

NATIONAL QUALITY FORUM

Moderator: Ashlie Wilbon
July 15, 2011
2:00 pm CT

Operator: Welcome, everyone, to the conference. Please note today's call is being recorded. Please go ahead.

Ashlie Wilbon: Hi, good afternoon everyone. This is Ashlie Wilbon from NQF. Just wanted to...

Male: Hello.

Ashlie Wilbon: Hello? Wanted to welcome you to the Resource Use Steering Committee conference call this afternoon. We're going to be - just going to go ahead and dive right in. See if I can get my slides to turn here.

Today we're going to just do a brief project update and follow up to our in-person meeting last week and we'll review the Scientific Acceptability Rating Process, just kind of, again, in follow-up to the in-person meeting.

We will complete the Feasibility Evaluation for the two Ingenix measures that we started discussing at the in-person meeting, which was the 1599, the non-conditions specific resource use measure, and 1595, the ETG-based diabetes measure.

We'll also review the 1591, which is ETG-based CHF, 1594 ETG-based CAD, both with Ingenix. The cardiovascular diabetes TAP evaluated these measures yesterday, so we've done a quick turnaround of their ratings and discussion. And we will have - we'll share that with you today.

And you will have your discussion today and we'll hopefully wrap up all of the cardiovascular diabetes measures today and the next time we would have you guys together we'll be evaluating the measures from Cycle 2.

So just a brief update, particularly for those - I think all the steering committee members are on the phone were probably at the meeting. But for the public and anyone else who is on the phone, as a result of the meeting last week, or two weeks ago now I guess.

The committee recommended four measures, 1598, which was a total resource use population based index and a complimentary measure which was a total cost of care population based index as well, both from Health Partners, were endorsed.

There was also two measures from NCQA that were recommended - I'm sorry, recommended, 1577, a relative resource use for diabetic patients and relative resource use for cardiovascular patients. So we've got a total four recommended at this point, and we're forging forward.

Is Helen on the phone? Okay. Heidi, are you there?

Heidi Bossley: I am here, yes.

Ashlie Wilson: Okay, I know Helen wanted to talk a little bit about the ABMS measures. Did you want to talk or do you want me to go ahead and (discuss those)? I can do that.

Heidi Bossley: Why don't you do that? I'll jump in. And let me see if she's around, because I know she (wants to be a part of this process).

Ashlie Wilson: Okay, so we had a bit of a development this week from the ABMS Measure Development group. They have decided to withdraw their measures from the project at this time. They felt that they needed more testing before they were put through a rigorous evaluation process.

So what we want to do for the evaluations that have been done up to this point is to provide that feedback back to the development team so that they can use that in their continued measure refinement and continued testing process.

So what that does for the project at this point is it does significantly cut the number of measures that need to be reviewed in terms of rating and recommendations.

We will be asking, though, for the remainder of the measures that haven't been reviewed yet by - particularly by the pulmonary TAPs and by the steering committee, that you do still provide some feedback on them and we will have some structured forms in which we will have some discussions on those measures so that they can benefit from that input.

And I think through the review process so far we've found that everyone was really excited about the measures and really liked the measures and the concepts that they had put together and would like to see those have some work on them in order to be put forth again.

So again, we've been - we appreciate everyone's effort in reviewing those up to this point and your input is definitely valid - valued.

Helen Burstin: And Ashlie, if I could just add in, this is Helen Burstin. I've had numerous discussions with Kevin Weiss over the last week or so and, you know, he very much appreciates the positive reinforcement for the important measure concepts where those measures are trying to go.

But he also recognizes that there's, you know, additional work needed to be done in terms of testing. So I think our thought would be we're not going to have you actually rate them or approve them or disapprove them.

But I think it would be worth, particularly given all the effort on both sides, both the committee side and ABMS's side to have an opportunity for the committees to offer feedback to Kevin and his team so that before they go test the measures, if there are some opportunities for improvement that he could take advantage of that shared learning.

So that's our - what we've worked out with Kevin and ABMS to date. Happy to answer questions if anybody has any.

Ashlie Wilbon: Okay. (Just talk) if there's any questions on that. We do have our final TAP meeting with the pulmonary TAP next week on Tuesday. So that will be the last TAP meeting for the project.

As a result of the ABMS measures coming out of the process for formal review, at this point we do have some updates to the conference call schedule that we had just sent out.

So please make note on your calendar that we're going to be canceling the July 28 call that was scheduled. And I know you guys are all very disappointed that you have two more hours to open up on your calendar.

We will keep the September 14 call and we'll use that time to have an open discussion on the remaining ABMS measures so that you guys can provide some input that we can forward on to the ABMS development team.

The Tuesday, September 20 call will be canceled as well, so please make note of that on your calendar and we will also be sending out updates for those - a follow-up e-mail for this conference call (so for those) who aren't on the call, they can...

Male: What was the second one that was canceled?

Ashlie Wilbon: The Tuesday, September 20.

Male: Okay.

Ashlie Wilbon: Okay? So just a quick sneak-peak into what we're thinking for the next steering committee meeting in August, that you would complete the evaluation and recommendations for the remaining seven measures.

We've got a couple bone joint measures for Ingenix. And we've got some pulmonary measures from Ingenix and NCQA, so that steering committee meeting will be focused on completing the review of those measures. For - we'll have some time to hopefully do some input on the ABMS measures.

And we'd like to have some time also on the agenda for this committee, particularly having been through this two year process at this point in time to offer us, you know, the kind of brainstorms and lessons learned and recommendations for next steps and how we might move forward as we continue to bring - resource these measures into our portfolio and continue to evaluate them in the future. So we're looking forward to that.

I won't spend time on this but these are the timelines now for the Cycle 1 and Cycle 2 review of the measures. And I'm going to transition quickly into the process that we'll use today to review the measures. But I would like to just pause for a second. Does anyone have questions about the kind of logistical project stuff before we move into the measure review process?

Tom Rosenthal: Ashlie, I wonder if it would be possible to take one minute and identify who all we have on the call.

Ashlie Wilbon: Sure. We can do that.

Male: (Thank you, Tom).

Tom Rosenthal: You have the committee membership so can we just maybe run down from the top?

Ashlie Wilbon: Sure, I don't believe we have (Ruth), Tom. Paul Barnett? Are you there? Jack Bowhan, I'm pretty sure he said he couldn't make it. Jephtha Curtis?

Jephtha Curtis: Yes, I'm on.

Ashlie Wilbon: Okay, great. Kurtis Elward.

Kurt Elward: I'm here.

Ashlie Wilbon: Bill Golden. I think he said he couldn't make it. Lisa Grabert?

Lisa Grabert: Here.

Ashlie Wilbon: Ethan Halm? Ann Hendrich? Tom Lee? Jack Needleman?

Jack Needleman: I'm here.

Ashlie Wilbon: Okay. Doris Peter? Mary Kay O'Neill? David Penson?

David Penson: (Here).

Ashlie Wilbon: Steve - oh, I'm sorry. Dr. Penson? Okay. Steve Phillips?

Steve Phillips: Here.

Ashlie Wilbon: David Redfearn?

David Redfearn: Present.

Ashlie Wilbon: Hi. Bill Rich?

Bill Rich: Here.

Ashlie Wilbon: Great. Jeff Rich? Barb Rudolph? Joe Stephansky?

Joe Stephansky: I'm here.

Ashlie Wilbon: All right. Jim Weinstein? And Delores Yanagihara? Okay.

Sally Turbyville: And Ashlie, this is Sally, I'm also on the call just for your records.

Ashlie Wilbon: Oh, okay, thanks Sally. Any other questions before we move forward? Okay. So we'll start with the feasibility discussion for 1595 and - I'm sorry, yes, 1595 and 1599. And I'll help - Tom and I will guide you through that discussion.

For 1591 and 1594, which will be in some ways kind of new review for you guys, you haven't started reviewing these measures yet, we'll have the developer do brief introductions on them and then we will share - we have the visual - the graph that we showed similar to the in-person meeting that will show you how the TAP has rated each of the sub-criteria for each of the overall criteria.

So we'll do that for importance in scientific acceptability, usability and feasibility. We also have Dr. Jephtha Curtis on the phone who was on the conference call yesterday with the cardiovascular diabetes TAP and he will help to summarize and give rationale for some of the ratings that the TAP provided.

We also had assigned, particularly for 1591 a lead discussant for some of the sub-criteria. So if those individuals are on the phone and would like to offer input, we can do that as well.

And for usability and feasibility, I think some of the discussions that we've had from the previous discussions on Ingenix measures will probably apply to this and I can - Tom can guide you guys through that discussion to see if there's anything new or different for these measures that would need to be discussed while we're on the call.

I would also like to point out that we did send you links via e-mail to have you vote and submit your final recommendations on the measures. So we'll have you - you can have that up during the call and submit them during the call.

We won't be able to show them in real time just based on the technology that we have at this time, but we would encourage you to do it while you have time to sit down and do it during the discussion if you're able to and you're at a computer. So those are the links. And we can resend them if anyone needs them.

So again, we're going to go through a systematic - use a systematic review process, we'll sequentially address the criteria and we'll be evaluating the measure as submitted.

So I just wanted to do a brief overview of the rating of the reliability and validity. I know the rating system that we use and the process that we use at the in-person meeting, and the use of the algorithm and that table that we were referring you guys to, was probably some confusion over how that applied and how your rating for reliability and validity fed into the table.

So what we're going to have you do going forward is to have the steering committee rate the overall reliability and validity on each measure. Even though the TAP has done that, we want the steering committee to do that as well. And then we'll ask the steering committee to, again, vote on overall scientific acceptability, which is a yes/no, kind of pass/fail criteria.

And then what we'll do (as staff) will also apply our algorithm that we implemented in NQF process and when we report - when we draft the draft report, we'll be including both what the steering committee voted as well as what the algorithm result would have been based on your ratings. So just to kind of have a process clarification there and how we're going to be doing that going forward, so again...

Tom Rosenthal: Can I just clarify one thing on that, Ashlie?

Ashlie Wilbon: Sure, Tom.

Tom Rosenthal: What that means for the group is that that grid will drive an independent score for scientific acceptability, so we might need to keep that in mind as we rate reliability and validity and in fact in some sense weight the elements within it ourselves, because if somebody thinks the thing isn't scientifically acceptable, then it will be more important to indicate that by either the reliability or the validity score as you score that; otherwise, we'll end up with a kind of mixed scoring. Does that make sense to people?

Male: Yes.

(Taroon): Tom, this is (Taroon). Ashlie and I will go through the gridding - the grid so the committee will be aware of exactly how the grid will work. We just wanted to clarify this just so that we have it going forward.

Tom Rosenthal: Right, I think this is a perfectly reasonable way, but - and the back-story to this is that in the interim from the meeting there was some consideration that what we should have done as a committee was simply have applied the grid for the reliability and validity scores and not voted on scientific acceptance.

But of course we did vote on scientific acceptance. And at least in one instance our overall scientific score didn't exactly match the way the grid would have come out. And so we can - this is a good compromise to have both elements to go forward with our recommendations.

