NATIONAL QUALITY FORUM

+ + + + +

RESOURCE USE STEERING COMMITTEE

+ + + + +

WEDNESDAY
AUGUST 31, 2011

+ + + + +

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

PRESENT:

THOMAS ROSENTHAL, MD, Co-chair
BRUCE STEINWALD, MBA, Co-chair
PAUL BARNETT, PhD, VA Palo Alto Health Care System
JACK BOWHAN, Wisconsin Collaborative for Healthcare Quality
KURTIS ELWARD, MD, MPH, FAAFP, Family Medicine of Albemarle
LISA GRABERT, MPH, American Hospital Association
JACK NEEDLEMAN, PhD, FAAN, University of California Los Angeles School of Public Health
DORIS PETER, PhD, Consumers Union*
STEVE PHILLIPS, MPA, Ortho-McNeill-Janssen Pharmaceutical, Inc.
DAVID REDFEARN, PhD, WellPoint
BARBARA RUDOLPH, PhD, MSSW, The Leapfrog Group
JOSEPH STEPHANSKY, PhD, Michigan Health and Hospital Association
DOLORES YANAGIHARA, MPH, Integrated Healthcare Association

NQF STAFF:

CARLOS ALZOLA, Consultant
TAROON AMIN
HELEN BURSTIN, MD, MPH
LAURALEI DORIAN
SARAH FANTA
KAREN PACE
SALLY TURBYVILLE, Consultant
ASHLIE WILBON

* Participating by teleconference

TABLE OF CONTENTS

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

8:46 a.m.

MR. AMIN: As we get started this morning, just some logistics. We sent out -- Sarah sent out this morning -- or Lauralei sent out this morning the updated PowerPoint. That was based on some of our discussion yesterday. So we wanted to have the most updated PowerPoint for everybody to see -- Just this morning, maybe 10 minutes ago. I think we also have some printed versions coming for everybody.

So there are some cheat sheets that are in the NQF folder that you received yesterday that will probably help us through this discussion.

So the first one is labeled Resource Use Measure Specifications. It looks like this document with the five modules and the submission -- or the overall topics that are included in each of the five modules.

This will be the format of the

discussion today. So it might be helpful if you have this document out, just to see where potential topics will be discussed later on in the afternoon.

DR. PETER: Hi. This is Doris. Could you email that to me?

MS. WILBON: We will email it out, all the paper attachments that we have. Thank you.

DR. PETER: Thank you.

CO-CHAIR ROSENTHAL: And, Ashlie and Taroon, I wonder if it would be helpful, or maybe it would be helpful to me at least, if we tried to lay out what is the goal, what are we trying to accomplish generally. That will help, I think, all of us, and maybe even Helen could weigh in on this, of what is our goal overall about this. What are we trying to do?

MR. AMIN: Okay. So there's a few objectives -- and, Ashlie, please feel free to jump in, as you see fit. There's a few

objectives for today's discussion.

The first is to look at the criteria in the way that we have set up the modules to assure that, as we move forward in the next evaluation of these types of individual measures, whether or not the submission items were sufficient and the criteria we use to evaluate the submission items were sufficient.

The second is also to look at, as we look forward -- so a second piece of this exercise, which is quite linked -- is to look at the next phase of work, which really will be to evaluate episode groupers and to see some of the challenges and potential guidance moving forward in evaluating groupers.

Now, granted, this exercise that we have gone through over the last few months has not been to evaluate groupers, although the TAPs and the Steering Committee has been very close to a product that is essentially a grouper system, and we have had some

challenges of looking at that product in the lens of an individual measure.

So to get some additional guidance along these different modules for potentially additional materials that we would need to evaluate a grouper or just -- I wouldn't say that we might need additional criteria, because I think the criteria would probably be sufficient to evaluate a grouper, but additional guidance on how to really evaluate a grouper.

So to summarize, there are two specific objectives for today: First, to evaluate our overall -- as we move forward in evaluating individual measures, do the submission items and the criteria that we have to evaluate individual measures -- are they sufficient based on our first run at this; and secondly, as we are looking at the next phase of work, which is really looking at episode groupers, is there additional guidance along the lines of the modules and additional

criteria that would the group would offer.

CO-CHAIR STEINWALD: Isn't there -- trying to think of putting the work that we are doing here in context. You know, we are coming up with recommendations for a small number of measures, and yet we think that the process that we have gone through might be illuminating to a broader audience than those that have developed those measures or those that developed some and didn't get them approved-- sort of along the lines of advancing the state of the art of resource measurement -- and also getting us at least partway down the road to where, in our first face to face meeting, we discussed that we wanted to be, which is to develop real efficiency measures, measures that bring resource use and quality/outcomes together. Right?

I saw some of that reflected in the materials you distributed. So we have this -- I am adding this as a third objective,

if others think it is reasonable: That at some point in our report, maybe the more reflective part of the report, we will depart from talking about specific criteria and certainly depart from talking about grouper evaluation, to talking about the state of the art of resource and efficiency measurement and how it could be advanced, what we have learned from the process we have been through for over a year that could advance that state of the art.

MR. AMIN: Definitely in agreement. We will see some -- There's a few slides toward the middle of the day that present the patient-centered episode of care framework that NQF has engaged upon, and thinking through some of these questions of how we would link potentially quality measures to the measures that we have been looking at on resource use. What components of the measures would have to be aligned to truly develop efficiency measures in the future --

so those types of discussions. Please, Helen.

DR. BURSTIN: In addition to that broad comment, I think that is exactly where we want to go. I think, because this is one of the first times we have actually taken the criteria and kind of morphed them a bit to fit an emerging area of measurement, it would also be helpful for us to reflect back about are the criteria as they stand really, for the most part, seem fairly applicable; and we go through this exercise in the future, how much of this sort of intensely customized criteria building do we need to do?

We are about to do population health, for example, going through very similar kinds of issues. So what is the testing of a measure that compares counties or communities as opposed to a provider? So there are different angles on this, but I guess my goal would be: We went, I think, to the nth degree of being very detail oriented, every single module, every single component

of these measures.

As we reflect on those criteria and how many of them mattered, it would also be helpful to think as we sort of go prospectively, is there a way maybe to simplify some of this for the next time we do this, especially if we bring in quality measures, and it gets even more complex.

So we would really like your good thinking there.

MR. AMIN: To take that and go back a few steps, in the NQF folder, the second material that I would suggest that you refer back to is the side-by-side table that is titled "Evaluating Resource Use Measures."

So the goal of this two-by-two table is to look at the criteria and describe the specific elements in the measure submission form, just to bring us all back.

It might be helpful to just spend a few minutes here. I don't want to take up too much time, but I think it will help in the

framing of this discussion.

So again, it looks like -- It is a two-by-two table like this. If anybody doesn't have it, please let us know, because it is going to be quite a particularly important piece of reference material.

MS. WILBON: I think it is at the very back. I think it is the last.

MR. AMIN: Yes, the last, last but not least.

DR. PETER: Can you email that to me, too?

MR. AMIN: Yes. So as we are thinking through the evaluation criteria is on the left. The evaluation criteria is on the left, and the submission items are on the right, and I will move down to scientific acceptability, because this is really where the modules interact with the criteria.

So as we are looking at 2(a)(1), the measures precisely specify: That includes the measure specifications as to the general

approach; as (s)(2), the general approach; (s)(3) the type of resource use measure; (s)(4) the target population; and (s)(5) the data dictionary.

It also includes the data protocol module, which we will spend a good part of -- or we will spend the first session reviewing the data protocol module, which includes all of the data inclusion criteria, the preparation for analysis, the data exclusion criteria, how to handle missing data, which was a robust discussion we had yesterday with pharmacy claims and behavioral health issues, and the data source.

The clinical -- The logic module fits into this criteria, including the clinical framework, comorbidities and interactions, the clinical hierarchies, and then also the construction logic, which is Module 3, which looks at the construction logic, the trigger and end mechanisms, complementary services. The risk adjustment

model, stratification approach and costing methods, also includes the reporting module which is the attribution approach, which -- the reporting module, and then the measure score.

Then we have the reliability testing, which then looks at the testing results and how the reliability testing was done. Carlos will be joining us later on this morning to offer some of his insights as part of this process, as he has been part of the process helping us think through this.

So 2(b)(1): Also the measure specifications are consistent with the measure focus. So a lot of these modules are repeated for this criteria to see whether it is consistent and -- consistent with the measure focus and intent. This is in contrast to 2(a)(1) which asks whether or not the measure is precisely specified for these different modules. So that is how it approaches both of the different modules that we were addressed

here.

The validity testing is 2(b)(2), and 2(b)(3) looks at the exclusions, how the exclusions are handled. 2(b)(4) looks at the risk adjustment approach, whether or not it is based on clinical factors. So again, this is a discussion that Paul brought up yesterday on whether or not issues that are occurring within the measurement period should be allowed to be part of the risk adjustment approach.

2(b)(5) looks at the scoring, the scoring approaches, whether or not they address clinically and statistically significant differences in performance, and this would address some of the larger issues that we have discussed of whether or not the O to E ratio offers tangible, actionable results for providers on the front line, which also begs the question of who the intended audience is.

We will get to a lot of these

bigger questions as we move along. Then the multiple data sources, which was not necessarily a major issue for resource use measures, considering the fact that they use the administrative data sources; and 2(c) looking at stratification.

Then usability really addresses a lot of the issues of the current use, interpretability and transparency, and then we will go into some more discussion around feasibility.

So as a framing device, this was really sort of how we thought it would help frame the discussion. As the course of the day goes, what we will do is we will take each of these individual modules and explore some of the larger questions that were part of the discussion for the TAPs and the Steering committee.

DR. REDFEARN: Is this all in the context of episodes, episode groupers, or generalized?

MR. AMIN: We should think about it broadly.

MS. WILBON: And to piggyback on Taroon and Helen, what you will notice from the table is that currently we only have two criteria evaluating the modules and the specifications. So I think it would be helpful to think about, kind of to Helen's point, are those two criteria sufficient or are there additional criteria that would be focused specifically on things within those modules that we discuss that need to be evaluated specifically.

So I think, as we got into some of the discussions, you know, there is a lot jammed in to resource these measures, and the specifications are so expansive that you jam a lot into those two criteria.

So kind of as we are going through this, think about how we might be able to either reframe the criteria or potentially add or whatever that may be to address those

modules in a better way. I don't know. It is just something to think about.

