NATIONAL QUALITY FORUM
+ + + + +
NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR ENDORSING PERFORMANCE MEASURES FOR RESOURCE USE: PHASE II STEERING COMMITTEE

+ + + + +

THURSDAY
JUNE 30, 2011

+ + + + +

The Steering Committee met in the Capital Room of the Venable LLP Conference Center, 575 7th Street, N.W., Washington, D.C., at 8:30 a.m., Tom Rosenthal and Bruce Steinwald, Co-Chairs, presiding.

PRESENT:

TOM ROSENTHAL, MD, Co-Chair
BRUCE STEINWALD, MBA, Co-Chair
PAUL BARNETT, PhD, VA Palo Alto Health Care System
JACK BOWHAN, Wisconsin Collaborative
JEPTHA CURTIS, MD, FACC, Yale University School of Medicine
LISA GRABERT, MPH, American Hospital Association

ETHAN HALM, MD, MPH, University of Texas Southwestern Medical Center (via phone)
ANN HENDRICH, RN, MSN, FAAN, Ascension Health
JACK NEEDLEMAN, PhD, FAAN, University of California, Los Angeles School of Public Health
MARY KAY O'NEILL, MD, MBA, CIGNA HealthCare

DAVID PENSON, MD, MPH, Vanderbilt University Medical Center
DORIS PETER, PhD, Consumers Union

STEVE PHILLIPS, MPA, Ortho-McNeill-Janssen Pharmaceutical, Inc.
DAVID REDFEARN, PhD, WellPoint
JEFFREY RICH, MD, Mid-Atlantic Cardiothoracic Surgeons Ltd.
WILLIAM RICH, MD, Northern Virginia Ophthalmology Associates
BARBARA RUDOLPH, PhD, MSSW, The Leapfrog Group
JOSEPH STEPHANSKY, PhD, Michigan Health and Hospital Association
DOLORES YANAGIHARA, MPH, Integrated Healthcare Association

NQF STAFF:
TAROON AMIN
HELEN BURSTIN, MD, MPH
LAURALEI DORIAN
SARAH FANTA
CAMILLE PRESBURY
SALLY TURBYVILLE, MA, MS
ASHLIE WILBON, MPH, BSN

ALSO PRESENT:

DAN DUNN, Ingenix (via phone)

BEN HAMLIN, NCQA
CHAD HEIM, HealthPartners
SUE KNUDSON, HealthPartners
TODD LEE, ABMS (via phone)
TOM LYNN, Ingenix (via phone)
JEN PEARSE, Ingenix (via phone)
JAIME ROSENZWEIG (via phone)

ARJUN VENKATESH, Brigham and Women's Hospital
KEVIN WEISS, ABMS (via phone)
CHERI ZIELINSKI, Ingenix (via phone)

TABLE OF CONTENTS

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

8:32 A.M.

CO-CHAIR STEINWALD: Now it says we're supposed to recap yesterday and we'll do that in the briefest way possible.

One thing that Tom and I would like to do and if anyone protests this, please speak up. We think it would be beneficial to finish the discussion of the HealthPartners measure. We got through scientific acceptability at the end of the day yesterday and we still have usability and feasibility.

We believe that the issues related to those two criteria have been discussed at some length already. And we're hopeful that we could finish up those two criteria fairly quickly and then get to the diabetes measures.

The chair of the Diabetes TAP will be joining us by telephone and we understand that he's available only through the morning and so we want to make sure we are able to get to the diabetes measures as quickly as we can.

And then once we're through the diabetes measures we have the wrap-up of the Ingenix measure to complete our agenda before we're able to adjourn.

Sally, are there any administrative or other issues, or Ashlie, that we need to talk about?

MS. TURBYVILLE: Just thank you for showing up for Day 2, rather than delay. So I understand this morning there's been confusion, somehow miscommunication on NQF's side about the hotel. So please be sure to clearly expense your hotel bill, if the hotel didn't already have it already covered back to NQF. If you have any questions, you can contact Ashlie or Sarah or anyone at NQF and we'd be glad to help you.

DR. BURSTIN: Brief item. When we get to the measures that have associated costs with them, we'll take a brief pause and I'll go over some issues of how we're going to handle some of that information.

CO-CHAIR STEINWALD: The associated costs that enter into our feasibility discussion?

DR. BURSTIN: Exactly. So for proprietary measures for which there's an associated fee, that becomes a subcriteria under feasibility and as we get to some of the follow up, it will be somewhat relevant under the ACG's use and HealthPartners, but especially for Ingenix.

CO-CHAIR STEINWALD: Okay. All right --

DR. ROSENZWEIG: Hello?

CO-CHAIR STEINWALD: Yes, hello?

DR. ROSENZWEIG: This is Jaime Rosenzweig. I'm calling in.

CO-CHAIR STEINWALD: Thank you. Dr. Rosenzweig, this is Bruce Steinwald. We didn't quite complete a discussion of one of the non-condition-specific measures yesterday and we're hoping to complete that fairly quickly this morning and then very quickly

after that go on to the diabetes measure. So if you can bear with us for a few minutes, that's what we'll do. Is that okay?

DR. ROSENZWEIG: No problem.

CO-CHAIR STEINWALD: All right, thank you.

DR. ROSENZWEIG: Yes.

CO-CHAIR STEINWALD: So the measure is 1604. The HealthPartners measure. And we completed scientific acceptability. Once again, the Steering Committee is acting as its own TAP, which means that we have to evaluate the subcriteria individually.

And the first one is usability. I guess we need to have on the board -- the first one.

MS. TURBYVILLE: Usability 3a is the measure performance results are reported to the public in national community programs by the time of endorsement maintenance review -- and so this is initial endorsement. So as we give you the context is if there is an

ability for them to be reported for public reporting or accountability models.

CO-CHAIR STEINWALD: Questions or discussion? Hearing none can we proceed directly to scoring. ?

MS. WILBON: Does everyone have a remote from yesterday?

CO-CHAIR STEINWALD: All right, let's take time out and make sure everyone has their remote. All right.

MS. WILBON: Okay, thanks.

MS. TURBYVILLE: So again this is high, moderate, low or insufficient. So go ahead and vote.

CO-CHAIR STEINWALD: And Sarah, we're ready to go? Oh, you've already got a platform. Very good.

Tell us when you're ready. Go now? Okay.

Paul suggested that the metallics in the case might be interfering. Do what you did yesterday.

(Laughter.)

MS. TURBYVILLE: If you vote high on Usability 3a, please raise your hand. Eight.

Moderates, please raise your hand. Eight.

(Laughter.)

Clearly, since staff can't count in a consistent way, so it's not reliable, I think you're right.

(Laughter.)

CO-CHAIR STEINWALD: Okay, so we're starting over electronically. We're going to start over. One, two, three, go.

MS. TURBYVILLE: So we have nine high and seven moderate.

Moving on to 3b, the measure performance results are considered meaningful, understandable and are useful to the intended audience for both public reporting and informing performance improvement.

An important outcome may have not

an identified improvement strategy, still can be useful for informing quality improvement by identifying the need for simulating new approaches.

(Pause.)

Four high, eight moderate, and four low.

Moving on to 3c which is that the data and the result details are maintained in such a way that the resources measure including the clinical and construction logic for defining of measurement can be decomposed to facilitate transparency and understanding.

So if you're ready to vote, go ahead and start, sir.

(Pause.)

So here we have seven high, six moderate, and three low.

And then 3d is not applicable. We actually -- it's a harmonization question, given where we are, this being the first resource use effort, at this time we didn't

ask developers to harmonize. Later on, if similar measures are endorsed, we'll either -- if there are opportunities for harmonization, we'll discuss it at that time, but right now we're not there.

And then for usability overall, this is not a yes/no. It is high, moderate, low overall and this is the Steering Committee vote. Please go ahead and start.

(Pause.)

I think one of you may have pushed your button before, sir. There we go. Two more, I think. There we go.

For the overall usability of this measure, we have six high, seven moderate, and three low.

CO-CHAIR STEINWALD: Any objection to moving right on to feasibility? Hearing none, so Helen, do you want to wait until we get to the subcriteria? Go ahead.

DR. BURSTIN: No, I'll just point out that I believe Ashlie said it's in your --

yes, part of -- NQF a couple of years ago allowed proprietary measures to come through the process, but part of that was to incorporate the associated fees with using the measure into the overall endorsement process, into that consideration.

So we requested, since the HealthPartners measure uses ACGs, that we actually provide for you the actual costs of using the ACGs. I know the submission indicated other potential tools are available. It had been tested with ACGs. So we thought it was important that you see this. It is one consideration as a subcriterion under feasibility. We at least wanted to have the chance to just point --

DR. CURTIS: I thought I saw on the application that the ACGs are now publicly available. Is that not correct? Okay.

MS. TURBYVILLE: And we would want to vote on it how it is now, even -- there might be some future efforts, but right now

this is -- just in case you're having a hard time reading it up on the screen, if you go to the thumb drive folder that you were provided by our team yesterday, there's a measure developer response sub-folder. If you click it open, you'll see each measure developer listed there.

There is the HealthPartners' folder, and if you open that, there is an Excel document which you can open up and will be this particular document right here.

CO-CHAIR STEINWALD: Let's move ahead.

MS. TURBYVILLE: Is everyone able to -- I just want to make sure people can read this. Yes, please.

MS. WILBON: As everyone came in, Sarah went around to everyone's computer and downloaded. We didn't get it until a couple days before the meeting, so we didn't send it out. It's called HP Price Table Proprietary Fees.

DR. BURSTIN: While everybody is looking for it, we'll just read it.

So -- it's simpler -- so the ACG's price that they provided to us were based on the client size, based on the number of covered lives. So under 50,000 covered lives on the commercial side it was $33,000. They specifically indicated that for other noncommercial entities there was no -- right, Sally, no associated causes?

And then it rises from there. So greater than -- less than 500,000 lives is between 42,000 and 159,000, and over 500,000 it raises from there tp 2,000 to 299.

CO-CHAIR STEINWALD: We'll wait until we get to the right criterion. We have 4a on the board.

MS. TURBYVILLE: Does everyone -- when we get to the subcriteria we'll just make sure everyone has had a chance to review the fee. So 4a does not involve the fee structure; 4a for feasibility is about the

clinical measure for the measure. The required data elements are routinely generated, and used during care delivery.

So thinking about this being based on administrative claims data, we would then request you to rate this particular subcriteria. Is it high, moderate, low, insufficient.

CO-CHAIR ROSENTHAL: Can I ask for clarification?

MS. TURBYVILLE: Please.

CO-CHAIR ROSENTHAL: Do we consider -- this talks about data elements routinely generated during clinical care. We wouldn't consider claims data or claims in this.

MS. TURBYVILLE: Yes, we would.

CO-CHAIR ROSENTHAL: Okay, all right. Claims. Okay.

MS. TURBYVILLE: It's meant to be broad, okay?

DR. BARNETT: And the consideration of costs, which subcriteria will

that apply to?

MS. TURBYVILLE: Four-D, and we'll be sure to remind you that that's it.

CO-CHAIR STEINWALD: I do notice that Sarah is not there. Are you going to stand in? Okay. Are you ready?

Let's go.

MS. TURBYVILLE: So for 4a we have 11 high and 7 moderate.

Moving on to 4b, the required data elements are available in electronic health record or other electronic sources. So are the required data elements available electronically is the question for 4b.

CO-CHAIR STEINWALD: Question. Turn your mic on and go ahead.

DR. NEEDLEMAN: I've been a royal pain in the ass asking about carve-outs.

(Laughter.)

But I think it is important if we're talking about capturing total resources, mental health services for a whole variety of

things that we're dealing with, including dealing with depression associated with chronic illnesses, and pharmacy costs are major components of our resources.

Historically, when they've been carved out, health plans have had a lot of trouble getting those back, and we're looking at a total cost of care measure here which means all I need is the total pharmacy costs. When we begin look at -- but when we also look at condition-specific costing algorithms that say we're going to get pharmacy costs from claims data, this will be an even bigger issue.

There are several ways this can be done. The worst is to take the per member/per month charges that are being paid to the pharmacy benefit managers and just bringing that back because that makes the charge or the costs associated with the HIV positive person and the kid who had an ear infection the same in pharmacy for a given year.

The most precise way would be to get all that claims data back and at least do a standardized pricing even if you didn't get the actual prices from the pharmacy benefit manager. I don't know, and I keep asking, is that what you're doing with these folks and sometimes they said no and sometimes they -- there are other options in between those two including doing some kind of imputation based upon the health condition of the patient, historic costs, and the average amount you're paying to the pharmacy benefit manager.

Every time you do an imputation or -- there's an element that imprecision, if not biased, introduced into these measures. So in the long run, if we're really going to have resource-based measures, the core plans need to figure out how to get some usable data back from the pharmacy benefit managers.

In the interim, we need to be sure there's a reasonable imputation method for approximating the pharmacy costs associated

with each patient. And that's why I've been a pain in the ass about this.

I had a conversation with the HealthPartners people yesterday about how they do this. And I think I got it. Unless you want to amend what you said yesterday, which involves doing some weighting of the average cost they're paying to the pharmacy benefits managers when they have a carve out around case suggesting that, which is good enough -- I'd say it's good enough for right now, but the long term future of resource-based measures has got to be to increase the precision of that and our report to the Board should reflect that concern.

MS. TURBYVILLE: Just a question for you, Jack, and the rest of the Steering Committee, because I do think this is an important point. I wonder if this is in 4c and 4d. My interpretation of 4b is that the data are available electronically, so I wonder if these sources aren't in electronic format,

but perhaps there are operational barriers, so they're not being collected. They're being carved out and it creates susceptibilities for error.

DR. NEEDLEMAN: It could be a 4d issue.

MS. TURBYVILLE: Just a question, okay, so it's not so much that they might not exist currently in electronic format, it's that they're, due to contractual arrangements in carving out, that they're not available for a total resource-used calculation.

DR. NEEDLEMAN: Right.

MS. TURBYVILLE: Okay.

CO-CHAIR STEINWALD: Are we content then to hold off this issue for 4d? Is that okay?

CO-CHAIR ROSENTHAL: I don't know. It looks to me like it's relevant. Available. They're not available.

MS. TURBYVILLE: Well, the longer read is that there are existing electronic

sources. I mean I'm not trying to split hairs. I just want to make sure it's captured -- I want to capture the conversation.

DR. NEEDLEMAN: Yes, so my understanding is the pharmacy benefits manager, the behavioral health managers, when you've got a carve out, are being paid a flat premium or payment per member/per month and the risk is shifted from the health plan that is the initial health plan for the beneficiary to the subcontractor.

And the subcontractors clearly have this because they're getting claims data, but historically the health plans have not been able to get detailed data back from them and have not gotten it back. So somebody has it, but if you're asking whether the -- whether Mary Kay has it, then the answer is no at the moment under most of those contracts.

So if somebody has it in electronic form, the prime insurer or health group does not have it if they've carved out,

is that a b or a d issue? I don't know.

DR. O'NEILL: I mean, I would -- just from an industry perspective, this is obviously an issue in evolution and so I think historically when there is carve outs for PBMs that they ran their business and like you say it was a financial arrangement. However, the whole world knows that this access to data is increasingly important for everybody's business in having a comprehensive view of what utilization looks like is increasingly important.

So basically from our industry's perspective, we in our contractual dealings with these organizations or through our self-insured employers who are choosing to opt out of our own benefit plan, increasingly those contracts have data-sharing language in them. It's usually the two legal teams that are getting in the way.

And sometimes the IT guys who don't have their databases talking, but I

think that this is the horizon for this becoming less and less of an issue is fairly close to us.

CO-CHAIR ROSENTHAL: I think the only question is really not whether it's a valid question, because I think Jack has made a compelling case. The only question is do we vote that issue in this one or in 4d. What does 4d say, if you give us guidance, then I think we can go ahead and vote.

MS. TURBYVILLE: So 4c, I'll talk about both, is susceptibility to inaccuracies or errors and then 4d is the data collection and measurement strategy can be implemented as demonstrated by operational use.

It's in here. So 4b really is that they exist electronically. And then we talk about barriers to getting these data that might hurt feasibility. Okay?

CO-CHAIR STEINWALD: So on the table is the notion that we'll address this issue in 4d and go ahead and vote on 4b.

We're back to 4b.

(Pause.)

MS. TURBYVILLE: Can we have the results? So the result is 11 high, 6 moderate and 1 low. Interesting. Okay.

Moving on to 4c which is susceptibility to inaccuracies, errors, or intended consequences related to measurement are judged to be inconsequential. So high would be that it's inconsequential. Or can be minimized through proper action or monitored and detected.

CO-CHAIR STEINWALD: Yes, Paul.

DR. BARNETT: So I would just remind that the two ideas, as a possible subject to unintended consequences, one is just by excluding anybody that doesn't have any visits or not setting some higher threshold for more than one visit to include people. There is this unintended consequence that the provider or plan that's being evaluated would want to get everybody in for

at least one visit a year so that they're in the denominator.

The other issue is because of the attribution rule that the visit with the primary care provider doesn't have to happen before the care is provided for that to be attributed to that primary care provider. There's going to be a disincentive for primary care providers to take on people who have had high healthcare costs who haven't had a primary care provider, and maybe even those that have.

And then the third is the fact that people are -- that only people in primary care specialties are counted as providers in this, that some specialties that act as primary care providers, and we gave the examples of cardiologists or people with serious cardiac problems or the infectious disease specialists who's caring for patients with HIV, the care that they provide will be, could be attributed to a primary care

provider.

An example might be someone that's receiving antiretrovials and being managed in an infectious disease clinic, goes to a primary care clinician for some Zyban to stop smoking and all those costs then get attributed to that primary care provider and the infectious disease specialist is not considered as a provider.

So that's going to provide some sort of disincentive for providers to get involved with these patients who are generally being managed in special clinics.

CO-CHAIR STEINWALD: Thanks for that. Yes, Mary Kay?

DR. O'NEILL: I know we discussed all of those yesterday and I mean the second part of what Sally read was can these otherwise be detected and managed. And so I guess even though we had a robust discussion amongst ourselves, I am not sure if there's any information from you about how those

issues have historically been handled within your organization?

I know that there are different percentages of primary care sort of driving the ship in different markets, but these issues can't possibly be new.

CO-CHAIR STEINWALD: Do you want to respond, HealthPartners?

MS. KNUDSON: Sure. I'll make a couple of comments and ask Chad to comment as well. I think on the non-user component, at the plan level, if that's the unit of measurement, that non-users could be brought into play at the plan level, if that's the unit of analysis. I think for the issue of assignment with the clinic visit being done after a hospitalization, a key element to recall is that risk adjustment for the acuity of that hospitalization will also come through. So in terms of mitigating and making that comparable, that's the whole design there.

And in terms of attributing just to the primary care for now, you know, so the premise that we operate on is that primary care is viewed as an opportunity to really enhance care coordination, partner with specialists to smooth transitions of care and improve, really, Triple Aim outcomes for the patients and members. And so we see a role in that for primary care.

We understand that in other areas of the country, perhaps that is not as strong right now, but again, reflecting on the discussions from yesterday as to whether or not there's opportunity for use of this measure in ACO development as it relates to understanding those models and how the care designs adapt in different areas of the country is a potential opportunity. So that's kind of where we've organized it.

I'll just see if Chad has comments.

MR. HEIM: The only other thing

I'd probably add is in our experience actually working closely with the providers, regarding the inheriting a case, I guess I'll call it, and actually what this measure does is kind of promote that coordination outreach, working with the specialists knowing that there's handouts and also intakes. They know who the referral partners are and so they want to reach out to them before so there's a smooth handoff from a specialist to a primary care and then vice versa.

And then also in terms of if there's, I guess, I'll call it opportunity to maybe to game, if you're going to go out and try to get everyone in to just get in for a wellness visit or a preventive service might come in and actually you find out they're diabetic and you've just inherited someone that you have to do some more care for them. So you get the ACG with that.

CO-CHAIR STEINWALD: Okay, thank you. Yes?

DR. J. RICH: So speaking of the flip side of this unintended consequence would be primary care provider who is now an exaggerated gatekeeper because he does not want to send his patient who needs a hospitalization for something to the hospital because those costs will fall to him.

Have you seen a reduction in services to any of the patient population that's not justified?

MS. KNUDSON: Actually, when we measure Triple Aim results for our care systems, in a lot of cases, and just use our own care system as an example, we see improved health outcomes in both process as well as clinical outcome measures, improved patient experience, and an improved total cost of care performance and we've tracked that over time for several years.

CO-CHAIR STEINWALD: Bill.

DR. W. RICH: One other issue I think you didn't get a chance to address

yesterday, I think this is a great measure for the region of Minnesota. You have lots of groups and you have a large patient population that you can assign to those groups, primary care groups. But again, this is a national metric.

How do you address the discussion we had yesterday about the difference in composition and the difference in patient population of Minneapolis versus Memphis. I hate to go back to that. One is about 60 percent African American with huge disparities in diseases compared to Minneapolis. How do you compare -- the measure doesn't say this is just regional. It's national.

So how do you address that? Because that fits right into the feasibility here, but perhaps unintended consequences, a change in a two-man primary care group that is taking care of 80 percent African Americans in Memphis versus a very large group in Minneapolis.

MS. KNUDSON: Right, and you know, actually that may be the traditional understanding of Minnesota and Minneapolis, but we have a very growing diverse population. We have one of the largest Hmong populations in the country as well as several other diverse populations.

Several of our clinics serve a large proportion of patients of color and we track those measure results, as I had discussed yesterday, on a quality and a patient experience realm and stratify those measurements in order to close disparities gaps.

We covered yesterday that this -- that we're not segmenting this measure in that regard, but so just to somewhat update the understanding of really the cultures in Minnesota and how growingly diverse it is. So that's one component.

And then I guess the other piece that I would add is reflecting on that

discussion from yesterday and some of the other discussions as well. Many of the measures that are endorsed by the National Quality Forum need community adaptation. So there's several of the clinical measures, for example from the Forum, that we can't implement in Minnesota because they're not endorsed locally by practice either by the Institute for Clinical Systems Integration, ICSI, and Minnesota Community Measurement, which is similar to the Wisconsin Collaborative for Community Measurement, uses a slightly different definition.

So I just raise as an example of community adaptation to making the measure work. So it's not always a one size fits all, but you know we work in sort of an imperfect world, but really with the goal of improving all those outcomes for patients and members.

CO-CHAIR STEINWALD: Thank you for that. An interesting discussion. I think some of it overlaps with yesterday's. My

preference would be to -- I appreciate that, but I think we should -- no, you're not prepared to do that, move to the vote?

DR. O'NEILL: I just think that some of the discussion here is based on a premise that if we endorse this measure somehow it will be nationally rolled out by some national entity whereby people are in every corner of the country are going to be compared with each other on the same measure. And I guess that's not my understanding of how these measures are utilized.

And so what I'm afraid of, sitting where I sit and based on the discussion yesterday, is that we have a very powerful measure here that measures something that the other things that we have considered so far do not measure which is what this stuff costs people and what it costs businesses and what it costs individuals out of pocket. And that somehow this discussion on this measure which is the same as the other measure we discussed,

but had standard pricing, suddenly all of these concerns which are really the nature of the measure in both points is a bigger deal. But it seems to me it's a bigger deal because there is not standardized pricing. That seems to be the biggest thing.

So I think that this measure is very powerful, very actionable and something that people in a local community and setting can get their heads around. And from a feasibility standpoint, if they're in Memphis and have primary care delivered by an endocrinologist and cardiologist that there's nothing in this data that would make that difficult to understand, and that this has the kind of power that the other measures don't. And just because it has real pricing and not standardized pricing, I'm just very worried that this whole measure is going to be jettisoned.

So I think if you take this measure and put it on a system or on a

community, there's a value in it. And even if the community looks different than the Twin Cities, you know, it still will have value locally. I actually, I'm not sure I haven't heard anything in the measure design that would tell me that -- say if my community has a particular medical community structure with heavily weighted to specialists, I don't see any reason why they couldn't be considered primary care. Maybe not.

I just don't want to abandon this opportunity. Thank you.

CO-CHAIR STEINWALD: We hear you. My sense of the chair is that we've covered the issue as well as we should feel that we have not left anything out. Can we move to the vote, please, on this sub-criterion.

MS. TURBYVILLE: Let's wait for Mary Kay to get back to the table.

CO-CHAIR STEINWALD: She has a whole minute.

(Laughter.)