But it does mean that if you really think the scientific acceptance is going to be a no, then it probably would be good to also indicate that via which element it was that caused you to rate it that way, either reliability or validity. And it requires a low score in order to do that.

Ashlie Wilbon: Right, thank you Tom for the clarification.

Tom Rosenthal: Okay.

Ashlie Wilbon: Any questions on that? I realize - we realize it can be confusing and we're doing our best to make it as clear as possible. So if you have any questions along the way, you know, please ask, because if you're wondering, there's probably someone else who has the same question. So, okay, thanks.

So for the high rating for reliability and validity, what you're saying, if you rate something high is that they're - the specifications are unambiguous, anyone who would apply the - I'm sorry.

All the measure specifications are unambiguous. And they're likely to consistently identify who is included and excluded from the target population, also that they have evidence of reliability of both the data elements and the measure score, okay?

For validity, a high rating means that the specifications are consistent with the measure intent and with the measure focus, that they identified any important criteria and that they have evidence of validity of both data elements and the measures score and that the measures score is - they've used the appropriate type of validity testing, and the result is within acceptable norms.

So that also includes if they've identified threats to validity and that they've addressed those threats, okay? So for moderate rating for reliability, you're saying that all the measure specifications are unambiguous and that there's empirical evidence that the reliability testing has been done on either the data elements or the measure score. They don't have to have it on both.

And then similarly for validity, the measure specifications reflect the intent that they've cited and that they have evidence of validity on either the measure score or the data elements, or that they have systematically assessed (face) validity of the measure score, okay? So - and that also includes that they've identified threats to validity and addressed those.

For a low rating on reliability and validity, it means that one or more of the specifications are ambiguous, there's potential for confusion on who's identified and the target population; how the resources are - or costs are being measures; or how to compute the score, all right; or that the evidence - the empirical evidence they provided is unreliable for either the data elements or the measure score.

And similarly for validity, the measure specifications do not reflect the intent or the evidence they provided on validity is - they provided an invalid approach to proving validity of their data elements or of the measure score. And they have not identified threats to validity and they have not been empirically assessed.

For an insufficient evidence rating, it means they you don't have the information in front of you you need to determine whether or not it would be high, moderate or low, or that they've used an inappropriate method for validity testing or inadequate assessment of (face) validity.

So that is how we would frame your ratings of high, moderate, low. And I'm going to let (Taron) kind of walk you guys through the grid.

(Taron): Okay, so...

Steve Phillips: (Can I just) ask a question about insufficient evidence? This is Steve Phillips. I mean is that considered basically equivalent of a low rating? I mean if it's not there then it's kind of a fail or is there an opportunity then to go back and submit more evidence?

Helen Burstin: This is Helen. Certainly it's always a possibility if there's additional information to be gotten, to try to bump up an insufficient evidence. That's certainly something that we can try to do in the timeline of the project.

Tom Rosenthal: And Helen, I have one. This is Tom. The validity and reliability criteria were taken to a large degree from the quality review process that's been going on for years, right.

Helen Burstin: Yes, so the way reliability and validity, at least with scoring mechanism, came out of the work our testing task force did last year, yes.

Tom Rosenthal: Okay, and the only point to that is that what did not exist in the quality arena was this attribution question of is the measure being properly attributed to the group that it's purporting to report on, which I don't think really ever existed in the quality world.

I mean if a hospital is doing heart attacks they're responsible for aspirin. And there's no confusion about who the measure's being attributed here to. But in the cost - these cost things this attribution question does come up. And so I assume that attribution questions would fall into this reliability and validity thing, even though there's no explicit statement about attribution in here.

Helen Burstin: Yes, it's definitely tricky, the whole attribution issue and quality measures overall is pretty unclear. In some ways the way we handle the level of analysis or shared accountability for like a readmission measure is somewhat analogous.

I think what's different in these measures is an actual quantification of the degree of attribution to the different providers, which I think is different.

Tom Rosenthal: Right, and it's - that's what I'm saying. I think it's almost unique to this cost world that we're trying to grapple with. And therefore we probably - in fact in our statements about reliability and validity, probably should have added something specific about attribution. But I think the committee has been doing that as we analyze the ones we did in the face-to -face meeting.

Helen Burstin: Yes, Sally may want to weigh in on that, because I know there were a lot of discussions in the early phases of the sub-criteria in scientific acceptability where, thought through, of at least asking for the information and attribution but not necessarily considering it. Sally, do you have any insights there?

Sally Turbyville: Sure, there's a couple of things in thinking about attribution. There's some similarity certainly in quality measures to some extent. And I think Helen already touched on this.

So whom you're going to hold accountable, even in the quality world, there's usually a request by the experts that who the measures are being attributed to, or which level has some kind of control over it. As your example, Tom, aspirin in the hospital, that seems to make sense.

In the resource use measure I think some of the discomfort also becomes - is because it's new. But also there's this other level of attribution, which is how do you split up the resource use measures among physicians or health plans.

So there's the level, so - as I just mentioned - physicians or health plans. But there's also are you identifying, you the measure developer, the appropriate peer group? So is this research use measure appropriate for a peer group of cardiac surgeons.

So there's a layer of questions that come up in resource use measures that are new in many ways. I say new without meaning they're totally new because they do come up in quality measures.

When we were thinking about how this would play out in the evaluation process two things came up. One was that the steering committee last year stated that they wanted the measure developers to have the ability to provide options, that in different parts of the country, they may want to take that resource use measure.

They may be fine with the peer group but they may want to change the attribution a little bit. But that they wanted the measure developers to provide clear guidance that had some logic behind it, it made sense, given the measure they were proposing.

They didn't want it to stay silent. We've allowed for some flexibility in the collection of that information from the developers.

As far as when you're evaluating what the measure developers have submitted, whether it's specification so they only give one approach and that's it, or guidance where they're providing options, you are right, Tom, that we would expect you to think of that in terms of validity.

So are the attribution approaches, are the peer groups that they've selected for the resource use measure under consideration valid? Do they make sense? So if it's a resource use measure for pediatric asthma and they're saying the peer group is cardiac surgeons you might say that hurts the validity of the measure.

Tom Rosenthal: Okay, so even though it's not specified in these criteria, et cetera, it's appropriate to consider the attribution questions in the validity realm?

Sally Turbyville: Yes.

Tom Rosenthal: Okay.

Male: But to take that extreme example, I mean that's a case where I would say the attribution is clearly wrong. Now if we think that they've got the resources correctly measured, that they've done an incredibly good job of identifying asthma care resources - so that part is right. And they patient level stuff is right. And the attribution is wrong.

Is that low or is that medium? Is the attribution so central to these measures that if we don't believe the attribution algorithm we should reject them?

Tom Rosenthal: Well, I think that's sort of - that's certainly a question. And we can debate that in theory, or we could wait and see how it applies in specific cases. I guess in my mind, again it falls back to, is this being used for improvement efforts or is this being used for public reporting.

And the public reporting piece is what for me makes it - if somebody's going to publish this group of pediatricians versus that group of cardiac surgeons on pediatric asthma, it's going to be invalid. What is produced for the public will in fact be invalid.

So for me personally the attribution thing is a critical factor, but mainly because of the hurdle that we have - that we all agree is appropriate around - that this is for public reporting and not just quality improvement.

Male: Okay.

Tom Rosenthal: But that's just my opinion. I mean people are clearly free to have their own opinions and others may differ about that.

(Taroan): Well, I would just offer one piece of guidance, I guess. This is (Taroan) from NQF. I just want to remind you that the attribution is - we allow the option of either a guideline or a specification.

So if the developer specifies as a guideline, it is an option in the tool. And thus it should be evaluated in that sense, because we did offer them the option of a guideline versus specification. So the specification would be built into the measure as opposed to a guideline which would allow some degree of flexibility.

Tom Rosenthal: I guess, (Taron), the thing I'm confused a little bit about that is the specification seems to me to be so central to the thing. If there's merely a guideline but the guideline has brought ambiguity - and let's use the kind of extreme and ridiculous example, but a good one for the sake of trying to sort this out.

You know, we would prefer if you compare only pediatricians to pediatricians. But basically you could, in fact, or pediatric pulmonologist to one pulmonologist, but in fact you could use any pediatrician to any pediatrician.

In the actual use of the thing you could have somebody in fact applying a cardiac surgeon against a pulmonologist and you'd have a really completely invalid outcome.

(Taron): Tom, I agree with you 100%. I think one of the challenges though, as we want to, you know, as we sort of try to maintain some internal consistency in the process, is that because the process was set up with the guidelines and specification, a developer who offers the option of multiple attribution approaches and doesn't really give a lot more detail beyond that would be - should be judged similarly to a measure that offers sort of a detailed description of one option of an approach and also specifies that as a guideline, because they did not select it as a specification, the current criteria that we've established as a steering committee.

So agreed, all those principles are important to consider. And I would encourage you to consider them as you see fit.

But as we sort of think about the criteria, because they've been specified a lot - sorry, because they've allowed the flexibility in the specifications (first) guidelines, we have to consider that the - allowing it as a guideline allows a certain level of flexibility, or should be evaluated in that sense.

Tom Rosenthal: Okay, well my suggestion is is that beat this up enough and we'll probably continue with every individual measure to revisit our own notions of what constitutes scientific validity in the face of measures that don't, you know, that don't have an extensive peer review basis to them. But rather than continuing to talk about this at 30,000 feet, Ashlie, should we kind of move to the measures?

(Taroon): Yes, that's fair. I'm just going to - just want to take a quick second on the matrix here just to understand how our internal NQF algorithm works. So if we have a high validity rating and a higher or moderate reliability rating that would pass scientific acceptability. While a high validity rating and low reliability rating would be a no in recommendation.

Moderate - and this is - I do want to spend some time making sure we all are clear about this. A moderate validity rating and a moderate to high reliability rating would be recommended for endorsement whereas a moderate validity rating and a lower reliability rating would be a no for recommendations for scientific acceptability.

And finally, low in validity and low in reliability would be no for scientific acceptability as, Tom, you said right from the beginning of this conversation. So really the threshold here is low - you really - if there is a low rating then that's really the only, you know, that's really going to drive this to essentially a no, you know, on passing scientific acceptability.

Are there any questions about this? Okay, thanks.

Ashlie Wilbon: So to Tom's point I think we'll just go ahead now and transition into the actual measure evaluation process. And why don't we start with the feasibility discussion. And what I'll do is bring up the feasibility sub-criteria and we can walk through each of those.

And, Tom, I was going to run by you, is it - do you think it's - rather than discussing each of the measures individually we could just kind of discuss feasibility...

Tom Rosenthal: Yes, it's applicable to all four of these Ingenix measures.

Ashlie Wilbon: Okay, okay, that sounds good.

Tom Rosenthal: But we can shortcut this a little bit by having basically one feasibility study, because the questions I think will be the same for all four of these.

Ashlie Wilbon: Right.

Tom Rosenthal: So if people will recall the 1599 and 1595, we did a full discussion at the face-to-face meeting on important scientific acceptability and usability. And we tabled the feasibility discussion because we did not have the cost. And now we have the cost.