MR. AMIN: And before Ashlie goes into some of the overarching discussion around the resource use measures, there was an additional component that I want to throw out there, and it is a little squishy, but it is the sense of the interactive nature of some of the criteria, which we have talked about in many ways, where the level of analysis is at the individual provider level, what that means to the risk adjustment model.

So I will just keep those -- I will just throw that out there, the interactive nature of some of these different criteria and the submission items and how well that is articulated through the way that we have set up the evaluation process.

So maybe we can just go through some of the overview of just how we have structured these sorts of things.

MS. WILBON: Sure. I think you

guys are pretty familiar with a lot of this stuff. We have shown it multiple times, our definition of resource use measures.

It might be helpful, too, I think, for those of you that are able to bring up the slides on your computer that we sent out, to kind of be framing your thoughts through some of the principles that we came up with for resource use evaluation.

This was, obviously, a year ago, but as we focus this discussion, if there are additional principles that we think need to be added to this list or maybe they need to be changed -- I think they are pretty broad and still applicable, but just to kind of help frame that discussion as well.

I won't spend time on this, and we can work on getting copies for everyone, for those of you that don't -- can't bring it up on your computer.

I think, actually, we will go ahead and kind of jump right into some of the

discussions that we had outlined. The questions that we are going to discuss for each of the modules are also listed on the updated agenda that I sent out on Friday, and I think it is also printed as a document in your folders. So you have the questions in front of you as well that we will have for each section.

So for the data protocol module, if you recall, it is not showing up very well on the slide, but we have got the different components of the data protocol module in the blue bubble here at the bottom.

What we have done for each of the modules, similar to how we set up the draft report, is we pulled out some of the overarching themes that we heard through the discussion of the measures across the TAPs and the Steering Committee, and to come up with some -- have a discussion about how these issues might be addressed in the future by developers and how we might want to approach

the evaluation of those items as well in the future.

So the two things that we pulled out from the data protocol module, which includes the data preparation, the data inclusion criteria, data exclusion criteria and missing data, was obviously the implications of using administrative data.

All these measures use administrative data, and there are certain limitations in that in itself. So sometimes the measure seems to be limited, because it used administrative data, but it is really more so limitation of the data that limits the measure's -- the ability of the measure to measure certain things.

DR. REDFEARN: What alternatives are there? What other data sources are there other than administrative claims data for this?

MR. AMIN: Let me just -- So the point of this -- I mean, there may not be, and

that is the current state of where we are right now. One of the challenges, though, as we are evaluating measures is are we holding the measure -- When we are evaluating the measures, how can we not hold the measure responsible for some of the limitations of the underlying data? So the ability to risk stratify when the administrative data is not -- doesn't allow that level of risk stratification.

So it is not to say that -- It is just something that was discussed. I mean, it is not to say that we have an option here, but it is a limitation and some of the implications of using the administrative data that was discussed at length through multiple TAPs.

MR. REDFEARN: The question, I think, it boils down to is, if you know you have significant issues with the only data you have available -- say, administrative data -- do you proceed to produce a measure that

incorporates that imperfection, and acknowledge it, or don't you do the measure?

Somebody has to make a decision. If the data is so crappy that you really can't get any measure produced using that data is going to be potentially misleading and not informative, then I think the decision you have made is you don't do the measure until the data improves. But you could also argue, well, it is imperfect; we know it is imperfect.

For example, all risk adjustment we are doing right now at best accounts for 25 percent of the variation in cost. It is, by definition, hugely imperfect. Yet we do it all the time.

So I think the consensus, basically, is you do the best you can. You label the limitations, and you proceed. You could decide not to do it at all.

CO-CHAIR ROSENTHAL: I think that is spot on point and, obviously, we made those

adjustments here on the fly. Perhaps if the next group could be a smidge more explicit about it, it might be easier.

I, for example, was impressed with the NCQA presentations about the attempts on their part to be as thorough and complete and accurate about collecting the stuff as possible, and I had a much greater sense of that than I did on a couple of the other groups that presented where it felt a little less clear-cut as to what was what, but I would offer up one option of a non-administrative data set.

It as interesting to me to watch this, in that much of the public focus, both on quality and a variety of things, ends up being at the hospital level. Now maybe it is a good thing that we have moved past that and that we are talking about ACOs, etcetera, to the extent that such a thing exists in the universe.

You could imagine a hospital

comparator that was not based on claims data, because we have the hospital reports that are at a very high level and actually CPA certified statements, and it might be possible to dissect hospital costs in a way that doesn't rely on claims based data.

So that is at least one thing to offer up, although we saw absolutely no measures put forward that attempted to compare hospitals, and maybe that is a good thing or maybe that is not a good thing.

CO-CHAIR STEINWALD: We saw no measures based on electronic health records, even though the prospect of having them is before us, and we know a lot of organizations are developing them. One would hope that not too distant future that someone would come forward with a combined claims and patient record based analysis.

The other alternative is external data collection. I was thinking of what Jack said yesterday about, you know, we are limited

in our analysis to what claims data have, but then again in a research setting you can often supplement claims data with independently collected data like, let's say -- We talk about comparing entities. You can use claims data to determine the rate of infection, let's say, in different institutions.

You can use external data to determine whether those institutions have installed programs to reduce infections or not, that sort of thing; but this data -- that art, I think, is pretty limited as well as electronic health records are.

CO-CHAIR ROSENTHAL: Other thoughts from the group?

MS. WILBON: So the other issue, obviously, that we have talked about a lot is this issue about carve-outs and outsourcing of mental health, mental behavior health, and pharmacy data or pharmacy coverage arrangements and so forth.

I think our main question for that

is -- We have talked about it a lot, and maybe we have exhausted all discussion about that. I don't want to beat a dead horse here, and we developed these slides before our discussion yesterday, but do we have guidelines or suggestions for developers on what ideally would be a good way to address this in measurement or are there certain kind of principles about how to address this, particularly with the missing data issue when they are developing measures?

CO-CHAIR STEINWALD: Jack, you are up.

DR. NEEDLEMAN: I think this actually relates very much back to the discussion we were just having about administrative data. I think part of the issue is what -- on both of these, is what is the responsibility, obligation, expectation for the measure developers, for the measure implementers, for organizations like NQF, to be pushing to get the data better?

Who has the responsibility for pushing to make sure the data are more complete, more complete in terms of collecting everything that is already being collected, that is relevant for what you are doing, like the pharmacy data, complete in the sense of identifying important resource elements that we are not currently collecting data on, and figuring out how to build information about them into routinely collected data?

So I think those are some of the general issues that go beyond is this measure a good one, is this measure an adequate one, that somebody needs to think about and, I think, becomes part of the broader context that the report needs to discuss.

A question I would ask is: We are trying to do resource measures, and Kaiser Permanente, for example, has many regions. Are any of these measures relevant to them, given the way their data is collected?

The VA, another major integrated

health care system, has many regions, has many individual provider groups within those regions that are trying to deliver integrated care. Are any of the measures that we have talked about over the last few days implementable, given the VA data resources?

How would either of those systems go about trying to figure out whether the resource use across all the units within them are comparable or different? What resources do they have for doing that in terms of their data and data collection?

That, I think, is relevant to thinking about where the data collection needs to go and what are short term limits of what we collect versus things that could be collected, and long term limits to what we collect in the sense that we have to build new data collection systems. I think all that is part of the administrative data issue.

On the carve-out, I think it is particularly relevant, because it gets to

David Redfearn's issue of do we have enough data here that we are confident that we have got a good measure of the resources that are being used, if we are trying to measure resources?

I think you could ignore the mental health or the pharmacy carve-outs for services where either those are truly minimal de minimis elements of the care we expect the population receiving that we are looking at for a specific episode, like knee, or that the care is so standard that the cross-differences that we would expect to see if we have the data are narrow, so they are not going to explain very much resource variation -- you know, post-surgical antibiotic treatment regimes, for example.

I am making this up. Remember, I am not a clinician. But where there are substantial parts of the expected costs, and we expect to see variation, you can't have a measure that purports to have present

resources without having those in, and that was my concern about the COPD measure and the asthma measures that we were discussing yesterday from Ingenix.

So Ingenix is a measure developer, but they are also a data aggregator. Right? They collect data from all the insurance companies that are basically subscribing to their service.

NCQA is also a data aggregator, but as a data measure developer and as a data aggregator, they have been far more aggressive in saying there are data elements like pharmacy we need to get in here, and we won't certify you unless we have got them.

So the insurance companies which are looking at carve-outs, I think, ultimately need to figure out a way to call that data back in some way to enable resource use measurement, if we are serious about doing resource use measurement.

The question is who creates the

pressure -- the incentives and the pressure on them to do that call-back of the data? Is it payers? They are the payers. So is it the folks who are subscribing to the service who say we want full resource measure use; we are not contracting with WellPoint unless we can get that. So you go deal with the PBMs and get the carve-out data. Not sure that is realistic.

Is it NQF? Is it CMS? Is it -- You know, where is the pressure going to come to encourage the core insurers that are the sources of the data for places like the Ingenix measures to actually go about having responsibility and feeling pressure to get the data that is often other places about resource use, like the PBMs, like the mental health carve-outs, to get back. I think that is the challenge here.

Once we have got the data, we know how to incorporate it into these measures. Even with standardized -- Even if we do

standardized pricing on pharmacy, because the PBMs don't want to share with you how much they have negotiated to pay for each drug that is in their formulary, that's fine. But NCQA isn't getting that. What they are getting are the counts of different drugs, and then they are using standardized pricing.

I think that is good enough for what we are trying to do here, but you have to have the data. So where does the pressure come for the call-back?

DR. REDFEARN: It is actually something the reverse in our case. There is one entity you didn't mention. The PBM problem -- The carve-out for drugs, the biggest problem for us is ASO groups, large ASO groups.

So the employer groups themselves have to do this, and I will tell you, I was struck by the NCQA guy yesterday saying, oh, they put pressure on people, and they find the data, and they submit the data. Boy, we have

maybe been doing a crappy job of it in California, but we have a hell of a time to get that data.

We go to the PBMs, and they say, no, you can't have it; it is proprietary. The only way to get the data out of them is go to their customer, the ASO group, and have the ASO group insist that they provide at least the NDC codes to us to do that, but it is really hard. It is really hard.

CO-CHAIR ROSENTHAL: Well, I think there is a public policy obvious question here, and the other issue is none of the stuff is comparable and, if you take the macro, for example, the macro costs on the commercial side and the macro costs on Medicare side don't always match up by geographies and, frankly, you would want a kind of all payer system, I think.