MS. TURBYVILLE: So there are susceptibilities to inaccuracies or errors, high, moderate, or low, please vote.

(Pause.)

So we have four high, six moderate, and eight low.

Moving on to 4d, we have the data collection and measurement strategy can be implemented as demonstrated by operational use. And external reporting programs or testing did not identify barriers for operational use; barriers related to data availability, for example, timing, frequency, etcetera.

And 4d includes the consideration of the proprietary fees which Ashlie has kindly put back up on the screen so you'll want to take that into account as to whether or not that would constitute a barrier for feasibility for implementation.

So again, it's a high, moderate, low. And I'll leave it to you to decide.

CO-CHAIR STEINWALD: We also decided that this was where we would consider the carve out issue that Jack raised. Any discussion? Yes.

MS. YANAGIHARA: This relates to the carve outs and the availability of data. I think ultimately we're all trying to move toward having all types of data available, but we're not there yet. And so I'm just wondering in the interim if there's a way that when -- that there's an option to say either mental health is in or out, you know, pharmacy is in or out. And as long as it's clearly stated what is in or out and it's used consistently within -- wherever it's being measured, it can -- you can use it then without it.

For example, our total cost of care measure, we don't get mental health data. So it's out. But it's out across the board for all the groups in California that we're measuring. And so as long as it's clearly

stated, it might be a way in the interim to kind of allow us to keep moving forward without letting this always be a barrier to moving forward and a barrier for people using the measure, but it's clearly stated what's in and what's out, especially for those data sets that are known to be problematic.

CO-CHAIR STEINWALD: Further discussion? Okay, let's move to vote.

(Pause.)

MS. TURBYVILLE: So for 4d, we have 13 moderate and 5 low. So we will now vote for feasibility overall.

(Pause.)

So in thinking about those subcriteria and how you would weight them in an overall score, I don't know if there's any discussion needed, but we can go ahead and start the vote.

(Pause.)

(Vote: 3 high, 8 moderate, 7 low.)

CO-CHAIR STEINWALD: What the

discussion up front here is do we need an overall vote on the measure.

DR. BURSTIN: And I'm saying there was enough discussion that I think it deserves a final vote and especially because I think how partners would likely want to respond following this to some of those concerns raised about scientific acceptability.

The other thing is it is very common in these days, especially with controversial measures, that sometimes that NQF puts out a measure indicating there was lack of consensus and gets comments. So I don't want to lose the chance. This is so important. I think it deserves to go out for comment, so I would move forward.

CO-CHAIR STEINWALD: The only consensus I see is that there's lack of consensus.

DR. BURSTIN: There you go. That's my point.

(Laughter.)

CO-CHAIR STEINWALD: So we have --

MS. TURBYVILLE: So this is a vote for a recommendation of endorsement and as already noted, it's a yes/no and abstain.

(Pause.)

One more out there. Maybe somebody voted before -- there we go.

(Laughter.)

So as was stated prior to the vote, we have kind of a lack of consensus here. Nine, yes; eight, no; and one is abstaining.

DR. BURSTIN: Yes, all measures go out for comment. We actually invite comments on any measures that are not recommended, but there will be a specific section in the report, the draft report that will indicate this one did not reach consensus, very close votes. But in the interim though, you'll still have a chance to have the measure developers respond, so again that may sway you ultimately, but I think this does happen

fairly commonly.

CO-CHAIR STEINWALD: HealthPartners, thank you very much for enduring this discussion over a two-day period and we look forward for our issuance and your response to some of the concerns that have been raised. Thank you.

MS. KNUDSON: Yes, thank you to the NQF staff as well as the Steering Committee.

CO-CHAIR ROSENTHAL: So I think we're going to move to item 1557, relative resource use for people with diabetes and this measure is from NCQA. Welcome back.

DR. ROSENZWEIG: Hello, hello? Can you hear me?

MS. TURBYVILLE: Jaime, this is Sally and we should have thought of this earlier as you weren't here yesterday. So the order that we have found that works successfully is we'll ask the measure developer to provide an introduction to the

measure. Then we'll hand it over to you as the co-chair of the top four diabetes measures and ask you to provide input as we go through the criteria on what the top discussions were, as well as offer your expert opinion, but try and help us understand when it's TAP and when it's your input.

And then open it up to the Steering Committee starting with the Steering Committee folks who were assigned to leave particular components. So I will be sure to signal you and make sure we're opening up the phone here and there throughout the discussion for you to provide input.

DR. ROSENZWEIG: Okay, very good.

MS. TURBYVILLE: Jaime, also, I'm not sure -- we just logged into the webinar. I'm not sure if you were able to do that, if you're at your computer, but we'll be displaying slides of distribution of the task ratings if that helps you kind of summarize your feedback as well.

DR. ROSENZWEIG: Okay, I'm now on.

CO-CHAIR ROSENTHAL: So I think we have the order of the morning here now set on the three diabetes measures and the first one will be from NCQA so if you give us a little quick summary and then we'll have at it.

MR. HAMLIN: Thank you very much. So our relative resource use measure for diabetes is a very similar methodology to the RCA measure you reviewed yesterday. So all of our resource use measures are a standardized price, use standardized prices to assign effectively standardized utilization across a number of service categories with a predefined eligible population for people with diabetes, using a multi-year denominator that's very similar to our HEDIS quality measures.

Really, the only difference in this from the approach the other day is that the population is different in a sense. The service categories are identical. There are also service frequency categories that were

reported alongside of these for in-patient procedures, as service frequencies for this same population. So I'll just leave it at that. We went through the other measure in detail, yesterday, and again it's the same methodology applied for this population as was for the cardiovascular population.

CO-CHAIR ROSENTHAL: Jaime, would you give the TAP review and again, one other piece, we'll do this -- we'll vote on importance, then scientific acceptability, usability and feasibility in that order and we'll take the same -- this is for Jaime's benefit, that the importance -- we can't to vote on importance, but I've got a feeling that this one is going to pass the importance hurdle without a lot of discussion.

So Jaime, if you just give us kind of the first piece of the TAP which is the importance part, and then we'll get into the meat of the thing with the scientific part.

DR. ROSENZWEIG: Sure, this

particular measure was developed by NCQA and basically, their rationale for the importance of the measure was very well done, talking about the increase and prevalence of diabetes in the general population and the economic burden of diabetes which is very substantial in the general population as well. And they gave a number of good citations for the importance of high impact. Is that how you want me to do this?

CO-CHAIR ROSENTHAL: Yes, that's perfect.

DR. ROSENZWEIG: So in the voting in the Steering Committee nine voted, or all nine people voted to support the high impact part of this.

CO-CHAIR ROSENTHAL: Super, then let's take a moment and vote at the Steering Committee for the importance of the measure, given that the TAP recommendation is strongly positive and in this, one is yes, two is no.

(Pause.)

One vote is not getting tabulated every time. Oh, even better. Okay, passes the importance hurdle.

MS. TURBYVILLE: Seventeen high.

CO-CHAIR ROSENTHAL: Seventeen to nothing. So Jaime, now I think is the time to get into the discussion of the scientific merits. And so if you would give us the TAP view of that and then we'll open it up for discussion.

DR. ROSENZWEIG: So we have already covered 1b as well.

MS. TURBYVILLE: So for the Steering Committee, they rate on the overall, so they're not rerating the subcriteria that the TAP did, Jaime.

DR. ROSENZWEIG: Okay, so we're already moving on to Section 2.

MS. TURBYVILLE: Exactly, exactly.

DR. ROSENZWEIG: Okay, I understand. All right. So basically, the measure specifications really utilize the same

measure set that is being collected by the HEDIS effectiveness of care measures from NCQA. So it's been fairly consistent and they rely on -- they report on the total use of the diseases by service category and standardized prices related to service units for each measure.

And it has the advantage of being able to look at their quality measures in combination, their existing quality measures in combination with the cost-of-care data that they're collecting at the same time.

CO-CHAIR ROSENTHAL: Yes, and to whom -- if I could summarize and see if I've got it right, this is a total cost of care measure for people with diabetes using a standardized pricing methodology and a roll up of total costs and then indexed. But I missed to whom is the cost attributed? I missed that.

DR. ROSENZWEIG: It's attributed on various levels as far as I can understand,

but primarily on the per capita -- primarily the per capita, but it's population based for the most part.

DR. CURTIS: It's the same as yesterday, it's still specified by the payor at the health plan --

CO-CHAIR ROSENTHAL: At the health plan level, right. Thank you. And higher, yes. So it's the health plan.

DR. ROSENZWEIG: For the most part, HEDIS measures are not reported per physician.

CO-CHAIR ROSENTHAL: Right.

DR. ROSENZWEIG: They're reported per plan for the most part.

MR. HAMLIN: So the same criteria apply having multiple years of communities in enrollment, minimum sample size at 400 members in your population. Again, it's all the same. We attribute the health plans for our health plan support.

CO-CHAIR ROSENTHAL: I was just

trying to clarify so there was -- we had a common starting point on the discussion. So we've heard from the TAP. Now our technical scientific reviewer, Sally, help me, who? Who on the committee did scientific?

MS. TURBYVILLE: Carlos is not here.

CO-CHAIR ROSENTHAL: Not here. Well, I'm going to open it for a discussion then. I think we have a pretty good idea from yesterday's conversation what, if any, of the scientific issues are. So I'll open this for discussion.

Jack?

DR. NEEDLEMAN: I had a question because I'm not sure -- to the developer because I'm not sure I understood the answer from yesterday. But before I ask the question, with all these claims-based measures, I don't think this is a deal breaker. It is just inherent limitation of moving forward on measuring resource use right

now. But it's important for us always to remember and keep in mind that we're only counting resources that are billed, that health plans or groups which have unbilled services they make available to their patients, care coordination, nurse educators, diabetes nutritionists that are not billed services, those resources are real resources. We think they make a difference in the effectiveness of the care.

We have no way of measuring whether -- how -- we may not have ways of measuring what's there and how that's done. To the extent that groups have negotiated differential prices to pay for that because they said look at the additional things we're doing, we need a higher physician fee or whatever, standardized pricing wipes out those differences.

CO-CHAIR ROSENTHAL: Comment on that?

MR. HAMLIN: Right, so our

approach is we're measuring utilization and we feel that because we're attributing these to the health plan, the health plan, through their various programs, their DM programs, their wellness programs, other incentives for participation, if you will, all affect utilization. And so by looking at the high level utilization across specific service categories, we're basically giving them a snapshot of their utilization for a specific period of time. And they can go back and look and see how these specific programs may affect the utilization results, if you will.

So all these programs, we feel, affect the utilization of the plan members when you're looking at it in the aggregate. We do not have ways of measuring specific care coordination components at this time, so therefore, we're measuring what we can measure at this current time and giving that back to the plan as here's how your utilization compares to other plans when you risk adjust

it and when you standardize it.

CO-CHAIR ROSENTHAL: I think Dr. Needleman would probably agree. You're agreeing.

DR. NEEDLEMAN: Yes.

CO-CHAIR ROSENTHAL: His point that was still valid in the sense that costs are being expended for certain things to get that utilization and consequently if your total cost of care will underestimate because it's claims based will underestimate the actual cost that was necessary to deliver those volumes of services and it's just a weakness of the extant methodology.

Somehow we have to figure out --

MR. HAMLIN: The plan is not able to capture that because they will be able to plug in their actual prices for each individual service categories based on this methodology. They can also -- I'm assuming, will be able to roll up as some sophisticated plans like how partners can do, show their

total costs, be it total actual costs for all services for these categories and they will be able to then make those arguments to each of their stakeholders, if you will, about the differences between these and why they look this way. But at NCQA's level, we only get the utilization level data. So we can't, as a measurement organization, measure that, but I think there are ways to measure it using the same template, if you will. It just requires an additional drill down into the data by the plan themselves.

DR. NEEDLEMAN: Just again, I don't think it's a deal breaker on moving forward with measures, claims-based measures of resource use. We just need to understand there are certain kinds of services and certain kinds of clinicians that are simply invisible in these measures and we just -- that's a general issue that we ought to just keep in mind and any reports that come out of the Committee ought to acknowledge that.

The question I had and I'm still -- I'm confused. We heard -- I asked about how in-patient pricing was done in terms of the standardized pricing, and I thought I understood your answer yesterday. And then one of the other measure developers said they were using your standardized in-patient pricing method and it was different from the way you described it.

So can you, once again just try to help me understand how in-patient pricing is done and therefore how and what kinds of variances of in-patient use beyond admissions we're actually capturing in our measures of relative resources.

MR. HAMLIN: So we use a number of different resources to generate standardized pricing tables which are again, the Medicare fee schedule, we have a large research commercial database that Ingenix has maintained for us over the years that that feeds into this. Our pharmacy data comes from

First Bank. Primarily, we're capturing about 33,000 of the prescriptions that are written frequently enough so that we can actually feel like we can standardize, price these things.

We then annually publish the standardized pricing table which is down to the code level. So CPT, there's a standardized price assigned. We make this freely available on our web site. Anyone can use it any way they wish to. We use it specifically in the section that's detailed. I believe it's 9.7 in your materials for our measures of measuring health plans against each other.

ABMS, I know, uses our standardized pricing tables and they use them in different ways, but again, but to use the standardized prices as sort of a leveling ground for removing the proprietary fee schedules and contract specifics out of the equation and we're fine with that. That's why we published these. We make these freely

available because we have spent considerable resources to generate these tables every year and we feel like we want to get more out of it than just five measures worth.

So we know that they do that. That's why they make them freely available. They're again, just a standardized, effectively national price index for these services that we can identify and we feel that we can price effectively because there's adequate utilization or there's adequate information that allows us to assign a standardized price to each of these individual components. But again, different measures. Stewards for different measures may use these prices in a different manner.

CO-CHAIR ROSENTHAL: Are there other questions either for the developer or for our own TAP chair?

Yes, Paul?

DR. BARNETT: Yes, I just noticed that looking at the TAP's scores that there

were a few concerns on the 2b and also 2b3 exclusions. I guess the exclusions had to do with the exclusion of people over 75. That was the concern. That was expressed by the TAP and maybe our TAP chair can explain why those votes happened.

CO-CHAIR ROSENTHAL: Jaime, did you hear the question?

DR. ROSENZWEIG: If you're talking about the voting for the 2b1, 2b2, and 2b3?

CO-CHAIR ROSENTHAL: Yes.

DR. ROSENZWEIG: They're still basically mostly high with only some of the people giving a medium rating. I'm not sure -- it doesn't mean they were against the --

CO-CHAIR ROSENTHAL: Jaime, no doubt. He is just asking what was the basis even for some people having only rated those three moderately.

MR. HAMLIN: I actually do remember the specific conversation because it applied to our mandatory exclusions for ESRD

transplantation primarily because in this population, the TAP felt that those two actually were things that could really contribute to the cost of care. And so what we have done is then take that back now and we're reinvestigating that now as a measure update.

So maybe our four mandatory exclusions for active cancer, transplantation, ESRD, and HIV may not be applicable across all five measures because they are particularly relevant to the diabetes population.

CO-CHAIR ROSENTHAL: Thank you.

DR. ROSENZWEIG: Who is this speaking?

MR. HAMLIN: This is Ben.

DR. ROSENZWEIG: Oh, Hi.

MR. HAMLIN: Dr. Rosenzweig, how are you?

DR. ROSENZWEIG: Yes, he just described that pretty accurately. I think the main issue was the ESRD and the fact that a

lot of those people were excluded from this population because they go to Medicare.

CO-CHAIR ROSENTHAL: And it's clearly just a modest concern because none of these were ranked low in the event.

Other questions? Yes, ma'am.

MS. HENDRICH: I have a comment. I just wanted to build upon the point that's already been made for just a moment. I think one of the most difficult questions we have to answer in the future is the issue of care management, disease burden, and readmissions back into acute care and from being on the acute care side, as long as we continue to bundle these care models within these large process measures where there are intermediate level providers, especially in the area.

And I was going to bring this up around the congestive heart failure measure where we have really some of the strongest evidence around the cost effectiveness of that. I think we're not going to be able to

answer the question. I think it's a yes and I'm hearing that the developer is saying that through the different groups, we're going to be able to unbundle that and perhaps answer it.

So my philosophical comment to this group and challenge is at what point though do we start to challenge developers and say we have to be able to code in such a way that we can start to identify the actual care model that lies beneath the cost structures we're looking at.

MR. HAMLIN: And our new measure development in the EMR realm that actually have and include measures of care coordination because that data is available, I think will strengthen our utilization approach because we're reporting the quality of care with the utilization by strengthening our quality side. By understanding how these specific components of care coordination and patient satisfaction will affect the quality results, we can then

link those more directly to the utilization.

But in the absence of quality measures around care coordination and management of patients and patient satisfaction, and the ability to access care, I don't think we as NCQA don't want to dive too deeply on the utilization side because we don't have supporting evidence that those indeed do make a difference. And again, we have very high threshold for tolerance in that arena as far as what we will use to report and rank plans for our results.

CO-CHAIR ROSENTHAL: I'm not sure either hers or Jack's comments really were directed at this measure as they are kind of the general feel. I think there would be widespread agreement and in fact we had a measure yesterday that excluded everybody that was sent to a SNF. You know, when you think about that, it's insane. And yet the SNF costs are very hard to capture.

And we're on this crusade now to

dump people at a "Uwe Reinhardt" of dumping people out of acute care hospitals as if that's the salvation of the healthcare cost system without any notion that where they're being dumped to is going to really and truly and unequivocally be a lower cost proposition. I think that's frankly still an untested hypothesis. But anyway, I don't think any of these comments are really directed at this measure as much as they are kind of a general --

MR. HAMLIN: That was an FYI.

(Laughter.)

DR. BURSTIN: I do think it's an important thing to consider into the final report. That's again, the exact kind of thing we want to make sure the Committee emphasizes the fact that the broad scope of cost codes are going to be really important to consider across the board to really get a full handle on who is doing what, what works, what doesn't work as it relates to quality and cost.

MS. HENDRICH: At the risk of being redundant, since we know that comment is going to be inserted, I would also stretch our thinking beyond just intermediate level care providers because this really goes into the issue of home health care aides, right? Which I think that need will probably outstrip everything we've looked at thus far based on what we're seeing in doing the deep dive into readminissions around chronic disease. So thanks for considering that.

CO-CHAIR ROSENTHAL: Mary Kay.

DR. O'NEILL: This is a comment from the carrier industry. When you're -- this measure is designed to compare our industry, not really compare practices and delivery systems. And so there's a lot of variability. And in fact, one of the things I think we compete on is our ability to support our members to various different care episodes. So we have huge infrastructure on disease management, case management, health

advisor, integrated behavioral health data, predictive modeling, preference-sensitive care.

I mean the amount of money that our company spends on this aspect of management of our specific population is what we sort of put our stake in the ground around and why we are active in NCQA and have been for a number of years. So this isn't really even getting at the codes that will allow practices to bill for care coordination.

There's other entities within the larger healthcare world that are providing this level of service. And so when you look at some folks that are coming in from Medicare or Medicaid, the robustness of their carrier in these areas is not comparable.

So anyway, I mean the cost and the benefit of this kind of activity resides different places and we're not going to ever get -- we don't have any claims data. That's our business investment.

CO-CHAIR ROSENTHAL: That needs to be done. Let me bring us back to this measure and the scientific -- I have one last question and then I think it's looking like we'll be able to bring this to a vote, which is you do have a truncation, as I recall and would you share with us the logic behind the truncation scheme?

MR. HAMLIN: Again, because in the population we want to avoid a small proportion of members driving the standardized costs up beyond a comparable level so when members reach that cap and there's a table of caps for specific costs, they're basically just truncated at a cap and they're not excluded from the population, but it prevents small spikes from skewing the results in one direction.

CO-CHAIR ROSENTHAL: And the number was 100,000 or something like that?

MR. HAMLIN: I don't remember the actual current number. We adjust it slightly

every year.

CO-CHAIR ROSENTHAL: It's not important.

MR. HAMLIN: Yes.

CO-CHAIR ROSENTHAL: It's not so important. Any other questions, comments, or discussion on the scientific merits? If not, then I think we're prepared to vote on the scientific portion of this and if I recall, this is yes or no. One is yes, two is no.

(Pause.)

MS. TURBYVILLE: So after a lot of sending signal we have 18 yes, Jaime, so we'll move on now to usability.

CO-CHAIR ROSENTHAL: So Jaime, if you would, give us the TAP view of this, please.

DR. ROSENZWEIG: Right. Can you hear me well?

CO-CHAIR ROSENTHAL: Yes.

DR. ROSENZWEIG: Basically, the usability part of this was generally -- the

analysis was pretty well received by the TAP. I'm looking to the section here. Because of the fact that they're collecting all of their data through HEDIS that this could be -- that they could be able to understand it fairly clearly and be able to use it for decision making because it was coordinated well with measures that -- of quality of care. So it can be use for quality improvement and public reporting and quality improvement with external benchmarks.

So for those reasons I think we gave them high scores with the exception of the issue of the harmonization part.

CO-CHAIR ROSENTHAL: But that's not applicable. Right.

Doris, I think you were our internal reviewer.

DR. PETER: Yes, I don't think there's too much to add. It's publicly reported. The plan obviously uses -- the aggregate results were reported for the public

and annual reports that they put out and I think we discussed it with the other measures.

CO-CHAIR ROSENTHAL: All right. Is there any discussion of usability? Boy, are we getting good.

I think then that this is ready for vote.

MS. TURBYVILLE: And this one is a high, moderate, low, insufficient.

CO-CHAIR ROSENTHAL: So one, two, three, and four.

(Pause.)

CO-CHAIR ROSENTHAL: Okay.

MS. TURBYVILLE: So 12 high and 6 moderate.

CO-CHAIR ROSENTHAL: All right, and then finally, feasibility. So Jaime, if you'll give us the TAP version of this.

DR. ROSENZWEIG: Yeah, here again, the TAP felt that this was quite feasible to be able to collect the data. There was a really -- it was uniformly agreed that NCQA

was able to be able to collect the data that they wanted to and be able to correlate it well with the measures. The only issue where there was disagreement was in the area of susceptibility to inaccuracies and unintended consequences. I guess there was some concern that there might be some issues related to the data audit process that might make it occasionally a little more difficult to be able to collect accurate data.

CO-CHAIR ROSENTHAL: Ben, do you recall what that was in specifics or an answer for that?

MR. HAMLIN: No, I mean, we do -- all of it is submitted to NCQA. I'm sorry, closer. All of the data submitted to NCQA has to go through a certified auditor before it's allowed to be reported. So we do reduce the amount of errors in the data through this, and each auditor must be certified through a very extensive process and recertified every year, like a licensing agreement kind of thing.

Again, there's a lot of data points. We're working to improve that process and automate much of it, so there will be automatic validations, so next year there will be a number of additional automatic validations that will again reduce any kind of misrepresentation of the data, but it's 50,000 data points. There's possibility for some error somewhere along the line.

CO-CHAIR ROSENTHAL: Okay, and Paul, you were our Committee reviewer, I think?

DR. BARNETT: Yes, I don't have anything to add.

(Laughter.)

CO-CHAIR ROSENTHAL: Okay. This is now open for discussion. Questions? Hearing none, I think we are ready to vote on this and -- I'm sorry, Jack.

DR. NEEDLEMAN: I'm going to vote somewhere between high and moderate, but I think NCQA is a well-established measure

developer and measure producer. A couple of times during the conversation, we've basically heard reliance upon the auditing function of the individual health plans. I think there's an issue of transparency there in terms of exactly what's being done and so forth.

So I just want to note that, without initially saying it indicts the measure, I think it's very feasible, but we ought to think about transparency here for understanding exactly what data is coming forward from the plans.

MR. HAMLIN: The audit is an independent audit. It's not a health plan audit. The auditors are certified as independent auditors of the data, and the health plan will contract with them to comply with the audit. But they're not health plan employees or have any other relationship other than their --

CO-CHAIR ROSENTHAL: Okay. Any other -- Delores?

MS. YANAGIHARA: I just wanted to comment that I think the data elements are available. It can be collected. But once the data are collected, there's still a lot of analysis that needs to happen before the measure can really be meaningful and you get a result back.

So what the health plans actually do is submit a whole bunch of data to NCQA. NCQA has to crunch the numbers and come up with all of the benchmarks and the results for each plan. So it's not something that an individual plan or individual organization could do on their own. It's all about the data are submitted and the data together needed to be calculated. So it kind of lays into feasibility here. It's a great measure, but it does rely on that very sophisticated analytic analysis after the fact, after the data collection. It's not just like here's the numerator. Here's the denominator. Here's the rate, you're done. It's quite

complicated.