And I think that was the major issue that needed resolution, possibly among some other issues. But I think that's mainly what's on the table here in relationship to feasibility.

Helen Burstin: And, Tom, this is Helen. So, you know, the cost data was the one thing we were missing the last time we met. So it's one of the - it's a sub-criteria under feasibility. But you should now be able to look at the whole criteria - the whole criterion including the specific one around how the cost effect would be factored in.

Ashlie Wilbon: Okay, thanks. So for those of you that are on the webinar, I have up on the screen the four feasibility criteria. 4a and 4b are ones we've probably not spent a lot of time on.

4a asks whether or not the required data elements are generated routinely - which obviously these measures use admin data so we don't spend time on that. The 4b focuses on whether or not the elements are available - the data elements are available electronically, again for admin data, not really an issue.

So we'll just ask to focus your time on 4c and 4d. 4c looks at whether or not the measure has addressed - or there are susceptibility to inaccuracies, errors or unintended consequences, and whether or not the developer has identified those and how they can be minimized, monitored and detected.

Tom Rosenthal: All right, so this is open for discussion. Do we have the TAP report on this, by the way, Ashlie, just to get us started?

Ashlie Wilbon: Well, there's...

Male: Which (code) do we start with?

Tom Rosenthal: 4c.

Ashlie Wilbon: 4c.

Tom Rosenthal: The data's not inaccurate or, et cetera, basically.

Ashlie Wilbon: Yes, I have - well - you guys have done a discussion of 1599 so you've done an in depth review of 1599. You can kind of start off with your knowledge of that and I can pull up the TAPs ratings for 4c for 1595 if you want to do that.

Tom Rosenthal: Well, that's fine. I was just wondering if we had a TAP thing on feasibility for everything other than the cost, or not.

Ashlie Wilbon: Not specifically, no.

Tom Rosenthal: Okay, then...

Ashlie Wilbon: TAP rated it but it's not probably going to help much (in this).

Tom Rosenthal: Okay, so this is on us. So it's open for discussion.

Male: Well, we've got the same issue of the carve-out here for, you know, and the availability of carved out data and mental health and pharmacy costs.

The Ingenix people at least - my recollection is the Ingenix people did not speak to mental health carve-outs, did speak to drug carve-outs, and said the way they were addressing that in this measure was by stratifying by whether or not the pharmacy data was available in the estimates of resource use.

Tom Rosenthal: Jack, do you think that these issues are sufficiently well addressed in this to enable a positive determination around 4c?

Jack Needleman: Well, you know, in contrast we'll say the NCQA folks who said, we go back, we audit, we know these are supposed to be included, our plans are supposed to get it, you know. So we had their assertion that the numbers were always there.

Here we've got, you know, the Ingenix people are not saying the numbers were always there.

What they're saying is, at least with respect to pharmacy you can stratify and look at the resource estimates where we've got pharmacy and where we don't.

I'm a little concerned about the mental health side of this in a non-specific condition thing, given how many people suffer from depression in the country in any given year and the role of those drugs, both drugs - not just drugs but therapy in that area. But, you know, the question is how much error does that produce in these estimates, and how systematic is it?

Male: And what's your judgment?

Jack Needleman: I'm at the moment floating between low and medium. I certainly wouldn't go higher than medium on this particular item, given...

Male: Given those issues.

Jack Needleman: ...folks have not seem to have solved - thought about the carve-out issues enough to solve them.

Tom Rosenthal: Okay, all right. And then the other issue that occurs to me - and I can't remember whether we talked about this last time - and this one is literally almost a matter of one's philosophy - is what are the various inaccuracies in administrative data generally, coding inaccuracies, coding variation, this, that, and the other.

Male: (And you plan to add) that to the list of stuff.

Tom Rosenthal: Right. But it's almost an act of faith - it's almost an act of faith as to whether you believe that's a serious impediment or just the nature of doing business using administrative data. Other comments on this?

Bill Rich: One other issue here, Tom - this is Bill Rich - is did he say he'd talk about sample size, and since small sample sizes are customarily dealt with there's no discussion like we had last time of what statistical criteria apply to small sample sizes, how they're dealt with, how they define them.

But I'm favorably impressed that at least they addressed the issue. But it's not clear in my mind what they actually do with small sample sizes, whether they get published or not.

Tom Rosenthal: Okay, other issues related to susceptibility to inaccuracies, errors or unintended consequences?

Male: So what did the TAP say, I'm still struggling to find that document, about this issue?

Tom Rosenthal: Actually, did the (TAP) weigh in on 4c?

Ashlie Wilbon: For 1595 they did. If you recall 1590...

Tom Rosenthal: That's fine, because it's the same on all of these. So did they weigh in on feasibility for any of these Ingenix measures?

Ashlie Wilbon: Yes, and I can bring that up. But again it'll be - give me one second here.

Jeptha Curtis: This is Jeptha, so I'm - well actually looking for the formal documentation. I mean, I think we struggled with it from start to finish in that, you know - and it overlaps with what we were discussing with validity for these specific measures.

You know, there is a wide range of opinions as to how accurate they are at measuring what they're purporting to measure. And there are some people who believe that they were and there are some people who believe that they weren't.

And there I think was a fairly wide range on the votes, although I think overall it was sort of a moderate. There was some good faith that in fact it is, in general, measuring what it's supposed to be measuring, at least for diabetes.

Male: That's sort of a validity...

Jeptha Curtis: What's that?

Male: That's a validity issue, not a feasibility...

Jeptha Curtis: Yes, I know. But it had to do with, at the end of the day in terms of, I guess, feasibility and the susceptibilities to errors, that question of, you know, if it's not valid is it feasible, or is it susceptible to errors?

I think the other part was, you know, as you addressed, or someone brought up the issues of the small volumes, basically there's really just the guideline as to how to approach it but there was really no content that supported that approach or demonstrated how that approach could be used, although they may have submitted some documentation thereafter that does support it.

Tom Rosenthal: Alright, well here's the vote on the screen for those of you that have it you can see it, but for those of you that don't the TAP on 1595, and again I think these were, for the purposes discussion, considering these to be fairly similar if not identical.

The vote was four low, two medium and two high on 4c susceptible to inaccuracy. So I think that probably reflects the discussion we're having here as well although I haven't heard anybody, at least on this conversation yet, speak to the high side on this one.

Male: Yes, I just wonder what those four low votes were, is it about the sample size and attribution stuff or...

Jeptha Curtis: I think it had to do with mainly about the attribution in that even though it's a guideline and again I think (Taroon)'s comments were useful and might have been helpful earlier but we weren't really sure what to do with attribution and where to put it in the equation here.

But if you take the worst case scenario, which I think the TAP did in this particular discussion of the diabetes measure, if you were to apply to a single physician, are you getting a useful result or is it an accurate result?

Male: Okay and in that regard it does overlap the feasibility conversation.

Jeptha Curtis: Exactly and I mean they're hard to differentiate.

Tom Rosenthal: They are, particularly on 4c.

Bill Rich: I think what (the issue is here), Jeptha, this is Bill Rich again, is the reason it's not specified is that different people take these and use them in different ways whether they're self-assured or planned so they each define what a low number is.

I think that's probably why it's fairly generic. Maybe the ((inaudible)) staff could comment on that. But that was my major issue was low sample size and that's been our evaluation in the past we've looked at.

Tom Rosenthal: Okay, and then any other comments on 4c? Then let's do the last part on the feasibility which would be 4d. And just as long - Ashlie, leave that on the screen for two seconds because we can give you the TAP vote on this was five high, two medium and 1 low on data collection strategy can be implemented.

But they didn't - the TAP did not have the benefit of this cost question, and Helen, correct me if I'm wrong, but this is where the question about the cost of access to the data becomes relevant to the feasibility point right?

Helen Burstin: Exactly, so the one sub-criteria.

Tom Rosenthal: Right so, Ashlie, bring up 4d back on the screen if you could.

Paul Barnett: So Tom, just to back pedal, this is Paul Barnett, just a moment, I would just observe, in terms of 4c, and the unintended consequences that I do feel a little bit better about this episode-based method than something that could be, seems less prone to gaming than we were talking about some of the other measures.

That is, there's not much you can do too much to manipulate the start or end of an episode that would seem to be beyond, in order to gain the measure. So unlike, something like the per person, per, you know, capita, per year, you could game it just by bringing in people for well checks and game a measure

That - this kind of sophistication of this, would seem to preclude that, so you're saying nobody seemed to have too much positive to say. I just won't point that out. There may be other drawbacks, but there is, because of this sophistication, it does have that advantage.

Now the question is, is it worth that much money, does that make it feasible, is the next issue, yes.

Tom Rosenthal: Well thanks for the positive comments cause if in fact the vote turns out positive the staff will be wondering, all we heard were these various negatives - how'd the vote turn out that way so that's helpful. So if anybody would like to further comment on 4c?

Okay, then let's go to 4d, and again the real issue here I think is the cost. And again I'm trying to get us to that and now on the screen for those of you that have it, the webinar is the small medium and large client range for MD's and small medium and large for covered lives if it were applied to a health plan.

And for physicians it's small is \$70,000, medium is \$90,000 and large is \$110,000 annual cost. And large is over 2000 physicians and small is under 800.

And then for health plans it's small, medium, and large is \$90,115 and \$135,000 per year. Small is less than 400,000 covered lives, and large is over a million covered lives. And then there's a little governmental discount for the health plan piece.

So that's the general frame of reference on what this thing costs. And, (Ashlie), the one relative comparator that I think we have in terms of trying to be internally consistent was the Hopkins

Male: ACG.

Tom Rosenthal: Yes, ACG thing and what do you recall approximately the range of cost on that?

Ashlie Wilbon: I don't off the top of my head, although I believe the distribution of pricing was a little bit wider so it started at a lower amount and so between low and high. (It's like) between a small

practice and a medium sized, it started at a lower amount and then the moderate was a little bit higher.

So it's a little bit - it's hard for me to say, but again, I think, to try to steer away from - I understand your point internal consistency. But as much as possible we want to try to make sure we're evaluating the measures, or sorry, evaluating the measures individually and try not - you know, it's a fine line. I understand your dilemma but just the kind of, you know...

Tom Rosenthal: (Well is) the frame of reference anywhere the same? I don't think anybody on the group is going to go, wow that one was 35 and this one is 50, therefore, this one's too expensive.

But I personally think there is some notion of internal consistency that if that one was 2000 and we passed it and this one's 100,000 we might have a discussion. But if that one was in fact in the same general range and we approved it, I'm not sure on what basis we would have then to disapprove this one based on cost. That's all at least for me.

Ashlie Wilbon: Right, I understand.

Tom Rosenthal: Because I do think we are trying to achieve some level of internal consistency and fairness about our thing and we shouldn't one day do something that's exactly in contradistinction to something we did the previous day because we forgot what we did.

David Redfearn: This is David. I think the prices were about the same.

Tom Rosenthal: That's what I think but I just I'm just trying to...

David Redfearn: In my opinion you're both way, way too high. I don't think there was any material difference between the two products.