The only way you are going to get at this is through a public policy thing, and I don't think the commercial world is going to

have the same sort of ultimate imperative to get comparable information. I think this becomes a public policy thing.

For me, the disappointment about the process -- and I think each of us, and I know Barbara has had her disappointments about what we have done or haven't done, but to me the disappointment is we can't answer the question you pose.

I was reading in a magazine the other day somebody being interviewed about medical policy, basically saying the VA is clearly and unequivocally the least expensive health care delivery in the entire country with the best outcomes and the highest quality. I asked myself how could anybody -- I mean, maybe it is true, but how could anybody actually know it for sure, because you can't -- There is no basis for comparison, and we haven't accomplished anything at all, I don't believe, in the measures that we have produced, because we basically endorsed by

necessity sort of fairly proprietary oriented measures, and it would sure be nice to see some measure that was much more like one that you would say by necessity you would be measuring every hospital in the country, every physician in the country, every somebody in the country about something.

CO-CHAIR STEINWALD: Paul, do you want to weigh in?

CO-CHAIR ROSENTHAL: That wasn't an anti-VA thing, by the way.

DR. BARNETT: Maybe we just have better publicists.

CO-CHAIR ROSENTHAL: That's good. There is nothing wrong with that.

DR. BARNETT: No, but the issue of how do you compare -- you know, how do you benchmark -- is a big one. So VA has the data. We are not comparable in many ways, because our benefit package is quite different from other folks.

I think the other issue, one of

the things that makes us seem very low cost is that many of our members have dual coverage with Medicare and get part of their services from Medicare. So that is a little bit -- That may explain some of the efficiency, something not widely acknowledged, I think. But that question of how would we compare across systems -- it is a difficult one.

I like your idea of having at least one global measure.

DR. ELWARD: Yes, I agree. One other thing, I would agree with you about one of the opportunities being employer groups. I know in Virginia the Chamber of Commerce has finally said, we really want to be involved in this, much more active than I have ever seen them before.

Essentially, if they are self-insured or if they are paying the premiums, ultimately they can get the data or they can go someplace else. But I don't think they have been utilized as effectively as they

could be.

CO-CHAIR STEINWALD: So what I heard Tom saying, to put it in economist terms, is that a comprehensive database that is at least adequate for developing the kinds of measures that we want should be seen as a public good. Right?

CO-CHAIR STEINWALD: A lot of states have them.

CO-CHAIR ROSENTHAL: Yes, but I guess at some point, for the benefit of the staff, we are going to have to come to a principle or conclusions that they can write up. Barbara?

DR. RUDOLPH: Yes. I had a couple of comments about the last thing you mentioned, the all payer claims databases. If you look at New Hampshire, New Hampshire has done a really nice job of putting out costing of health care services on their website. I can't find it, for some reason, today, but I have seen some of their work.

You know, they didn't come -- None of the states have come forward who have these, because they really don't have the resources to go through the NQF process. It is really very time consuming, and the staffs are small, and they don't have the resources that a NCQA or Ingenix has.

So I think some of the more creative work being done now is out in those smaller kinds of arenas where there just aren't the resources to do this process.

For those of you who have never gone through the measure endorsement process, it is very time consuming. So that is one avenue of, you know, maybe there is a way other parties could go out and actually do the work to write up the measures and things like that. I don't know, but it is an issue.

The other thing that I wanted to bring up that sort of happened yesterday, and I am wondering if it is the idea of the measure developer's role and the end user's

role, and whether or not like that piece of auditing is really part of the measure or is it part of the user's responsibility.

Auditing is very expensive. Again, there is another -- Unless you are a membership organization like NCQA who can charge for the services like that as part of either -- That one wasn't part of accreditation but it probably will be -- you do not have the resources to do that kind of auditing, of sending -- you know, having a cadre of trained auditors who go out to 480 plans or whatever to get that information to assure it.

I am just thinking that we need to be specific. If that is actually going to be a requirement, then it needs to be stated as such, that don't come forward unless you have an auditing piece in your measure, because to hold other people sort of at -- You know, it is kind of not have it specifically mentioned, but then to not approve measures where it is

not available. It is a problem. So it has never been part of the requirement.

CO-CHAIR STEINWALD: I was going to suggest holding that thought until we resolve the issues that we discussed earlier. Okay, you guys are good with that?

So just to see if we can close the loop on the issue as the staff originally raised it to us, what I heard on the issue of carve-outs and missing information was, first of all, to strive to make the databases as comprehensive as possible and not be satisfied with the usual administrative reasons why we don't have pharmacy data or we don't have behavioral health data, if there, in fact, ways with a some additional effort to obtain those data.

Second, we don't make comparisons of entities with and without data. That could be a hard and firm principle. Then the third one, and I think Jack has really been helpful in elucidating this principle, is you are also

not satisfied in comparing with to without when there is a good reason to believe that that comparison could be biased by the absence of the missing data.

MR. AMIN: Let me just clarify on the last point, I think, a question that, I think, was clarified for us in this process from Jack's point yesterday, and I will offer this as a question again.

I think the clarification from the first meeting was that we don't do comparisons with and without, but I think where we got to yesterday was that, if a measure is intending to measure a clinical condition that has a predominant portion of its costs in pharmacy claims -- so asthma was the example yesterday -- then is it fair to even look to measure asthma resource use without pharmacy claims?

CO-CHAIR STEINWALD: Well, but it is the responsibility of the measure developer to justify that the comparison is still valid.

DR. PETER: But you have to look

at the variation in the cost of the pharmacy. If there is not a lot of variation in that cost, then it really doesn't matter. Right?

CO-CHAIR STEINWALD: Well, but the point is that the pharmacy is missing, and maybe the upshot of it is that they would need to do an independent analysis to show that the fact that the pharmacy is missing really doesn't invalidate the comparison.

CO-CHAIR ROSENTHAL: She was giving a counter-example. The pharmacy costs don't matter. Therefore, it could be valid to have it without the pharmacy.

CO-CHAIR STEINWALD: Right, but you have to have a way of determining that the pharmacy --

DR. PETER: That they prove that that is the case.

MR. AMIN: As a quick point of --

DR. PETER: With other databases or other knowledge that was by utilization of the pharmacy.

MR. AMIN: As a quick point of introduction, Carlos arrived, just so everybody knows, our statistician, and Karen Pace, our NQF methodologist, also joined us for the day.

DR. BARNETT: Well, the issue of data: So we have good national hospital datasets, and what we lack is the pharmacy and the outpatient claims database. I think we can wring our hands and say, gee, wouldn't it be great if we had comprehensive national data.

I think there are lots of reasons to think it is not going to happen anytime soon. First is the whole issue of patient confidentiality in HIPAA. The payer's groups, providers, everybody has an interest in keeping things secret that is proprietary to rates they negotiate, secrets to their efficiency. It is going to be very hard to create such a dataset.

So it is great to put that on your

wish list, but I actually think, kind of back to where we started, which was we were thinking about how do we link efficiency with quality and that we really ought to be thinking about some small but more specific measures that have to do with appropriateness of care and quality of care that are HEDIS-like, one by one, that people can look at.

So some of the quality measures that we have now like readmission rates are implicitly resource measures. Right? If you avoid inappropriate readmissions, if you avoid central line infections, you are going to save money, and you could actually talk about how much money you save.

In a larger sense, there are a lot of things that we do, maybe 30 percent of care, according to some estimates, that is inappropriate. We ought to have metrics of appropriateness and try to apply those. A very hard thing to do, I realize, and there are like maybe 10,000 things on that list, but

there are such lists.

NQF, in fact, has in the National Priorities Partnership created such a list. The Institute of Health Care Improvement has a list of inappropriate stuff. The NICE in the United Kingdom, the National Institute on Clinical Effectiveness, has a Do Not Use list that has about 600 items on it.

The advantage of that approach, I think, rather than these observed to expected, is it is actually something very actionable. If you are doing too much low back imaging, well, you know exactly what it is to tell the clinicians what they shouldn't be doing.

So I think we ought to think about these alternate ways of going, rather than putting -- You know, wishing for something that we are not likely to get, I think it would be better to focus on measures like the quality measures that have already been developed that are about appropriateness or about quality where efficiency is implicit.

CO-CHAIR STEINWALD: Jack, then Dolores, then Kurt.

MR. BOWHAN: I was going to respond to the idea about the carve-outs and these large data aggregators that have tons of data. They could maybe provide some guidance in standardizing a process to how to find within your own data with an organization that something is wrong, and here is a red flag, and here is the algorithm you use to find out, so they could set some minimum thresholds on things where you know there should be lots of behavioral health or drug data.

MS. YANAGIHARA: I wasn't going to say this, but I echo that. I mean, just in our work, aggregating data across seven health plans, the data aggregators have access to lots of data and can give you benchmarks in terms of data completeness and things like that.

So I don't know if there is room there for setting standards for data

completeness and things, but I know that we do a lot of data quality checks before we even use the data.

I was going to comment on the submission of measures. I know that the measure developers are supposed to put in there what data is required, but it doesn't seem like in our current format it is very clear, like we had to kind of dig through to see if pharmacy data was actually required for Ingenix measures, and then it was sort of the worry was kind of like, well, it is recommended that you don't use -- you know.

So I think it just needs to be explicit: Does this measure require pharmacy data to be a valid measure, or whatever kind of data, just to be really clear on what is required and what is not required, or could it be done without it, but it has to be either everything with compared or everything without compared, and not trying to compare the two. So I think that it just needs to be really

explicit.

Then secondly, I think that, if it is okay to be with or without, we just need to have a standard way that, when users are using the measures, that they can check off was it included or was it not included, so that it is very clear which ones you can compare and which ones you can't compare, so that it is not sort of like hidden in some line of a description of methodology, but it is very explicit -- here is what is included in this measure.

DR. ELWARD: I just had a comment about groupers, if it is the appropriate time to do that or if we can do it later. Okay.

One of the things that -- I sent some information about this. This is a really challenging field, to be able to look at episodes of care, and that is a huge challenge.

One thing we might use for guidance is what they have used in Europe for

about 20-some years, which is ICPC, which is the International Classification of Primary Care. They use that as their ICD-9 in Europe, and it actually crosswalks to ICD-10.

I am not suggesting we use that as-- It wasn't designed as a resource tool, but there is good data in a number of studies that show how you could track episodes of care over time, which is, in fact, how they do it in a lot of the world.