CO-CHAIR ROSENTHAL: I see, so your point was different than Jack was making.

MS. YANAGIHARA: Different from Jack's, yes.

CO-CHAIR ROSENTHAL: And it relates to the fact that without NCQA --

MS. YANAGIHARA: You need some kind of data aggregate or body to collect the data and do all the analysis and spit out the results. It's not just an individual organization that can do that because it's all about how you compare it to others.

CO-CHAIR ROSENTHAL: That was a good question, but the algorithms are in the public domain.

MR. HAMLIN: We post them on our website.

CO-CHAIR ROSENTHAL: Pardon me?

MR. HAMLIN: We post all our methodology on our website. If another data aggregator wanted to do this, they would be

able to do so.

CO-CHAIR ROSENTHAL: I think that's a really -- it's not a black box, but it's not trivial. I couldn't do it.

MS. YANAGIHARA: I think the other comment related to the audit, the audit also is not black box. The audit manual is posted on the web. That's all transparent as well and exactly what's done in the audit, and their prophecies and what they're looking for is all on the website as well.

CO-CHAIR ROSENTHAL: Other comments or questions? Was there another hand up over here?

All right, then I think we are prepared to vote on feasibility. I'm breaking my back checking to see how much time. This is one, two, three, and four. So high, moderate, low, and insufficient. Okay?

(Pause.)

MS. TURBYVILLE: Eleven high and seven moderate. So I think we're ready to

vote on the measure overall.

CO-CHAIR ROSENTHAL: That's next. So now it's time to vote on the measure overall and this is yes/no and abstain, unless there's any further discussion about the measure in toto. Hearing none, let us vote.

(Pause.)

MS. TURBYVILLE: So we have 17 yes and 1 abstain.

CO-CHAIR ROSENTHAL: All right, that concludes the discussion on 1557.

MR. HAMLIN: Thank you very much and thank you, Dr. Rosenthal.

CO-CHAIR ROSENTHAL: Thank you.

DR. ROSENZWEIG: That was -- I hope all the others go half as smoothly.

CO-CHAIR ROSENTHAL: Well, when they go smoothly, they go smoothly. I think that's about all we can say.

DR. ROSENZWEIG: Yes, I shouldn't editorialize. Okay. I'm sorry.

CO-CHAIR ROSENTHAL: We all share

that hope that the rest of them are as smooth as that.

So next is 1576, which is episodes of care for patients with diabetes over a one-year period. This is an ABMS measure. And is Kevin on the phone?

DR. WEISS: Yes.

CO-CHAIR ROSENTHAL: Kevin, good morning. Did I hear a yes?

DR. LEE: I'm not sure if Kevin is on, but Todd Lee is here I can --

DR. WEISS: Kevin is on as well.

CO-CHAIR ROSENTHAL: Good morning, gentlemen. If you would not mind giving us a brief summary of the diabetes measure.

DR. WEISS: This is Kevin. I'll give a short intro and then I'll ask Todd if he'd like to add to it. But essentially, as we have proceeded with the work group on this measure, the question was what would be a look at a person who has need for management of a diabetes that was stable in the time period of

what's recognized as a long disease process, recognizing that it's at different times in the disease process, particularly towards the advanced stages, that it has a very different set of complexity and thereby a whole set of different expectations than it does for much of the time of the person who has diabetes.

It shows a one time period to measure this and we're very reflective on the fact of how it would eventually match up to quality measures since there are quality measures in this area, so we're so well advanced, recognizing that in the one-year period what resources really could be attributed to the provider, recognizing in that context that there are a number of activities that may be associated with resources that may alleviate the long-term supply, then really what could be looked at in one year.

So rather than look at total costs, we look for anything adverse to that in

the context of this, but rather looking at diabetes-specific costs as it relate to the type of activities one would expect in a one-year aspect of care.

We had the work group ask the question as to what primarily would one want to look towards attribution, and they thought that this could be attributed to an individual provider or based upon a provider group that you had.

Those are the basic elements of this diabetes measure.

Todd, would you like to add anything to that?

DR. LEE: I'll just add how this is similar or different than the other measures that you all have reviewed from us. Unlike -- this is probably more similar to the CAD measure in that individuals are identified during a 12-month period. And then we look at the resource use in the following 12-month period, so we have an identification and a

measurement year, unlike our AMI measures which are triggered by an index event. So we're taking an approach from a chronic disease standpoint and looking at resource use over the 12-month measurement period of individuals who had identified in a previous year.

CO-CHAIR ROSENTHAL: All right. Great. Thank you very much.

Jaime, will you discuss the TAP discussion?

DR. ROSENZWEIG: Sure. This was the 1576 on the ABMS measure and -- hold on a second. Let me just get my notes. Yes, as was just discussed, this measure discussed resource use and costs associated with management of diabetes over a one-year period. It identified patients in a management phase of diabetes by including people in the year prior to the measurement year and resources use and cost during the measurement year and patients with new diagnosis of diabetes and

those with end-stage disease, which was not exactly clearly defined, were excluded from the measure and resource use was attributed at the level of the individual provider as opposed to the last measure set.

So the measure type was per episode, but really it is over a year period and the level analysis was at the clinician or individual level. So with respect to 1a, everyone clearly agreed that the measure had a very high impact and high importance. Eight thought that it was high and one thought that it was medium.

However, there was a sense that there might be some resource use or cost problems. The TAP discussed that the submission provided evidence of gender and only racial disparities and did not address the other areas of disparities, including socio-economic issues, and the TAP discussed that this may be due to a lack of literature in the area. But I think there is some

literature in the area, and there was some suggestion that if in the future that the Steering Committee would have to give guidance with respect to how this resource measure should be used.

And then with respect to the purpose clearly described there again, there were six highs and three mediums, and the concern among some of the people was that they needed more detail about whether the measure is paired to other quality measures. It's discussed later on in the section, and then there was also some concern about the resource use service and categories being consistent and representative. And here again, there was disagreement over this issue.

CO-CHAIR ROSENTHAL: All right. Thank you very much. I think it sounds as if the TAP on this one looked at this importance question in the dimensions like we talked about yesterday. Does the importance apply specifically to the measure? Just to clarify,

we've been considering the importance question more broadly, and so I would suggest that we go ahead and vote on that and then we can get to the scientific questions on this thing. And this is one, yes; two, no.

(Pause.)

MS. TURBYVILLE: So we have 18 yes on importance, so we can move on to scientific acceptability.

CO-CHAIR ROSENTHAL: So Jaime, if you all would discuss the -- if you discuss the TAP discussions --

DR. ROSENZWEIG: Okay, going to 2a1, which is whether or not the measures are well defined and the specifications were precise, there was actually a sense that this was not fulfilled. And some of the issues that came up included as to -- it was unclear as to why renal failure codes 585.3 and 585.2 and 585.4 were excluded from the measure. The codes apparently that they listed were not updated.

Bariatric surgery was not included. The TAP required rationalization for the specific drug selections, in particular, why the uses of only oral hypoglycemic or injectable medications are in the inclusion criteria and others should be considered.

They requested a clarification for the lower age band of 30 years for Type 2 diabetes that was being specified. The developer responded that the measure was supposed to be focused on Type 2 diabetes. However, Type 2 diabetes is being seen at earlier and earlier ages, as most of you probably are aware.

And there was some issues, if that's the case, then the TAP said that the title and measure description in 10 should clearly state the focus on Type 2 diabetes rather than just on diabetes as a whole.

I don't know if that's been changed since we reviewed it or if the actual

text has been changed. And then the TAP required clarification on Type 1 exclusion and how you would exclude it, considering it's very difficult based upon data from the chart or data from administrative data. The distinctions between Type 1 and Type 2 are listed in coding, but they're often not used correctly by physicians.

And there was also a sense that the inclusion and exclusion criteria needed to be tightened up, at least as written in this protocol to be sure to exclude patients with Type 1.

And also, there was a question as to how new diabetes would be excluded. New diabetes diagnosis would be excluded since there's a fairly high proportion of patients who are diagnosed, who are under-diagnosed and may be diagnosed in one place and they're not listed elsewhere. So those are a lot of the issues related to definition, which caused a lot of debate.

With respect to reliability testing, people felt that that was, indeed, very sound and has been tested in large database by ABMS.

With respect to the issue of specifications consistent with resource use and the cost problem, and here again, there was a lot of people who had concerns about this particular part of the protocol because the specifications were not always clear. Issues related to the time of entry into the target population, how that would be determined, and how that would be counted with respect to resource use. There was some concerns that they were not listed precisely enough and that they should also -- the costing method -- it was felt that the costing method should require more clear clarification.

And there were also concerns about the issue of exclusion and exclusion criteria, which it was felt at least within the respect

of this particular protocol that they required more clarity and specific rationale. And in addition, the target population identification, as listed earlier, needed to be more precise.

In general, there was also some issues that were raised about the validity testing. The general sense was there was insufficient information provided on the validity testing, testing analytic methods, and results.

With respect to exclusions, also, there again, almost all the people voted medium, with one low and here this was -- largely the rationale for this was that they were -- we didn't know whether they were going to be consistent inclusion and exclusion criteria across the measures that were relevant. And the measure as was written didn't provide clear rationale for measure exclusions.

Also, there was some disagreement

as to the score for the risk adjustment. The TAP wanted confirmation upon which risk adjustment approach would be selected and the methodology that they listed there appeared to be based upon the widely-used CMS HCC approach, which TAP liked, but the TAP couldn't assess risk adjustment because some things were missing, including fit testing and the RSQ value and the rationale and list of selected covariates.

So we felt that they needed to be more clear on how to instruct the users how to apply risk adjustment to this measure. And then it was felt, however, that in general that most people felt that the identification of statistically significant and meaningful differences could be done with this measure set. The minimum sample size for reporting implementation was not provided, and that's important because this measure is being selected on the -- is actually looking on the physician level, and many physicians don't

take care of that many patients with diabetes in their population, and there may be a large percentage of them that would have a small number of patients that could not be sufficiently compared with other physicians.

CO-CHAIR ROSENTHAL: All right, thank you very much.

DR. ROSENZWEIG: Am I missing anything? Yes, the multiple data sources thing was not applicable, and stratification for disparities, there was no real stratification listed there.

DR. CURTIS: The only other thing I thought was the issues of attribution that we discussed, the difficulties of assigning particular physician. I think you raised the concern about --

DR. ROSENZWEIG: Oh, yes. I'm glad

-- who is that, is that Jeptha?

DR. CURTIS: Yes, that's me.

DR. ROSENZWEIG: Yes, yes. That

was an important issue was that typically the care of these patients is very shared, okay, among various providers and how you would attribute the overall care to which provider with respect to the costs, how the costs would be, would be sort of clarified, becomes very complicated in the diabetes population.

DR. HELM: This is Ethan. You know, the attribution part, the complexity was one of the things that caught my eye in reviewing this as well. There's sort of a tiered algorithm of costs of attributed to sort of the primary diabetes provider, based on three criteria. One is that that provider did 70 percent of those visits in a year. If that's not met, then it's the person who did 30 percent of the visits, and then there's a third tier which is kind of like we don't know who attribute it to, and one of the things empirically which was kind of striking but probably not surprising was that 55 percent of the patients that they identified, they could

not attribute to a provider, so it's slightly over half could not be attributed to a provider. So it brings up some generalizability concerns.

CO-CHAIR ROSENTHAL: All right, thank you very much. A lot of work went into doing that analysis, and this Committee much appreciates the effort that you guys put into thinking that through and doing such a careful and thorough evaluation.

Our internal primary reviewer for scientific acceptability is not with us. I wonder if, out of order, but Steve, feel free to say no, but you did look at the usability. And I imagine that perhaps you might have read this a little more closely than some of us based on having to look at it from a usability point of view. Do you want to make any comments with regard to the science?

MR. PHILLIPS: Well, just a couple and they were actually touched on in the TAP review. Looking -- coming at it from the

usability review, I guess the biggest question is just as was touched on linking this with some sort of quality measures because just with the resource use, I'm not sure what I would make of it, given the outputs, this higher resource expenditures, good or bad. But you know, we've already talked about the need, eventually, that these will have to be meshed with quality measures.

I guess the biggest issue that I saw, again, was just touched on as far as attribution. I can think of situations where, basically, you're not able to attribute a patient because they're really not being very well managed and so then they end up getting spread across or attributed to someone else or multiple providers. To me, that's a significant problem because that's exactly what I think we're trying to get at here is identifying -- you may not have much resource use because you're not really managing the patient very well, and then you lead to these

downstream expenses in the overall health system. I think those were my main comments.

CO-CHAIR ROSENTHAL: Okay, open for discussion. I thought I saw a hand up over here earlier, maybe out of sequence. Jack?

DR. NEEDLEMAN: Yes, a question for the TAP. When you looked at the services and procedure codes that were being included in this measure because it's diabetes-specific care rather than a total cost of care for patients with diabetes, were there any important exclusions with things you felt should have been on the list that weren't?

DR. ROSENZWEIG: Well, there were a number of things that were excluded that we felt shouldn't have been. I mentioned that earlier. The renal failure codes which are closely related to diabetes, bariatric surgery, things of that sort.

I cannot recall whether there were very many specific codes that were missing --

that were just missing by accident. They looked like they were including most of the diabetes-related codes that -- that was my recollection. I don't have the actual list up in front of me right at the moment, but I don't think that there were many problems that were raised related to that.

DR. ROSENZWEIG: I can imagine with patients with a chronic disease like diabetes, some codes that are not specific to diabetes -- you would expect to be part of the diabetes management would be included in something else. So that's the kind of things that aren't are on the list that I'd be a little bit concerned about.

DR. NEEDLEMAN: Yes, there are a lot of things that are hard to sort out. Patients get admitted for an acute infection, but it's also maybe associated with uncontrolled diabetes, and sometimes -- that would probably be listed as something that's not diabetes-related in most settings, but in

fact, it is diabetes-related.

I think we had some discussion about the issue about how in many cases length of stay, hospital length of stay may be increased fairly significantly in some of these situations, but they're not really dealing with most of those issues. They're dealing with in this particular measure set with the diabetes-related admissions, which mostly includes either a hyperglycemia or a hypoglycemia, those kinds of things.

CO-CHAIR ROSENTHAL: This is a question out of ignorance, but they exclude polycystic ovary disease explicitly. Why that? I'm sure there must be some reason, but it's not specifically apparent to me.

DR. ROSENZWEIG: Well --

CO-CHAIR ROSENTHAL: Kevin, could you answer that?

DR. WEISS: I think it's because in order to define the patients with -- in this data set with diabetes, they use

medications that are normally associated with treatment of diabetes. And what's happening nowadays is that some of these medications are being used in situations other than diabetes. So metformin is being used quite commonly to treat polycystic ovarian disease.

So if you're using some of these medications to identify patients with diabetes by the use of medications, then you have to exclude -- I mean, some patients with polycystic ovarian disease have -- a significant number of them have concurrent diabetes, but the issue is that you can't -- if they're being identified by the use of metformin or thiazolidinedione, which are drugs which are usually used to identify patients with diabetes, some of them are being used to treat PCOS.

CO-CHAIR ROSENTHAL: I think we got it. Thank you. I didn't get that. That's an interesting confounder.

DR. WEISS: It's going to occur

more and more in the future because some of these drugs may be used actually to prevent diabetes. There are varying studies that have shown that some of these drugs actually can decrease the risk of getting diabetes by a certain amount. Whether or not they're cost effective is a great concern and there's certainly no uniformity in terms of clinical guideline.

CO-CHAIR ROSENTHAL: But the trigger in almost every one of the diabetes measures that's extant is the pharmacy identification of drugs being prescribed, right?

DR. WEISS: It's a combination of that and then the diabetes-related codes.

CO-CHAIR ROSENTHAL: Right, right. Okay, are there questions? Mary Kay.

DR. O'NEILL: I just would like to say I think it's great that DME is in here on this and that I actually didn't notice if it was missing on the other ones, but it's

significant for chronic management.

CO-CHAIR ROSENTHAL: I had one question. In the culling out of eligible cases, one of the slides has the Market Scan enrollees started off with about 1.4 enrollees with a diabetic indication. By the time the various exclusions are applied, the cohort is down to 212,000. Does that cause the TAP any concern or am I missing something about that?

DR. WEISS: I can't answer this question. I missed it. This is in the actual protocol?

CO-CHAIR ROSENTHAL: No, it's in the slide set that accompanied the measure, Slide 4.

MS. WILBON: Jaime, we'll try to pull it up on the webinar for you.

CO-CHAIR ROSENTHAL: Half of them were because of coverage issues, and I think this came up yesterday in some of the discussions about we're missing a whole chunk of people because of the rule set. And again,

this is not a critique of the developers. You've got to have some rule set and that's not a ridiculous one. But the net effect is if you apparently start with 1.4 million potentials and you get it down to 200,000.

So Kevin, the question I would ask is, and it's the question we asked several times yesterday, which is if you've got 200, approximately 212,000 episodes, how many episodes per physician then did this end up being attributed? Do you know that number for the diabetes one?

DR. WEISS: I'll ask Todd that, but I just also remind the Committee that this is to some degree a feature of this data set as well. We need to be mindful that in a data set that may have more information or more pharmacy coverage, those numbers will dramatically change and so I want to be careful that we're not looking at data sets specific concerns or that we have random or nonspecific biases introduced by trying to

look at this now.

Now Todd, I don't know, do you have the average numbers close at hand?

DR. LEE: I don't. The attributable issue is exactly the same as we have for other measures where about half of our final episodes indicates that it's actually a little under half, end up being attributed at the physician level because of the missingness with provider ID that you all heard about yesterday and talked about.

DR. ROSENZWEIG: Now, looking at this slide, I remember when we reviewed this. I don't have any real big problems with them picking a well-defined cohort and eliminating patients that they may have problems with, even though it is a relatively small percentage of the initial -- but they're talking about people with any diabetes indication in 2006, and they're eliminating people with discontinuous coverage and with -- and a variety of other issues, patients who

hadn't had any visits during the previous year. So I guess this was not of great concern, at least it didn't come up as an issue of great concern as far as I could tell during our discussion, as far as I can recall.

CO-CHAIR ROSENTHAL: All right, thank you. Doris?

DR. PETER: I just had a question about why is it just the first half of 2006. It excludes a huge number of people between the eligible enrollees and the cohort one. Why not the whole year?

DR. WEISS: Is that for us, the developer?

CO-CHAIR ROSENTHAL: Yes. Thanks, Kevin.

DR. WEISS: Because we are trying to focus on a group of people that are not newly diagnosed. If they had a diagnosis in December of the identification year, we didn't want to bring in a cohort of patients that had a new diagnosis. So the definition, was let's

make them have a diagnosis in the first six months of the identification year to ensure that the resource is not going to be a function of trying to manage a patient with a new diagnosis.

CO-CHAIR ROSENTHAL: So you were trying to make it established diabetes, as opposed to new diagnosis diabetes in the index year.

DR. WEISS: That's exactly right.

DR. PETER: Could you look at the prior visit though? Maybe -- you still can't tell because it's the second half. The first half -- to me, it doesn't make sense. You'd have to look at the prior visits still and see if you can see diabetes even before.

DR. WEISS: Well, there could be a new diagnosis in this first half, but then they're going to have six plus months of experience of management of their initial diabetes care that won't be counted as part of the episode because we don't set the clock

until at least six months later or start counting resources until at least six months later. And they have to have a diabetes visit in the follow-up year, so we know that they're continuing to be managed for their diabetes.

CO-CHAIR ROSENTHAL: Thank you for that explanation. Just to clarify though, of the cohort that was very precisely defined, you're not sure, though, then how many per physician that ends up being. Again, I think for the group that's a somewhat relevant concept because if this ends up being five cases per physician attributed, it's hard for me then to know how would that be validly distinguishing one from another, but you don't have that number.

DR. WEISS: I don't have that answer for you.

CO-CHAIR ROSENTHAL: Okay, thank you. Jack?

DR. NEEDLEMAN: Yes, Kevin, quick question on these no prescription drug

coverage, which is the way you've characterized it in the slide set. Is that people with no coverage or people with PBM coverage that was not -- where the claims were not submitted to Thomson Reuters by the primary insurer. Do you know what proportion falls into that category?

DR. WEISS: We don't know what proportion falls into those two buckets, but it's likely that the larger of the two is the latter, where the PBM is not part of what's submitted to Thomson Reuters and contained in the MarketScan data.

CO-CHAIR ROSENTHAL: Thank you. Just again, this illustrates this issue of carve-outs and its impact to identify disease specific and general.

CO-CHAIR STEINWALD: An issue that's come up several times is the issue of exclusions of certain procedures and diagnoses from the assessment of resource use.

Recall yesterday, there was one

measure where they adopted a very inclusive strategy. Once they identified the population, they basically let everything in, knowing that there would be a lot of noise coming in with certain events that occurred that were unrelated to the underlying condition.

But I'm personally more comfortable with that. I'm thinking back to a previous conversation I had with a clinician who said, told me, that if you exclude fractures, for example, you may, because they don't look like they're relevant to the underlying chronic condition, you may actually be missing some important information about the management of the chronic illness because the likelihood of a fall and a fracture may be related to how well that chronic disease is managed.

So I guess when I hear this discussion of why did you exclude this certain diagnosis, it makes me think of this issue all

over again, and my preference would be personally to be much more inclusive than exclusive.

CO-CHAIR ROSENTHAL: Kevin, would you guys want to comment on that in relationship to diabetes of sort of lumping, rolling in, basically, all things that happen to the diabetics and how you guys thought about the question that Bruce just posed?

DR. WEISS: It's a great question and it's one of those that can get easily debated. There's no right answer. I think the issues parse off into the signal to noise. We threw in all these other things, including bumps and lumps and skin tags and twisted ankles and stuff. You just threw in a lot of information that may or may not be relevant, and that creates noise which makes it very hard to detect signal. That's just a technical concern explained very non-technically.

The other is what is a reasonable

expectation for in one year's care of a person with diabetes and in that section of care where they don't have advanced sequelae and they're not newly diagnosed, what can you really attribute to a physician to say what you should be spending money on. And for that, you know, and what should the patient's adverse complications where money is being spent be attributed to that one year of care?

And we heard in the work group a clear recognition. These are great diabetes experts and they were clearly cognizant of all of the relationships that exist over time, and if this was a five or seven year cost-of-care measure, they would probably have looked at it very differently, adding cardiovascular and adding in a huge amount of issues related to end-stage -- end-organ damage that starts to develop. And it's even the five-year cost-of-care measure that would have been difficult for them, but they would have been more comfortable. But over one year, both on the

issue of the technical signal to noise and the second issue related what really in terms of cost can you hold a provider accountable for. It didn't make sense to that work group to go for total cost at this degree of granularity of measurement.

I hope that's helpful in terms of a response.

CO-CHAIR ROSENTHAL: No, I think that was a perfect response. And I think the discussion demonstrates how difficult this whole field is when you capture all of the complexity and trying to get it right in all of the dimensions that we're asking these measures to perform against.

Yes, Jeffrey.

DR. J. RICH: A point of clarification. This slide says cohort 1. There's a cohort 2 on the next slide which even has a much lower number, 4 percent at the final. I think -- I went back to look at the inclusion, and cohort 2, I think, are the

people who are on insulin and not on oral hypoglycemics. So it begs the question, I thought this was a measure for Type 2 diabetes in the introductory remarks. So why is there a cohort that involves insulin only?

DR. ROSENZWEIG: Just to answer that, this is Jaime Rosenzweig. Just a very large percentage of patients with Type 2 diabetes are on insulin. It's 20 or 30 percent.

DR. J. RICH: Without being on oral hypoglycemics?

DR. ROSENZWEIG: No, some of them are on oral hypoglycemics.

DR. J. RICH: Well, if you go to the inclusion criteria in this document, the including criteria says no oral hypoglycemics, but insulin, way back, on some earlier page.

DR. ROSENZWEIG: Yes, but some of them are on insulin alone, so there are a group of patients within that population that are on insulin.

DR. J. RICH: Okay, thanks. And then the second question --

DR. ROSENZWEIG: One thing that did come up is there was a question as to why people who have Type 2 diabetes and are on no -- either oral or insulin medications, why they were not included in this population.

DR. J. RICH: So that comment comes from page 11, top of page 11.

