was the question around, this issue around, is it considered to be an, I was taken around this issue around understandable and useful to intended audiences.

It is a measure that would be useful. I think the issues around, can it be implemented in these health plans in a very user friendly fashion. So if it can be, I think that Ingenix has come out with a ways for them to actually do that, then it would, you know, increase its usability.

But in terms of what we're asking for 3b, I think it does lend some useful information for the health plan in and of itself. Now across health plans, it's difficult to know because of the issues around

standard costing.

MS. FANTA: Six moderate, two low and one insufficient.

MR. AMIN: And I think the low ratings there for asthma reflected a question of the interpretability of the actual point estimate.

MS. FANTA: So you can go ahead and vote on whether or not it's usable for quality improvement and public reporting. Just waiting on one vote. Seven moderate.

DR. STANFORD: Right, so the last, 3c is around, can it be, can you decompose it and look at other measures? According to what they've showed us they are able to just aggregate and look at individual components of the costs.

MS. FANTA: Three high, four moderate.

CO-CHAIR ELWARD: Yes, I'll talk about feasibility. The required elements I think is exactly the same as we had before,

they are routinely generated, used during care delivery so I would, I think that's a high. Let's go ahead and vote.

MS. FANTA: Go ahead and vote on that. All right, we have five high and two moderate.

CO-CHAIR ELWARD: Feasibility in terms of all the data elements are available on electronic health records, or I would say not, but in electronic sources. So I would, I think this is similar to before. So we can go ahead and vote.

MS. DORIAN: Yes, and you previously voted seven high and two moderate.

MS. FANTA: And we have seven high.

CO-CHAIR ELWARD: And I think you can foresee from my recollection of what we did with asthma. It looks like they can identify inaccuracies and errors, and they have a program to address that. I don't think there's, other thoughts about that, some concerns?

MS. DORIAN: You did vote one high and eight moderate last time.

CO-CHAIR ELWARD: Okay. Go ahead and open the vote.

MS. FANTA: Three high and four moderate.

CO-CHAIR ELWARD: Okay, and the last one I think it's clear that it has been implemented in an operational use by certain health plans. So, okay, go ahead and vote. I think we did all seven before, didn't we? Oh, all high?

MS. DORIAN: You did four high, four moderate and one insufficient.

CO-CHAIR ELWARD: Oh, yes, right, thank you.

MS. FANTA: Three of six high and one moderate.

DR. LYNN: I think you guys are exactly on time.

MS. TURBYVILLE: Impressive, Tom?

MS. WILBON: So operator, Katie,

can you open the line to see if there's anyone there for a public comment?

OPERATOR: Certainly. And if you have a comment, please press star 1 on your telephone keypad at this time. And we have no comments at this time.

MS. WILBON: Okay. You guys finished in record time, I think from that last measure, which is great. You're going to probably get out of here 15 minutes early.

So again, I just wanted to thank everyone for your efforts today, really great work. We got through all the measures.

Staff will be going over the next few weeks, we've had several meetings, we'll be going over all of our notes, compiling everything, doing follow-up with the developers.

The items that you guys have questions about we will be following up with the developer on, and we will forward your ratings based on the measure as is.

But we will forward that additional information to the Steering Committee so they can address those gaps that you identified, along with the information that they submitted.

We will have a follow-up conference call on August 2nd. We did actually schedule an additional call for the 17th, so you can take that off your calendars.

We did two calls up front because we weren't sure how, we knew we weren't going to get through everything at the in-person meeting, but at the time we still had the ABMS measures on our plate so we were kind of anticipating that additional discussion.

What we're going to do is we're going to contract, or smush all the ABMS measure feedback. My brain is really compressed.

CO-CHAIR ELWARD: Smushed.

MS. WILBON: Thank you. Shortened, compressed that discussion into one conference

call. We'll go through the measures and by the criteria, but probably not in the detail that we do here. Just so they have some feedback to carry forward as they refine their measures.

CO-CHAIR MAURER: So we're not going to wait for ABMS to revise whatever they felt like they needed to revise before we see it again?

MS. WILBON: No, they're actually taking them out of the process. So the review that we're doing for them on the August 2nd call is really more of a professional kind of courtesy for the effort they've put in, and the fact that we do actually want them to bring their measures back to NQF at some point.

They've put in a lot of work and they do have a really good concept for these measures and we want to try to help them, you know, get them to the level that they'd be ready to bring back into the process.

DR. BLAKE: And would this group be involved with that at that later date?

MS. WILBON: Yes, so the people who have already, you guys have already reviewed those measures so that input will be, that you entered into the Survey Monkey tool, we'll be compiling that and submitting that to them in addition to any of the verbal kind of discussion about the measures.

And the team will be thinking about how we can kind of, along with the co-chairs to figure out how we can kind of truncate that discussion in a useful way to get, I think it's like six measures, in a two-hour time frame.

So we'll kind of do some thinking about that and hopefully structure the call to kind of get through that in that two hour.

CO-CHAIR ELWARD: So we should be prepared. If we've reviewed the ABMS measure, we should be prepared to talk about those?

MS. WILBON: Yes, particularly

those that were assigned as a primary reviewer for those measures.

We'll kind of be leaning on you to give your input on those, and hopefully that will generate some discussion and feedback from others. But yes, we are still interested in your feedback on those measures.

DR. BLAKE: So after August 2nd, will we still be involved with ABMS as they move forward?

MS. WILBON: No. No, they're going to be removing them from the process. Staff will take on the responsibility of compiling that feedback and getting it back to them. And then they're going to kind of continue on their own path too.

MS. BOSSLEY: And I would just add that we know at some point we will have another project that it focuses on resource use, and our hope is that at that point, they'll be able to bring the measures back.

And then we'll figure out if we're

using the same structure as we did for this first project or not, but yes.

MS. WILBON: The Survey Monkey will still be available. You're welcome to do that, but the call on the 2nd, I mean if you would rather just kind of give your input verbally, we'll be prepared to take that.

We're going to record the call and all that so, but if you have enough that you feel like you need to write it down and you won't be able to articulate it verbally, we'll take either one. So we're open to whatever is convenient for you on that. Any other questions?

I want to thank our co-chairs for getting us done 15 minutes early. You guys did a great job today. Way to push through everyone, and we'll be in touch, okay. Thank you.

CO-CHAIR ELWARD: Well, we really want to thank all the members, but particularly you all as the staff here, really

amazing, you know. Sally, Sarah, Lauralei, Ashlie.

(WHEREUPON, the meeting in the foregoing matter was concluded at 4:50 p.m.)

G

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Pulmonary Technical Advisory Panel

Before: Kurtis Elward and Janet Maurer, Co-Chairs

Date: 07-19-11

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