So I would suggest perhaps what we could do is look at that methodology and how they structure that to get a leg up on how we would approach our own measures and what we are expecting.

There are a number of people in the United States who know how to do this, Larry Green, Mike Klinkman, Wilson Pace, a number of good people who have actually done that. Mike actually uses a cross-walked ICD-9 at University of Michigan to do some of their internal analyses.

So in terms of not reinventing the wheel or be able to refine one that is already there, we might want to consider that.

MR. AMIN: I just wanted to also throw out the question, which is actually the first one, as we are talking through this, which goes to Dolores' point, and I think it was brought up a number of times.

The question of this module can be submitted as guidelines or specifications, and this may be some of what is going on, but let me just clarify the difference between guidelines and specifications.

So specifications allow for user options, but must be specifically adhered to; and guidelines are well thought out guidance that allow user flexibility.

So these components allow that degree of flexibility in guidelines and specifications. So it continues to -- So one of the questions is posed as we are discussing this is: Is the option of guidelines and

specifications appropriate for these four submission components, which would be the preparation for analysis, the inclusion criteria, the exclusion criteria, and missing data.

CO-CHAIR ROSENTHAL: Well, I think the experience would suggest that that is what got us into trouble in several areas. So we could go through them one at a time and make a determination that sort of guidelines would still be okay, but to the extent possible, they ought to be specified, because that is what got us into trouble in at least several of those areas through the course of the thing.

CO-CHAIR STEINWALD: I don't remember being -- What trouble did we get into?

CO-CHAIR ROSENTHAL: Well, where we had to reconsider the whole Ingenix ones, because the original go-through had been a guideline and not specified. I think there

were a couple -- and I don't mean we got in trouble, but I mean we had to go back and reconsider something, because several people thought the thing had been specified, and then we found out that it hadn't been specified.

CO-CHAIR STEINWALD: Right. That was troubling.

CO-CHAIR ROSENTHAL: It created trouble for us, and that is all I care about.

DR. REDFEARN: The contrast I thought was really dramatic between the way NCQA approached this and Ingenix approached it: NCQA has all these standards and rules and formalities, which I think fit in what we were trying to do beautifully.

Ingenix presented this -- and I can speak as a customer. They presented it as though we were some sort of general customer of their solution, and it just was rife with, well, you can do this if you want, you can do that, and the system supports this, and it supports that.

That was not what we needed to hear, and I think Tom is right. It caused problems, because then we had to spend extra effort: Well, what is your recommendation? What do you really think we should be doing?

So I think the way -- The developers should be told to be specific. I think that is really important.

DR. BURSTIN: And, actually, just to build on that, that is exactly right, and I was smiling at Karen, because we have faced this in the past around some other sort of measurement systems, measures that emerge from broad based measurement systems in the past.

It is probably just a broader issue we need to talk through about when there are sort of customizable measure options in a broader measure system, and we are trying to get a standardized measure for NQF. There is sort of an inherent tension there that it would be helpful for us to think about as well.

MR. BOWHAN: This may be a repeat, but it sounds a little bit -- It depends on what the intent of NQF is. Is to have a measure that allows organizations to compare themselves across the country or is it just something else?

To the extent that it is comparable across the country, then it should be specified.

CO-CHAIR STEINWALD: Just to play devil's advocate a little bit here, I appreciate the point that was made, and especially that it fits into the NQF process. The other organization that you don't actually compete with has a process that is more consistent.

On the other hand, don't we value flexibility? If the Ingenix approach is one that really is better tailored to meet the needs of their customers, what is to do?

DR. BURSTIN: I will try one volley back, and Karen may want to engage on

this.

So I would argue that customization is perfect for internal quality improvement, and that is ideal, and they should continue to do that. God bless them, you want to make it work for your individual system, but at the end of the day, if you really want to be able to compare apples and oranges, we kind of need some standardization that allows us to do the comparisons. But I agree, you could still do whatever you need to do for internal QI, but it is probably not what we are talking about in terms of standardized measures.

CO-CHAIR ROSENTHAL: And I think we meant well when we put the specifications out, because nobody really had done this kind of work in this space before, and I think we opened it up with just that idea of let's, in fact, allow a little bit of inclusivity, because if we were too specific in the first round, maybe we would "scare people off,"

because there is no frame of reference against which to evaluate the specificity. But now we have one round, and again the outcome was allowing the flexibility, I think, created more difficulty for us in evaluating the measures than it did by helping move the thing along.

CO-CHAIR STEINWALD: Jack?

MR. BOWHAN: I think you can do both. There is no reason that you can't have a very specific measure that we can use nationally, but Ingenix still has tons of flexibility in it, and the user can use it the way they want.

CO-CHAIR ROSENTHAL: Yes, but the question is what are we going to put in the directions to measure developers? Are we going to say, dealer's choice, or are we going to say, you know, you have got to specify these things, and it has to be precise? My vote, based on the experience, would be more specificity.

CO-CHAIR STEINWALD: That sounds like a principle or something you guys could work with.

MS. WILBON; Yes. I just wanted to add, another reason why we ended up having this module and the last reporting module as guidelines or specifications is those are things that NQF has not typically required as specifications on the quality side, in terms of how you aggregate your -- how you collect your data, how you clean it.

So in terms of trying to be consistent across the organization -- but I think that is also something we probably need to talk about internally on what we tend to require, and I don't know if you want to --

CO-CHAIR STEINWALD: I think this might be -- Go ahead.

MS. PACE: I just want to say -- I mean, just because you call it something different doesn't mean we don't have comparable things on the quality side. There

are very specific identification of cases for the denominator and numerator and the measure logic.

MS. WILBON: So this module that we are talking about is actually before you even -- So when you have like a database of data, how do you figure out which data even gets pushed into the measure. So it is a little bit even before that. Right?

MS. PACE: It is still comparable. It is still identification of the data that you are going to use in the measure. So I think you -- I would consider that part of it.

CO-CHAIR STEINWALD: Steve, and then Jack.

MR. PHILLIPS: Yes. I was just going to make a comment that I think, and tying it back, I guess, to the conversation preceding this about the issues around getting pharmaceutical data, and that there we need a lot of flexible thinking about how to get some of this data in.

So I think it is a process in thinking about, well, what do we need to be specific about. Personally, I was willing to live with more flexibility around the pricing, but be that as it may. So I think that we just need to hash out, or NQF folks need to has out, I guess, what areas is it really important to have specificity around, and where can there be some benefit for seeing some framework allowing for flexibility to have innovative thinking.

CO-CHAIR STEINWALD: Jack?

MR. NEEDLEMAN: Yes. I think the conversation we are having underscores in part the dual nature of the NQF process here. On the one hand, there is a measure endorsement process, and that is, I think, an inherently conservative one. We don't want to endorse measures that still have ambiguity, that we are not quite sure work.

On the other hand, we want the measure development process, and we want NQF

basically creating -- I think, creating a vision of where we need the measurement world to go and, in that sense, to be supportive of innovation, to be supportive of identifying the directions that things have to move in, and how those two roles play out in the work of a committee like this or the Board as it reviews measures, I think, is a challenge for the organization.

To take one concrete example, I know a number of people in the room voted against a lot of measures, because they didn't trust the attribution module down to the individual provider level on a number of the measures that were promoted. That is fine. If we don't think the attributions are right yet, people haven't solved the problem of doing that correctly, it shouldn't be endorsed.

On the other hand, folks like Dolores' members, folks like UCLA Health System as they analyze their resources are

trying to think about which of our doctors are delivering care efficiently, which do we need to have some conversations and some education with and maybe pair up with some of the folks who -- and collectively, as an organization, what do we have to do to figure out how to learn to use resources more efficiently.

That does require some degree of attribution, and people are going to be experimenting with how to do that, how to deal with the weaknesses and limits of attribution as they think about what conversations are taking place internally, and how the data is presented and how the data is used, and how the conversations around the data are structured.

If we don't believe the attribution algorithms work yet, they shouldn't be endorsed, and some of us feel that way. But NQF needs to think about how it creates a vision of the agenda for further development, which is not a conservative

agenda -- it is a very aggressive agenda -- and how that balances with a very conservative what are we going to endorse agenda.

Those two, I think, have come up repeatedly in the discussions we have had, and has been a tension in the Steering Committee. It would be nice to have had a little guidance on how to balance those, but that is part of the conversation we need to have about any set of measures that are being approached.

CO-CHAIR STEINWALD: I would like to pursue that a little bit, but first, Sally, you wanted to say? No?

MR. AMIN: A lot of this discussion is -- I don't want to stifle it because of the nature of the day, but it falls into the reporting module section that we will be discussing this afternoon. We can keep going with it, but I know that issues of attribution and sample size could take over.

CO-CHAIR STEINWALD: Well, actually, with the prerogative of the Chair, I

would like to -- Since Jack raised it, it was something that was on my mind as well. I think we have this kind of unfortunate cul de sac in the logic of the specifications as they exist, and I think it is largely on reporting, which is a discomfort with reporting at the individual provider level.

Yet as Jack points out, if you develop a measure, and let's say it is at a higher level of 400 providers, you as an organization want to know within that 400 what are the performance variations.

You may not have any desire to report, but you have a desire to drill down and discover what the data can permit you to discover.

So I was often very uncomfortable with what I described as a cul de sac, and I know you pointed out repeatedly that, well, we can't do public reporting at this level, but that doesn't mean we can't do analysis and feedback and things of that nature.

CO-CHAIR ROSENTHAL: This tension goes all the way back to the very first meeting we had, and I think, Bruce, you and I were both -- We even posited the idea that could there be measures that we would sort of semi-endorse, and the answer we got was absolutely not. In fact, I think we took it on about three times before we both finally got the message of stop it.

Yet at the tail end, it comes up again, that tension. And, yes, when something is going to be a nationally endorsed measure, I personally believe it has got to meet all of the criteria, which is why so few of these in the first round got through, and yet I am with Jack.

There needs to be a way for us to make a statement about the necessity of moving some of these forward and figuring some of these things out, so that the next round is more successful, and I don't know what that is, but maybe it is the white paper. Maybe it

is the report. I don't know what other options are available to us.

CO-CHAIR STEINWALD: I think it also relates to the issue that Barbara raised. What is the responsibility of the measure developer and what is the responsibility of the user, and does the measure developer have to take on some of the user's responsibility? Maybe you have a different issue.