The second question I had related to the risk adjustment model. And this looks very complicated and robust and I don't think I heard anybody from the TAP discuss it. I was wondering if there was some discussion from the TAP or reflection on it. It seems very complicated.

CO-CHAIR ROSENTHAL: Jaime, I think you did mention that briefly in your presentation, but would you mind --

DR. ROSENZWEIG: Yes, I think we did think it was complicated. I thought I had mentioned that.

CO-CHAIR ROSENTHAL: Just reiterate just briefly again, if you would.

DR. ROSENZWEIG: Yes, the enrollment criteria was kind of confusing and they used in one case, as I said, some who were on insulin and some were on oral medications. And there are people exactly as you're mentioning who are on both and there are also people who have neither. And then in addition, there are patients who are on medications that are neither insulin nor oral hypoglycemic medication.

CO-CHAIR ROSENTHAL: Thank you.

DR. ROSENZWEIG: Like the GLP-1 agonists, which are injectable, non-insulin medication. It becomes kind of complicated, and the definitions were not very well clarified.

CO-CHAIR ROSENTHAL: Thank you. Lisa, are you following the rules from the last time where we said if you wanted to speak you turn your thing up? That's amazing. From

one year, she remembered and she's the only person who did that.

(Laughter.)

MS. GRABERT: It's just easier than holding your hand up.

CO-CHAIR ROSENTHAL: I'm trying to be respectful of that. So you're next, and then Bill, and then David.

MS. GRABERT: Are we supposed to comment on price standardization and scientific acceptability? Is that the right category for that?

CO-CHAIR ROSENTHAL: It is. It has not come up yet in the discussion, but if you'd like to.

MS. GRABERT: I just have a question for the developer, since this is specified for both commercial and Medicare data. How did you price standardize, specifically, the Medicare data?

DR. WEISS: We didn't. We only tested this specific measure in a commercial

population and created a standardized price file from that commercial population.

CO-CHAIR ROSENTHAL: Does it say Medicare in there somewhere?

Well, the clarification is, it's not Medicare, so Bill is next, and then David.

DR. W. RICH: I'd like to go back to Jeff's point. I had this same concern and actually I queried the staff to see if the TAP had Carlos' scientific acceptability and had some comments that I'll read. And specifically the risk adjustment methodology, is it described completely inaccurately? No. Six models were tested and the most parsimonious chosen.

Carlos is the technical statistician's evaluation of the -- was it adequately described? The answer is no. In general, the only descriptive process is the threshold of a P value of less than .1 was used for variable selection. And general selection process based only on significant

testing is not reliable, blah, blah, blah. So there were some significant comments, and I had the same concerns that Jeff did about the lack of clarify of the risk-adjustment model.

CO-CHAIR ROSENTHAL: David?

DR. PENSON: So I wanted to add to that and ask the instrument developer a question, and then I wanted to ask the TAP chair a question.

With regard to the risk adjustment, I mean, this is the same thing we've seen with all the ABMS foundation measures, and in the end, the same basic methodology is used, but we run into problems with how extensively it's tested. So I wanted to ask the instrument developer if it's been -- if this has been tested the same as the other ones, if not at all, if it was just the Delphi process and preliminary testing because you know, Bill Rich's comment reading the statistician's comment is important. It seems like there's a problem with the risk

adjustment. So that's the first thing.

The other thing I wanted to ask the TAP was about the accountability piece and I think we've sort of touched on it a little bit, but I'm just -- when I looked at the accountability piece, ascribing it to a provider who had more than 70 percent of the E&Ms associated with diabetes, it just made me a little nervous because you could have a patient that has an acute event one year vis-a-vis has a below-knee amputation or some sort of heart event which is coded with the diabetes.

And suddenly, the cardiac surgeon or the vascular surgeon is now held accountable for the E&M care of the diabetes which he or she had nothing to do with. So it's a two-parter. Sorry, Tom.

CO-CHAIR ROSENTHAL: I would elect to do them separately. Okay, so the first one was?

DR. PENSON: The first one was the

instrument developer with regard to risk adjustment. Is this basically the same as the other measures we've seen? The same very basic testing but still a lot of questions left to be answered?

CO-CHAIR ROSENTHAL: Kevin, that one is for you.

DR. WEISS: If I could -- Todd will take that partly, but I just wanted to note, we did submit a substantial additional information with regards to testing after the TAP meeting. I don't know if you had the chance to receive that and review it. It sounds like that may not have happened.

But with that in mind, I just want to ask Todd if he wanted to say anything additionally?

DR. LEE: Yes, sure. We went through a process, as is described in the submission forms and maybe not extensively enough, a process of asking our work groups what conditions from the HCC list they felt

were important in terms of adjusting diabetes-related costs. We then compared those models to models that were derived via standardized statistical fit, you know, looking at T values.

And I understand the statistician's comments. It's an issue. However, when we compare the performance of those two models, we're stuck in a spot where if we pick the model that's derived wholly from the Delphi process where our clinical work groups selected it, we're going to say your model doesn't fit very well because it doesn't predict the tail.

So we tried to predict or select a model that fit our distribution, wrote the best, and was the most parsimonious. That was our strategy. And as Kevin noted, we've submitted additional documentation about the performance of our models for I think all, maybe not all of the conditions, but the majority of the conditions that you've

reviewed in the last two days.

CO-CHAIR ROSENTHAL: So Ashlie, let's clarify, was the additional information factored in to Carlos' review when he wrote it?

MS. WILBON: Yes, he did -- the information that ABMS sent we did package and send out to the TAP and to the Steering Committee. We also sent it back to Carlos for him to update his original analysis, and I believe, I have to double check. We sent a couple of versions of the analysis out, but I'm 99 percent sure.

CO-CHAIR ROSENTHAL: Can we be sure that the thing that Bill Rich quoted is --

MS. WILBON: I can't right now because I'd have to check emails and stuff.

CO-CHAIR ROSENTHAL: Okay, I would say for the purposes of discussion, let's -- I don't know quite what the right idea is, but let's put -- the reading of Carlos' report

into

-- on hold because I don't think we can be sure that it factored in all of the elements.

DR. LEE: Again, we have Carlos writing that it was inadequate. They respond, here's why it's really okay. Unless Carlos looked at it and said it's okay or not okay, I'm not sure what to do with his analysis. That's all I'm saying.

DR. BARNETT: It says 616 in the file name.

CO-CHAIR ROSENTHAL: Okay, so it appears that it was all factored in. Okay, all right. After, after, after. Okay, thank you. Unfortunately, Carlos not being here, we can't easily clarify.

Now David, your other part and then we have a couple of questions.

DR. PENSON: So my question was to the TAP members because again, it's hard for me, not doing research in this particular condition, whether or not the accountability

technique described is appropriate and, frankly, the word I would use is fair.

CO-CHAIR ROSENTHAL: You're thinking about attribution.

DR. PENSON: Attribution.

DR. ROSENZWEIG: I would say that your concerns -- these were concerns that were raised during the TAP, that there was a sense that maybe this method of attribution might be somewhat arbitrary and then might also actually interact with a question related to risk adjustment because typically in a particular year, you might find a patient would let's say develop a retinal hemorrhage or a vitreous hemorrhage or something like that. And during that particular year, over 70 percent of their visits might be attributable to the ophthalmologist who would see the patient. It would be a high-cost year and a lot of issues related to that.

On another year, in addition, patients who are more severely affected with

diabetes might be on insulin would more likely be seen by an endocrinologist, as opposed to a primary care doc, and therefore might be seen at more frequent intervals and therefore could very well be higher, much higher resource use and also would require a lot more diabetes education.

CO-CHAIR ROSENTHAL: This is why we need multi-specialty group practices because then you could attribute this to the multi-specialty group and who cares which one of them didn't do their job. They all need to do their jobs. But unfortunately, that's not the world we live in.

DR. ROSENZWEIG: And the other issue was that certainly the endocrinologist might be seeing a larger proportion of their patients having diabetes, so they would have much larger ns than a lot of the individual primary care docs who might only see 20 or 30 or 40, maybe 50 patients with diabetes that would be part of this particular group. So

there were a lot of concerns that were raised about this particular system.

The other issue was that the identification -- they used their stratification model, the hierarchical -- the HCC model, hierarchical condition categories model and we -- there was a lot of concern about that, the way they were using it with respect to individual physicians that might not be as successful. And if it was used in much larger groups like the plans which was used in the case of NCQA.

CO-CHAIR ROSENTHAL: Okay, I think we have a few more people who want to pose questions or make comments and maybe we want to start trying to make sure that it's on, perhaps, a new topic because I think we've been round and round on several of these, so hopefully, we won't need to go back. But Paul and then Jack and then Bill.

DR. BARNETT: So I'm also thinking that we have three other measures to get to

before we finish. And if you look, I think there are nine criteria for scientific acceptability and for five of them, this measure didn't get a single high rating. So I think in reality, this is not going to fly and I think we ought to vote on the scientific acceptability. I think that the measure developers have a lot of skill. They've got some great ideas here. It's just not far enough along yet. So you and I were beginning to think alike on this, that we probably heard most of the issues. So if there's something new or --

DR. NEEDLEMAN: In the spirit, I agree with Paul. In the spirit of thinking about next directions, I'm a little concerned about the risk-adjustment model, but not because of it being regression-based.

I'm concerned about your concern about parsimony in your model. That you've got tens of thousands of cases that you're using to project your -- to try to project the

expected costs of care from. You've got lots of degrees of freedom here. I would much prefer to see you do a standardized, get all of the HCC categories into your regression model so we're not picking and choosing which ones we're doing. And the only reason for dropping one would be you have such a small cell in your data that you're going to have trouble fitting that one or there's a clear risk of overfitting.

But treat as your default not going for parsimony, but going for inclusion of these measures as a way of standardizing what you're doing across all your measures and doing it in a way that minimizes the work. Parsimony is not a high value here in terms of getting your risk adjustment right.

CO-CHAIR ROSENTHAL: A free consultation.

DR. WEISS: Can I respond to that? Because that's exactly what we did. I mean that was the second set of six models that we

fit. I mean, we fit the first several based on input from our work groups. The others were data driven. It was all the HCCs. And where we had very, very small sample sizes, we dropped those. So we then selected the model. If they fit similarly, we opted for one that was more parsimonious. If they didn't, that's why you see in our diabetes submission this long list of coexisting conditions that are used in our risk adjustment. It was not the list that the work group told us. It was the methodology that was just described.

CO-CHAIR ROSENTHAL: Okay, well, thank you for that clarification.

Bill, maybe the last comment and then we'll --

DR. W. RICH: Last thing. One of the intents of this type of measure is to identify outliers, and again, as a result of the risk adjustment and the exclusions and the data where you have that tremendous compression issue, it overestimates the

observed by 100 percent at the low end. And underestimates the observed by 60 percent, so the implication of this is that you're not going to be able to differentiate anybody who is not right in the middle, so it doesn't work.

CO-CHAIR ROSENTHAL: Okay, well, I think we've pretty thoroughly covered the various issues. Let's look at the TAP scoring. So again, for me to remember how we did this. We separately will vote -- we're going to vote scientific acceptability, but along the dimensions of the grid of reliability and validity and as the drivers of scientific acceptability, and if you recall the grid, if either one in your mind is ranked low, then it fails. If validity is ranked low, it fails automatically. If reliability -- you've got the grid.

(Laughter.)

I'm sorry. I thought I could do that --

MS. TURBYVILLE: These are high or moderate --

CO-CHAIR ROSENTHAL: I thought I could do that out of my head. I really thought I had that in my head.

You can have moderate to high on reliability and still get it passed. The grid is the same one we used yesterday. But fundamentally, if either one is ranked low, then it doesn't pass scientific acceptability, but here is the top vote on reliability, which was nine, high; seven, medium; and two, low.

Wait --

MS. TURBYVILLE: These are the number of ratings because there are numerous subcriteria. So when you look at the subcriteria, there were nine high ratings on subcriteria for reliability; seven moderate on the subcriteria ratings.

CO-CHAIR ROSENTHAL: Oh, summed up.

MS. TURBYVILLE: Summed up.

CO-CHAIR ROSENTHAL: I couldn't

make it work. On the validity --

MS. TURBYVILLE: I know.

DR. BURSTIN: It's supposed to give you a visual detection of how they all fit together. You can see the size of the bars of high to moderate versus low.

CO-CHAIR ROSENTHAL: I do think though for the sake, and maybe this is worth spending one minute on, because this discussion was, I would say, just trying to broadly weigh it, was, I would say, substantially more negative than the TAP vote.

So maybe Jaime, you've listened to this whole discussion, hopefully, and hopefully have been able to follow it. I know it's difficult at times when you're on the phone. But do you want to make a comment about the discussion you heard here versus your TAP discussion and make, maybe, a final comment or recommendation to the group?

DR. ROSENZWEIG: Are we still on scientific acceptability or have we moved on?

CO-CHAIR ROSENTHAL: Pardon me?

DR. ROSENZWEIG: Are we still on the scientific acceptability section?

CO-CHAIR ROSENTHAL: Yes, yes, yes.

DR. ROSENZWEIG: Our votes weren't all that high. Reliability testing was high, but most of the others were not so good. There were a lot of lows in the specifications consistent with resource use section. Validity testing was all in the medium range. So I don't know.

And then in addition, there was -- the measure set didn't really -- wasn't able to stratify for disparities.

CO-CHAIR ROSENTHAL: I don't think this has to do with inter-related reliabilities.

DR. ROSENZWEIG: I think there's a big disagreement between what I'm hearing here and what we were discussing.

DR. CURTIS: Jaime, I think it has

more to do with how the group discussion in the TAP got rolled up. I don't think the rolling up necessarily is effective for this because our concerns about this individual measure were mainly expressed in 2b1, which is the specification consistent with research use and cost. I mean that was, I think, kind of where we got into a lot of the issues that have been raised.

And so when you roll all those up into one mega vote, I think you lose that. But I don't think anything that's been said here was substantively different than the tenor of the TAP's recommendation.

DR. ROSENZWEIG: Yes, I agree.

DR. W. RICH: The other thing was that Carlos was there verbally. They did not have the advantage of looking at his final statistical analysis that we sought today. So that's why I think some of the other ones -- I was struck also by the --

CO-CHAIR ROSENTHAL: When you look

at that roll out and there is a kind of interrelated reliability question in the sense of what was a medium score in the TAP on this one versus a medium or low score on the other one. So I think it's just --

DR. BURSTIN: Just one comment, again, for initial endorsement, we are only requiring that they demonstrate pilot testing on the data source they have. So it's a little difficult to compare a measure that's been out in use for four and five years and extensive testing to this. I think you have to keep that in context.

The overall ratings of the testing they had done were still moderate or high. I just want to point that out.

CO-CHAIR ROSENTHAL: All right, so I think we've heard plenty on this. Again, I think the time was well spent in really trying to understand this thoroughly, and it's very complex and it would have been a mistake to sort of just gloss over the details because

the details matter. But I think with that and we have the TAP and we've had this very thorough discussion, I think it's time to vote, and on this one, this will be yes and no on scientific acceptability. Right, Ashlie?

(Pause.)

DR. ROSENZWEIG: Hello?

CO-CHAIR ROSENTHAL: We're just tabulating the vote.

DR. ROSENZWEIG: Okay.

CO-CHAIR ROSENTHAL: Five, yes; 13, no. Can we take her word for it that that's what it was?

(Laughter.)

Or is it only valid if it's projected on the wall? Now we all feel comfortable. It is 5, yes; 13 no.

So I think that concludes the discussion on this issue. We won't consider usability, feasibility.

Yes, Paul?

DR. BARNETT: Doesn't this mean

that we no longer consider item 3 or 4 either, right?

CO-CHAIR ROSENTHAL: That's what -- this concludes the discussion on this. Again, I would say a huge vote of thanks to ABMS on their careful consideration of this issue. And hopefully, some of the conversation will provide some ability to make some adjustments to this because again, there's no doubt that trying to be able to figure out how to attribute diabetes at the physician level and the cost issues is an enormously important task and this was an unbelievably good first run at it. And so I wouldn't let the rather intense criticism be a barrier for going forward.

Sally, did you have something you wanted to add on that? Okay, so we have an operational announcement, and then I think it's -- we're going to have a break and then we'll do the Ingenix measure or are we moving forward? We have the diabetes and Ingenix.

Are we scheduled for a break?

(Laughter.)

I'm ready to keep going? How many people want to keep going? No, that's not open for discussion.

Sally, you've got a housekeeping tooling and then we'll take a break.

MS. TURBYVILLE: So for those of you who were charged for the room at the hotel, our meetings folks have contacted the hotel and is requesting that they refund your credit cards, so your statement should show the credit back for the rooms and that the bill should come directly to NQF. So let us know if that does not happen. The hotel should be crediting to whichever credit card you gave them today. Thanks.

CO-CHAIR ROSENTHAL: All right, so a 15-minute break. Thank you.

(Whereupon the meeting recessed from 10:51 a.m. to 11:07 a.m.)

CO-CHAIR ROSENTHAL: All right.

We are ready to deal with Number 5795, ETG-based diabetes resource use measure from Ingenix. And do we have somebody from Ingenix on the phone?

DR. LYNN: Yes. Tom Lynn is here, as well as Jen Pearse and Cheri Zielinski.

CO-CHAIR ROSENTHAL: All right. Thank you very much.

And, Jaime, you're still on the phone for the TAP?

DR. ROSENZWEIG: Yes, I am.

CO-CHAIR ROSENTHAL: All right. Well, terrific. Then, we will go ahead and do this, and we will start with a brief overview from Ingenix, and then we will vote on importance, and then we will get to scientific acceptability in that sequence.

So Ingenix?

DR. LYNN: Thank you. My name is Tom Lynn. I'm a Medical Director working with Ingenix. This rule is based on our ETG methodology. That's their treatment group's

methodology. And it starts with creating an episode of diabetes that's a year long by examining administrative claims and putting claims in diabetes episodes back -- that sit in the episodes.

And then, once the episodes are created, identifying severity of the diabetes using clinical diagnostic-based markers and then evaluating expected costs and observed costs for diabetes based on those -- the different severity level and looking at the observed cost of a physician or a physician group or a health plan compared to the expected cost based on the severity level.

We were asked to respond to a number of issues from the TAP. I'm just going to hit the highlights. One of the concerns was that some of our labels were confusing in that they used "other" but didn't really explain what was underneath that. And we tried to update those labels to make it clear what was included in those categories.

We also included data from our large data set, benchmark data set, that examined the grouping of diabetes and the severity assigned to diabetes in a case where we dropped the fourth diagnosis curve off of the claim, so we compared grouping using all of the four diagnosis codes that we had versus grouping only using the first three diagnosis codes.

We've done smaller sets of that in the past and noted relatively small differences, because there was a question about whether the grouper should set some diagnosis codes or not. And that was concluded in the -- in the work since the TAP met.

In addition, there is a more detailed description of how we take into account members that don't have a pharmacy benefit during a diabetes episode, and basically we stratify those cases, those without pharmacy benefit and those with

pharmacy benefit.

And, finally, we were asked to look at some statistics around how well our severity level works inside of diabetes. We did show some data for total costs as well as the different categories of metrics showing progression across the different severity levels and calculating R squareds for the different measurements, the total cost R squared being 0.22.

That's all I have.

CO-CHAIR ROSENTHAL: All right. Jaime, a quick summary on importance from the TAP.

DR. ROSENZWEIG: Sure. I'd just like to say that, you know, we looked at this measure very carefully, and we recognized the great effort and extent to which the measure developers put into this -- into this effort.

With respect to the importance, obviously, here again, they were able to make a very good case for sufficient support for

the high impact of diabetes in the population and the importance of looking at resource use in this population. So everyone agreed on that.

There was -- the question was whether or not there were some issues related to resource use and problems with cost, and there was a very large discussion, at least in the text that I have.

I don't know if it has been changed by the time that you have had a chance to look at it, but there was a lot of discussion about how in their database there was a lot of variation in resource use and cost between various geographical regions, but there wasn't much discussion of other types of variation, such as socioeconomic differences or the severity of illness and those kinds of things, in this particular group.

And there was -- it was also felt that the -- that the -- that they could have tried to more clearly describe the purpose of

the use of the measures than they did in this particular summary.

They did think -- they did have a very large and extensive resource use list, and that was the -- looked like it was very adequate.

CO-CHAIR ROSENTHAL: All right. Any discussion on importance from the Committee?

(No response.)

Hearing none, I think we should vote. And we have the TAP scores on the screen, and this is one yes, two no.

DR. ROSENZWEIG: I'm not sure I understand that.

CO-CHAIR ROSENTHAL: It's okay.

(Laughter.)

The vote is 18 to -- 18 yes, that this is important, zero no.

So with that, we will move to the scientific acceptability portion of the discussion, and, Jaime, if you would share

with us the TAP review of scientific acceptability.

DR. ROSENZWEIG: Yes. There was -- with respect to the specifications and the precise specifications of the measured properties, there was some disagreement about this.

Five thought they were highly specified, and the rest were either medium or low, and it was basically felt that the specifications of the various comorbidities were not totally clear, and it was especially unclear if the severity of ratings were weighted based upon services of comparable cost. And only costs that are mapped back to the diabetes code were accounted in the episode, so they weren't considering a lot of the other kinds of costs that occur with patients with diabetes.

And then, with respect to reliability testing, it was felt that there was internal consistency and reliability in

this patient population.

With respect to whether or not those specifications were consistent with resource use, and if there was a problem related to the cost, it was unclear in the text as to whether or not diabetes education was included as part of the specifications and whether or not any of the education codes were included, and that obviously would add a certain amount of resource use that is very important.

DR. LYNN: Just to interrupt -- I apologize -- this is Ingenix. There is one thing that they asked us to address that I didn't mention. Between the TAP and the steering committee alert, there is a list of diabetic education procedure codes, and they are eligible to group the diabetes.

DR. ROSENZWEIG: Okay. I just should mention to the Steering Committee that I am looking -- I am basically looking at the document that we were looking at when we had

the TAP meeting. I don't think I have the updated version, as far as I can tell.

So, and validity testing -- it appeared to be that there was some -- a little bit of disagreement on this, but in most cases people felt that validity testing was adequate with the information that was provided. And it was unclear as to -- at least to us in the TAP as to how the exclusions were identified, at least in this protocol.

And then, one of the big issues that did come up was the issue of risk adjustment. And I believe they use -- Ingenix uses a proprietary model of risk adjustment, and it was, at least it was felt by some, that it was kind of a black box, that it may be valid, it has certainly been used by -- in a large patient population already.

But to the individual providers who might be graded on how well -- how, you know, effective, you know, they are in controlling costs, the black box aspect of it

was of concern to a number of people, so that there were a lot of lower scores with respect to the whole issue of risk adjustment, largely because it was not transparent.

And then, identification of statistically meaningful or significant differences. Here again, there was insufficient evidence, at least that was presented to us, that the sample size threshold and analysis at the physician level was meaningful.

They were talking about a 30-sample size as being important to distinguish between different physicians, 30 patients I assume, and it was unclear how they came up with that number or how that would necessarily be adequate to compare individual physicians.

And then, they didn't use multiple data sources. They were using their own data source, so it was felt -- that issue was felt to be non-applicable, and there was also not evidence about whether or not they actually

stratified for disparities.

So I think that pretty much summarizes what we discussed with relationship to this particular section.

CO-CHAIR ROSENTHAL: All right. Thank you very much.

David, your comments on scientific acceptability.

DR. REDFEARN: Yes. The first thing I will say is this definition of diabetes is the standard episode treatment group definition of diabetes that has been in the ETG model out there for years, used pretty widely. So it has a lot of experience in use and practical experience with whether it makes any sense and holds together.

The definition of the episode, I'm not a clinician, and I can't comment on that, but you have to look at it and there is reason -- people vary in terms of what goes into an episode, and particularly what should be a comorbidity complication and add it into that

episode, or what might be something different.

I know the CMS, when they looked at ETGs -- the Medstat MEG Program a while ago -- they differed in terms of what the two methods captured into diabetes. So you have to look at that clinically, I think, and make some sense about whether you think it does make sense. It is a definition.

The only other comment I'll make is about the risk adjustment methodology. This is a little different from what all of the other measures have been doing. This is a risk adjustment specific to that episode of diabetes. It is not a global patient risk characteristic.