Tom Rosenthal: Okay, well that's what I think too but now...

Male: (I think the) major concern was the lack of transparency rather than the pricing with the ACG things.

Tom Rosenthal: All right, well let's discuss though the issue about the cost on this because again, Helen in particular, has brought this one up in particular in relation to Ingenix. I think one other difference between this one and the ACG one was, this is license fee to Ingenix to be able to use their proprietary system, which I think is slightly different than the other one.

But let's open up the general question, what do people think about the cost issue here and on this?

Cheri Zielinski: Can this be from Ingenix? If I could just caveat that last comment about the cost, the cost that we provided is not only for the licensure of the proprietary software, it's for all the measures. There's over 558 ETGs in our product.

The software documentation, support, control reports and other output, it's quite - it is not just the cost for that one single measure and the proprietary nature of that. I just wanted to be clear on that.

Tom Rosenthal: Alright, thank you for that verification. Point well made. Open for discussion.

Male: I'm just curious what do the three different tabs represent? One says ETG, one says ETG/ERG, and the other one says ETG/PEG.

Cheri Zielinski: Sure, some of the measures that we submitted for endorsement considerations are related to ETG only, for example diabetes, is an ETG specific relationship. We add our non-condition specific, population based measure which is composed of both ETG and ERG, which is our risk-assessment grouper and our episode treatment grouper.

And then we have another measure that we submitted for - in consideration that involves our PEG and ETG methods and that would be our hip and knee replacement.

Male: But the comment says that the pricing is for the entire suite, so you have to buy the whole thing to do any part of this, in essence.

Cheri Zielinski: But the pricing is for all the measures, so we have six products within the Symmetry suite. ETG is one product in that suite, ERG is another, PRG is another. So depending on which measure you would have to either select one product from the suite or two, ETG and ERG, or ETG and PEG.

Male: So you add together if there's more than one measure then you start adding together these amounts that are under the different tabs?

Cheri Zielinski: Again, ETG has multiple measures in the product, we're talking about products, software products. And so, ETG has over 500, you know, measure, quote unquote, measures or episodes so diabetes, CHF, CAD, depression, fractures, all those are inclusive of the cost.

Male: But if I could clarify, your point is well made. You get the whole package. But in fact if you were an entity that wanted to use this particular measure, and this particular measure only, you would get a whole bunch of other stuff that you might or might not want to use.

Cheri Zielinski: That is correct, we do not provide pricing for one single measure, we do not provide our software for one single measure. It is - the measure is a component of the whole.

Male: And can they be used without, could any one of these be used without availing yourself of all of the technical support of Ingenix in order to be able to implement it?

Cheri Zielinski: We do not license the measure without licensing the software itself. It's a comprehensive package.

Male: Well, but I'm going one step further, in order to actually use the software would you not need the resources of Ingenix to basically consult and help you figure out how to use it?

Cheri Zielinski: You could go that route if you want, I mean that's included in the cost. We don't break that out.

Male: Is it usable without that help?

Cheri Zielinski: It would be complicated, but sure it would be useable.

David Redfearn: This is David Redfearn. Having gone through this process many years ago, it is technically possible to take their documentation and work through it. It's challenging though, but you could, you don't have to have the handholding. You could actually read the documentation and probably figure it out, but it's difficult.

Male: Well but this gets to feasibility a little bit. I mean is that really feasible?

David Redfearn: It's probably not ideal.

Male: Is it feasible? With whatever definition of feasible we want to apply here?

David Redfearn: It's possible but not very feasible.

Male: Right. So my question was, if you needed, if you had two, you chose two measures and one was ETG/ERG and the other required ETG/PEG you'd actually have to pay, if you were a medium-sized payer, \$150,000 for one and \$150,000 per year for the other.

Cheri Zielinski: That is not correct. The PEG and ERT are percentages that are added on when you buy ETG. So it is not an additive effect, it would be a percentage of the cost of ETG.

Male: Well but just to clarify, it is additive, it's just not geometric.

Cheri Zielinski: That is correct, I apologize for that.

Male: So you would pay more, it's just it doesn't double or triple.

Cheri Zielinski: You wouldn't ((inaudible)) ETG twice. In the tables that are provided those are independent instances where you would buy ETG and ERG and then you would buy ETG and PEG. You could also buy in one package ETG, ERG and PEG and that would be not additive of those two costs.

Male: Well it seems like the ERG adds \$35000 a year for the medium payer.

Cheri Zielinski: Right.

Male: And PEG would add another \$35000 per year...

Cheri Zielinski: That is correct.

Male: If you needed both.

Cheri Zielinski: So it'd be 10 plus 35 plus 35.

Tom Rosenthal: The way this would work is for 1599 you have to license ETGs and ERGs, because you have to get to that ERG to get the risk score. For 1595, all you need to license is ETGs because you're only interested in diabetes.

But what you're getting when you license ETGs is you get a module, a grouper that will generate all the various 500 different ETGs, not just the diabetes one. And the one can sign - I know, Cheri has indicated, that Ingenix doesn't license just individual ETGs, but all you're interested in in measure 1595 is diabetes.

Yet you're buying a product, an expensive product that does everything, that has the whole thing. I just would ask Ingenix whether they have any, have ever thought about the fact of creating a pricing module in which the model runs and it just spits out the diabetes ETGs, if you're not interested in anything else, and whether that would have an effect on price.

I mean, I think frankly, my opinion on all of this, and the ACF stuff too, is that this is so expensive that it would be prohibitive. If you don't already licensing Ingenix, maybe you're licensing the Medstat grouper.

Why would any big carrier that has, you know, a million plus members want to pony up all this money when they already have an existing ETG - existing episode methodology in place that's not ETGs just to run the one measure? That's my question.

Male: Well and another question about this is, if you're going to do it just to run the diabetes, wouldn't you have to run all your claims data through the system and get it all cleaned up so that it can run through it?

Tom Rosenthal: Oh yes, you'd have to do that because you have to be able to distinguish diabetes from all the other things.

Male: But it's only about, less than 10 % of it probably has to do with diabetes so that kind of makes it not so feasible. It seems a little bit cumbersome. In other words...

Male: It is. It is...

Male: So (including) 100% of the data in order to look at (18%) percent of the data seems...

Dan Dunn: Actually if I can - this is Dan Dunn from Ingenix. But you would have to run all the data to find the patients who have diabetes and then to sort out which services are related to diabetes or not. I don't think there's any way around processing the entire population to do this measure.

Tom Rosenthal: But it would be fair to say that this is not a trivial exercise.

Dan Dunn: I think any of these measures that you're likely considering, (well), trivial exercise, any of this is work. But any of the measures you're considering, you know, if I was a health plan or a provider organization, I would be preparing all of my data for processing to, you know, to be able to identify episodes for certain conditions or describe a patient from a, you know, population based perspective.

Tom Rosenthal: Is there any distinction among the group about the pricing and the feasibility vis-a-vis application of this at the health plan level versus the application of this at the physician level? Is there any differentiation or is it the same issues equally applicable?

Male: Well, the small, definition of a small physician group is very large.

Male: Yes.

Male: So, I mean, whether you're looking at a group of 2 or 10 or even 100, you're looking at the same \$90,000.

Tom Rosenthal: Has it been, my one question for Ingenix, has it been used at a small, and let's - (for the best description) small being might be more like a couple of hundred physician group. Has it been applied in that setting?

Cheri Zielinski: We do have a physician groups that utilize the product. The challenge is those physician groups having all the data they would need to run a successful run of the product. And that would include as you can see from our specifications, you know, lab data, hospital data, pharmacy data, so I think...

Male: Well and the use of other physicians right? If they got their care from a different outside that group, then it's not very comparable would it?

Dan Dunn: No, it would not be.

Cheri Zielinski: So I would say that the product can be used. Larger physician groups have more accessibility to that level of data and therefore, can run the product successfully or more

successfully. But we do have smaller employer groups that are more self-contained that could have access to that data.

Tom Rosenthal: Well, I get it on the employer side, or even on the health plan side the size isn't such an issue there because you have all the data. What's the - and frankly you probably have the IT infrastructure to, as Dan said, process your data and get it through the system.

Most physician groups - many physician groups don't. And so can you share with us an example of - without, I don't know whether this is proprietary or whatever, but what - an example of a physician group that's using the tool?

Dan Dunn: Do we need to name them Tom, or can we just describe them?

Tom Rosenthal: Yes, I think describe if you're not comfortable naming them. Just describe them and what it was to get it set up and go on.

Dan Dunn: Yes, and maybe the tricky thing here which I know people keep bringing up is often when this is implemented, you know people are looking for not just a diabetes measure. They're looking for you know support across the clinical breath of medicine, especially if it's a multispecialty organization. So you know, there aren't any obviously that are implementing just to support the diabetes...

Tom Rosenthal: Right understood.

Dan Dunn: ...measure. Okay, a couple of examples and I'll skip the large provider entities like in California for example. But you know there are some you know multispecialty organizations who have you know decided to take on risks or do improvement, who are you know taking - you know

licensing the tool and working with either because they have global risk arrangements, they have all the data or are working with their health plan partners to pull information together.

There are - there's not a lot of those. They're not the norm but there are those organizations, probably a little bigger than us, 100 physicians and maybe a ((inaudible)) on the other side. I don't know of any sort of whatever, 20 physician multispecialty practice or even your primary care practice that's taking this on.

Tom Rosenthal: Okay well thank you. Are there questions from the group?

Male: Yes this one is to NQF not Ingenix. And I understand you know. We're looking at specific measures and the cost of implementing specific measures, but Tom's question about comparability to (atg)s was one example.

But what's been the experience in terms of endorsing proprietary measures with price tags and other measured sets? Are there - where did the numbers we're looking at here fall onto the range of what's been considered and either accepted or not accepted as acceptable cost?

Helen Burstin: Yes this is Helen. We've had a very limited experience with bringing in proprietary measures. We had a couple measures we looked at in the past from a couple of proprietary groups in our outcome safety projects a couple of years ago.

The measure didn't make it through, so that there wasn't really for other reasons more on the sides of acceptability side, so there wasn't really - it didn't come down to as much the feasibility discussion.

And I don't know what the comparable data is but we're literally talking about you know a handful of measures. We currently don't have any measures endorsed that I know of that are proprietary

with fees, but again we knew walking into this researchers project that this was a very different space.

Tom Rosenthal: So we're in sort of uncharted waters here a little bit?

Helen Burstin: Yes.

Male: Helen this is...

Tom Rosenthal: So the nine of us are going to get a...

Male: ((inaudible)), to respond to my questions before about the lack of specificity about having handled small sample sizes?

Tom Rosenthal: Small sample size question. I don't - when (Bill Ray) raised that earlier, I don't think that you - the Ingenix folks might - it might or might not have been on the call.

Dan Dunn: Yes I was, Tom. I can help here. I believe we followed up with - what I think I get from our general response which you probably have in front of you was, you know, the issue around small sample size. And obviously a source of potential misinterpretation and accuracy was the use of confidence intervals when making comparisons either to a benchmark or across providers.