DR. RUDOLPH: I just think we should in the submission actually clarify what the standard is going to be, because it is really a lot of work to go through and submit measures, and information in white papers, briefing papers doesn't really get translated when you are out in the field into what should be actually included in your submission and what the measure should or shouldn't do.

These are very concrete decisions by the measure developers to go one way or another, and if they don't -- If they haven't gone through the process before, if whatever,

they may not know sort of the fact that we want people who do auditing or we want this or that.

So I just think it really has to be very clear. That way, the measure developers know up front that their measure is going to stand a good chance of being passed. It is too much work to go through.

If sort of the attitudes and beliefs and values of the Steering Committees are in a certain direction, we should acknowledge that and say this is -- you know, in order to get passed, this is what you have to do.

CO-CHAIR STEINWALD: Why don't you give us some guidance on what to talk about next?

MR. AMIN: Oh, okay. We will move to clinical logic. I know everybody wants to get there.

CO-CHAIR ROSENTHAL: We have hardly figured out the agenda yet.

MR. AMIN: Yes, sorry. We will keep reporting for the afternoon, so we get everybody's thinking around. That was a joke.

So for Module 2, the clinical logic, looking at the overall issues: The clinical logic includes the steps of identifying the condition or event of interest, the comorbidities and disease interactions, the clinical hierarchies, the clinical severity levels, and the concurrency of clinical events.

At this point, we are also going to have a little bit of discussion -- I mean, clearly, we will have discussion around the same concepts, but I think this will be an introduction into how we could start to think about this for the Medicare population, because all of these steps become infinitely more complex when dealing with that population.

So the two major issues that were discussed, that were brought up in the TAP and

the Steering Committee were along the lines of exclusions and clinical severity levels, and specifically for exclusions it was a question of ensuring -- what we have heard from the TAPs, ensuring that exclusions weren't -- patients weren't driven out of measurement by care that could be potentially related to poor -- that could be related to poor care.

For example, creating risk stratification approaches on subsequent revascularization for patients with PAD post-revascularization had the potential -- if that was the criteria that drove patients into higher risk strata, it had the potential for creating unintended adverse consequences or potentially having these patients removed from the measurement was a concern that was brought up many times.

Secondly, the issue of clinical severity levels was around the complexity of the methodology of linking patients to the severity level. For example, in the Ingenix

measures we heard many, many times around the complexity and lack of clarity.

While detail was provided after asking -- after requests on additional information from the TAPs, there is still lack of clarity on how these patients were actually assigned.

I don't know if there is additional information or additional -- definitely, this was part of Carlos' evaluation in many of the measures. So specifically, some of the questions that are posed here: What are the appropriate characteristics to exclude patients out of measurement populations?

Again, another example that was used in some of the TAPs were excluding patients with AMI who were discharged to a skilled nursing facility or exclusions of patients that died during hospitalization. That, clearly, has the potential for bias in the measure score.

Then the second question that we will pose here and that will continue to be a theme in the rest of the discussion is: What special considerations should be made for considering the clinical logic for patients who are over 65 with multiple co-occurring conditions?

This could be thought of in two different frames, first when looking at individual measures, but also as we are looking in the future to actually evaluating groupers. What other information might we need to actually start to evaluate this?

I recognize that is a big question to be asking, but we can break it up into chunks as we sort of think about this, and we could think about each of them individually, but I sort of pose that to the group and the Chairs, and there may not be answers to this, so to give everybody that out also.

CO-CHAIR STEINWALD: Steve has an answer.

MR. PHILLIPS: I just had a comment, that what I would have found helpful, I think, looking back on kind of how the request for information was laid out, would be to actually have the requester include some information about kind of this under 65/over 65 break within the data, so that it is kind of clear and up front, I guess.

Particularly since CMS is such a big user and looking to use this data, I think that would be a good way of kind of informing, okay, is this a measure that can be adopted in the Medicare population or is it based on private pay patients primarily in terms of the analysis that has been done.

MS. YANAGIHARA: I think one of the things that, I think, is really challenging is, when you have measurement systems that are developed to parse every claim into a particular bucket so that then you can roll up all of those different buckets to get an overall, when you are trying to look

at a measure for a particular area using that kind of a system, it may not be complete; because you may have parsed some of the costs or some of the resources that were used for this particular thing into something else, because it also applied to something else, and their logic applied it to something else more than this other thing.

So it is a tension of these systems that are created to do one thing being applied in a different way in an individual measure. I think that that kind of plays into this whole multiple co-occurring conditions.

It is like, if you have all these different conditions and you are trying to just measure one of them, if you are using a system that parses everything into just one bucket, you are not going to be getting everything that has to do with that particular condition.

So I don't know what the answer is. It is just something that I have been

struggling with as we go through this. It is like this is not a complete measure on its own. This is a complete measure when you are looking at in the context of the whole thing, but not necessarily on its own.

CO-CHAIR ROSENTHAL: Well, what struck me is what is the definition of an episode, because we had episode grouper people, and some of the episodes -- Frankly, even Ingenix has got four or five hundred or several hundred for sure definable episodes, and yet they only chose to bring a dozen forward. Then when we looked at it, only a small number, at least to some of us, then made sense to carry on.

So, really, what is an episode, and which one of the episodes actually works for this kind of comparative thing seems to me to be an unanswered question. I think we are in the process of answering it.

I wanted to follow on Steve's point just for a second, because I thought

that is a really interesting notion of the episodes that we saw that were in commercial populations, we never asked the question, I don't think, which ones of them should Medicare also ought to be considering, because they would easily and clearly work in the Medicare population.

Ultimately, you want this thing across the whole commercial Medicaid/Medical spectrum for completeness sake, and I don't think we ever posed that question.

MS. WILBON: There was an item on the submission form for checkboxes where we asked them to identify which population they tested the measure in, and it was commercial, Medicare -- I don't remember what the other -- Medicaid, and then Other option. But that was only for what they actually tested it in versus -- I don't know -- I see what you are asking for. It is what they tested in, which is versus what it could be used in.

DR. REDFEARN: I am really

concerned about how you are going to deal with the complexity of the episode grouper models. For example, just to give you another example, Thompson and Medstat wasn't here doing any proposals, but if you wanted to try to understand the Medstat severity adjustment, you have to understand their disease staging model, which is pretty deep and complicated.

That just kind of brings up another point to me, that I thought the forms that were used for the measure developers to use to submit worked better for NCQA and their kind of an approach, and didn't work very well for the episode models.

Frankly, I got really tired of reading that stuff. It was a lot of repeated stuff, a lot of complexity that didn't really illuminate what I wanted to know. I know the Ingenix stuff pretty well. I was hunting to try to find things. I didn't think that way of submitting it worked.

Frankly, what I expected Ingenix

to say was, go to our transparency and read all the documentation we have about our clinical model, because that is where the details of the clinical model is, and it didn't -- That is the legitimate way to evaluate Ingenix clinical models, is to look at that level of detail, not what they put in that form.

So I really don't know what you are going to do when you go forward and start comparing clinical groupers. Let's say you are looking at something that Thompson proposes and something in Ingenix proposals and try to make sense of them. These forms are not going to -- It is not going to work. They are going to be confusing. They are not going to contain the data details that you need.

So I think that way of submitting those proposals is not going to work very well for that model. That is my concern.

CO-CHAIR STEINWALD: Do you have

advice on how they should be advised? Maybe later?

DR. REDFEARN: Maybe later. Frankly, I don't think Ingenix did a very good job of dealing with the way it was said, and so that is part of this. But I don't think the model fits episode methodologies very well because of the inherent underlying complexity.

Related to that is, if they had said, well, go to our transparency site and read it -- I mean, who here on the committee has time to do that? That is the other thing. Even the TAP -- I don't think the TAP members have enough time to go through that.

It is literally a matter of months of studying that kind of stuff to try to understand it, and so the inherent complexity of the measure is so deep that how can a committee that meets, you know, five or six times a year for maybe a total of 20 hours going to deal with that and discuss is? That is my concern.

CO-CHAIR STEINWALD: Jack, and then Steve. Barbara, do you want to respond to this specific thing first?

DR. RUDOLPH: No.

MR. NEEDLEMAN: The issues that David raises -- we are looking at a major issue here, and what we get are the end results of, frankly, a lot more hours of a lot of very talented people trying to solve the problem in the measure development.

I think it is helpful to go back to what problem we expect to solve here and to think about the clinical experience of the patient and the clinician, and to see how well the treatment of the multiple conditions is doing.

It strikes me that we have got two issues with multiple -- patients with multiple conditions. One is, if we are dealing with a specific condition, we are trying to understand how much it costs to treat, what the resources are that are being used to treat

coronary artery disease, CHF, diabetes.

There are some of these comorbidities that directly add to change the way the treatment is delivered and, therefore, affect the resource use. If you got a patient with dementia, what you do and what is prescribed, what you do, who else you have to deal with is very different than if the patient is fully competent.

If the diabetes already has vascular complications associated with it, the way you -- what you are going to do as a clinician when that patient is in is going to be very different than if they are relatively -- if the disease has been relatively complication free in terms of how far advanced it is.

All those are direct factors that affect the treatment of the disease, because they are directly related to the treatment of the disease and the treatment decisions of the clinicians involved.

There is another group of things that we have got with these comorbidities which are to ask a very different question, which is: If you have got a patient with diabetes who also has asthma, and they have come into your office, Helen, what are our expectations about the time and attention you are going to give to the other things on their problem list beyond the disease and the immediate complications and factors associated with that, and what do we expect there, because those are also going to affect resources that are used?

We need to understand how well the risk adjustment and the dealing of the comorbidities and the complications, how the clinical logic translates to what we expect in terms of resource use.

Right now, people are basically being very crude empiricists. They have got their list of things, and they are running regressions or their equivalent, producing

groups that they think are similar, and then doing -- How much does this add to the cost of treatment, and that is how much we are going to risk adjust.

It seems to me that to get closer to the heart of what we are trying to do here and to be comfortable with it, we need to think about these comorbidities and complications in terms of these two issues: How much do we expect the direct complications or the disease staging, whatever, to affect the resources that are used and, therefore, the costs of treatment; and what else about this patient with multiple comorbidities, multiple chronic conditions, do we expect to also be taking care of when they are in there for their diabetes treatment or their asthma treatment, and how will that affect the resources used and the cost of care?

If we go back to those core visions of what resources we expect the system to consume, then we can think about how well

the grouper or the resource use analyzer is doing in effectively taking those into consideration when it tries to answer the question, how many resources we use for this patient.