So they build it -- if you look at the methodology, they build it -- they build levels -- four levels inside the diabetes episode based on the specifics of the risk associated with that definition of diabetes, which is a little different. And you can argue that you might want to use global

because that captures global risk, but this has some -- this is designed specifically to be used in that measure, in that way.

DR. ROSENZWEIG: One aspect of the issue of the diabetes-related episode of care was that, at least the way it was explained to us at the time of the TAP meeting was that this diabetes is a chronic disease and really reflects overall care during the course of this particular period.

It is hard often to define specific episodes of care within a 12-month period, so that for the vast majority of these patients the actual episode is the full 12 months.

Perhaps a measure developer would want to comment on that, but it appeared to -- at least it appeared to -- that was the way it was explained to us, that it was a -- even though it -- these costs are concentrated around what is called an episode of care, it just seems like it was reflecting a full 12-

month period.

And I think some questions were raised as to what happens if the patient, you know, enters the plan in the middle of the 12-month period or if -- you know, if it hasn't been seen for a while and then starts being seen later. Am I explaining this correctly?

CO-CHAIR ROSENTHAL: Yes, I think so.

DR. LYNN: If I can comment on that, we do divide -- diabetes is chronic. We do divide it up into year-long episodes. However, if the member is not eligible during that entire year, then the episode is marked as incomplete, and the method that we used is not the method everybody uses.

But the method that we presented excludes those episodes where the member wasn't eligible for the entire year. There are some methods where you make an adjustment for the partial year, but that's not the method we presented.

In addition, I wanted to make one comment about the minimum 30. We do set a minimum of 30, but we try to make the point, especially in response to the TAP that was provided since the TAP meeting, that what really matters is that you only -- that you only show the statistically significant differences and that you measure that.

And the method we presented does measure that, which is a requirement of PHQ 2008 when you are measuring for resource use. The minimum number was once used as a proxy for statistical significance, and in that case it is really important that you get that right, and it may be sort of impossible to get it right.

But in the case where you are actually measuring statistical significance, that minimum number is not as critical. Our opinion.

CO-CHAIR ROSENTHAL: Bill.

DR. WILLIAM RICH: A couple of

questions for the developer. Is this risk-adjusted methodology that you outline different than the one that was up on your website about two years ago?

And, secondly, what was the total -- to get to 30 per provider, what we have seen in others -- other groupers that have come before us is that problems with the physician ID dramatically decreased the eligible number of assignments to a physician. What was the total -- what percentage -- if you take 30 by your number of total physicians, what percentage of that is the total number of claims and physicians that you evaluated? Does that make sense? Probably didn't verbalize --

CO-CHAIR ROSENTHAL: Tom, I think that was addressed to you guys.

DR. LYNN: I'm sorry. I was on mute. I think it makes sense. I'm not sure I know. Were you -- are you asking if we looked at diabetes and measured a physician

and found that that physician had more than 30 cases, what percentage of the dollars for that physician would we have captured on average?

DR. WILLIAM RICH: What percentage of the claims were you able to identify at the physician level?

DR. LYNN: Well, okay, so the question is --

DR. WILLIAM RICH: And the first part -- I'm sorry you were on mute was -- is this risk adjustment methodology different than the risk adjusted methodology that you had up on the -- on your web page about 18, 24 months ago?

DR. LYNN: No, it's the same. The second question is: are we -- since the data that we use -- maybe I'm being dense, I'm sorry, but the data that we're using, are we able to successfully identify which physician was -- which physician was responsible for the claim?

DR. WILLIAM RICH: Yes. And the

third part is, where did you get the number 30? I mean, 30 is a statistically significant sample for surveys and things like that, and then people that have published in this area, like Dr. Thomas, Bill Thomas, says 100. How did you get to the number of 30 and decide that that was statistically significant?

DR. LYNN: The first question, you know, we -- we certainly have challenges with matching physicians, but, you know, we are -- we only sort of bring in -- use data where those challenges have been sufficiently resolved amongst identifying physicians across.

Our data doesn't, in fact, include data for multiple health plans, but we do the best we can to make sure that we have valid physician IDs that work across the health plans. If we don't, then we don't include that in the data that we use. That's -- I think that answers the first question.

The second question is about the

number of cases, and, you know, honestly, the number of cases -- 30 -- comes from one of the original NCQA documents about doing resource utilization required a minimum of 30 cases.

But, again, you know, I think the important thing is that you -- when you're doing measurement that you identify statistically significant differences from an expected benchmark, which is usually the average across the peer group, and only report statistically significant differences. And if the differences are not statistically significant, basically, say that you can't tell a difference.

And we think that it's more important to use a valid statistical method, which is a method that's -- we use a method that has been published and used by RAND in some of the work they do. And then, the number of cases is not as important, because if the number of cases is too small, then you won't have cases that are statistically

significant.

CO-CHAIR ROSENTHAL: So to that point, Tom, with the attribution model and the 30 cases, what percentage were either higher than expected or lower than expected?

DR. LYNN: Oh, that's a good question. I don't have that off the top of my head. I would -- we would have to go back into the data and look at that.

MS. TURBYVILLE: This is Sally.

DR. LYNN: If you looked at the number of doctors that had some range of cases, what percentage of them would be statistically significant or different? I -- we would have to go back and do that work.

MS. TURBYVILLE: Just for point of clarification -- it may or may not change the request for that information -- similar to what the HealthPartners measure, remember that for sample size, because we knew this might vary, we allowed it be a guideline.

And in their submission, they do

state that a valid statistical test is preferable, and then -- and we can -- we pulled it up for you -- and then they talk about 30, but that actually, as Tom said, what's more important is that you can demonstrate to statistical differences.

CO-CHAIR ROSENTHAL: Right. I think I'm saying the same thing.

MS. TURBYVILLE: Okay.

CO-CHAIR ROSENTHAL: And, therefore, it would be really interesting to know in this cohort what is statistically significant and what fell out above or below based on the 30. So, okay, but they don't -- they don't have --

DR. LYNN: Again, we could say that we'd provide that to you. We just -- I don't have it. I'd have to go back and calculate it.

CO-CHAIR ROSENTHAL: Okay.

DR. REDFEARN: I can comment about -- we use the same methodology, using episode

data, across all types of episodes when we do provider profiling. And we do confidence intervals around the observed-to-expected ratio, and you get about 50 percent of the docs that fall into the middle "don't know" category. That is, there is no statistically significant difference, but about somewhere around 25 percent efficient, 25 percent inefficient. That's rough, and it varies across a lot of episodes.

But if you do confidence intervals, you get a huge "don't know" category. You get a lot of cases in which you can't make a determination that the doctor is efficient or inefficient, costs are higher than expected or lower than expected.

And that -- to reinforce what Tom was saying about sample size, you see that for doctors in which we have assigned 300 episodes to the doctor. Even in that case, there is so much variability, we say we can't say it, but we do assign efficiency to doctors that have

had 10 or 15 episodes, way below the 30, because they are absolutely rock consistent in terms of how they perform, either high cost or low cost.

So you always have that big category, but you can make this determination with this data.

CO-CHAIR ROSENTHAL: But it clearly does -- sample size matters, and there is no doubt that with a small number, if somebody has a gigantic Six Sigma outlier, that it will be statistically significant. But generally, the smaller number of attributable cases, virtually everybody falls into an indistinguishable one. That's certainly true of the transplant data, where we have that for years.

The one that --

DR. LYNN: Yes, I think that's absolutely right. I think, you know, what Dave just said, that the percentage -- the detail depends upon what you said is your

minimum number of cases.

But what -- to me, the important thing is making sure that you are doing some sort of test to make sure even someone with 100 cases, the difference is statistically significant, or even, as Dave pointed out, 300.

CO-CHAIR ROSENTHAL: Absolutely. No doubt.

The one other area that I don't think we have touched on at all is the attribution model here went on for a page and a half and was rather complex. Perhaps could we -- could you discuss your thinking about that? And, well, let's just leave it that open ended for the moment. Tom, could you guys sort of talk about your attribution methodology?

DR. LYNN: Yes. I think in the case of diabetes, the attribution methodology -- and if I said something wrong, it may have been because I believe we presented

alternatives that the attribution methodology is to identify counts of contacts between physicians and members for diabetes that were grouped in this diabetes episode.

And then we assigned that episode to the physician that has the most number of contacts -- the highest number of contacts as long as the number of contacts is greater than 30 percent of the total number of contacts.

So even if you have a provider who has -- let's say if you had a member with 10 contacts for the episode of diabetes, and they were all assigned to different doctors except for two of them, then that doctor would not be assigned the diabetes. Nobody would be assigned to diabetes because no doctor met the 30 percent threshold for attribution.

Then, I think a lot of the complexity may come from, you know, what happens when you -- when there is a tie. When there is a tie, then you use the cost as a tiebreaker. And there may be a third

tiebreaker, but I'm not -- I don't have it right in front of me. I could pull it up here, if someone would remind me what section it was in. Is it in S8?

CO-CHAIR ROSENTHAL: It's in Section 11. And just to clarify, though, you are proposing options for attribution. Did I get that correct? So an entity that would want to use this measure could pick one of the attribution methodologies and apply it?

DR. LYNN: Right. But the one I describe is the one that was used in this analysis.

DR. REDFEARN: Yes. The -- my notes indicate that they sort of leave it up to the user. There is a lot of different ways you can do this. They have a suggested one, but there's lots of different ways of doing it.

CO-CHAIR ROSENTHAL: Well, assuming a different -- I mean, we have struggled with the attribution on these

specific measures, you know, diagnosis-specific measures. And we have been critical of the people who sort of narrowly picked one.

It might be internally inconsistent to be, then, critical of somebody who says, "Pick whatever one you want," but to pick one -- "whatever one you want" appears to have been the alternative strategy here to the really challenging difficulties in, particularly, diabetes that we talked about in the last one of -- does the primary care doc get it? Does the endocrinologist get it? Does the surgeon get it who happens to do the amputation, or the ophthalmologist who ends up with the eyes?

And, I don't know, does anybody have a comment on sort of the fact that they have taken the opposite approach to this and whether that's a better way to do it, or not a better way to do it?

DR. STEPHANSKY: I much prefer the flexibility. In practice in Michigan, we have

been dealing with a lot of different health payers and a lot of different attribution models, and I guess the best way to describe it is in the meetings it gets very contentious. And while there has never been an actual murder in one of the meetings --

(Laughter.)

-- the homicidal ideation is so high that the --

(Laughter.)

-- so I think we are much better off being as flexible as we can with reasonable attribution models, leaving it up to how it's going to get used for a local community or a region.

DR. LYNN: There is an attribution method. That varies depending on what you're using it for.

DR. STEPHANSKY: Right.

DR. LYNN: Absolutely.

DR. NEEDLEMAN: Got a question because one of the bases for attribution

you've got are, you know, who is the assigned PCP under the plan? You know, who somebody picked. And I'm just wondering, when you look at that, have you seen -- have you looked how consistent the PCP that is formally assigned is -- how close -- how often that matches who you wind up attributing the diabetes care to? So have you done any cross-checking of what the different attribution models --

DR. LYNN: No, that -- that particular attribution method, you know, would be used in a case where that primary care provider was acting as a gatekeeper, unless I'm not sure it would be the best one.

DR. NEEDLEMAN: But your data should be able -- would answer the question posed.

DR. LYNN: Yes, we could go back and say, "Of the physicians that were identified as primary care, and the members that had diabetes for those physicians, what percentage of the time did the primary care

physician get attributed to a diabetes episode?" We could answer that question.

CO-CHAIR ROSENTHAL: All right. But you didn't --

DR. LYNN: I don't have the answer to that question, but we --

CO-CHAIR ROSENTHAL: -- you didn't -- don't have it, okay. So the answer is don't know.

DR. LYNN: Right.

CO-CHAIR ROSENTHAL: Paul?

DR. BARNETT: I have a question about a different part of it. Is that all right -- okay?

CO-CHAIR ROSENTHAL: Absolutely.

DR. BARNETT: So trying to think about how this differs from the NCQA diabetes measure, which takes all costs, so the costs here, as I understand it, are attributable to the episode which, in the case of a chronic disease, is the entire year. But it's things pertaining to diabetes visits.

And so what I was wondering about is, so if somebody has diabetes and they get, you know, ischemic heart disease and they become a CHF patient, or they become -- you know, they get eye problems and become an ophthalmology -- that's an ophthalmology episode, and then that cost is not attributable to diabetes anymore, or is it -- it's end stage renal disease, and that cost is no longer attributable to diabetes anymore, is that right?

DR. LYNN: Yes, that's a good question, and that's a good description. You know, this episode is part of, obviously, an application that groups claims to all sorts of diseases, not just diabetes, and this particular thing that we hold out looks at the cost of the direct treatment of diabetes and not the cost of the sequelae of diabetes.

We know that's important, but inside of our applications it's easier to pull things -- it's easier to put things together

than it is to pull things apart.

I mean, another approach, which we did not present here, would be to say let's look at -- especially when you're measuring, say, a system for a primary care doctor that maybe has the responsibility, let's look at the entire -- let's look at all of the episodes for diabetes and its sequelae together.

We call that an ETG family. And we have a method for doing that and an opinion about what -- if these should be grouped together to do that sort of thing. But if you're measuring the ophthalmologist who is taking care of the diabetic retinopathy, most of our customers would want that to be pulled out separate.

So the application pulls them out separately, and, you know, basically the philosophy there is it's easier to put things together than to take them apart.

DR. BARNETT: So the retinopathy

screening, would that be a different episode, or is that part of diabetes care?

DR. LYNN: The retinopathy screening would be part of -- it would be part of diabetes care if the diagnosis was diabetes. If they had some diabetic retinopathy, then it would go into a diabetic retinopathy episode.

DR. BARNETT: That makes sense, and so then the -- I guess the big concern is is that the -- how do you tease out diabetes from, say, coronary artery disease? Because those are really, you know, linked together. It's kind of a chicken and egg thing, I think, isn't it? So I'm just --

DR. LYNN: I think the diabetes probably comes first, but that doesn't really matter. Yes, I mean, it is challenging to tease it out, and, you know, we try to -- we look at the diagnosis code, we look at the procedure code and how -- and the clinical information contained in the procedure code to

help us make the determination, and we have extensive tiebreaker logic, which, you know, honestly is a little bit hard to read to try to figure out what is the best place for a claim to go that could be eligible for different episodes. But I -- you know, I think we do as good a job as you can do given the limitations of the claims data.

You know, the point being made, which I think is a valid one, is that, you know, in some situations don't even try -- just look at how much it costs to take care of the patient or limit some of that choice of which episode it goes to by using a more expensive ETG family. And, you know, we certainly agree that there are times when that may be the better approach.

CO-CHAIR ROSENTHAL: Let me ask one last quick question about the risk adjusting. I hate to go back to that, but it wasn't completely clear to me. If somebody has a four-vessel CABG and diabetes in that

year, do they get risk adjusted differently than a normal person who only has a little bit of hyperglycemia? Is that accounted for in the risk adjustment?

DR. LYNN: Right. So there's two things -- there's two parts to that answer. The first part is the coronary bypass graft surgery and the coronary artery disease that was directly responsible for that surgery, would be captured in a separate episode. It would not be part of this episode.

After that occurs, then the fact that the person did have coronary artery disease at the same time as diabetes is taken into account in building the severity model for diabetes, recognizing -- and the models do recognize because it's mathematics, that the cost that is increasing cost for diabetes is indirect. It's a direct cost of coronary artery disease, and the CABG is captured by a separate episode.

CO-CHAIR ROSENTHAL: Okay. I've

got it now.

DR. LYNN: It did not --

CO-CHAIR ROSENTHAL: I've got it.

DR. LYNN: It did not look at whether there was a CABG or not.

CO-CHAIR ROSENTHAL: Okay. I got it, I got it. I got it. The answer was yes. The answer was yes.

DR. LYNN: All right. I'll try to limit my answers. Sorry about that.

(Laughter.)

CO-CHAIR ROSENTHAL: No, I appreciate it. You have the unfortunate thing of not being able to read the body language, and so -- Jeptha.

DR. CURTIS: I understand that approach, and it makes sense on one hand. I just want to raise the issue that down the road, if we're trying to match these resource use measures with quality measures to get to value, this is leading to a potential paradox where you will have quality measures that are

specifically set up that you risk adjust on things that are present before the estimation of quality, and you have resource use measures that are adjusting for things that could potentially be complications or consequences of care.

And so I'm not sure how we should involve that at this stage, but I can definitely see that becoming a major issue down the road.

CO-CHAIR ROSENTHAL: Okay. Thank you. Jack, I think you had your hand up.

DR. NEEDLEMAN: Yes.

CO-CHAIR ROSENTHAL: And then, we'll perhaps try to sort of maybe bring the scientific part of this to a vote.

DR. NEEDLEMAN: I've got two questions for the developer. One is, you know, you talked about a family of diabetes ETGs, and I'm just wondering, can you explain what you see as the scope of what this ETG is trying to measure and where it fits into the

family of other diabetes-related ETGs that you also have that we're not looking at as specific measures? So that's question one.

DR. LYNN: This is -- this episode captures the direct cost of diabetes. If there is a complication of diabetes, that -- you know, that is basically a disease in and of itself, it is captured in a separate episode.

Our concept of the diabetes family is not a separate episode, but a way to combine multiple episodes to come up with one cost -- you know, the cost of -- the cost of the diabetes, coronary artery disease, the congestive heart failure, the renal failure, are some examples of what we included in the diabetes family.

So it would be calculated by basically summing up costs in the separate episodes for diabetes as well as complications of diabetes.

CO-CHAIR ROSENTHAL: And your

other question, Jack?

DR. NEEDLEMAN: And my second question -- you talked about dealing with the pharmacy carve-outs that exist in much of your data by basically stratifying your cost analysis for patients where you have that data and where you don't.

There are also carve-outs in behavioral health. And if you were looking at the direct cost of diabetes, obviously, depression as a diabetes-related comorbidity is not in the cost -- the cost of this. But I'm wondering if it's part of your risk adjuster and -- because somewhere you talk about psychosis as part of your risk adjuster.

And how are you dealing with behavioral health carve-outs for your risk adjusters, to the extent that they include mental health services or mental health conditions?

DR. LYNN: Yes, we do have -- you're right, we do have severity markers that

are comorbid that are based on mental health issues. And we do not stratify based on mental health -- whether mental health is a carve-out. So we don't deal with that probably the way you want it, you know, dealt with.

You know, hopefully that is mitigated by the possibility that some of these diagnoses may be included in medical claims, because they are relevant to the treatment.

CO-CHAIR ROSENTHAL: All right.

DR. LYNN: So that's the answer to the question.

CO-CHAIR ROSENTHAL: Okay. Thank you. I think we have pretty well run the gamut of the issues around this. Jaime, maybe I'd like to give you, on behalf of the TAP, kind of the last word.

And in turn, for having participated in this conversation -- and we've got your -- the TAP scores up in terms of

reliability and validity -- and apropos of sort of the last observation -- the last review of this where -- give us your summary of what you think these numbers mean on the screen. Can you see them?

DR. ROSENZWEIG: Well, you know, I think there was -- there is certainly internal consistency that was demonstrated, but the other validity measures were highly debated, and there was a lot of variability in the scores between the various members of the TAP.

In general, I think the -- one of the bigger issues was this proprietary nature of their risk adjustment score, and I think there was sort of a difference between the clinicians on our Committee and the people who were more in tuned with health plans with respect to their ability to trust the data with respect to that particular aspect, and that's why there was kind of a mixed review.

CO-CHAIR ROSENTHAL: All right. Thank you. I was --

DR. CURTIS: Can I follow up on that?

CO-CHAIR ROSENTHAL: Please.

DR. CURTIS: These ratings broken down into the individual elements show that the lows in this case are different than for the last measure, which was more on specifications. This is more about the risk adjustment than the identification, and that, as I recall, was directly our concern, that you were adjusting for things that were happening during the measurement year, as well as the difficulty of attributions.

DR. ROSENZWEIG: Correct.

CO-CHAIR ROSENTHAL: I think with that clarification, does the group feel prepared to make a judgment on scientific acceptability? It appears so. This is one yes, two no.

MS. TURBYVILLE: We have 10 yes and eight no.

CO-CHAIR ROSENTHAL: All right.

So we will move now into -- that at least gives us the opportunity to discuss usability and feasibility, and so, Jaime, would you give us a TAP rendition of usability?

DR. ROSENZWEIG: Yes. With respect to the usability issue, the measured performance results are already publicly reported, but the usability information that was submitted was the same -- was really not specific to diabetes, but really for all of the Ingenix measures.

And there was some concern in the TAP that -- about the -- with the availability of this data to the public and requested -- and we requested clarification from NQF as to what would be required for public reporting.

So I think this is an issue that came up, and, as a result, with respect to 3A, most of the people felt that the data was insufficient.

And there was also some -- here again, on 3B, the usability information

submitted was not specific to diabetes and for Ingenix measures. It was for all Ingenix measures.

The usability -- and it was felt that diabetes presented specific problems that had to be addressed. So we had some concerns about this here, and that applied to 3C as well.

There was also some -- felt that there was -- that it was difficult to assess the extent to which this particular -- the individual measures could be evaluated. And then, I think basically the major issue was that the whole section here that was put in place with respect to usability was fairly generic and could apply to a whole variety of different measures other than diabetes. I'm repeating myself.

CO-CHAIR ROSENTHAL: All right. Thank you. That's very thorough. Jack, you're our rep on this.

MR. BOWHAN: Yes. I mean, it's in

the notes and the description, and I wouldn't have anything to add. But these are all complex measures, and I don't know that -- with any of the groups that it's any easier to figure these out than this one. But it is complex.

CO-CHAIR ROSENTHAL: Do you want to comment on the 3C?

MR. BOWHAN: Other than knowing that -- you know, trying to decipher down to the level of figuring out where you fit in and how you got to your rating, you know, it's complex.

CO-CHAIR ROSENTHAL: But I think the issue -- what I'm hearing from the TAP is more than complex. Is this -- or maybe I'm misunderstanding it, but is this not the place where the fact that the methodology is not at all transparent is the issue? This is the black box issue. I mean, or am I missing it? I'm hearing some yeses so, but if somebody wants to --

DR. BARNETT: Yes. So one important -- another important way that this differs from the NCQA diabetes measures, I looked at the NCQA thing, and I was thinking as -- I could have one of the programmers in my center do this. I could read it, and I could have them do it.

And so it's partly the complexity issue, yes, that it's simpler. But, you know, we don't -- so I was just looking at the submission. I don't see any further documentation that I can go to, unless I have overlooked something in there, that explains, you know, all the codes.

So I think that their -- what they have done is great. The issue is, if we are going to start judging plans and providers on it, we really have to understand exactly how it is constructed. And it's a dilemma because they can't give this -- build it and then have something so complicated, and then give it away. So I'm -- it is a dilemma.

MS. ZIELINSKI: This is Cheri Zielinski. I have a couple of salient points I think would help this discussion.

Number one, due to the fact that the diabetes measures for usability are cited for all of the other measures that we have -- due to the fact that, you know, we way we package our software -- ETG -- those measures are grouped together, you know, and can be recorded publicly. You know, so several chronic conditions can be reported on publicly -- and are reported on publicly by our clients and users.

And so we envision that to be widespread usability using all of our measures, which we feel is an advantage.

And then, secondly, in terms of the black box technology, we do have a website -- it's ingenix.com/transparency -- that people who -- people who have questions about how the episodes are constructed who want to know how the coding -- how the coding maps

and, you know, what codes are included in diabetes, ETGs, and so on.

Anybody that is open to the public do not need proprietary measures. You don't have to have a license in order to see our -- how our codes are constructed and how our episodes are constructed, so that's something that can be accessible to the public as well.

CO-CHAIR ROSENTHAL: Okay. Other discussion, then, on the usability criteria?

(No response.)

Hearing none, I am assuming that that means the group is ready to vote.

MR. PHILLIPS: Just to -- I mean, I guess a follow up --

CO-CHAIR ROSENTHAL: I thought so.

MR. PHILLIPS: -- a follow up on the point about the transparency website. And so, I mean, is the point, then, that a provider could work back using this information to decipher its score?

MS. ZIELINSKI: That was exactly

why we constructed that website, because people who are being measured with these tools need to understand the measurement being used. And so it's primarily -- well, it's for anybody who is interested in the construct of the episodes, but, yes, it's especially for providers who are being measured.

CO-CHAIR ROSENTHAL: Bill?