I believe we followed up with the actual formulas that you would use for a confidence interval that we were proposing.

And in terms of - we actually did a couple of that with a minimum number of observations, but you know we could certainly do that. But you know I think if you add a sample size of one, obviously a

very significant confidence interval is not a very useful number and wouldn't use - or anyone put any weight on it.

Male: I appreciate it ((inaudible)). Is the calculation - the confidence interval for all sizes ((inaudible)) cut off where it's done? Is that part of the software package or is it up to the discretion of whoever leases the software?

Dan Dunn: Okay. And maybe this is getting off track here a little bit, but you know if you think...

Male: I don't think so.

Dan Dunn: I'll try though. Yes, one of the things with -- and maybe this is up to us to decide -- but one of the things with ETGs is there are you know folks that will you know license the ETG software, which basically gives you, you know episode results at which then people will take and use for different applications including physician measurement, and those folks you know could make decisions on sample size, et cetera.

The same thing that there are other software vendors out there who embed ETGs within their ((inaudible)) and you can use that for their purposes.

So there's some flexibility in how they are doing in our own application of ETGs which was guiding the measure here, and we provide a threshold of 30 observations for a certain measure of which we are arguing now, and you probably don't even want to put any weight on it, and on top of that we'll suggest these confidence intervals.

So there's a level of which is ((inaudible)) this thing that describes the practice of other physician. Then on top of that meaning that you even should be putting something out there, but on top of

that the assessment of, you know, what's the sort of confidence interval around whatever you're going to propose.

Male: Thank-you, sir. That was very helpful.

Dan Dunn: Yes.

Tom Rosenthal: Okay other questions on feasibility, 4D in particular. All right. Well if not, we're not voting the way we voted in the face-to-face because that's just - as I understand it we don't really have the set up to do that and then report the results back, but we do have the ability to go on the survey monkey.

And I think the suggestion of staff was that - is everybody at a computer and can do this? That now is the best time to render a judgment on the thing or you can write it down and go later, but I would suggest we take 30 seconds and go ahead and do that. Is there any objection to that?

Male: ((inaudible)) question for staff. This is (Bogen), Tom. You know on 95, there is only an up and down vote. It doesn't have the four elements. Is that intended?

Ashlie Wilbon: Correct.

Male: Well 99 does have all four elements of...

Male: That is correct.

Ashlie Wilbon: Well for 1595 you already rated - for 1595 and 1599 you rated at the in person meeting on important scientific acceptability and feasibility. So you should only need to vote. So...

Tom Rosenthal: I will wait. There are two votes that are left as I understand it for both 99 and 95.

Ashlie Wilbon: Right.

Tom Rosenthal: Feasibility, high, moderate, low, insufficient, not applicable and then we will have to vote overall...

Ashlie Wilbon: ((inaudible)).

Tom Rosenthal: ...thumbs up or thumbs down to recommend acceptance. But what's on the table right now is the feasibility and there should be a feasibility screen on both 99 and 95, Ashlie, is there?

Male: They're different.

Ashlie Wilbon: So I said - there's a link for - there's a separate link for each measure.

Tom Rosenthal: Right.

Male: Right.

Ashlie Wilbon: So you should go - in the email you should have a link for 1595 to rate your feasibility, and so the difference is if you recall 1599 was the non-conditioned specific measure, so you're acting as a TAP. So we ask that you vote - you rate each of the sub criteria for that measure with overall feasibility and then an overall recommendation.

Male: Right. That's why the screens look different.

Ashlie Wilbon: Right.

Tom Rosenthal: Oh, I see. So 1599 has the sub categorizations A, B, C, and D.

Ashlie Wilbon: Right because TAP - the steering committee was the TAP for that measure.

Tom Rosenthal: How come? Was there some reason?

Ashlie Wilbon: Because that was a non-conditioned specific measure.

Tom Rosenthal: Oh.

Ashlie Wilbon: Right.

Tom Rosenthal: Okay. Okay.

Ashlie Wilbon: I know it's complicated. So for...

Tom Rosenthal: I think it is.

Ashlie Wilbon: ...1595, the TAP already rated the sub criteria which is what we have on the screen, and we're only asking the steering committee to do the overall feasibility and then a final recommendation for that measure.

Male: Yes.

Tom Rosenthal: All right. Well let me ask you one last question is, does the committee feel that the issues are sufficiently saying between feasibility on the non-conditioned specific one and the

diabetes one that they do not need to be separately discussed or is there any sense in which they're substantively different?

Ashlie Wilbon: I think they're the same.

Tom Rosenthal: I do too but I just - well thank you.

Male: Great I agree. I agree.

Tom Rosenthal: Okay. So we can - but it is the case that 1599 requires a rating of the sub measures...

Male: Right.

Tom Rosenthal: ...and then an overall rating of feasibility. So people want to take a minute to do that and then we'll have a quick discussion on how - what we want to do, how we want to discuss overall approval or not.

Male: Absolutely.

Tom Rosenthal: Okay. So I'm going to do it and I'm going to judge how fast I can do it and by when we need to start talking again.

Male: So which one are you going to start with?

Tom Rosenthal: I'm starting with 1599 but as we've just discussed, the issues we're viewing the same, so presumably people would vote the same way on both of them. Again if we value internal consistency, that should be the outcome. There was a big "If," there in front of that.

Male: And the cost issues come up in 4E?

Tom Rosenthal: The cost issues come up in...

Ashlie Wilbon: There should be a 4D as in Dog. If there is an E that's probably a typo.

Tom Rosenthal: No there's A, B, C, D and E on the 1599.

Ashlie Wilbon: A, B, C, and D. I have to put it back...

Male: ((inaudible)).

Ashlie Wilbon: Is there four though? The last four - the last one that's E should probably be a D, so that's four...

Tom Rosenthal: No. No. There are five 1s on here. There's A, B, C, D and E. And E is the one that we've just been discussing with the cost issues. D was susceptibility to inaccuracies. C was - it says, "Exclusion should not require extra data." B is the data elements in electronic sources and A is that the data elements are routinely generated. I think C is new and doesn't jive with what we've been talking about.

Ashlie Wilbon: Yes, that's the mistake. Just put...

Tom Rosenthal: What do you want us to do with C, just ignore it?

Ashlie Wilbon: And put it back and not applicable.

Tom Rosenthal: Not applicable on C. Okay.

Male: Thank you.

Male: I'm sorry which one was this?

Tom Rosenthal: C. For 1599, 4C should be scored an A.

Male: Okay.

Tom Rosenthal: And then this also is slightly different because the scoring here is C, P and M, which I guess corresponds to high, medium, and low.

Male: Yes.

Male: Oh.

Tom Rosenthal: All right. So when you do that, you'll then get - and you click Okay. You'll then get to the endorsement recommendation screen.

Male: Wait, wait, wait.

Tom Rosenthal: I'm not going to move ahead until people are all ready.

Male: Yes well I'm looking at - you know it's high, medium and low but there are in fact four categories here -- completely, partially, minimally and not at all. So and minimally and not at all would both qualify as low under a high, medium, low vote?

Tom Rosenthal: Unfortunately I clicked to the next screen and I can't...

Ashlie Wilbon: So I'm sorry. So I just double checked and the surveys that we spent apparently have not been updated correctly. I apologized for the confusion.

Can we just have you guys check it - we'll have - do another pause in about 10 minutes maybe and have you go back in and it will look exactly the way you expected it to look.

So just give us 5 minutes to fix it and we'll revisit it. So if everyone could just X out and then we're going to fix it and again I'm sorry for any confusion on that.

Tom Rosenthal: All right. Ashlie this is a process check. I don't believe we're going to get through both 1591 and 94 today in the time we're - bad time management on my part but I've...

Ashlie Wilbon: Oh, that's fine because the discussion as you said it would apply - the next two measures that we're discussing are both from Ingenix, so it does somewhat overlap and bleed into the discussion time that we had allotted ((inaudible)).

Tom Rosenthal: All right. We still may not get to both of them but - and I think we want to do justice to all of these. But Ashlie there - I think there is one other staff work that those are going to have to be done to - for us to successfully vote and vote on our overall recommendation, which is can you go back and give us what our ratings were for importance, reliability, and usability...

Ashlie Wilbon: Yes for...

Tom Rosenthal: ...on each of these.

Ashlie Wilbon: Yes.

Tom Rosenthal: For 1599 and 1595.

Ashlie Wilbon: Yes. So we've got - let me just - we do have it on the screen here put up. So for 1599, what I'm showing on the screen is the steering committee's ratings that you submitted at the in person meeting.

Tom Rosenthal: Okay. So that's the sub units for one, and then what was the overall vote on importance?

Ashlie Wilbon: That's for importance. The overall vote on importance was that it's not important.

Tom Rosenthal: Okay but what was the vote? Just - can you get us the vote?

Ashlie Wilbon: Yes hold on one second.

Tom Rosenthal: Okay.

Helen Burstin: And Tom this is Helen. Just keep in mind obviously, you don't have the full committee, so it's just kind of a straw man really today because obviously everybody on the committee will get to vote as we summarize the discussion.

Tom Rosenthal: Right, right.

Helen Burstin: So you know...

Tom Rosenthal: Yes no I understand. But you know, I assume for again, those of us on the phone would want to be reminded of how we voted on the sub measures. If we voted low on all the sub measures, it would be a little silly two weeks later to then vote the thing or vice versa.

Helen Burstin: Oh, I agree. I'm just saying in terms of just, you know...

Tom Rosenthal: I get it.

Helen Burstin: ...in real time, right.

Tom Rosenthal: But I would also suggest if in fact as you're pointing out Helen, everybody gets to vote.

Making the sub unit votes available to the other people as they go to vote on overall acceptability would be a good idea.

Ashlie Wilbon: Okay. Yes, we're pulling it up now.

Male: So they will hear this conference in theory?

Tom Rosenthal: Assuming they're willing to sit through an hour of us.

Helen Burstin: Well yes and no Tom, because we would also actually summarize the discussion for them, so they shouldn't have to just listen, although captivating as it has been...

Tom Rosenthal: Thank you Helen.

Helen Burstin: You're welcome.

Male: Liar.

Tom Rosenthal: Yes, there was no sarcasm there. Now the discussion I think has actually been very helpful on all of these. I have a much better sense of this with - in the basis of the discussion.

Helen Burstin: I actually wasn't being sarcastic.

Male: You're well placed in your job then.

Ashlie Wilbon: So for importance, the overall vote for 1599 was 16 yes and 0 no?

Tom Rosenthal: Okay terrific. And then do you have scientific acceptability?

Ashlie Wilbon: Yes. The overall rating was 9 yes and 6 no.

Tom Rosenthal: Okay. And usability?

Ashlie Wilbon: Usability was 10 moderate and 5 low. And I'm also showing the - up on the webinar for scientific acceptability was a distribution of the scientific acceptability ratings and then...

Tom Rosenthal: Okay Oh, yes I see all six of them.

Ashlie Wilbon: Right. And then for...

Tom Rosenthal: Well validity testing was the major one that didn't score well. The other scored moderately well.