MR. PHILLIPS: I have just a general comment following up on some of the ideas about kind of helping up front kind of spell out the application. So I apologize. I am not answering your questions directly.

For me as a reviewer, I think having -- You know, we have got a brief description of the measure, a brief description of the clinical logic. I think understanding the episode clearly up front and some of these key issues that we have focused on about attribution -- having those up front, I think, would help. Then as you are going through the details really, okay, I want to dig into this a little bit more, but just to frame kind of the -- what the proposal is, I think at least those two would be something we

might consider putting up in the description section.

MS. YANAGIHARA: I will second that. I think, just like a two or three page, whatever, summary overview of all the key things, and then where to refer for the detailed information might be helpful, because I found, like I read the Health Partners description that was like a three-pager. Wow, I understand the measure way better than like digging through these 45 pages of submission, just because you have to have that framework. So I think that is a great idea.

My comment was going to build on David's -- my second comment, I guess, was going to build on David's comment. This is something I know that we weren't out to endorse a measurement system, but in reality, to try and -- to endorse one measure for a measurement system is basically endorsing the system, because you have to buy the whole system in order to do that one measure.

So I just -- I know that there is a next piece of work looking at public groupers and things like that, and maybe that is where all of this -- I think it really does make sense to look at which -- I mean, compare the systems, and pick a system, because otherwise you are really telling people you have got to buy all these multiple systems to get these individual measures.

So I think that is just something to keep in mind. Then you can dig into the clinical -- the underpinnings of how this was developed, and what are the differences between the different grouper methodologies, and which one seems to really fit and make the most sense, and there is an assessment of that, instead of trying to do it in the context of one measure.

DR. RUDOLPH: I wanted to just talk for a minute about the clinical severity levels part of this, and Jack was talking about that as well, I think.

One of the things that we saw at Leapfrog in looking at the length of stay measure was that some of the things that individual clinicians thought increased length of stay actually, when you put it into the models -- and this is maybe -- maybe these were cheap models -- that they really didn't have any contribution to length of stay.

So each time -- I am a little distraught about the fact that you assign a severity level, and that severity level goes into every measure, whether or not it is really contributing to the outcome.

So particularly in additive models where you just add up the number of points for that severity, you end up giving sort of -- giving the power of the measure away in the sense that people -- the measure is being kind of risk adjusted. Perhaps the errors or the problems are being risk adjusted away by these sort of additive models of clinical severity levels when, in fact, that particular item

might not be a contributor to the end outcome.

Does that make any sense?

CO-CHAIR STEINWALD: How does that translate into something that the staff might write about or modify the requirements to the specifications?

DR. RUDOLPH: I think it is what steps have you taken to assure that some of the items in the risk adjustment models are actually contributing to the outcome. In other words, particularly if you have additive models of risk factors -- I just feel like I know that was the case, I think, in the NCQA model -- or Ingenix. One of them had it where you just add up, in essence, the number of times if this particular diagnosis pops up, you count it for whatever measure you are doing.

So it just seems like there ought to be some specificity about the way that that works, that the method that they have used actually includes only those conditions that

truly contribute to the outcome that they are measuring, in this case resource use.

DR. ELWARD: Ingenix did a little bit of that for the pulmonary TAP, although it could have been better, and it was very helpful for them to say here is an example of how it would work, for example, if somebody has COPD and CHF. Give me an example of how you differentiate. Shortens breath. You know, the resource use associated with that.

You can lump them altogether and say, boy, if you have both of them, you are going to use more money, but we know that, but some specific examples of saying here is how it actually works in this situation, so we can tease those kind of questions out.

If they can't provide that, then I don't think they should even bother to submit something.

CO-CHAIR ROSENTHAL: Well, in the quality world I think how we would try to grapple with this is that we would look back

to the literature, and particular the peer reviewed literature, and at least that is often my guidepost, and you go back and you look at what they wrote in the Annals of Internal Medicine, and you go, oh, and here is the p value, and somebody has studied this, and you learn.

The challenge here is, with a few exceptions, there is very little peer reviewed literature about this stuff, and that is sort of a complaint. But I think, to the extent -- and the exception to that, actually, from my experience about this is the NCQA folks using the HCC methodology is well described in the literature, and you can go back and you can read that literature, and you can make sense out of, oh, I see the limitations, it has been well studied, and I know the people who have studied it independently, and now I can draw some conclusions.

With the stuff that is solely proprietary, they have had very little

incentive to publish and, therefore, we are left with a real paucity in relationship to the way we would normally evaluate stuff like this.

I am not quite sure what my point is, other than it is missing, but we should ask for it where it exists, because at least you would put some additional pressure on: So what is the basis of your saying that, when you count up these things, that that ends up with a risk adjusting that is adequate other than, as David has suggested, well, go back and read our entire website, and you will find it there somehow, if you are really very skillful.

DR. REDFEARN: Actually, to Barbara's point, one of the advantages of the episode methodologies is that the risk adjustment built into it is episode specific. So Ingenix risk adjusts that episode as very specific as opposed to NCQA using HCC which is the patient risk.

So the point about the relevancy of the risk factors to the condition we are looking at is implicitly a criticism of the NCQA approach, not the Ingenix approach. Medstat does it the same way. They risk adjust inside the episode. They don't use the overall patient, although Medstat occasionally uses the patient stuff things, both.

MR. ALZOLA: I agree quite a bit with Barbara's and Tom's points. One of my complaints, I would say, is that how little detail they put into the description of their risk adjustment models.

That went for all the submissions, and nobody really presented any technical detail on how they arrived at their models, what kind of models they were, any descriptions of how good the models were. So that will value for making these things part of a specification.

I am not -- and, really, it is not something that I like to do, because I like to

allow people creativity, and they can do things in a different way, and they don't have to evaluate it the way I would do it, but if we don't do that, it seems that we won't have the information to say, well, this is a good risk adjustment methodology or not.

CO-CHAIR STEINWALD: Across the board, you are not singling out any developer. You think that the submissions were inadequate in the extent to which they described their risk adjustment methodology and supported it through any sort of their own analysis or external analysis.

So does this sound like a group, something that we might, going forward, want to suggest could be an improvement in the future, that if there is risk adjustment, it needs to be described?

CO-CHAIR ROSENTHAL: Yes. The only reason I spoke positively about the NCQA one, though, is that they did use a risk adjustment methodology that was, in effect, in

the public domain. So in fact, their specification could have been an opinion: Here are the three articles on the HCC methodology, and anybody could look it up without them having to necessarily reprint all of it.

So I would say, if it is not in the public domain where it is independently verifiable, the methodologies were lacking. One or the other would seem to me to be acceptable, but I agree with Carlos. They weren't really there to look at.

CO-CHAIR STEINWALD: Paul.

DR. BARNETT: Yes, I agree with Carlos, too. That is exactly right, that we didn't really have that sort of good table of evidence showing how the models work.

I also think it is a little bit naive to think that we are going to be able to look at a submission like this and really evaluate what is going on with a system like this. What it is going to take is some study

where somebody takes one dataset, and they run the different groupers on them and compare them.

In fact, some of those studies have been done by McCurdy and Thomas and others, and seems like we ought to be looking at the evaluation that neutral third parties have done of these different methods and see what their findings are, because that is going to be stronger and more impartial information.

In terms of the presentation, I think it is worth noting that -- and this is something that I have learned from participating in this, is that the methods seem to fall into two broad categories, and I am not sure if anyone knows which is superior.

One is this idea where we take the claims data, group them into episodes, and then compare costs of episodes, and the other is where we are taking all data and trying to use the risk adjustment to predict all costs.

So each has its strength, and each

has its limitations, and I think the complexity of assigning care to episodes is going to be very difficult to look into how that is done.

One concern I have about that is kind of this joint cost problem where ultimately -- So every visit gets assigned to one episode or another episode, as I understand it, and yet multiple episodes are being managed in a single visit. Right? So somebody is getting their diabetes care and their hypertension care and their hip pain all dealt with in the same visit.

So how do you assign that visit? So I always worry about when people have a joint problem, and they try to assign the cost to mutually exclusive categories that they are engaged in an undoable activity, a fool's errand.

So I believe in econometrics. So we try to parcel that out with regression. So that tells you the nature of my bias, but in

any case, there are these two different approaches, and it is worth noting that and understanding that as a part of submissions.

DR. REDFEARN: You guys are probably really sick of hearing about Ingenix methodology, but I have to -- That just brings up a point which, I think, is really fun.

They have a concept such -- they call it phantom episode. When you have multiple diagnoses being rendered in a particular physician-patient contact, they do the best they can to assign that contact to an episode, but if there's multiple diagnoses and they think something else is going on, they create a phantom episode, and the phantom episode sits there waiting, looking at more administrative claims data until it thinks it finds a service that matches that phantom episode, and starts another episode. But phantom episodes can exist in your data and never get brought back into reality. They sit

out there, and they account for some diagnoses.

So this business of -- This is a critical issue in episodes, is how do you parcel what goes on in those contacts when you have multiple conditions going on, and every one of the groupers have a different clinical rule for how they do that.

I am just afraid -- I think the benchmarking idea is a great idea in terms of doing this, but I am very much afraid you do the benchmarking, and the conclusion is they are different.

CO-CHAIR STEINWALD: Good luck on writing up that comment.

So, Paul, the issue you raised -- do you think that is an issue that the committee needs to kind of address and take sides on?

DR. BARNETT: Well, I think the first issue is that we can't ignore the literature that is out there and the reports

that have been commissioned by CMS that have compared these methods. And of course, we have only -- So I know that Thomas and McCurdy have done these evaluations, and there are some reports out there on these products that are comparing them head to head.

We haven't looked at that literature, and that seems like we should have. Of course, there is this all historical accident that some of the people submit to us because they are submitting directly to CMS and all that going on, too. So I understand the comments there.

CO-CHAIR STEINWALD: I understand it.

DR. BARNETT: We rely on whoever submitted. That is who we are going to evaluate. So the evaluations they did include groupers that weren't submitted to us.

CO-CHAIR STEINWALD: To make sure I understand the issue, it is the are you inclusive about all of the resources that are

utilized by a patient in a given episode, and then try to use regression to identify the costs associated with that diagnosis, or are you less inclusive at the outset. You try to eliminate resources that you don't think are connected with the episode at the outset.

DR. BARNETT: Right. So I am just noticing, there is a kind of a broad taxonomy approach. That is a separate kind of disjointed, entirely independent comment that I made. But, yes. So there is this taxonomy, whether you episode group or not or look at all costs.