DR. WILLIAM RICH: Yes. In reality, a physician first has to get their report from the -- whoever put the report together using the Ingenix. They have to get the data. Then, they can go to the website. To look at one measure, if you will, takes about six to eight hours, but it is there, and you can go through and see how, as an ophthalmologist, I was assigned urograms and things like that, but you can go through everything and actually map it out.

CO-CHAIR ROSENTHAL: Is it an issue of usability that, in point of fact, the only people who could use this would be ones

who would hire Ingenix? Is that a usability issue? You don't think so?

DR. WILLIAM RICH: More feasibility.

CO-CHAIR ROSENTHAL: Well, no, I'm not talking about the money, that it is not feasible. I mean, the only people that are going to have -- be able to use this tool are going to be health plans that would engage them. Is that an issue about usability? If --

DR. RUDOLPH: Well, it could be others in the health plans, right? I mean, I'm thinking, you know, say a large employer with all of their own claims or it could be, you know, some other group -- a state, many of the states --

CO-CHAIR ROSENTHAL: Could also engage them on --

DR. RUDOLPH: -- who have health data, which is --

CO-CHAIR ROSENTHAL: Okay. All

right.

DR. RUDOLPH: -- just like all payer claims data, states could use this measure. So --

MS. YANAGIHARA: Big medical groups in California license the product.

MS. ZIELINSKI: Well, we have several provider organizations, NCOs, large employer groups, state Medicaid programs, all using it.

CO-CHAIR ROSENTHAL: Mary Kay?

DR. O'NEILL: Without playing an economist for a second, though, if it does take as many hours as Bill described to figure this out, that is actually a cost.

CO-CHAIR ROSENTHAL: If there is no other point -- Jack, last point.

DR. NEEDLEMAN: A number of the folks that we have been -- were using some variation of the HCC weighting system, that's documented, and the sources of it are documented. You've got your own comorbidities

that you are including, your own other weighting factors that you are including, and your weights.

Can you just briefly tell us a little bit about how those were developed, and what kinds of analysis went into them? I'm sure it's documented on your website, but I'd just like the slightly -- a long elevator ride explanation of where that is and how it contrasts to the HCC kind of development, which looks very similar in terms of the weighting up of each individual.

DR. LYNN: Yes, I think -- I'll try to keep it to the long elevator ride explanation. I think there are two components to it. One is the -- what is going on outside the episode that has an indirect cost that we capture, and that is similar to the HCC model, you know, which tells you what comorbidities the member has.

We felt like it was important, and I think some of our colleagues that have

presented felt like it was important, that you see what the specific comorbidity rate was for diabetes as opposed to taking some overall disease burden quote from an HCC system, using the individual markers. I think our colleagues have used the individual markers.

We use markers that are similar, and I think the HCC models that are used, and our models that are used, are similar. But that's looking at what's outside. We also looked at what is inside the episode for -- you know, that might explain costs, the clinical diagnosis, as opposed to procedures and use those markers as well.

So I think we felt like it was important to have a specific diabetes model that the comorbidity had specific effect on the severity of diabetes that you don't use on some sort of measure of overall disease burden, because maybe migraine doesn't has much of an effect on diabetes, and they -- you know, COPD, for example. But they might have

similar increases in the disease burden.

So I think it was the specificity built in a specific model for diabetes and looking outside for the indirect effects of cost and inside for the more direct effect of cost, and making sure it was all diagnosed.

The way they were developed -- you know, a lot of our comorbidities jive with our definitions of other diseases that we looked at. That worked well with what is happening outside of the episode.

And then, we modeled it. We used that large database that we have been working with, although it's a version a couple years old -- older, but -- and we modeled those markers, looked at sort of a -- cast a wide net clinically about what would be a marker and what wouldn't be a marker, looked at what effect those markers had, ran the model, looked at what is statistically significant, what wasn't, you know, and adjusted the model until we felt like we had the best marker for

diabetes.

And, of course, all of this is, you know, done as much mathematically as it is clinically. We tried to use the two together.

And that's sort of the long elevator ride explanation about how the models were developed.

DR. NEEDLEMAN: Thank you.

DR. LYNN: Diabetes models.

DR. NEEDLEMAN: We got to the 100th floor.

(Laughter.)

DR. LYNN: Yes. I'll stop.

DR. NEEDLEMAN: Thank you.

CO-CHAIR ROSENTHAL: All right. Okay. I've lost all train of thought now about usability. Hold on. Now, focus, focus.

I think we are ready. I think we have dealt with the various issues around usability, and I get the sense the group is ready to weigh in on this.

And if we recall, let's -- can we

look at the TAP scores on this? And then, this one will be high, moderate, low, and insufficient. All right. So we've got 3A, B, and C. Here D is N/A.

And, Jaime, do you want to just very quickly review -- just get the last word in on this?

DR. ROSENZWEIG: Yes. Here again, I think with respect to usability, the issue was that it wasn't specific to diabetes, that the data that was presented -- it might be usable for overall costs and other disease states, but there wasn't any data of their diabetes -- people of their subgroup within -- that they have already looked at with respect to diabetes.

CO-CHAIR ROSENTHAL: In terms of what has been publicly reported.

DR. ROSENZWEIG: And we are -- with respect to what has been publicly reported, yes.

CO-CHAIR ROSENTHAL: Yes.

DR. ROSENZWEIG: And there was a certain amount of concern about the fact that it was because of the lack of attribution -- if you go back to where we discussed the attribution, there were so many different options for attribution that physicians who would be judged by this might be judged compared to other physicians with respect to their resource use, whereas the -- with respect to the overall picture they may be saving money, keeping people out of a hospital, even though they were using more resources.

CO-CHAIR ROSENTHAL: Okay. And then --

DR. ROSENZWEIG: Does that make that sense?

CO-CHAIR ROSENTHAL: Yes, absolutely. And then --

DR. ROSENZWEIG: It's a big issue in diabetes, you know, obviously, because the proportion of actual resources that are

actually related to outpatient provider use is actually relatively small compared to the entire resource use picture.

CO-CHAIR ROSENTHAL: And then, 3C is the decomposition, the ability to decompose the data had some negative votes.

DR. ROSENZWEIG: Yes, yes. The TAP thought it was difficult to assess the extent to which the measure could be decomposed --

CO-CHAIR ROSENTHAL: Okay.

DR. ROSENZWEIG: -- as currently specified.

CO-CHAIR ROSENTHAL: I think this is very helpful, and the discussion on this point has been good. And I think we are ready to vote. It's 1 through 4, then, on this one with high, moderate, low, and insufficient.

MS. TURBYVILLE: So we have nine moderate, six low, and three insufficient.

CO-CHAIR ROSENTHAL: All right. So we will move on to feasibility. So, Jaime,

the TAP. Oh, yes, I'm sorry. Helen goes first on this one.

DR. BURSTIN: Yes. So I just want to have a couple of minutes to talk to the Steering Committee about this particular issue for feasibility. So as you saw earlier, when we considered the HealthPartners measure, we -- you have the ability to look at the fee schedule for the ACGs as part of that data.

To date, we still have incomplete information from Ingenix. We have not yet received the fee schedule. So at this point, you actually can't assess feasibility, so I think at this point we're going to -- we are having some continued ongoing discussions with Ingenix, but I think at this point we are going to table feasibility. You also can't make an overall assessment of the measure, because you won't have feasibility, won't have the benefit of looking at that.

It is important to note this is clearly, as part of the policy the Board

approved a couple of years ago of proprietary measures, the fact that they thought it was important that the Committee and the end users have a chance to see the fees involved and have that incorporated into feasibility.

Clearly, we can't -- you know, the measures may score well, as they have sort of done moderately on many of these other criteria, but if they are not feasible they can't move forward.

So at this point, we really just need to -- we will work with Ingenix to continue to get that information to share with you, and then we will continue at a later date on feasibility.

CO-CHAIR ROSENTHAL: Do we have some way of capturing the key points of this, so that we don't have to repeat the entire exercise when we are finally able to have feasibility, and then have the overall vote? Because otherwise we will have wasted an hour.

MS. TURBYVILLE: We're taking

meeting notes right now and summarizing the discussions and key points. And you voted on the first three, so hopefully that will be sufficient for all of you. You know --

DR. WILLIAM RICH: I would venture to say, Tom, we can't, because there's other issues in feasibility about adverse consequences of the reporting that we haven't addressed. So we are going to table this.

I would be unable to vote until we -

CO-CHAIR ROSENTHAL: I was not suggesting a vote today. I was suggesting a methodology by which we could have some of the discussion crisply summarized, so that we don't have to repeat it all --

DR. WILLIAM RICH: I'm sorry.

CO-CHAIR ROSENTHAL: -- when we do that thing. I --

DR. WILLIAM RICH: I thought you were talking of calling for a consensus.

CO-CHAIR ROSENTHAL: No, no, no,

no, no, no. I understand the absolute constraint on our freedom on this one.

MS. WILBON: Tom, when we are ready to bring it back to the Committee, we will provide you guys with a summary of what your votes were, what the key points were for the previous three prior criteria, so that you have an idea of where the discussion was --

CO-CHAIR ROSENTHAL: Okay.

MS. WILBON: -- before that.

CO-CHAIR ROSENTHAL: And then, as a point of order, the same will apply then to the other Ingenix thing, which we were supposed to do usability and feasibility wrap up, and which in fact we are going to have to rediscuss a little bit of the scientific thing, because we didn't really complete all the votes on that, and so we'll need a short conversation. But it's clear we will not reach a final decision on it either.

I assume people are beginning, though, to direct themselves in their heads.

And to the extent that you can do that so that you remember how you were at least leading up to this, so I'm just trying to, again, create some efficiencies for our group, so that we just don't -- because this is complex stuff, and the details are important.

And if we come back two weeks later or three weeks later on a phone call, and we have to reiterate every single point, that will be unpleasant for all involved, I think.

DR. WILLIAM RICH: Do we need a motion to table?

CO-CHAIR ROSENTHAL: Yes.

DR. WILLIAM RICH: So moved.

CO-CHAIR ROSENTHAL: Okay. Second?

DR. STEPHANSKY: Second.

CO-CHAIR ROSENTHAL: Any discussion? Not discussable. Motion to table as not discussable. Thank you, Robert's Rules of Order.

(Laughter.)

No, you can't speak. It's not discussable. Oh, you can eat. Oh, you said we can eat.

Okay. All in favor?

(Chorus of ayes.)

Okay. All opposed?

(No response.)

Motion carries.

We will take --

CO-CHAIR STEINWALD: We need to take some lunch, yes?

CO-CHAIR ROSENTHAL: Right. So half an hour for lunch?

DR. LYNN: Can I ask you a question?

CO-CHAIR ROSENTHAL: Yes, absolutely.

DR. LYNN: Will 1599 be discussed at 2:30, or will that be tabled as well?

CO-CHAIR ROSENTHAL: Oh, 1599 will be discussed in about a half an hour. Sorry,

we should have clarified that.

DR. LYNN: Okay. So we are going to do that right after lunch.

CO-CHAIR ROSENTHAL: Yes, right after lunch.

DR. LYNN: Okay.

CO-CHAIR ROSENTHAL: So about a half an hour break, and then we will go right into 1599.

MS. ZIELINSKI: Was that agenda item -- this is the first I've heard of it. I'm not sure if our resource is going to be available. I have him coming at 2:30. I was not aware of this agenda change.

MS. WILBON: Cheri, I think I sent you an e-mail yesterday afternoon about 1572 getting moved and 1591 getting removed.

MS. ZIELINSKI: And then we started today, but then those -- there were no changes to it.

CO-CHAIR STEINWALD: Well, the question is, do you think that you could get

the person that we need to have involved at 12:45?

MS. ZIELINSKI: Not in a half hour. I apologize. I have him coming at 2:30, which was what the agenda had said.

DR. ROSENZWEIG: I'm going to sign off here. Thank you very much. Bye-bye.

CO-CHAIR ROSENTHAL: We'll confer with staff here, and we'll get back to you on this.

CO-CHAIR STEINWALD: But as it stands, we're going to start the discussion at 12:45.

MS. ZIELINSKI: Ashlie, can I talk to you offline?

MS. WILBON: Sure.

MS. ZIELINSKI: I'm not going to be able to have a resource there for that discussion. I -

MS. WILBON: Okay. Give us some time to confer to see what we can do.

MS. ZIELINSKI: Okay. So you'll

send me an e-mail, then?

MS. WILBON: Yes, I will.

MS. ZIELINSKI: Thank you.

CO-CHAIR ROSENTHAL: We are going to adjourn now. We will confer and we will have an offline conversation with you about how we are going to manage this.

MS. ZIELINSKI: Okay. Thank you.

(Whereupon, at 12:16 p.m., the proceedings in the foregoing matter recessed for lunch.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

(12:53 p.m.)

CO-CHAIR STEINWALD: It's the sprint to the finish line, and the sooner we start the sooner we can go home.

If my co-chair would come and take the -- take his seat. All right.

Well, in the course of a number of challenging discussions, we have an extra challenge. The agenda says we are going to have the usability and feasibility wrap up of the ETG-based non-condition-specific resource use measure by Ingenix.

We have, in fact, had a conference call that many of you were present at, and we discussed the importance and the scientific acceptability criteria.

However, we did not have any official vote, even though some people attempted to access the monkey -- the monkey bars --

(Laughter.)

-- service, okay. All right. That either didn't get recorded or it's unofficial, and so we do have to go back and -- again, since we're acting as our own TAP, we have to vote on the individual subcriteria of the measures that we have already discussed.

We hope we are going to be able to leverage the discussion that we already had. Sally and staff will use their notes and remind us of main points that were made or conclusions that were drawn during our discussion of importance and scientific acceptability.

And the measure developer was present and did make a presentation of the overall characteristics of the measure. I guess there is some question about whether the appropriate person from Ingenix will be able to join us or not in this conversation, but we will forge ahead in any case.

Ingenix, are you on the line?

DR. DUNN: Yes. Hi. This is Dan Dunn from Ingenix.

CO-CHAIR STEINWALD: Good. Thank you. So you are available to respond to questions and points made.

DR. DUNN: Yes, sure. Happy to.

CO-CHAIR STEINWALD: Okay. But why don't we go directly -- all right. Ingenix, please say your name again, please?

DR. DUNN: Hi. This Dan Dunn, Ingenix.

CO-CHAIR STEINWALD: Okay. Would you give us an overview of your measure, please, for the members of the Committee who weren't present on the conference call?

DR. DUNN: Sure. Just to confirm, we are talking about the population-based measure for total cost, right?

CO-CHAIR STEINWALD: That's correct, the non-condition-specific measure.

DR. DUNN: Okay. Thank you. This is a measure, you know, based on a title which

is designed to be not condition-specific, but to be a measure of groups of individuals at a population or member level, if you will, so it's not looking at their resource use related to congestive heart failure or diabetes. It's looking at their resource use for all the services and all the conditions that they present with.

The measure includes total resources or total cost as one of the numerators of measures. It includes resources by type of service, the cost as well, and also includes some utilization measures, such as in-patient admits, days, and so on.

The risk adjustment approach is -- I'm sorry, just sort of step back, so there is -- you know, including all members in the measure, with risk adjustment based on their underlying risk as measured by episode risk groups. Each individual information is processed to identify all those numerator measures to all of the cost in use.

Also, the information for a 12-month period is processed through episode risk groups, which is a risk adjustment methodology that uses episode treatment groups and episodes of care as its foundation.

And ERGs is essentially looking at an individual mix of episodes of care and translating that into an overall risk core, and that risk core is then used to risk adjust the measures themselves.

CO-CHAIR STEINWALD: Remind us, was there a standardized pricing or costing technique used as well?

DR. DUNN: No, this has actually been applied using either approach. It will work either way, and we left that up to the user to decide which way they wanted to go.

CO-CHAIR STEINWALD: Steering Committee members, any questions for Dan before we proceed?

(No response.)

All right. Then, let's bring

importance up. Individual criteria.

MR. PHILLIPS: I have -- do we have a TAP?

CO-CHAIR STEINWALD: No. We're the TAP.

(Laughter.)

MS. TURBYVILLE: Right. So in this measure, like the other non-condition-specific measure submitted by HealthPartners, the Steering Committee is serving both as the Technical Advisory Panel, so you will be rating each of the subcriteria, and then, of course, as the Steering Committee.

We have a few notes about what we heard on the call, realizing that we would follow up. And we did feel confident that the Committee, acting as the TAP during the June 22nd call, wrapped up importance.

The notes that we walked away -- was that, in our sense, though this is without a final rating, so it's those who shared their sentiments that the measurement area is of

high impact.

There is a resource use and cost problem that the description of the purpose of the measurement was described well enough in the submission, and that they were in -- they were able to meet the criterion about the service categories that they are proposing as they are quite numerous and comprehensive. That's what we heard. Those were our walkaways.

CO-CHAIR STEINWALD: So are we prepared to vote on the subcriterion 1A? And it's -- this is one where we vote high, moderate, low, or insufficient. Are we ready? Okay.

MS. TURBYVILLE: Great. So there are 16 Steering Committee members here in the room, and there were 15 high and one moderate.

So moving on to subcriterion 1B, demonstration of resource use or cost problems and opportunity for improvement includes showing data demonstrating variation, et

cetera.

CO-CHAIR STEINWALD: Can we take the vote? All right.

MS. TURBYVILLE: Thirteen high, three moderate.

Moving on to subcriterion 1C, which is that the purpose and objective of the resource use measure is clearly described. Twelve high, four moderate.

Moving on to 1D, which is that the resource use service categories that are included are consistent with and representative of the conceptual construct represented by the measure. Eight high and eight moderate.

For our -- because this split is a little bit different, I would be interested if anyone who voted on moderate, if you could give us a little input on that, so we can capture that in our notes.

It's the 1D which is that the resource use service categories are consistent

with the conceptual construct of the measure.

DR. NEEDLEMAN: Yes, I voted moderate here, just because the pharmacy data is not required, the mental health carve-outs are not clear. Those are important cost categories and resource use categories, and I'm concerned that they're not always consistently present.

MS. TURBYVILLE: Anyone else who thinks it might be new information for us to consider?

(No response.)

All right. Well, then, let's just move on to scientific acceptability.

CO-CHAIR STEINWALD: No, we have the --

MS. TURBYVILLE: Oh, I'm sorry.

CO-CHAIR STEINWALD: This should be --

MS. TURBYVILLE: Clearly, I'm going faster than I'm supposed to.

CO-CHAIR STEINWALD: This is a yes

or no. Are you ready? Go.

MS. TURBYVILLE: Okay. For overall importance, the final tally is 16 yes. So now we can move on to scientific acceptability. Fantastic. Do you want me to recap some of the --

CO-CHAIR STEINWALD: Yes. I'm not sure if it's good to recap it all at once or parse it out? I leave it up to you. It depends, really, on how much recapping there needs to be.

MS. TURBYVILLE: These are very draft notes, because we hadn't yet tried to synthesize them for appropriate distribution at this time. We heard questions about making sure the -- which is the minimum threshold on face validity was explained, and I believe that Ingenix provided more clarity on that.

We did hear a lot of questions around the risk adjustment method and wanting more explanation of how it worked and what that meant for the ERG measure, and how the

weights were assigned to the ETG. And then, we did hear a request for a verbal description of the individual R squareds.

I do want to say that there have been a couple of questions to Ingenix about the ERG measures, primarily about -- and let me pull it up, because we documented it and they did respond verbally, and it gets the scientific acceptability. Just give me -- I think I remember, but because my brain is in crash mode, I do not want to inadvertently provide you -- you have notes, too?

While I pull up their response, Ashlie, do you want to -- just on scientific acceptability, in general.

MS. WILBON: So I'm looking through my notebook here. So for reliability, I remember that Carlos was on the call, and he was -- you guys had asked him for his input on reliability and validity, and he had thought that they had done a good job of their reliability testing, and that there was I

think a 99 percent match in the way they had compared their results and their reliability testing, that they -- he didn't find any results for face validity in the submission that was given.

There was some discussion about the risk adjustment and how the risk adjustment assigned severity scores, taking into account comorbids, and some explanation of how -- the ERG grouping of ETGs and how they assign weights using the ETG risk score.

CO-CHAIR STEINWALD: These are questions that were raised that Ingenix responded to?

MS. WILBON: Yes, it was more of -- I think I was writing down more where the discussion --

CO-CHAIR STEINWALD: Okay.

MS. WILBON: -- was going, and I think there was definitely some -- I remember Dan -- it might have been Dan that was on the phone, and he can clarify if he also remembers

from that phone call. But I know there was some questions from the Steering Committee about -- for him to kind of explain how the risk models work and how the ETGs feed into the ERG in determining the risk adjustment.

DR. WILLIAM RICH: All of a sudden my brain is clearing. I do remember the construct was that -- you heard it addressed, how they do it for a measure. But how you do it to a population-based thing didn't make a lot of sense to us at face value, so that's why we asked them to.

MS. TURBYVILLE: And then, Ingenix had provided some written input to us, a couple of things specifically, and I think Taroon also may have something to add. But there was a question about what happens to records or claims that do not match to the ETGs and what is the implications for the then total cost of care that the ERG measure is putting forth.

Ingenix did respond that, as far

as identifying the members who are in ERG -- and, Dan, please correct me if I'm not representing your written response back accurately -- that they might not be included in the measure, but all costs are.

So even if a claim is not being grouped by the ETG when they are estimating the total cost, they go back and make sure the claims -- whether or not they made it into the ETG that is helping support the risk adjustment. They are still including those in their total cost, and they provided some statistics on those implications, but I think that gets to the heart of the question on that one.

So they, while not included in the ETG and risk adjustment, they are included in total cost.

And then, they did respond to us formally about more adequately describing how the face validity, at minimum, was vetted through their process.

Did you have something to add to what you have or --

CO-CHAIR STEINWALD: Then, why don't we go to the subcriteria, and then if there's more discussion --

MS. TURBYVILLE: Walk through?

CO-CHAIR STEINWALD: Yes.

MS. TURBYVILLE: Okay. So 2A(1), if you recall, having just gone through this earlier today, is about the precision of the specifications that are provided, such that it could be implemented consistently. So it includes, as you can see, many components. So how well defined and precise are the specifications?

CO-CHAIR STEINWALD: Questions?

(No response.)

Okay. Then, let's call the vote.

MS. TURBYVILLE: We have 10 high, five moderate, and one low.

And so moving on to 2B(2), which focuses on the reliability testing

demonstrating that the results are repeatable was the -- is 2A(2).

CO-CHAIR STEINWALD: Call the vote.

MS. TURBYVILLE: Nine high, seven moderate. Okay.

Overall reliability of the measure as submitted. Eight high, seven moderate, one low.

I'm tempted to --

DR. BARNETT: Well, I voted low.

(Laughter.)

So, you know, I don't -- I don't see -- so one thing that the -- some of the measures that we've done actually have some measure -- have some indication of how well the case mix measures perform, and also how well they repeat in different years for the same providers. And so I don't see that sort of reliability testing in this submission.

DR. REDFEARN: Well, Paul, the ERGs was in the Society of Actuaries paper,

along with all of the other ones and tested in all those same ways, and it performs about the same as the others.

DR. DUNN: This is Dan. We did submit some R squared measures as well. Actually, we did reference the SOA study in our internal testing. We did -- you're right, we did not cover the year over year for the same provider issue. We did not comment on that.

DR. BARNETT: So I'll just observe, you know, the Adams paper which was -- you know, you distributed to us before we started our meeting last year -- that was sort of their -- the key issue was is that -- for them was is that you would want a provider to be judged the same way or similar ways.

You would want them to flip-flop around from year to year. You would expect that they would be doing things reasonably the same. Or another way of thinking about it was with different cohorts of patients that their

practice would -- style would end up showing the same result.

If you split the sample, say, and half their patients, and then compared that to another half of patients, you would expect them to get rated about the same. So that's the kind of measures of reliability that I hope we would be looking at in these measures.

DR. DUNN: And then, we did see that clarification, the Adams paper and others, in The New England Journal. That was relating to episode-based measures, and this was a population-based measure.

But the point is still valid -- you would want the quality of the -- one of the qualities of the measure is that, you know, time over time consistency -- and just as a note, you know, I think the conclusion of that was you need a reasonable sample size to support, you know, that type of reliability.

And we did provide some guidance in our response around both that issue as well

as tests of statistical significance, which should take into account, you know, appropriate sample size as well as the general precision of the measure.

CO-CHAIR STEINWALD: Okay. Let's, if we can, move on to validity, and keep in mind that there are six separate subcriteria for validity.