Ashlie Wilbon: Right. And then for usability, I'm now showing the distribution of the sub criteria for 1599 on usability and the - and your overall rating for the usability criteria was 10 moderate and 5 low. That was your roll up of usability based on what you see on the screen.

Tom Rosenthal: Okay perfect. And while you're - I assume you're still trying to fix the screens on the survey monkey?

Ashlie Wilbon: Yes.

Tom Rosenthal: Do you want to give us the...

Ashlie Wilbon: We'll let you know when they're ready. Yes.

Tom Rosenthal: Yes, and do you want to then while we're waiting give us the scores on 1595?

Ashlie Wilbon: Yes. So again, this is the importance rating. This was from the TAP what I'm showing on the screen now. The - you know the importance and the steering committee's overall rating of importance at the meeting for 1595 was 18 yes and 0 no.

Tom Rosenthal: Okay.

Ashlie Wilbon: The distribution for scientific acceptability on 1595 from the TAP is what you see on the screen. It maybe a little bit behind but it should be showing up shortly.

Tom Rosenthal: All right. So these are the sub breakdowns -- one, two, three, four five, six, seven.

Ashlie Wilbon: And for the steering committee's overall rating on the scientific acceptability for 1595, it was 10 yes and 8 no.

Tom Rosenthal: Okay. And then usability?

Ashlie Wilbon: And then usability, usability again you see on the screen is from the TAP and the steering committee's overall rating of usability at the in person meeting for 1595 was 9 moderate, 6 low, and 3 insufficient.

Tom Rosenthal: Mm. Okay.

Ashlie Wilbon: And we can send these slides out to everyone and we will also have the overall ratings for each one as well, especially for those who aren't on the call today, we want to make sure they have the benefit of this information as well, so...

Tom Rosenthal: The detail may be fascinating to people and they want to plough through all the detail and it's nice to have it prescreened. We - the other folks won't have the benefit of you there predigesting it when we say, "Show us ((inaudible))," it appears. So the summary thing though I think really would be helpful to people and to highlight that.

Ashlie Wilbon: Okay great.

Tom Rosenthal: All right. So how close are we to being able to do this?

Male: I think we need a few more minutes on that. Any chance we can start with 1591?

Tom Rosenthal: Yes.

Male: Yes.

Tom Rosenthal: All right. Let's start on 1591 which is ETG based congestive heart failure resource use measure again from Ingenix. And I don't have anything further to summarize it. I think we should move quickly to the TAP review on this and Jephtha can you do that for us?

Jeptha Curtis: Yes, I'm trying to do it justice. We got it yesterday and still probably processing as Jack ((inaudible)) said. And we had I think a smaller TAP for sure. You know we didn't have as wide a range of opinions on that. I do not - I have not seen the TAP scores. Ashlie, you have that for us as we go along, correct?

Male: Yes, we're going to bring them up on the screen.

Jeptha Curtis: Okay. So I think you know, it'd probably easiest thing is to dispense with importance. I think that everyone agree that heart failure and resource use within heart failure population is important, so really...

Tom Rosenthal: Yes, great just move on to scientific...

Jeptha Curtis: I'll do that.

Tom Rosenthal: Yes.

Jeptha Curtis: In terms of scientific acceptability, you know the application of broad Ingenix measure is really similar in terms of describing the techniques, and so there's nothing fundamentally different about any of the data handling, processing, et cetera that's different for this than any of the other measures, so again I'm not going to spend a lot of time going through that.

What is different and what probably warrants some consideration and certainly what was on the discussion for the TAP yesterday was that this is a measure of systolic heart failure. And there's a paired measure of diastolic heart failure that they did not submit but there is within their portfolio of measures you know, the two measures -- for diastolic and systolic.

So the greatest amount of discussion was really in specifying or how clearly specified these codes were and how good they are at capturing systolic heart failure. I think we considered it from the standpoint that we could only consider the information that's available to us for evaluating the measure.

We couldn't take into account the existence really of the diastolic measure, so there was some concern as to the completeness and accuracy with which this measure would capture systolic heart failure.

So I can give you some specific examples of that, but if you go within the data dictionary, the primary diagnosis codes they're really specifically limiting themselves to the 428ths that used the word, "systolic" in them. They don't use some of the 404s and 402s that other measures have used to capture the larger heart failure population.

So I think there were some concerns that that was giving a limited view of a heart failure population and that was the decision that they made. There was disagreement on the TAP or individuals on the TAP expressed concern about that approach.

Tom Rosenthal: And Jephtha, in the scoring on the scientific either, where was the concern around that reflected in the scoring?

Jephtha Curtis: Because I - again I have not seen the scoring, so I cannot tell you what that is, but let me - is it up on the screen?

Tom Rosenthal: It's up on the screen.

Jephtha Curtis: Okay sorry. Let me go back here.

Tom Rosenthal: Because the only one I'm seeing that didn't - that scored low was 2B5 which is identification of statistical - significant meaningful differences, which doesn't sound like the coding concern that you're raising.

Jeptha Curtis: Yes, no I mean, I'm just saying - I'm trying to reflect back on what the conversation was. And I think there are more scores here than there were people on the call, so again there may be opinions that are in here that weren't necessarily reflected on the calls. You know, there are people who submitted their scores after their own review.

Tom Rosenthal: Right.

Jeptha Curtis: You know and so I can just speak for myself and my own scores.

Tom Rosenthal: Yes okay. That's fine.

Jeptha Curtis: And that was really in the validity testing of it. And I think we also spent some time being concerned about the risk adjustment approach, again using the stratification and the severity levels.

And there were TAP members who were particularly concerned that they were adjusting for comorbidities identified during the measurement episode as opposed to comorbidities identified prior to and there was concern that you would be potentially over adjusting or adjusting away for significant differences. Yes.

Female: Nqf, why is the count of the votes varied by question? 2B1 there's only five people voting. For some of the others there's nine people voting.

Tom Rosenthal: Does anybody have an answer for that?

Jeptha Curtis: I see. It's the vagaries of survey monkey would be my guess, and the completeness of which people...

Male: Yes, this is just what we have from survey monkey, so there were people who did not answer the questions that could be reflected here.

Tom Rosenthal: Okay that's the answer to that. I think our process requires us to consider reliability first and then validity and then the combo. So Jeptha can we - should we open it up for discussion?

Jeptha Curtis: I think that's meaningful but let me just...

Tom Rosenthal: Oh, I'm sorry. I didn't ((inaudible)).

Jeptha Curtis: No we'll eventually get done then we'll eventually get to 2B which is there was a wider range, the 2B ((inaudible)).

Tom Rosenthal: Okay.

Jeptha Curtis: Okay.

Tom Rosenthal: If you're all right with that...

Jeptha Curtis: Oh, yes of course.

Tom Rosenthal: Okay. So let's talk about reliability and it's open for discussion.

Jeptha Curtis: So to start with you know, the TAP reflected that the measure is precisely defined and it's defined in a way that it you know and has been used over a large period of time, and that they did demonstrate that if you do the measure in the same population at the same time, that you'll get the same results. So from the standpoint that the codes are precisely defined, yes we've felt it. It measured or met that metric.

Tom Rosenthal: All right but that is a question. So to the extent that the codes are defined but that was an area of some concern, but it wasn't reflected at least in a TAP vote of anything less than a moderate...

Jeptha Curtis: Correct.

Tom Rosenthal: ...score on the thing.

Jeptha Curtis: Mm-hm.

Tom Rosenthal: Okay discussion? Everybody ready to move on? In the absence of complaint, Jeptha go ahead then with validity.

Jeptha Curtis: So for validity these measures all as the other measures have, they rely heavily on face validity I believe. There's not a lot of validity against the gold standard. And there are - I believe we're - you know some - there were definitely some concerns about the validation.

They weren't paramount as reflected in the voting here and you know I'll leave it at that, but you know it was a fairly muted discussion about the validity of the measure.

Tom Rosenthal: Okay. Open for discussion on validity.

Male: Still doing the stratification for with and without drugs?

Jeptha Curtis: Yes.

Male: And no mental health. Mental health not explicitly dealt with?

Jeptha Curtis: I think everything is identical to what we've discussed.

Male: So okay, so let me ask. With diabetes I'm concerned about mental health because of the high level of comorbidity of depression with diabetes. What - as a non-clinician, what should I understand about mental health and CHF?

Jeptha Curtis: I think that it would be identical. Certainly there's lots and lots of literature about the high prevalence of depression ((inaudible)).

Tom Rosenthal: Depression. It'd be virtually identical.

Male: Well except that the mental health treatment causes diabetes were not necessarily CHF.

Male: Right.

Tom Rosenthal: I'm sorry. I missed that. Say that again?

Male: Well some of the antipsychotics cause metabolic disorder, right?

Tom Rosenthal: Oh, in terms of being an etiology for the disorder.

Male: Yes I'd say - right.

Tom Rosenthal: Yes, yes.

Male: Right. Right. It's like there's more concern there.

Tom Rosenthal: You don't have a causal relationship here.

Male: Yes.

Male: Yes, I'm just trying to understand how - you know how much the measure could be missing important costs if - without the carve out being explicitly dealt with.

Tom Rosenthal: Right. You know and it seems to me Jack at the health plan level, which again is one of the attributions of this, comparisons within the health plan are going to be probably valid because the carve out will be exactly the same or virtually the same or likely to be the same. It's comparing across health plans or across physician groups or something like that where these issues are quite relevant.

Jack Needleman: Yes. Yes.

Male: Okay Tom, this is ((inaudible)).

Tom Rosenthal: Yes.

Male: Is it appropriate for me to ask a question?

Tom Rosenthal: Absolutely of course, I'm saying this is completely open for discussion.

Dan Dunn: You know I thought Jack has brought this up before. I do agree with him on the importance on two different perspectives of the mental health data, but as specified, you know we're specifying that you have mental health data available to support this measure. So is this more of a feasibility issue or than validity?

Male: Yes, from my perspective we've been dealing with it as a feasibility issue, not a validity issue. But since validity is about what's measured, I thought I'd ask a science question here.

Dan Dunn: Yes, I know it's the right - I agree with your question. Just in terms of - you know I would ((inaudible)) same, and based on the way we specified it and we want people to use it, you would have the mental health data available.

Male: Okay.

Jeptha Curtis: Ashlie in terms of - this is Jeptha. Just on the scores that we're seeing, were these summarized scores over the two phone calls under which we discussed the measure? Or did you zero out the ones from the first call and redo it based on yesterday's call?

Male: I think if any member had resubmitted a score, we used the scores from yesterday.

Jeptha Curtis: Okay so there's no double counting.

Male: There's no double...

Jeptha Curtis: I just wanted to confirm. Okay.

Tom Rosenthal: Okay. Is there any further discussion on either reliability or validity?

Male: No.

Tom Rosenthal: All right. As a point of order, how do people want to handle going on survey monkey? Do you want to plough through all of usability and feasibility now and then sort of jot down your own notes and then go to survey monkey or do you want to quickly do a survey monkey on these as we go?

Male: I got it open. I'm working on this filling it out. I hope we get through it before the...