CO-CHAIR STEINWALD: Okay. Kurt, and then Jack.

DR. ELWARD: Yes. Two comments. One is that, again, I think the idea of these phantom episodes are what -- is a headache a headache or is a shortness of breath, shortness of breath, and for what, are very important.

Again, I would make a plug for us

looking at how other people have done it throughout the world. But also I do want to reiterate what -- or support what David was saying.

One of the things that did come through in Ingenix was the concept of how they handled that question, which I think is very well done. While there are other problems with Ingenix, I think they came across with at least one very solid approach of how to keep from having everything piled onto one diagnosis, and the phantom concept is really -- It sounds a little weird when you first read it. You go -- you know, supposed to see the Green Hornet next. But they handled it pretty well.

DR. BARNETT: If I may, it doesn't entirely solve the problem, because the phantom can pick up, yes, so you know this lab test is really about diabetes, because it is a hemoglobin A1c. So that is not the hypertension episode. So, yes, you can add

that on, but ultimately that visit is only being assigned to one or the other.

Maybe it works out. I mean, it is an empirical question whether it works out, but whenever you produce two products simultaneously, in this case a diabetes episode and a hypertension episode, you have to make some rule about how to divide the costs, and this one is where we are going to assign this visit to one or the other.

I am not sure how it works out with hospital stays. That could be a pretty profound effect on what you think an episode costs. So if that episode gets entirely assigned to diabetes or it gets entirely assigned to hypertension, that is going to markedly affect your results. In fact, both products are being produced simultaneously, diabetes care and hypertension care. So that is the joint products problem. There is a lot of literature on it in economics.

The classic one is hides and

tallow, as I recall from my undergraduate course.

DR. NEEDLEMAN: So a joint problem. So a patient walks into a doctor's office, and the patient has COPD and diabetes and a bad knee. Okay? This is the joint production problem. So if they walk into their primary care doc's office or an internal medicine office and they have COPD and diabetes, we don't expect the hip to be a large portion of that visit, but Paul's question is which of those conditions is that visit being applied to? Is it the COPD episode? Is it the diabetes episode?

Well, if they have walked into their primary care doc's office, they are probably getting both problems discussed. So suddenly that short visit, because it is routine, turns into an intermediate visit or the intermediate visit turns into a long visit, perfectly appropriate for discussing both.

The risk adjustment methodology basically says, yeah, we expect longer visits in a COPD case if the patient also has diabetes, because we expect something else to be going on in that visit. So the risk adjustment says more resources are appropriate.

If we have thrown that patient into the diabetes episode grouper, because they had come in for diabetes, we also expect more time because of the COPD. So more resources we would expect to use.

The issue Paul is raising is do we expect that visit to go into only one of those groupers, into only one of those episodes, or is the visit really about both of those conditions and when we throw them into the COPD category, that visit should be counted in the COPD grouper; and when we throw them into the diabetes episode, that visit should also be in the diabetes episode?

So we got this issue of are we

saying only -- that visit only goes into one episode or not? That is one issue that Paul raised. The other is, given the comorbidity, do we expect them to go into the -- do we expect more resources to be used, because we expect those other things to be treated? That is a matter of risk adjustment to the resource use, and that is a different issue than are we only counting it in one episode or are we counting it in multiple episodes.

If that same patient takes their bad knee, talking about joint production problems -- takes their bad knee into an orthopod's office or into an anesthesiologist's office and they got the COPD, do we expect that to be affecting the way in which the discussion of treatments for the bad knee is taking place? Orthopod, I don't know, but for sure the anesthesiologist is going to want to take into account the COPD as they think about anesthesia options.

So we've got all -- So when do we

expect the -- But is that a risk adjustment model? It is clearly not the same problem of attribution. This is a visit about the knee. It is not a visit about the COPD per se, but there is a potential need for risk adjustment as the orthopod or the anesthesiologist deals with a more complex patient.

So we have got two different issues here, and it is important to understand how the groupers deal with them. Do you count the same visit in multiple buckets or do you try to arbitrarily assign it to one bucket and not the other, and do you risk adjust to the complexity of the patient where we expect other problems to be dealt with in a visit with a primary care doctor or an internist, and how does that logic apply when they are going to see a specialist about something like a knee or, frankly, a specialist about the pulmonary problems?

Those, I think, are two different issues. We need to understand how the logic

deals with it, and then the committee should be thinking about whether there is some logic that we prefer in dealing with those two separate problems over others or whether, as long as the logic is convincing, we will let the grouper deal with it and let the customer decide which of the logics they prefer.

DR. BURSTIN: Yes. As I say, in addition to the work Taroon is going to show you shortly the patient focused episode work that we have been trying to conceive over the last few years of not being so episode grouper specific, but really in a patient centered context, what does an episode look like.

We have actually got a group now working as well on a multiple chronic conditions framework, just recognizing the reality that this is such an artificial distinction of figuring out -- I mean, well, Kurt, I live this all the time in practice.

I mean, my patient routinely walks in with five to seven comorbidities. So even

one to two to three isn't relevant for most of the -- as a general internist. It doesn't compute at all.

So I think the other thing that might be interesting is, as that framework emerges, we will share it back with this group, and perhaps you can reflection how you are able to think about episode based cost measures. How does that fit in that multiple chronic conditions framework?

MR. AMIN: There is -- This is very helpful. So there is a lot of very important things that are being discussed right now. So I just have a few different topics.

Before I go on, I do feel the inherent need as a disclosure to say that, before I joined NQF, I was working on the public sector episode grouper work with Brandeis and Prometheus.

So some of my sort of orientation comes form that background. But I think this

issue, Paul, that you are bringing up around how the claims are attributed to an individual episode -- there are different methodologies out there that can actually, as Jack is pointing out, attribute that claim to multiple episodes occurring at the same time.

So this idea that it needs to be assigned to a specific is, I think, a residual of the fact that that is one approach that we saw through this process, but there are other approaches out there that are, as Helen pointed out, trying to conceive the unit of analysis as the patient, looking at it across the patient centered episode of care, not necessarily creating these episodes as forcing binary decisions in some way, and better understanding how that works, I think, is a clear take-away that there is a level of specification that we really need to think about as we sort of look at the Medicare population.

Additionally, the question that

Carlos brought up around the risk model: We do have some submission questions in our current submission form that asks the question of defining risk adjustment and variables and describing conceptual, statistical and relevant aspects of the model.

The question I would ask the group, and Carlos also specifically: What other characteristics are we looking for? From what we have heard through the TAPs and the Steering Committee, it seemed like, clearly, the R-squared or the goodness of fit of the actual final risk model was really important, but also the question of how specific variables were included into the risk model, whether they were just based on statistical significance or if they actually had some question of clinical validity in inclusion into the variable seemed to be another of specification that was needed. But is there additional information that we think we need to evaluate the appropriateness of the

risk model would be a question that I would again frame to the work group?

CO-CHAIR STEINWALD: We are due for a break. Could we ponder that question for 15 minutes, and then reconvene? All right, 15 minutes, and it is about -- So 10:45 reconvene? Okay.

(Whereupon, the foregoing matter went off the record at 10:33 a.m. and went back on the record at 10:49 a.m.)

MR. AMIN: So I will just reframe the question that I sort of posed to the group also. I don't want to break the flow of other conversation that may need to occur in this area.

We talked a little bit about the risk adjustment model of what other information would potentially be needed, and basically what I heard from the group was that some justification of the variables that are used in the risk adjustment model need either clinical evidence based on literature or some

justification of how they were entered into the model, not just that they were statistically significant, but they have some clinical relevance.

Additionally, all models should provide goodness of fit information through R-squared, but if there was any other additional information for the risk adjustment model -- and this will be discussed again in another module -- and also, if there is some guidance -- and this is a totally different topic, but if there is guidance on how claims for patients with multiple co-occurring conditions should be assigned to an episode, if this issue of the binary logic that it has to fit into one particular episode is limiting, and the committee feels that this is not an appropriate approach, I think that is another area of guidance for when we are looking at Medicare populations, it would be helpful.

CO-CHAIR STEINWALD: I encourage you to continue doing what you are doing now,

is repeat what you think you heard. The committee should think of what the staff say as being in print and being representative of our views, not the staff's. So please keep doing that.

MR. ALZOLA: With respect to the question of what are things we should require in terms of evaluating the models, one thing that I think is crucial, actually, is the calibration of the model. That means how well the model predicts at different ranges.

So for patients who have low resources, they would have a low prediction. Same for the middle and for the extremes, the ones with high resource use, they would predict a high resource use.

That is usually pretty difficult to do, but most models already would have predicted means, but the real interesting cases are the outliers, the ones who are very expensive to treat, because you could have a situation where we -- For the very low, you

predict very high. So for the very high, you predict very low. So you have something -- and they predict perfectly for the mean.

So you have a situation that connects like this. So although on average the model is going to do very well, but at the specific cases where we are interested, it will not. So that information is really important to have.

DR. BARNETT: I think what Carl has said, we should underscore that whole idea, and it is especially important in costs, that usually where most cost models fail is in predicting the top decile, and I am very uncomfortable with the idea of eliminating the high outlier costs, which I have seen everybody does. I wonder if it is just me, but I don't understand this.

I understand, you know, data has got problems with etcetera, but you worry that, if providers or plans are going to -- you know, are the results sensitive? Rankings

of providers and plans, are they sensitive to the threshold of where you are doing this truncating of the high cost outliers, because it is those train wrecks that we care about, and maybe it is outside the provider or the plan's control, but maybe not.

So I sorry about that. There should be some sort of sensitivity analysis about that outlier trimming.

MR. ALZOLA: Yes. Sensitivity analysis was missing by a lot in all the submissions. So that is something we should ask for.

CO-CHAIR STEINWALD: You wanted to say something?

MS. PACE: I was going to say, I don't know if it got on the resource use form, but in our general measure submission form, we do ask for risk model metrics in terms of discrimination and calibration, and specifically ask for the risk decile plots or risk decile information. So I think that

would be comparable.

MR. AMIN: Yes. It was asked for, but whether or not it was -- I think that some of the take-away is there is a translation issue, and we will have to think about that internally at NQF about how we are able to garner that information.

DR. REDFEARN: Maybe you need -- When you have your specs and you get an initial submission from the measure developers, take a quick look at it and say, sorry, folks, you missed it; you are not doing what we are asking, and give them another chance before we see it?