MS. TURBYVILLE: Okay. So two -- thanks, Dan. So 2B(1) is about the measure specifications being consistent with the evidence presented to support the focus of the measurement under criterion 1B. So is it consistent with what was presented under importance for its purpose, as specified?

CO-CHAIR STEINWALD: Just a footnote here. My recollection of the conference call is that there was much more discussion and some disagreement or different kinds of concerns raised in the validity area. So this is an area where we may want to raise some of those concerns again for the benefit

of the entire group.

Should I call the vote? All right. Go ahead.

MS. TURBYVILLE: For 2B(1) we have seven high, eight moderate, and one low.

So moving on to 2B(2), which is about the validity testing itself, demonstrating that the measure data elements are correct, or the measure score correctly reflects the costs of care or resources provided and adequately distinguish higher and lower.

DR. O'NEILL: And it looks like Carlos had a lot of comments on this particular one.

CO-CHAIR STEINWALD: Comments or questions?

(No response.)

Okay. Let's call the vote. Oh, Carlos's comments.

(Laughter.)

DR. BARNETT: I was trying to find

-- I tried to find Carlos's comments in the --

CO-CHAIR STEINWALD: Okay.

DR. BARNETT: -- it's in the other file.

CO-CHAIR STEINWALD: We'll pause, then. My recollection was that he said that there wasn't much in the way of validity testing. That was what their summary --

DR. O'NEILL: They said they were going to follow up, and they did, right, follow up?

MS. TURBYVILLE: Right. They followed up and provided a little information to demonstrate face validity.

Dan, did I get that correct?

DR. DUNN: That's right, Sally.

MS. TURBYVILLE: So, and then just as a reminder, as we discussed yesterday from the NQF testing task force report that face validity would be the minimum threshold to demonstrate validity, that it is something that we allow to come through to demonstrate

validity, but it is kind of the minimum.

CO-CHAIR ROSENTHAL: I'm trying to play by the rules now. What would a statement of face validity be against this measure? What would be an articulation of that? I'm not asking that as a challenge. I'm asking it because I'm unclear in my own mind of what a statement of face validity would be against this.

I get the definition, but I'm trying to -- I'm posing the question for myself of, what would that mean in relationship to this measure? Open to anybody to help me with.

MS. GRABERT: I don't have a response to your question. I have a question for the developer. As I read this definition it says, "Validity testing demonstrates that data elements are correct, or that the measures score correctly reflects the cost of the care."

So if you accumulate all of the

episodes into one per capita measure, how do you account for the fact that an individual -- whoever this is attributed to, some of the episodes may be high cost and some of the episodes may be low cost, when you look at the total resource utilization.

DR. DUNN: Yes. So maybe this is a clarification, so -- and I apologize if this wasn't clear in our submission. So think of it as a numerator and denominator type of concept.

In the numerator, we are capturing all of the costs for an individual, you know, whether or not they grouped episodes, no matter what episode they grouped to -- you know, it's very similar to some of the other population-based measures. You see like the NCQA or you measure, for example, all the costs for the individual identified.

And then, where the episode, then episode risk groups, come in is categorizing individuals based on their relative risk. So

it is going to capture all of the costs, you know, from high, low cost episodes. Even things that didn't group to episodes are part of the measure.

But getting back to I think the point you're getting at is, what do we do with outlier patients? And part of our specification was a guideline that you would, you know, develop an approach for outliers on the higher side, and we had proposed as the guideline there that you would Windsorize or truncate the costs for high or, you know, really outlier patients at some level.

So, you know, for example, say that was $50,000. We would count the first $50,000 towards the measure and ignore those other dollars above that threshold for the patient.

MS. GRABERT: Do you Windsorize for outlier episodes as opposed to outlier patients?

DR. DUNN: It's outlier patients.

I'm sorry. Did I say episodes? It's a patient-based measure. The only real episodes at play here is in trying to estimate that -- an overall level of risk for the patient.

CO-CHAIR ROSENTHAL: Is the answer to my question that, in fact, it is ipso facto valid because it sums up total cost, and the total cost is the total cost? So it, by definition, is -- has face validity? I'm trying to figure out what the criteria is for answering this question, so I can figure out how I should vote.

CO-CHAIR STEINWALD: Okay. Total costs are generated through claims. Claims have their own adjudication process, so that's an element of it as well. But then, it is subject to all the other kinds of problems of the kind that Jack and others have raised. Jack?

DR. NEEDLEMAN: Yes. Well, I'm going to raise it again.

(Laughter.)

Again, we have got -- you've dealt with the pharmacy carve-outs by acknowledging some of your folks don't have pharmacy data. And pharmacy isn't a problem in terms of identifying ETG groups, because pharmacy claim cannot be the trigger event for an ETG. So it doesn't affect your risk adjustment.

But you have also got carve-outs in some of your populations for behavioral health or mental health benefits. And you've got ETGs, which will be affecting your risk adjustment that are mental health based, and there are clearly costs associated with treating various kinds of behavioral health issues.

So you've got some groups with carve-outs and some groups without. Can you tell us what kind of bias is being introduced into the measure to not have -- to have carve-outs and whether -- what kinds of steps you take to adjust for that in your analysis.

DR. DUNN: Again, that's a good

point, and this did come up in the earlier meeting. And maybe even take a step back, you know, if -- our specification was assuming we at least had complete medical services, and then there was the option that we could risk adjust, you know, for the difference, or they would measure adjustment for the difference if someone had pharmacy data available or not.

So missing mental health or behavioral health claims or lab claims or anything else, that bets against our, you know, guideline and specification for the measure.

I'm not sure that helped, but, you know, obviously, if you want me to answer the question if someone didn't have information, I certainly wouldn't compare, say, one organization against another where one had that information, one didn't.

You potentially could argue the measure could still work if you were able to equalize, you know, the fact that you didn't

have behavioral health services on either side. But that wouldn't be my recommendation. It would be, you know, that you would have, you know, complete and consistent medical service claims as a minimum, and we -- and if you have pharmacy data, we were able to -- measure data to use on it, and it adjusts appropriately for people with and without.

CO-CHAIR STEINWALD: Tom?

CO-CHAIR ROSENTHAL: Jack, I think this is an issue of, to some extent, possibly of how these measures have been used in practice, which is, as it was described either at a health plan level or a state level or whatever, where the differences in -- and then applied it to physician groups.

So if it's Blue Cross of Ohio, the carve-outs are going to be basically the same. And so when they're saying physician group A, physician group B, physician group C are different, they have accounted for, in general, the pharmacy benefit or lack thereof,

but that could be different for, you know, the State of Wisconsin, if they chose to do it.

So there's internal consistency when it's used by one set of people. But it raises the question, which I think will come up in 2B(2), of, as we talked about yesterday, how comparable are these results across entities that might be using this, and then the challenge of some having pharmacy in and some having pharmacy out, some having mental health in, some having mental health out, is going to render those comparisons to be --

DR. NEEDLEMAN: Yes. And if you've got physicians with -- you know, who are serving patients in multiple health plans, some of which have carved out mental health benefits, and some of which haven't, and you're getting data from all of them and you want to pool it, so you get a richer vision of what the experience of this -- of physician A is, you've got real problems if --

CO-CHAIR ROSENTHAL: Then you've

got real validity problems.

DR. NEEDLEMAN: You've got both risk adjustment problems, and then you've got cost problems.

CO-CHAIR ROSENTHAL: I think that's -- I mean, I've been thinking what I'd vote on this one, and I think it's 2B(2). And the same issues will come up about cost adjusters across geographies, which I don't think are here, because, again, they are not using standardized pricing on this.

DR. DUNN: This is Dan. Maybe I can -- in my mind, I don't think this relates to the validity of the measure, because the measure itself, you know, includes the behavioral health services or any carve-out as part of the specification. It maybe relates more to applicability and challenges in measurement using this measure.

CO-CHAIR STEINWALD: Okay. Any further -- this is still 2B(2) -- on validity testing? Can we call the vote?

DR. BARNETT: I just wanted to -- we have mentioned Carlos's work.

CO-CHAIR STEINWALD: Right.

DR. BARNETT: And then, but we never -- we never actually said, "What does it say?" And so he says, "Has measured score validity been shown?" and his response is, "No."

"Description of the approach used to test validity lacks detail and clarity." I'm just reading from this, right? "It is mentioned that the process described above to test validity included a review by clinical analysts to assess face validity, but no details are provided."

And then in the fourth paragraph, "Finally, there is an attachment in the submission labeled 'Reliability Validity Testing' consisting of several tables describing resource use and its components for different peer groups stratifying by the presence of pharmacy benefits. Unfortunately,

there is no description accompanying the tables that explain how they relate to validity -- to reliability and validity." So that's Carlos's --

MS. TURBYVILLE: Right. And so then there was -- thank you, Paul. That's absolutely right. And then, there was follow up by Ingenix. The table that was attached wasn't intentionally supposed to be a part of their reliability testing.

What I understand and what was submitted was face validity was established for this measure. So as far as the table having empirical results and validity, it was really focused on the face validity, which they did provide a more detailed review of how they vetted it.

CO-CHAIR STEINWALD: Anything further?

(No response.)

Okay. Then, we'll call the vote.

MS. TURBYVILLE: For 2B(1), we

have eight moderate and six low.

Moving on to 2B(3), which is exclusions are supported by the clinical evidence, they are supported -- yes, 2B(3).

CO-CHAIR STEINWALD: All right. This is the one that addresses exclusions. Jack has already weighed in. Anyone else like to raise a question or make a comment?

(No response.)

Well, there are no exclusions of resource use. That I think was explained pretty clearly. And, yes, I don't have -- I mean, there is certainly the carve-outs issue, but I don't have anything else to raise myself. Anyone?

MS. TURBYVILLE: Dan, do you want to briefly describe what exclusions the ERG measure specifies?

DR. DUNN: Is the question patient exclusions or service exclusion?

MS. TURBYVILLE: Patient exclusions.

DR. DUNN: There are no patient exclusions other than -- which I guess there are patient exclusions -- the handling of patients with extremely high cost due to outlier methodology, but there are no patient exclusions.

CO-CHAIR STEINWALD: Anything further?

DR. O'NEILL: What Carlos refers to as low outliers, that is not excluded.

CO-CHAIR STEINWALD: That's for you, Dan.

DR. DUNN: Yes. No, there would be no outliers, because there is not a lot of -- different from an episode, a lot of people have no services, no costs in any given year. But a low outlier doesn't apply to this type of measure.

CO-CHAIR STEINWALD: Okay. Any further?

(No response.)

Okay. Let's call the vote.

MS. TURBYVILLE: Nine high, four moderate, and two low.

CO-CHAIR STEINWALD: Okay. Then, on to 2B(4), which is where risk adjustment is addressed. I have -- I am going to, with some trepidation, raise the same question that I raised before on the conference call. I don't understand, once you accomplish risk adjustment through ERGs, why the patients then subsequently need to be grouped into ETGs as well.

DR. DUNN: Yes, they do not -- I apologize again if this wasn't clear in the submission. Think of the ERG risk adjustment having two steps. One step is to categorize the patient's ETG, so what episodes were observed. If they have congestive heart failure episodes, diabetes, episodes, and even within diabetes episodes, what level of severity are they? That's step one, and that is where ETGs play a role.

Second step is taking those

episode results of the presence or absence of episode -- of certain ETGs for an individual and translating that into an overall risks core. So think of ETGs as giving is the risk markers, if you will, similar to the way, you know, the CMS HCC model has diagnosis-based markers of risk.

ERGs also has diagnosis-based markers of risk, but that sort of diagnostic categorization is based on the episode of care framework. And once you have a member's episode of care, you know what ETGs mix they had, and where computing an overall ERG risk score.

But once you have an ERG risk score, you don't need to use ETGs at any point in either the measure numerator or denominator.

CO-CHAIR STEINWALD: Thank you. That helps. Any questions or comments?

DR. O'NEILL: My question has to do with whether or not a situation, you know,

with the Dartmouth Atlas that shows that there is different levels of utilization in different populations.

If you had a high level of utilization of a given set of procedures, would that set off more ETG identification of patients that would drive a higher risk score through the ERG, or would that look like overutilization? I'm just wanting to make sure that the frequency of a treatment that would trigger an identification of a patient is looked at independently of the utilization that is evident in the database. You know, I'm trying to make sure this isn't circular.

CO-CHAIR STEINWALD: You don't want to -- Dan, that's for you. Go ahead.

DR. DUNN: Okay. Good question. You know, ETGs are diagnosis based. So, you know, even though ETGs won't capture the fact that someone had a -- you know, with a CAD episode had a CABG surgery or catheterization, you know, that doesn't effect what ETGs are in

that, just kind of what is observed within the episode. And so then, following from that, given ETGs is driving ERGs, that wouldn't affect the risk scoring that should be observed as over -- a relative over or under utilization.

CO-CHAIR STEINWALD: Satisfied? Yes. So the -- even the frequency of episodes, then, would not generate a higher ERG score.

DR. DUNN: No, that could be -- so if they're diagnosis based, so if, you know, certain patients in an area had, you know, more -- had a higher prevalence of CAD or hypertension or congestive heart failure episodes, that does drive risk score for the area.

So whatever is, you know, triggering those -- similar to any of these models, risk models, things that are triggering more observable and usable diagnoses are going to -- you know, going to

generate a higher number of episodes or diagnoses and higher levels of risk.

DR. O'NEILL: Can I use as an example lumbar fusion where in some neighborhoods that looks like an automatic surgical case, and then other places it does not. And the fact that a higher percentage of people would get the surgery would not necessarily indicate that there is a higher level of more severe back pain.

CO-CHAIR STEINWALD: But it would generate a higher risk score.

DR. O'NEILL: That's right.

DR. DUNN: No, it would not, because what drives the risk score is the diagnosis of back pain, not whether there was lumbar fusion or not.

CO-CHAIR STEINWALD: You know, I think part of the -- and accepting what you say, the very fact that you call them ETGs, where the T stands for "treatment," is a source of some confusion. Paul?

DR. BARNETT: I was going to say, my understanding is that ETG is sort of a way of grouping the records, and then the ERG is based on the codes in there. And they don't allow the medical treatments to define the risk group, except they said one category, which was malignant neoplasm.

So I think that we're going to find that every time we rely on claims data, administrative data, to characterize the health of a population, that we are going to be in this -- in this -- have this same problem. And that actually it seems to me that this is as best we can parse it out from what's available to us, that this is a pretty elegant solution, that they are only looking at the diagnosis and not the procedures that were done, not the treatment.

But where all the risk models break down is for the people who don't get very much care. If they're outside the health care system, we don't really know what their

health state is, and so, really, at the low end risk models perform poorly. But, you know, there is not so many resources being used in that case, so maybe we are less worried about it.

So it seems like a very elegant and detailed method, and it makes me wonder how well this performs compared to some of the simpler models that are more transparent and free, like the HCCs, so this is presumably a pretty expensive product to go this route for a case mix.

And so what are we gaining? You know, what's the marginal benefit from this more extensive case mix measure compared to something that is free like HCC or maybe just a different product?

And we have that Society of Actuaries study that I have not read or seen, which seems to say they all perform equally well. So we have to think, well, do we need to spend millions of dollars, because, after

all, we are really using all of the complicated ETG/ERG stuff for is to get a case mix measure. The rest of it is just total costs in that year pretty much.

CO-CHAIR STEINWALD: Further comments, questions?

DR. BARNETT: Well, I'd actually like to know if the measure developer has an idea of what is the marginal benefit of their case mix measure in terms of variance explained compared to some of the others.

CO-CHAIR STEINWALD: Dan, you're up.

DR. O'NEILL: Sure. And maybe I'll note the free -- the free HCC model is the one that I'm -- at least my understanding is the one, you know, modified by CMS to support Medicare Advantage payments, built for an elderly population, and actually somewhat different than a model that you would likely use for measurement purposes.

You know, on purpose, the CMS

excluded some diagnoses and/or some HCCs in that model, so it-- I'm not sure I would use the free model here.

But, you know, in terms of the other approaches -- and there are -- you know, there are, you know, other, you know, methodologies to do this, like ACGs, the commercial HCC model which is licensed in the same way ERGs is, you know, we have always found we have done as well or better than the competitor.

So it -- the characterization that it is -- they are all in the same ballpark is probably valid, and, you know, I think there is reasons, though, people would use one of the commercial models, whether ACGs, you know, DXCG, HCCs, or ERGs, you know, versus what is free, you know, through the CMS HCC model, for example.

CO-CHAIR STEINWALD: Further comments or questions?

(No response.)

This is 2B(4). Can I call the vote?

MS. TURBYVILLE: Six high, eight moderate, and one low.

So moving on to 2B(6), which is --

CO-CHAIR STEINWALD: Five.

MS. TURBYVILLE: Oh, sorry. Thank you. 2B(5), which is that the data analyses demonstrate methods for scoring and analysis of the specified measure allowed for identification of statistically significant and practically or clinically meaningful differences in performance, or there is evidence of overall less-than-optimal performance.

CO-CHAIR STEINWALD: Tom, do you have something?

CO-CHAIR ROSENTHAL: A question I want to clarify, if I could, is, have there been instances where this measure has compared across geographies? So I think this is addressed to the developers.

DR. DUNN: Well, for this and measurement applications, I don't know of any application that goes beyond, you know, what geographic area, like a state or market. It could be used for that purpose, but I am not aware of any.

CO-CHAIR ROSENTHAL: Just to stay somewhat internally consistent, to my knowledge, it has not been used across geographies. And, again, I would submit that the -- a) it hasn't, and part of the reason it hasn't is that isn't the way it has been used, and that we get into the same set of issues that we talked about yesterday, which is this is a total -- this is about a dollar denominated number, not a resource use that has standardized pricing.

And so the aptness and accuracy of a comparison of applying this in one geography won't be -- for some provider group that you say this primary care doctor has 1.8X of the norm of utilization using the ETG grouper

won't be a meaningful comparison of comparing individual doctor internists in another part of the country, or a group of internists who might have .8. It's quite possible that those numbers will be -- would be incorrect.

CO-CHAIR STEINWALD: Bill?

DR. WILLIAM RICH: You also state, again, that this can be used -- this is a total resource use measure. But you also say it can be used for comparing physicians. We have the same issue. You avoid the issue of attribution if you just look at total costs, total resource use.

You also address that -- stated that you will eliminate some for low number. What -- is there a specific low number that is specified, or is that based on statistical significance of that individual provider? But you specifically say there is a low number. What does that mean?

DR. DUNN: So this is related to low number of patients, for example, for a

measurement entity, like a physician. That's the question?

Yes, we actually didn't specify when we specified a statistical significance test using confidence intervals or something similar. And we also I believe had submitted -- you can almost get two thresholds here. One may be more judgmental is -- what is the number of patients that it's even worth reporting a number, you know, even with a confidence interval?

And we had just suggested that number, you know, was 30, and -- but, you know, beyond that is, you know, some application of a statistical test that adjusts a sample size, and that that was our preferred approach.

CO-CHAIR ROSENTHAL: We're a little bit recapitulating the same thing we talked about in the diabetes measure, but I -- it feels to me even more troubling here of this balance between an attribution model that

is quite specific, but the specificity creates issues.

But it's not certain to me that you solve the specificity problems by making it totally more or less open-ended, that the individual user in an individual site can sort of decide how they want to attribute it.

I think, then, the comparisons from one place to the other, possibly now even within a state, could be quite challenging. And I'm personally trying to be internally consistent with our discussions in fairness, and our Health Partners friends are sitting here. So in respect to that, I am trying to be internally consistent.

But I also don't want to necessarily recapitulate the whole set of arguments from yesterday, but I do think they --

DR. WILLIAM RICH: My point is just to point out we have the same issue.

CO-CHAIR ROSENTHAL: The same

issue I believe resides here.

CO-CHAIR STEINWALD: It's the same issue. But I am going to call on Mary Kay, who is going to tell us why there are times when we want to -- we don't want to adjust.

DR. O'NEILL: Well, it does state in here you can use standardized pricing. It's not only the real dollar thing.

DR. DUNN: And maybe to comment -- that's correct, you can use standard pricing. And, you know, the State of Wisconsin, for example, uses this measure, and they use standard pricing to do comparisons across the state.

And maybe to point to attribution is -- you know, I think we provided the guideline there on what are reasonable ways to do that. I think the challenge is, you know, depending where you're using this, whether it's, you know, for an ACO or for an individual physician or places where -- I mean, there is still some gatekeeper type

arrangements.

You know, attribution is -- the right attribution approach often, you know, makes sense for where the application is applied. That's why we had provided it as a guideline with options rather than out of the specification itself.

CO-CHAIR STEINWALD: Yes, Mary Kay.

DR. O'NEILL: So to talk about this for a moment, I mean, if you -- if we are going to do our famous Memphis-Minneapolis comparison --

(Laughter.)

So if we ran standardized pricing in those two markets, which is a surrogate for utilization -- and I'm taking standardized pricing as a function of essentially weighting the different types of services, which is why use it and not just utilization numbers. So like you're a little off on your labs, but nobody is going in the hospital, it's less of

a big deal.

What I would really love to be able to see is to do the standardized pricing check and see where people are comparatively on their utilization, then run the real dollars and see where we're spending our most money. And the combination of those two analyses would be the most powerful thing that we could do.

CO-CHAIR ROSENTHAL: If in fact standardized pricing were specified, I would probably be in favor of the whole thing. Well, I get it, but, I mean, it's not clear to me -- first of all, when we asked them on the phone, that was the first question the Chair posed to the developers, "Was there standardized pricing?" and the answer was no. So --

MS. YANAGIHARA: They don't specify a particular standardized pricing methodology, but it can be -- this methodology can be used with standardized pricing. So you

can use the NCQA one posted on their website for -- I mean, there --

CO-CHAIR ROSENTHAL: Yes, but somebody has got to do it. This is the measurement. I mean --

MS. YANAGIHARA: Right.

CO-CHAIR ROSENTHAL: -- I don't know what the rule set is around, again, something being a guideline versus being specified, but it seems to me, again, this needs to be specified and not just kind of, "Oh, it's out there," and then they can sell it. I mean, they are a commercial entity.

If they can sell it to people who want to use it any old way, we are supposed to be adjudicating this for the country. So I think it has to be specified and not just kind of people can use it any old way they want. And that is a problem, because we've been told over and over, once we approve it, people can use it any way they want consistent with the actual written measurements.

MS. TURBYVILLE: So whether or not you think it's complete enough, they do explicitly state in their specification that the measure should use complete and valid financial amounts, or a standard price-to-resource cost amount. Is that -- that's as submitted by the spec, but so --

CO-CHAIR ROSENTHAL: I don't think those are equivalent. The first one of those is in fact -- if in fact you have total claim data from a particular plan, you've got -- you've met Category 1. And that does not invoke standardized pricing. In fact, they use standardized pricing in that articulation of it sort of in lieu of maybe we've got inadequate claims data, so we'll use some kind of standardized pricing to do it. Well, that's the way I heard it.

MS. TURBYVILLE: I don't want to unfairly represent what they wrote.

DR. BARNETT: So this kind of relates to which peer are you being compared

to, and in terms of the performance results reported, which is I think what we're -- the topic we're on right now. So there is in the submission packet a sample report for the year ending 12/31/2007 -- or two years ending December 2007, and it shows a cost and also utilization for this provider compared to the peers. And there it is magically on the screen in front of us. Very good, Ashlie.

And, boy, it took me a lot longer than that to chase this down. And so the question -- it says page 4 on the paper, but --

MS. WILBON: If you open the PDF packet, it's bookmarked. Item S12, Attachment Sample Score Report. If your bookmark doesn't open when you open the file, just click on the bookmark icon and you can --

DR. BARNETT: So is it up to the consumer to identify the peer, or is that something that comes with the product? And then, how do I know if I'm the primary care

physician and my pharmacy costs here are 32.78 versus the peers' -- let's see, I guess we're on a slightly different page. Next page maybe. So how -- my costs are actually different from the -- my peers' cost. Whether that's a really significant difference or -- it's a little bit further than -- is it?

CO-CHAIR ROSENTHAL: Is the one she's looking at close enough?

DR. BARNETT: No, this is the one I was thinking of, see? So here we have my actual and then my peers. And so I'm wondering, well, who is my peer? Is the user of this responsible for finding the peer to be compared to, and defining that peer, and how do I know that the -- my costs are significantly better or worse than the peers? I mean, is there some sort of confidence interval or statistical significance that comes with this?