Tom Rosenthal: Okay ((inaudible)).

Male: ...hour is over.

Tom Rosenthal: All right. Let's - then my suggestion is, let's go and do the scientific piece of this thing right now. And I would also note that I've looked back and 1595 and 1599 had been fixed. So let's take a minute and deal with all three then. Or yes, let's deal with the two -- 1599, 1595 -- and now scientific on 1591.

Male: Well does that make sense? I mean, should we just take ((inaudible))?

Male: No I don't think it makes sense.

Tom Rosenthal: People would rather just keep going?

Male: Yes, as soon as we've voted, we're done with the...

Tom Rosenthal: Okay all right. I'm fine with that and then we can vote at the end and people can do it however they want. I would note that the other two have now been fixed and so there is clarity

about the same scoring measure - the way that we have been doing. Okay, so we now have the question on 1591 of usability. So Jephtha, do you want to take us through that?

Jephtha Curtis: Let me just finish. I'm sorry, because we didn't really discuss the - you know back to a couple of slides there, Ashlie...

Tom Rosenthal: I apologize.

Jephtha Curtis: So there were - you know there was 2B5 which I don't think we really...

Tom Rosenthal: Oh, I'm sorry. You're absolutely right. Let's spend a minute on that.

Jephtha Curtis: But one of the concerns was that there really was in the application no demonstration of statistically significant and meaningful differences across groups.

So the alignment approach to it in terms of how this would statistically test that, but there was really no - absolutely no demonstration of how that would work in practice.

And I think there was - that's reflected in the lowest scores that you're seeing there, but it's really a black box at this point as to - you know when they're calling something statistically significant different is that clinically significant, and you know so there was a fair amount of concern about that particular issue.

Now they did say they would not be able to get this today but that they might be able to do these runs within the CHF measure and provide that at a later date.

Male: The same issue we had with the ABMS measure ((inaudible)).

Tom Rosenthal: Okay. Are there any comments on that? All right. It seems to be factually accurate.

Although did not we hear from Ingenix in talking about one of the earlier measures that they in fact were able to put confidence intervals around the measures?

Jeptha Curtis: I think we did - in a response, right, Tom?

Male: Yes, to Dan...

Tom Rosenthal: Yes Dan.

Male: Is Dan still on the line?

Dan Dunn: Yes I am.

Tom Rosenthal: Do you want to make a comment about that, about your ability to put confidence intervals around the various numbers?

Dan Dunn: Yes, that is - I believe that was a general response. I think we followed up with the formula that we used when we do that. In fact that's the question.

Tom Rosenthal: Okay that is the - I think Jeptha, that's the question, right?

Jeptha Curtis: Yes it's about - and just demonstration of what that would look like in practice.

Tom Rosenthal: Okay, so even though they say they have confidence intervals...

Jeptha Curtis: Right I mean, you're talking about confidence intervals ((inaudible))...

Tom Rosenthal: ...you've never ((inaudible))...

Jeptha Curtis: ...that's not the hard part.

Tom Rosenthal: Right. It's how do they get used and - okay right. I think that's correct. All right.

David Redfearn: And NQF folks this is David. I'm in that 1599 survey monkey and it just keeps looping on me. I fill it out. I say I'm done and it just goes right back to the beginning again.

Male: Yes, mine says 1591 and 94 voting at the same time.

Male: And there's no Question 4 in 1591 and 1594.

Male: Okay. Why don't we try this? Because we're obviously having some issues with the survey monkey. Could we just - I want to really use the real estate of the time that we have on the phone call to go through the discussion, and we'll try to do our best to have the survey monkey ready for the end of the call.

Tom Rosenthal: All right I think that's fair. So let's wrap up the discussion around scientific acceptability. Does anybody have any further conversation or questions for Jeptha, the TAP, or comments referable to either reliability or validity? Okay can we talk about usability?

Jeptha Curtis: So with regards to usability, there was really not a lot of discussion about the specific measure. We discussed it a fair amount in terms of the diabetes measures within the steering committee, the non-conditioned specific measure.

The rationale, the demonstration of the usability of the measure was identical across I believe all the applications. And so you know there wasn't a lot to add to that based on this particular measure of CHF.

So there is demonstration or statements stating how individual customers or consumers are using this data but not specific to heart failure.

Tom Rosenthal: All right. Can I ask you a question? The - so your TAP results are on the screen?

Jeptha Curtis: Mm-hm.

Male: Yes. Okay.

Jeptha Curtis: Okay. Yes.

Male: I'm sorry.

Tom Rosenthal: So a pretty wide divergence of opinion, it appears.

Jeptha Curtis: Right and I think that had to do - I'm not sure how consistent that is with our previous measured votes on the diabetes one, although I think it was probably a wider range. I think that - you know there's a lot of insufficient and I'm not sure - what do the blues indicate? Sorry not applicable. I think a lot of those are...

Tom Rosenthal: Those are N/A.

Jeptha Curtis: ...kind of just the - you know that it wasn't specific to the TAP there.

Tom Rosenthal: Yes I got it. And then was there a reason why 3B doesn't have any scoring?

Jeptha Curtis: I don't know. I know - I thought we voted on everything.

Ashlie Wilbon: I'm pretty sure you didn't. Probably we used the Excel tool to transfer the score, then it probably just got lost in translation. We can pull it up though.

Jeptha Curtis: Again there - you know we've been working this 24 hours to turn these scores around.
((inaudible)).

Tom Rosenthal: Nothing critical. I was just asking.

Jeptha Curtis: Yes I know but I think ((inaudible))...

Tom Rosenthal: I mean, there was no criticism implied. Other than the fact that this one is important in the sense that measure is also meaning for use - for public reporting and quality improvement.

Jeptha Curtis: Mm-hm. So I think it's basically - it's a transposition error. I think 3C is actually 3B and 3D is 3C. The - what's currently listed as 3D is the harmonized. They're - that's not applicable, so we weren't even asked to vote on that as I recall. Is that right, Ashlie?

Tom Rosenthal: And that's the one that's generally not applicable. We voted and you know every vote on that one has ((inaudible)).

Jeptha Curtis: Right. So I think if you just shift everything to the left, I think that would be how it should be transcribed.

Tom Rosenthal: Well here is the diabetes one.

Jeptha Curtis: Yes. So our test/retest is not great.

Tom Rosenthal: It's not grossly different but your point is still well made. Well let's open this for discussion about usability.

Male: We'll go back to the CHF slide?

Tom Rosenthal: Ashlie, can we go back to the CHF one? But I think she put the diabetes one because it is - probably these are similar enough that while the scores aren't identical, they're not the inverse of one another, so there's I think certainly some consistency -- pretty good consistency, but discussion about usefulness, utility, usability?

Male: No we're switching to 1595 right, Ingenix diabetes?

Tom Rosenthal: No. No. No. She just put the scores up on that one because we didn't seem to have complete scores for 1591 from the TAP.

Jeptha Curtis: I think it was Slide 13. I think you just passed by it again.

Male: Yes you're right.

Male: I see. We're using that instead.

Tom Rosenthal: We just use that as kind of a frame of reference of how they were thinking of usability in these various Ingenix ones given that the CHF one is conceptually similar if not the same idea.

Jeptha Curtis: Right.

Tom Rosenthal: Okay and so we just don't have the TAP score on 3B.

Jeptha Curtis: Well I think you do but I think it's currently grafted as 3C.

Tom Rosenthal: All right but then we don't have 3C.

Jeptha Curtis: Well I think that's 3D because...

Tom Rosenthal: I don't think - no I don't think that's right.

Jeptha Curtis: No it might be - You're right. You're right.

Tom Rosenthal: 3D is the one that's not N/A.

Jeptha Curtis: No you're correct. You're correct.

Tom Rosenthal: So we either don't have 3B or we don't have 3C. I mean, because I think D is the one that basically is N/A. Open for discussion. All right.

Male: So we're talking about which, feasibility?

Tom Rosenthal: Yes. No this was usability, and that's the TAP score that you can see on the thing other than we don't have the measures or meaningful or useful for public reporting, although we saw that same TAP score for the diabetes measure which again had a fairly wide spectrum of opinion from the TAP.

Jeptha Curtis: I mean, personally what I think is most useful is actually looking at the steering committee vote on the usability of the diabetes measure is probably the best barometer for consistency because again it is identical.

Tom Rosenthal: All right, well that's fair enough and if we recall on that one for diabetes it was 9 high, 9 medium, 3 low and 3 insufficient.

Jeptha Curtis: Right.

Tom Rosenthal: Okay, and that probably isn't terrible different than the TAP if we were to decompose those various scores.

Jeptha Curtis: Mm-hm.

Tom Rosenthal: Okay. Any discussion on the substrate or the meat of the usability question? Okay, I think we've hammered the themes pretty thoroughly on that, so let's move to feasibility. And Jeptha, I guess we could possibly shortcut this one a little bit. Are the feasibility issues any different than they were on 1599 and 1595?

Jeptha Curtis: 100% identical.

Tom Rosenthal: Okay identical. So does anybody want to spend any time rehashing that? Okay.

Male: I don't think it's necessary.

Tom Rosenthal: I don't either. Thank you guys. Then let us - I would say that we have completed this and at least the discussion. We still have the vote and hopefully the tool will be - have been fixed by

the time we're done and hopefully people are taking some notes as to what their thoughts are and how they're leaning towards voting.

So let's put up 1594 and this one is an Ingenix coronary artery disease measure. I assume it's fairly similar conceptually to the others only now on coronary artery disease. Let's stipulate that it's important and Jephtha if you could give us the TAP on scientific usability and then we'll again talk, get through those.

Jephtha Curtis: I think that the overall issues were identical, and in fact we basically deferred discussion on most of the individual criteria just saying that we'd have to conceptually vote the same way on both. The major approach to this one is that - the key to this one is that again in the identification of the target population and this is a little different than heart failure perhaps just because of the wide variation and the risk of the patients included in the measure.

They identified patients for - the primary incurring diagnosis codes are the 410s through 414s, and within that strata as you know, there is very chronic, stable coronary artery disease through to patients with cardiogenic shock complicated by a mitral -flail mitral leaflet.

And so there is a huge spectrum of risk of adverse outcomes in that population and would also carry with that the risk of different resource use.

Tom Rosenthal: Did your group then weigh in on whether that was reflected in some difference of opinion about risk adjustment?

Jephtha Curtis: I think it generally was reflected in the risk adjustment would be my guess.

Tom Rosenthal: Ashlie, can we bring up the Category 2 scores from the TAP on the screen?

Male: Yes, I think that should be it, right?

Tom Rosenthal: Yes.

Jeptha Curtis: And again you're seeing even smaller votes here because this was discussed just yesterday and it was not discussed over two settings which is think why you had seven or eight votes, so...

Tom Rosenthal: All right but I - it is. You've talked the statistically meaningful differences, although the risk adjustment scores of the three people who voted on it weren't all that bad. But I hear your discussion of - that that was a concern that was slightly different than the diabetes or the nonspecific measure.

Jeptha Curtis: It's something we suggested to the developer that they go back and demonstrate proof of concept that this is in fact accurately accounted for differences in the population.