DR. PETER: Yes, it says it actually at the first page of the report. It

has -- I think maybe the results of the sample were not -- on the first page of the report, which is page 80 of the PDF, there it says -- you see the little sort of footnote it says, "Statistical significance," I think if it were it would have an asterisk, but I don't think it is, so it's not -- it doesn't say "not statistically significant."

CO-CHAIR STEINWALD: Dan, can you tell us, for purposes of this analysis, how peer groups were defined?

DR. DUNN: Sure. That isn't part of the specification. You know, to be honest, it is usually something that the user defines given a guideline as -- usually, you know, same specialty within a homogeneous geographic area.

But you do get -- I'm not sure how you would ever specify that, to be honest, but, you know, I guess you could say only use for providers, physicians where their specialty is X or Y, but then the geographic

definition I think would be pretty challenging, and even you could go beyond that to be honest.

So in terms of peers -- and the related followup question was related to confidence intervals between many of the measures, and that is -- although not shown on that report, that was part of our, you know, recommendation was to use confidence intervals to assess differences between a provider and their peers.

And most -- I think as a -- all of -- and just as a note, that is just a representation of how this measure may be reported. You know, there are some users, in my experience, of this measure who decide to report by using a different format and template. So that is not necessarily tied to the measure itself.

CO-CHAIR STEINWALD: Thank you. Any further --

DR. BARNETT: I guess the question

is -- is have -- has this been done, or is this something that is hypothetical, that could be done?

DR. DUNN: Sorry. Your question -- has this measure been used? And the answer is yes.

CO-CHAIR STEINWALD: And I thought I heard you say that the kind of analysis that we are looking at here is likely to be confined to within geographic area, as I understood.

DR. DUNN: Right. That's not part of the specification, but my experience is it is always linked to a geographic area, and it is -- you know, this type of measure, when used for physicians, always linked to either a general adult permanent care concept, internal medicine/family practice combined, or those separately, and then pediatrics is a separate entity.

And, you know, increasingly now you will see organizations looking to use this

type of measure for ACOs or medical homes as well, but there may be more groupings of providers and individual providers.

DR. O'NEILL: So to Tom's point and concern historically that we have difficulty -- we get into difficulties in trying to compare delivery in different markets with different market structure and habits and all of those kinds -- and, you know, contracted rates, that that has not actually to date been an application. But some of your concerns are to make this measure able to be used in that fashion. Is that --

CO-CHAIR ROSENTHAL: That has been my concern.

DR. O'NEILL: Yes.

CO-CHAIR ROSENTHAL: And we did see one that had a very specified standardized pricing. I can't remember which one it is now. They are all running together. But we had one that had it very well specified -- a standard -- oh, yes, that's right, it was the

Health Partners resource use one, the initial one that we approved.

And they exist -- this is not, you know, impossible. It just -- so I -- you know, I think this is the same set of issues that we talked about yesterday.

CO-CHAIR STEINWALD: Bill?

DR. WILLIAM RICH: I think we're seeing a real issue with a commercial product that can be sold and adapted for many different ways. I don't know how we are going to ever address this issue.

At least they're honest enough to say that it -- that this -- if there is any attribution, they say what it is and people are free to use it any way they want. I don't know how we can control that except to express some concerns as we have as we go along. And we don't have any primary care docs here anymore, but -- oh, I -- I'm sorry. I have a right homonymous hemianopsia, but --

(Laughter.)

I actually got that out after lunch. But I don't think we are ever going to be able to control the -- they are very honest enough to say, if you look at their applicability, they list a whole bunch of different things.

And so we're trying to hit a moving target. Whether it's, you know, how the things are aggregated to the doc, whether it's standard pricing or not, I don't think we are going to be able to address that except to express some concern that if you are going to compare across areas, geographic areas, that there be standardized pricing.

CO-CHAIR STEINWALD: Any further --

(No response.)

Hearing none, this is 2B(5), let's call the vote.

High five, moderate seven, low three.

And now we go on to 2B(6), which

is possibly non-applicable.

MS. YANAGIHARA: Just a quick -- it's kind of related to the standardized pricing, because I think it's going to keep coming up over and over. But I wonder -- I think with attribution that is tied more closely to the business case than business use of it.

But with standardized pricing I am wondering if there might be an opportunity to select a standardized pricing methodology that would then should be -- I mean, as the endorsed standardized pricing methodology. It doesn't seem like there is much -- as much tie to the business use in that case. It's just really how do you cost a particular item.

So I'm just thinking -- I'm not saying we should adjust that today, but just for future thinking and NQF work, it may be worth it, because that would help standardize things and not just, well, pick whatever method out there, but here is the endorsed

method.

CO-CHAIR STEINWALD: This is for Helen's bucket list.

(Laughter.)

All right. So 2B(6) is non-applicable. And that means that we have an overall vote on --

PARTICIPANT: Did we do 2C?

CO-CHAIR STEINWALD: No, we have to do an overall on 2B? First --

DR. BARNETT: Can I just follow up something -- on what was just said about the standardized pricing?

CO-CHAIR STEINWALD: Yes.

DR. BARNETT: Which is this is not trivial. Every year there is dozens, if not hundreds, of new CTP codes and HCPCS codes to add. Two years ago they entirely revised the DRG system.

As head of a center that does this routinely, it costs a lot of money, and it's non-trivial.

CO-CHAIR STEINWALD: Duly noted. Somebody raised a question about 2B(6)?

MS. WILBON: Right. It talks about multiple data sources.

CO-CHAIR STEINWALD: Right.

MS. WILBON: We can open it up for discussion. But generally, since they are only using admin data, then it's not really right. Okay.

CO-CHAIR STEINWALD: It's non-applicable, which means, then, we move to an overall vote on validity and taken as a whole. Call the vote?

MS. TURBYVILLE: Two high, 10 moderate, and three low for overall validity.

CO-CHAIR STEINWALD: Okay.

MS. TURBYVILLE: Okay. So now to 2C, which is the last subcriterion for scientific acceptability, and it is concerning disparities in care, if they have been identified, that the measure specification scoring and analysis allow for identification

of disparities through stratification of results -- again, getting to that one exposed these differences. The examples are race, ethnicity, et cetera.

MS. WILBON: So I just wanted to interject quickly. I do have some notes from a call on your discussion on this, and it was talked about that, to the degree that low SES and different racial groups use more or less resources, that it is relevant.

But it's often not captured in the admin data, and that it can't be captured systematically right now, and that the Committee would recommend that this would be captured in the future in this measure to the degree that it's possible.

CO-CHAIR STEINWALD: Further discussion or questions?

(No response.)

Let's call the vote, keeping in mind that there are four choices.

(Laughter.)

MS. TURBYVILLE: So four moderate, two low, and nine insufficient.

CO-CHAIR STEINWALD: And now -- well, yes, we have an overall on scientific acceptability. This should be --

MS. WILBON: I can read it aloud. So keeping in mind that grid that we've been referring to for the overall scientific acceptability votes, your rating for the overall reliability was eight high, seven moderate, and one low. And your overall rating for validity that you just completed was two high, 10 moderate, and three low.

CO-CHAIR STEINWALD: So can we call the overall for scientific acceptability?

MS. TURBYVILLE: So for overall scientific acceptability for this measure, we have nine yes and six no.

CO-CHAIR STEINWALD: So now that we've finished our recapitulation of our conference call --

MS. TURBYVILLE: It was much

shorter this time.

CO-CHAIR STEINWALD: -- then we move on to usability, and we have had a number of discussions of usability. But we do have four subcriteria that we have to separately vote on. Would anyone object if we moved right to the first one?

(No response.)

Okay.

MS. TURBYVILLE: So 3A asks about the measure performance results are reported to the public in national or community reporting programs by the time of endorsement maintenance review. This is actually the time of initial review and is not being currently reported for public and accountability models, whether there is demonstration that it will or does benefit those models in which it is reported.

CO-CHAIR STEINWALD: A question for Dan. Has this measure been used in any published peer reviewed articles?

DR. DUNN: No, it has not.

DR. WILLIAM RICH: Is it currently being used to profile physicians on an individual level, group level?

DR. DUNN: Yes.

CO-CHAIR ROSENTHAL: Do we know, though, a public reporting of any of those individual analyses?

DR. DUNN: No. They are used for, you know, information-sharing with physicians where the users will present and discuss results with, you know, those folks who are deemed to be outliers, either that or used in some cases to -- you know, if there is some pay for performance type strategy, so both that information-sharing and there may be some financial compensation related to performance based on the measure. But taking those results and reporting them to the public, I am not aware of that happening.

CO-CHAIR STEINWALD: Further comments or questions? Doris?

DR. PETER: Sorry. I have a question about the document you submitted in the -- under public reporting. It says that the information was used to support public reporting initiatives. It looks like it is health plan related, but -- so was that not to the public? On page 32 of your submission, U1.1.

DR. DUNN: And so does that -- has that been put on our website, for example, and has Dr. -- whatever -- Jones, Dr. Smith, and so on, that was my interpretation of public reporting, and that was what I was responding to.

DR. PETER: Well, just in the phrase above it says public reporting disclosure to performance -- of performance results to the public at large. So in some form -- I guess it wouldn't have to be -- I don't know, it just -- this is your submission. It says current use for public reporting. Am I missing something?

DR. O'NEILL: It looks like it might be on a health plan --

DR. PETER: Health plan level, right, right. So to the beneficiaries of that provider, is that where they're reported? But it's to the public, I mean, to the people who are covered, the patients covered by the plan.

DR. O'NEILL: Right.

DR. PETER: Not the providers.

DR. O'NEILL: And it wouldn't be a publicly available -- you couldn't go on --

DR. PETER: No, but if I were a covered provider.

DR. O'NEILL: Right, right.

DR. PETER: Okay. That's what I wanted to clarify.

CO-CHAIR STEINWALD: Bill? Oh, go ahead, Dan. I'm sorry.

DR. DUNN: Again, I wouldn't change my response. I'm -- an inconsistency of my response in the -- what we've put -- to my knowledge, that has not been done, and I

think I'm being accurate. Not that it couldn't be done, but I do not know of that happening.

CO-CHAIR STEINWALD: Bill?

DR. WILLIAM RICH: Is the -- since you said it is used for payment between the payers and the individuals, is the tiering that results from the use of this tool, is that publicly available?

DR. DUNN: Well, I'll have to get back to you on that. I can't completely answer that. I do not know of that, but I couldn't be sure.

CO-CHAIR STEINWALD: Anything further?

(No response.)

Okay. Let's call the vote.

MS. TURBYVILLE: So for three -- yes, 3A, four moderate, six low, and four insufficient.

CO-CHAIR STEINWALD: Okay.

MS. TURBYVILLE: So moving on to

3B, which is about measured performance results are considered meaningful, understandable, and useful to the intended audience for both public reporting and informing quality improvement.

CO-CHAIR STEINWALD: Questions or comments?

DR. WILLIAM RICH: This is going to make this itself.

(Laughter.)

If you a four-person internal medicine group on the east side, which is fairly affluent compared to a four-doc east -- internal medicine group on the west side, which is probably 80 percent African-American.

These are not actionable reports, because of the different patient populations and the aggregate costs. So how is an internist going to find these reports actionable? That's a question there.

DR. DUNN: Well, the assumption is that you are making -- you cannot compare --

that it is not being appropriately adjusted for. Is that the point, that there are differences in those patients that has been captured by the measure, therefore --

DR. WILLIAM RICH: Right. And actually they are not actionable. Some of them are, some of the costs are actionable.

DR. DUNN: Maybe take -- if I could take it separately and at a general level, then comment -- I think there's two questions in that question. One is, you know, what is the composition of the peer group? And if it's heterogeneous in some way, does that compromise the ability to interpret results? And put that one aside for the moment. And the second -- and are these measures actionable?

You know, you have the total cost measure, which is a challenge in terms of actionability, which is why we also specified the major components of total costs, so they're broken into imaging of different

types, breaking out advanced imaging, you know, breaking out, you know, labs from, you know, sort of specialty and consultative and hospital, and so on.

You know, and I guess I would argue those measures have some level of actionability. If total cost is measured on its own, I think it's more of a challenge.

CO-CHAIR STEINWALD: Further comments or -- comments, questions?

(No response.)

Can we call the vote on 3B?

DR. WILLIAM RICH: We haven't answered the last part of the question about heterogeneity, if they --

CO-CHAIR STEINWALD: Okay. Hold off for a second. Go ahead.

DR. DUNN: Oh, I'm -- yes, I'm not sure how to answer that one. I think that must be more to the point of, have you constructed the right peer grouping? It also gets back to that question on, how do you

measure disparities and have them unveiled by this type of measure?

And, you know, if you measure just those differences in SES and other factors, then you are not going to observe them. If you decide that those were appropriate things to adjust for in a measure, then you would stratify the population when you create your peer groupings, you know, to support that type of difference.

I'm not sure I answered your question, but I -- it's -- you know, it relates to the objective of the measurement, and to what extent you'd want to be homogeneous in terms of the peers you are comparing a physician against or the organization.

CO-CHAIR STEINWALD: Paul?

DR. BARNETT: I was just going to say, those problems are going to exist for every measure we are going to look at for the whole course of this. And I think we are

unfairly putting them on the spot for those issues.

CO-CHAIR STEINWALD: Barbara?

DR. RUDOLPH: I agree, and in the old days when you did risk adjustment you included those data elements in your risk adjustment. You include race and ethnicity, and SES if you had it, so -- but then we wisely -- NQF took those out, so that we could stratify. But in -- and I agree with Paul that we are just not capable of doing that yet.

CO-CHAIR STEINWALD: Any further?

(No response.)

Okay. We will recall the vote.

MS. TURBYVILLE: Okay. So we have three high, six moderate, three low, and three insufficient.

CO-CHAIR STEINWALD: That's very symmetrical.

(Laughter.)

MS. TURBYVILLE: Okay.

CO-CHAIR STEINWALD: All right. Moving on? On to 3C.

MS. TURBYVILLE: All right. So on to 3C, the data and the results detail are maintained such that the resource use measure can be decomposed to facilitate transparency and understanding.

CO-CHAIR STEINWALD: Comments, questions? Tom?

CO-CHAIR ROSENTHAL: If we were going to be internally consistent, we would review our diabetes vote.

MS. WILBON: I can do that for you real quick. So this -- I don't have that vote. Did you guys write that down? I don't have it. Just bear with me for a second.

CO-CHAIR STEINWALD: Sorry. Anything further? I'm not hearing anything further except for people cranking up. Vote, let's call the vote.

MS. TURBYVILLE: Okay. So we have one high, eight moderate, five low, and one

insufficient. Getting a little bit more to the normal distribution there.

CO-CHAIR STEINWALD: Yes. Is 3D a non-A?

MS. TURBYVILLE: Yes. 3D on the harmonization of the measures we explicitly told the measure developers at this point not to try and harmonize it. As the project progressed, if that came up as an issue, we would work with them and you through that. So it's non-applicable at this time.

So that means that we would be ready to ask you all to rate usability overall for this measure.

MS. WILBON: And just a quick clarification -- the reason why we couldn't find the score, the usability score for 3C, is because the TAP scored that. That was a diabetes measure, and the TAP scored the subcriteria.

So that's why we couldn't find it in what the Steering Committee had done

because it was in the TAP notes. You guys just scored the overall criteria for that one.

CO-CHAIR STEINWALD: Right.

MS. WILBON: Okay.

CO-CHAIR STEINWALD: Let's call the vote, overall usability.

MS. TURBYVILLE: One more? There we go. So for overall reliability we have 10 moderate --

DR. DUNN: Usability.

MS. TURBYVILLE: It ends with a T-Y, okay?

(Laughter.)

I even got the L right. For overall usability -- sorry, Dan -- the score is 10 moderate and five low. And at this time, based on where we are, feasibility will not be assessed by the Steering Committee on the Ingenix measures.

CO-CHAIR STEINWALD: And the reason for that is the same as the prior measure. We don't have the information on

pricing. Is that correct?

DR. BURSTIN: Right. So once Ingenix shares with the Steering Committee the fee schedule for the measure, which we will then share with you, we will return to feasibility on both of the Ingenix measures so you can assess feasibility and your overall recommendation for the measure.

MS. TURBYVILLE: So I think, Dan, that's all the questions for you. We really appreciate you adjusting your schedule and providing responses and input to the Steering Committee today.

DR. DUNN: Okay. Thank you. Take care.

MS. WILBON: Operator, can you open the line to see if there is anyone who would like to provide a comment, a public comment to the Steering Committee, before we wrap up?

OPERATOR: Yes, thank you very much, and I will open it up.

MS. WILBON: Okay. Is there anyone on the line who would like to ask a question for the Steering Committee?

(No response.)

Anyone in the audience who would like to ask a question or make a comment to the Steering Committee based on the discussion?

(No response.)

Okay.

MS. TURBYVILLE: Clearly, we want to say thank you, and thank you for hanging in there for two days, in addition to all of the pre-work and conference calls that we have had all through this year.

I think we have a few next steps. I will ask Ashlie to speak to those, but I want to make sure I give the Co-Chairs an opportunity to provide any final thoughts on the past two days. Or is everybody done? Which is fine. We can just wrap it up.

(Laughter.)

CO-CHAIR STEINWALD: All right. Tom, you're up.

(Laughter.)

CO-CHAIR ROSENTHAL: I checked out about 30 seconds ago.

MS. TURBYVILLE: Anybody else?

DR. BARNETT: So the voting process and the being able to collectively see where we were at in polling was very helpful I think for the process. And if, when we're participating in conference calls, we had some way to do through the meeting, so I know that we -- Web Meeting and Microsoft Live Meeting, they all have poll functions, right? So we can use the poll function in the future. That would be great.

The other thing is, just as a consumer of all the materials, it was hard for me to keep them straight, and it would be great if we had some consistent way of naming, you know, like the consultant's document and the -- the submissions all pretty much began

with the submission number, and that was very helpful.

And if the attachments also began with the submission number, I think that would help us, you know, keep things straight. So, you know, that -- that was a struggle for me to -- you know, and I'm --

MS. TURBYVILLE: Yes, it's a lot.

DR. BARNETT: It's a lot of stuff, so to the extent that you can kind of help us by organizing it, that would be great.

DR. NEEDLEMAN: Likewise, you know, Carlos's submission were extraordinarily helpful, but some got labeled with numbers, multiple numbers, because multiple things -- one in a document. I'd rather have more documents that are easy to find.

MS. WILBON: Okay. That's helpful. We try to sometimes do things in a PDF and bookmark, because we don't want to send you like 50 files and then it gets hard to send them through e-mail. So, you know, we

try to think about when it's best to try to package things in one -- one PDF versus sending separate files.

So, you know, we do weigh that, but it's hard sometimes when we have so many things to send you, and we don't want to have to send you 10 e-mails to get so many documents to you.

DR. BARNETT: So the website is great. I don't know whether -- you know, are some of the things -- some of the things -- what do I want to say? Confidential, you can't put them on the website, but --

MS. WILBON: We --

DR. BARNETT: But all the submissions could be gotten from the website, and that was very helpful.

MS. WILBON: Yes.

DR. BARNETT: And all the meeting minutes and those sorts of things.

MS. WILBON: It's timing. Sometimes it takes us -- we don't have the

ability to personally post things to the website, so, you know, we may have something ready for you, but we have to wait two or three days for it to get posted to the website.

So rather than waiting for there to be like a hyperlink, we sometimes will just go ahead and send it to you, so --

DR. BARNETT: Well, there's Google Groups, there's SharePoint.

DR. BURSTIN: And SharePoint has just been put into place, so my guess is within the next --

MS. WILBON: We're working on it. We're -- but that's helpful.

CO-CHAIR ROSENTHAL: Doris?

DR. PETER: Yes, I was going to bring up the SharePoint site, because I know we had talked about that before. But I also wanted to say thank you for all your help. I know it's a lot of documents for you, too, so thank you.

MS. WILBON: Thank you.

DR. WILLIAM RICH: Ditto. I'd like to thank the staff, because I don't even know how to work a watch, so I really appreciate their help.

MS. TURBYVILLE: Is someone on the phone?

DR. JEFFREY RICH: Yes, it's Jeff Rich.

MS. TURBYVILLE: Jeff, hey, welcome.

DR. JEFFREY RICH: Thanks. I wanted to make a comment about the documentation, if I may. Sometimes the applications are so complex, and having them on your laptop trying to find bullet points, you know, while we're talking is a little bit difficult.

I might suggest that having an executive summary of the application would be very helpful, just, you know, basic things like where is the level of attribution, you

know, what time period are we talking about, it's easy -- I mean, because as you go through it and talk and have discussions, I am always flipping back and forth through a very long document trying to redefine that for myself.

And I don't know if I'm the only dysfunctional one in the group, but it would help me a lot, you know, with that. I don't know if anyone else feels the same.

CO-CHAIR ROSENTHAL: I had the same feeling. And, obviously, the applications have to be thorough and detailed and all the rest of that. But as we are beginning to figure out where the honing-in points are -- I mean, for example, I now know that the attribution thing is in Section 11.

But if it could either be highlighted or pulled out, those very key things, what the risk adjusting methodology is in a couple of bullet points, but I -- I hesitated to make the suggestion, because I think the staff has already got a ton of work.

But if it cold be done without making it a ton-plus of work, that really -- it would be helpful.

And also, trying to keep the thing straight, because to now go back and compare one to another is really a big challenge. And yet there are really only a half a dozen elements that are kind of the key ones that we are going -- seeming to come back to over and over now.

DR. BARNETT: And maybe we could just skip importance.

(Laughter.)

I mean, the reason that these -- that we even have these measures is because NQF has already decided that they're important. So I'm not sure we gain much by all of that.

MS. TURBYVILLE: I think the only time potentially it could come in is if it's really not a resource use measurement area, but I think that's right, so --

DR. O'NEILL: Could I just -- I mean, this has been a great exercise, and it is -- I think there is a couple of overarching things. I guess I'm a convert from Jack here. There are a few overarching things that have come up over and over again.

And as a Steering Committee in this area, I don't know if there is any opportunity to do something like advocate for things like pharmacy data to be included. I mean, there is ways of making this happen in the commercial world by having the data sharing be a standard part of contracting with the PBM and things like that.

But in terms of having the resources at our fingertips to analyze how well things are going, there are some overarching things that I think we have kind of learned here. I just hope we can capture them as a kind of policy or, you know, standard that we would like to see on a go-forward basis. And it's outside of each

individual measure, so --

CO-CHAIR STEINWALD: Jack?

DR. NEEDLEMAN: Yes, two things. First, in terms of process, I thought that Sally occasionally asking people who were at the extreme to explain their votes in a non-judgmental way --

(Laughter.)

-- was extremely helpful. And I actually think there may be opportunities to think about straw polls before the formal poll is taken -- vote is taken. Just have us do it once and see where we are, because there is at least one vote. If I had known it would have been as close as it was, I would have switched from one category to another, from a yes to a no.

So I just want to encourage Bruce and Tom to think about straw polls as a way of checking how much consensus there is, how much we need to discuss things, so that might be helpful to use the technology.

The other thing -- I want to thank the staff. The materials were great. Everything you did in real time as we were working was great. Helen, you are a very lucky person.

(Laughter.)

CO-CHAIR STEINWALD: Anything further?

MS. WILBON: I just wanted to thank the Co-Chairs for your efforts throughout the last couple of days. It really helps when you have two good Co-Chairs to lead you through a meeting as arduous as this, and we recognize that the materials are quite challenging, and we appreciate everyone sticking through.

And I also want to thank Jeptha and Jaime, because they have already had a TAP meeting and then they kind of had to rehash the whole thing again and be prepared to that level again. So I don't know if Jeptha is on the line, but I wanted to thank him for that,

and Jaime, who is off the phone, so --

CO-CHAIR STEINWALD: Okay.

MS. WILBON: -- thanks.

DR. BURSTIN: Please leave your voting devices.

MS. WILBON: Yes, don't take them with you.

DR. BURSTIN: They don't work on anything else.

DR. NEEDLEMAN: Did we just get finished an hour early?

DR. BURSTIN: Yes, congratulations.

CO-CHAIR STEINWALD: If we can -- if I can adjourn the meeting.

It's adjourned.

(Whereupon, at 2:32 p.m., the proceedings in the foregoing matter were adjourned.)

G

Q

S

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: National Voluntary Consensus

Before: NQF

Date: 06-30-11

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