Tom Rosenthal: Okay all right. Well there's a fairly small number of votes and I'm not sure putting the other - the usability and feasibility scores up are going to be terribly useful. But the - 1594 then is open for discussion?

Male: Jeptha, is your assessment firstly is chair of the issue for the same - for 91 and 94 pretty much...

Jeptha Curtis: I think that for all of the Ingenix measures, I think the overarching issues are really the same. I - personally and professionally I think that there is more concern for me on this one than there is on the diabetes and the heart failure.

As you get to a wider swath of differences in the underlying risk of the population, I think it's harder and harder for these measures to work effectively and I think that they're sort of the extreme example of that.

Male: Agree.

Tom Rosenthal: Okay. Other questions either for Jephtha or open comments for the group?

Male: So you're worried there's insufficient - you know it's a scientific question about whether they're really - these episodes are really hanging together as homogenous thing that can be compared.

Jephtha Curtis: I mean, they're homogenous in the sense that all these patients have obstructive - presumably obstructive coronary disease for the most part, but it's just that if I'm physician A and I take - happen to get a patient with cardiogenic shock who ends up with an LVAD, you know is it possible that it could - that their approach to risk adjustment could account for that difference and I would argue absolutely not.

Tom Rosenthal: Yes, I think we had the same issues. Wasn't there a nondescript CAD measure from the American Board Group and we had the same set of questions.

Jephtha Curtis: Right. But I think the risk adjustment and the Ingenix measures is a little - I don't know. I thought it was a little less robust even than that.

Tom Rosenthal: Okay that's not so good. Okay. Other discussion on...

Jephtha Curtis: Let me just clarify it. Simply because they're going through five different risk strata, or four or five different risk strata and that's all really they're adjusting - not adjusting for it but putting it into different categories.

Tom Rosenthal: Okay all right. Great. Do you want - are the usability or feasibility questions then any different either?

Jeptha Curtis: Absolutely not.

Tom Rosenthal: Okay. And then Ashlie, we had a statistical review on all of these that I forgot to bring up. Is there anything that you recall from that that's really relevant?

Ashlie Wilbon: I - is (Carlos) on the call by any chance? I think he was - it's either - no. Okay.

Tom Rosenthal: And unfortunately I think I've deleted the email that had it so I can't even comment, but I did read it this morning and at least on the statistical of validity part, I think they raised the same concerns that the TAP did.

Male: One more question, Ashlie. My survey monkey has concurrent voting for 91 and 94. Does anyone else have the same screen?

Ashlie Wilbon: Yes, there is a dropdown box where you would select which measure that you're going to...

Male: Well I started out just doing 91 but I went to the next screen and both of ((inaudible)).

Ashlie Wilbon: I think that's just the title.

Male: ((inaudible)).

Ashlie Wilbon: It's just the title of the survey that says that but the - depending on what measure you selected and the first two questions, that's what ((inaudible)) comes through.

Male: Yes I think this is - it would be good to be consistent and just have one measure per one survey, and not have this dropdown. I think it induces confusion but...

Tom Rosenthal: Right. Well we're obviously having a little trouble still with the voting. I'm seeing that we have five minutes left on the call. Does the group that's on the call feel that they have the information or their input to the discussion on these measures, that's sufficient to make a judgment?

Male: Yes.

Male: Tom this is true. And I have the statistical evaluation in front of me. Do you want us - do you want me to go through it and...

Tom Rosenthal: I think it was similar for all of them and if you could just kind of - and just quickly go through it and I think that would be helpful.

Male: Okay. So some of the summary for - this is - I'm using 1591 as an example but I think your assessment is correct that you can use this across. But the measure was submitted for implementation across various levels of analysis. Individual clinicians also were as (Carlos) noted. The sample size may be an issue and they just submitted guidelines.

Some additional considerations were that additional pigtailed on the statistical model, specifically the ((inaudible)) and the R2 values would have been - was provided, and the approach determining the high and low outliers and the choice of cut off points needed more explanation.

And the measure developers should have provided or should have performed some sort of split example of validation statistical model.

Some things that Jeptha pointed out also but I'll just add to is the risk adjustment methodology was described but there was in his opinion insufficient detail about the specific techniques that were used. And I think other than that, a number of the topics were discussed so I'll just leave it there.

Tom Rosenthal: Okay well that's helpful and I think it does validate the TAP concerns or reiterates the TAP concerns. Is there any further discussion on any of these and can - if not, can we consider our discussion on the four measures for today complete?

Dan Dunn: Actually Tom and just as a point of clarification, we did respond to that statistical report with some follow-up on the areas that were asked. I don't know if they found their way back. We're adjusting it but we did - I think we answered a lot of those questions.

Tom Rosenthal: (Sharone), can we make sure - was that follow-up available to the TAP?

Ashlie Wilbon: So the follow-up information - on which measure was that Dan? Was - are you talking about 1595 or 1599?

Dan Dunn: I'm sorry. I was - let's track the - well maybe let's say if you need - I know we followed up with at least a couple of them and I think...

Male: For diabetes for sure, Dan.

Dan Dunn: Yes, diabetes. I'm sorry. If you need exactly the same type of information and you don't have it, we're happy to do so. Some of them are more just a description of what we did which applies across all measures and things like the R2 would be measure specific.

Ashlie Wilbon: Right.

Tom Rosenthal: Okay all right.

Ashlie Wilbon: I mean, these information was provided. It was part of the packet that we - or information with private in person meeting for 1595. If you remember we did discuss the bulk of that at the in person meeting.

Tom Rosenthal: Oh, That's right. Of course we did.

Ashlie Wilbon: And 1599 as well. Now as discussed yesterday with Tom Lynn on the phone, we did identify some items that might - there may be some additional follow-up needed. And based on how the discussion went today we said we would regroup and see whether or not the steering committee want it and the TAP needed that additional information for the R2 and so forth. Is that correct Tom?

Male: Yes the R2 for CHF and CAD...

Ashlie Wilbon: Correct.

Male: ...could be provided.

Ashlie Wilbon: Right okay.

Tom Rosenthal: Okay so Ashlie, we are obligated to do a brief public comment?

Ashlie Wilbon: Yes let's go ahead and do that. Operator - (Gwen) if you're there, if you could just open the lines to see if there's anyone who would like to make a comment to the committee?

Operator: All the lines are open at this time.

Ashlie Wilbon: Okay, okay.

Tom Rosenthal: I think we have fulfilled our duty to the public comments.

Ashlie Wilbon: Yes, and if everyone as Tom said is okay with voting, again I apologize for the confusion with the electronic tools again. We were trying to do a quick turnaround on these and we encountered some errors there.

So I apologize for the confusion. But it appears now that we fixed everything and if you would like to go in now at the end of the call and do all of your voting, you should be able to do so.

The link just to clarify for the 1591 and for 1594, the evaluations for both of those measures in the one survey, and depending on which measure you select in the dropdown on the first page, that is what you're evaluating for the rest of the survey, and it loops you back to the beginning so that you can vote on the other measure in that dropdown.

So the survey is titled, "1591 and 94" but depending on what you select in the dropdown is actually what you're evaluating at that point until you get to the end of the survey, so...

Tom Rosenthal: And when I've done the other ones myself and I got to the end and I click next which normally would finish it, it circled me back to the beginning that I have confidence in my vote, that my scores were then counted or do I have to keep doing...

Ashlie Wilbon: Yes, yes.

Tom Rosenthal: Okay.

Ashlie Wilbon: You should vote confidence ((inaudible)).

Male: And then when you're finished the second - when both of them they say, "90%," I don't know what that means but...

Tom Rosenthal: All right. We'll look at the scores, make notes and they'll tell us...

(Crosstalk)

Tom Rosenthal: ...they'll send us an email...

Ashlie Wilbon: We will follow up...

Male: We could ask the staff to confirm and then our votes are complete on those?

Tom Rosenthal: That would be great.

Ashlie Wilbon: Yes. We will send a follow up email to everyone and let you know what the results of the voting was. Keep in mind though that we do have some people who aren't on the call today.

We're going to send a follow up email to everyone today to kind of recap and provide the information to them so they can actually submit their votes, and then we will submit - send out a follow up email to everyone when we have all the votes in with what the actual final recommendation was.

Tom Rosenthal: And you're actually...

Male: But ((inaudible)) is not the final recommendations just to make sure we have a record of everyone that thinks they voted here.

Male: Right.

Ashlie Wilbon: Right. Right.

Tom Rosenthal: And will you send us the results of our vote from today or not?

Ashlie Wilbon: Individually, is that what you mean?

Tom Rosenthal: No. No. Well I mean, will there be a tabulation of the nine of us who were voting today or will you not produce any tabulation until "All the votes are in"?

Ashlie Wilbon: We can produce a tabulation. Actually that's - we can do that in the follow up email that we send. It's just saying for the nine people that were on the call this was the preliminary results. We can do that.

Tom Rosenthal: I think that'd be useful understanding that it's completely preliminary.

Ashlie Wilbon: Okay yes. We...

Male: Actually I'm brain dead. Let me ask you a simple question. Do you have a record of my vote
((inaudible))?

Ashlie Wilbon: I can't - I don't have access to the - unless I change screens right now and show
everyone...

Tom Rosenthal: You'll tell us tomorrow or the next day, Ashlie...

Ashlie Wilbon: Yes, yes.

Tom Rosenthal: ...okay, so.

Ashlie Wilbon: Right.

Male: Ashlie give us a little time. I've got to digest a few more things before I vote.

Ashlie Wilbon: Okay that's fine. That's fine.

Tom Rosenthal: Okay.

Ashlie Wilbon: We are giving you guys until next week to complete it. So especially like I said, for the
people who are on the call we are giving people additional time to do it, so you don't - you're not
under the gun to do it right at this minute.

Tom Rosenthal: Well but there would be some virtue of the nine of us who are on the call trying to get
votes in by say tomorrow or Friday, so that in fact we could discern whether or not what we did

registered. And it's fairly rapid feedback based on the discussion, because again I find linking up the scores and how then the votes - how we discussed it to be instructive.

Ashlie Wilbon: Right. There's definitely some benefit in it being fresh. I agree.

Tom Rosenthal: Right. And as too much time, we forget the thing and we're trying to - again this is helpful in creating some internal consistency. I think we've done it. I - we got through our agenda and it's precisely 5 o'clock Eastern Time, 2 o'clock Pacific. So unless anybody has any other items, Ashlie, is there anything else we need to do?

Ashlie Wilbon: No that's it. We'll just follow up the email and our next meeting of the steering committee will be the in person meeting in August.

Male: Ashlie, can you send out a copy of the slides from this presentation?

Ashlie Wilbon: Yes I can.

Male: Thank you.

Tom Rosenthal: Okay.

Ashlie Wilbon: Thank you guys. Thank you for your participation today.

Tom Rosenthal: All right thanks everybody.

Ashlie Wilbon: Yes, have a good weekend.

Tom Rosenthal: Okay, bye bye now.

Ashlie Wilbon: Bye.

Male: All right thank you.

END