You can obviously do these reviews. If you ask for something, risk deciles or something, and it is not submitted, you immediately go back and say, you forgot this.

CO-CHAIR STEINWALD: I wonder if we are headed in the direction of asking for more information, and knowing, as we have been

told many times, it is very hard to prepare these submissions, are we going to be considering any ways in which we can make it easier for the developers to submit? It is just a global question.

MR. AMIN: I mean, I think the -- The answer is yes. I mean, the question is how, and I think we will have to figure that out over time. I think that is -- and I think there is a serious question here of developer burden. I mean, as we are sort of asking for this level of information, we also have to recognize that there are organizations out there that need to provide this information to us at a level that we are able to assess it, but at the same time we are not asking for undue burden.

So I think this is all a balancing game, and we will have to think about this as it goes along, but I can sort of outline maybe modules three and four, if we are ready for that. I don't want to push people too far.

They overlap with the conversation.

CO-CHAIR STEINWALD: We are hearing voices from above.

Why don't you forge ahead?

MR. AMIN: Okay. So moving to Module 3 -- and Dolores really set this up for us already, but we just want to pull it out as an additional consideration as we are thinking about this.

The way this evaluation process was set up was to evaluate individual measures, and some of the true challenges that we saw in the TAPs and, to a certain extent, in the Steering Committees were it is extremely difficult to evaluate some of the components of the measures, since they were functions of the episode grouper that were behind the actual measure.

Some of these includes methods of claim assignments to the episode, comorbidities, clinical hierarchies, and the handling of concurrent clinical events, as we

described, and a major issue, at least to me, during this evaluation was understanding this tie breaker logic when evaluating single measures.

So specifically, what this is referred to in a lot of the submission forms were individual and how they were assigned and their relative weightings or -- there was another term that was used -- their relative association to various episodes, which when you are evaluating a specific measure is very difficult to assess.

So some of the questions here are a little bit more overarching, but how can we better evaluate these individual measures when the select measure attributes are part of a grouper, and are we, in effect, just simply evaluating the grouper; and are there additional criteria that should be explored if we are going to evaluate the groupers themselves and, potentially, if we are looking at entire sets of measures, how they interact

with each other.

So there is a question of the actual episodes interacting, but then how do the measures interact, in some sense? I know it is a little bit conceptual, up in the sky, but bear with me.

Then Module 4 is looking at the adjustments for comparability. So one of the questions that was brought up in the TAPs was the appropriateness of various risk adjustment methodologies.

So a lot of the discussions relied on the Societies of Actuaries report of the appropriateness of various risk adjustment methodologies, and there is a legitimate -- There is a question of whether or not, if there should be additional evidence beyond that Society of Actuaries report of the appropriateness of various risk adjustment methodologies for various approaches, and should there be a way to assess the risk adjustment methodologies for the proposed

application.

I guess the question that we are asking here is: In what context -- Well, one of the questions here is are the risk adjustment methodologies specific to individual populations?

So we saw in at least one of the submissions -- this was ABMS prior to maybe even getting to the Steering Committee here -- was is it appropriate to use the HCC methodology in a population that is outside of Medicare.

So it was brought up many times that it is good that HCCs are used, because they are -- HCCs are used because they are peer reviewed, and there is a great deal of literature out there on the appropriateness of HCCs, but the question is: Are HCCs actually appropriate for the intended population within the measure?

So should there be some guidance here about not only appropriateness of the

risk adjustment methodology for its intended use, but also its intended population?

So I will summarize by saying there is just a general bucket of questions on the appropriateness of the risk adjustment methodology which goes beyond the type of detail that we would need to evaluate the measure -- or you would need to evaluate the measure.

Then a question of, really, evaluating the individual measures that are within overall groupers, and whether or not some of these aspects within the grouper maybe are outside of the evaluation. I don't propose that, but it is a question, or whether or not we really should be doing -- really evaluating measure episode groupers at all as measures.

CO-CHAIR STEINWALD: Don't all speak up at once. Make sure you raise your card. The last question as you posed it, whether we should be evaluating episode

groupers at all --

MR. AMIN: Let me clarify. Individual measures that are part of an episode grouper system, whether they should be considered as an individual measure, whether they should go through an endorsement process for individual measures at all.

DR. BARNETT: I think Dolores has said what I feel about it, is that it is kind of -- you know, they are trying to sneak something else into the tent, which is the whole -- you know, you have got to buy the whole product. Right? You got to spend a million dollars, basically, to get this product in order to do one little thing. It doesn't make sense.

DR. REDFEARN: Did you really say that?

DR. BARNETT: But you did say that it was -- you know, you have to buy the whole product to do one measure, and I am just observing that it was -- you know, if I look

at my health plan for a three-year contract, sounds like it is about a million bucks. So that is a lot to just figure out one outcome.

DR. REDFEARN: Well, I can tell you, HCC models developed by Verisk DxCG, and DxCG offers about 60 different flavors of the risk models, and a lot of the variation of the risk models they offer is the population that they are aimed for.

So at least in the opinion of Verisk, it makes a difference which model you use for which population. I can tell you informally, one of the things we are struggling with is: For some of the Medicare business that they say you have to risk adjust using HCC, and then we run some of the other DxCG models on the same population, and we get a different number, and we don't get a really good explanation back from Verisk about why that is happening, but we are certainly seeing that on an empirical basis.

So I think the issue of matching

the risk methodology, risk adjustment methodology, that you are using in this to the population of interest is a relevant question. I don't know if there is any published evidence that would tell you one would help you make the recommendation, but I think it is a legitimate issue. We have seen it empirically.

DR. BARNETT: I will just -- So, David, does that mean that what we are trying to do here is somewhat impossible? I understand what you are saying, and I think it is right, which is that, you know, one risk model doesn't fit all populations. So does that mean that we are never going to come up with one measure that is going to cover all possible cases?

DR. REDFEARN: I think it comes back to the question, is it good enough. My opinion tends to be it is good enough, and this variation, which I actually believe exists, I think, in general is low enough that

you can tolerate it in doing this kind of work. That is my personal opinion.

MR. AMIN: I want to take this question of good enough a little further. And it is okay. It is a question of our level of specificity in how we are analyzing this.

So are we saying as a committee -- or are you saying as a committee that it is good enough that these risk adjustment approaches that were outlined in the Society of Actuaries report that submitted -- state that they all perform equally as well for the populations that are under evaluation for the committee is good enough? So we will use these -- any one of these risk adjustment methodologies in application of these measures that are evaluation is good enough, or what other information would then be required in order to assess that, actually, is available?

CO-CHAIR STEINWALD: You wanted to say? Finish this? Okay. Then David, and then Joe and then back to Paul.

DR. REDFEARN: I am kind of amused at how everybody cites the Society of Actuaries papers. There have been a couple of them, and it serves a kind of a nice purpose, but it is very limited in terms of what they evaluated.

They didn't really think about any of the kind of issues that we are interested in, like what population are you running them on, because they are basically saying I am going to run it on a commercial population.

Basically, you can read those papers, and it boils down to R-squared, and I don't think that is -- and the conclusion was they all give you about the same R-squared, and if that is sufficient information for what we need, then fine, but I don't think it is.

I think we are interested in a lot more than just the basic R-squareds, and if you look at the papers really closely, particularly the second paper, they did a lot of phutzing around with the models and the

data that, if you go back and talk to the vendors that are involved, not a lot of the vendors were terribly happy about what they did to the data.

The one conclusion you can draw from the papers is all the models produce about the same power in terms of R-squared, but they don't address any of the other really interesting issues that I think we are struggling with.

CO-CHAIR STEINWALD: You can write that the committee is amused. Joe.

DR. STEPHANSKY: I am not amused. I am not amused. That Society of Actuaries paper, I think, has done us kind of a disservice. I think, when you consider the dollars that are at stake coming up in the next five to 10 years and all the risk adjustments that have to be done for, say, contracting for a population through an ACO and so on, all of these dollars -- there is a lot of work going on right now -- personally,

I know some of the work at University of Michigan -- in developing new risk adjustment methodologies for specific purposes.

I expect in the next five to 10 years a committee like this is going to be looking at a lot of new ones, and a lot of the ones that we have already started to use are going to be just abandoned. So we are going to have to learn to take a closer look at these things and not accept good enough.

CO-CHAIR STEINWALD: Paul, and then Jack.

DR. BARNETT: So the good enough is good enough for what, and the real question is, if you change your risk adjustment model, does it change the ranking of plans or providers? Is it sensitive to what risk adjustment method you use? And I don't think the Society of Actuaries addresses that issue at all -- their study addresses that issue at all. They are asking an entirely different question.

That is why -- you know, back to the people who have done some head to head comparison of some of these different methods, those are the studies that we need to read and probably need to commission some more.

CO-CHAIR STEINWALD: Good. Jack.

DR. NEEDLEMAN: Yes. Paul said a fair amount of what I wanted to say, that the issue with the risk adjustment is not per se what the R-square is. It is does it change your relative rankings? Does it change your absolute judgments about whether the resource use for a given provider, a given plan, is high or low, and that is the criteria against which things should be evaluated.

The other point I would make about the risk adjustment is it is driven by the data you have, and we've got two issues. One is we've got limitations on the data we have. So if you create enough categories and you tailor the weights to the problem you are using -- so you've got the basic categories,

whether it is the HCCs or for hospital stuff we have got Elixhauser comorbidities.

Whatever way you group the data you have, if you basically making -- using basically the same data to create groups, and then you are tailoring the weights that are assigned to that based upon the data you are going to get, you are going to wind up with about the same R-Square.

You may or may not wind up with the same rankings, but you are going to wind up with about the same R-Square, because you are using the same data, and you are tailoring the analysis to the actual cost or the actual resource use you are looking at.

So we need to think about things beyond R-Square and rankings. We also need to think about the data, that we are tending to think of this as a technical issue of analysis when it is a data issue. Do we have the right data to risk adjust effectively for the differences in resources?

We were talking during the break about the patient who has a spouse is going to get sent home with a bag of drugs and a spouse who supposedly knows how to handle that and the dressing changes and whatever else is taking place after the hospitalization, and the patient without a spouse is going to have a prescription for a home health agency, a visiting nurse of some kind who is going to come, and that is a difference in resource use.

You know, we come back to the data. Have we captured those resources, but the explanation for why one patient is having those resources consumed had nothing to do with what we see in the standard reports of the medical condition, the comorbidities or anything else. It has to do with the fact that they either have somebody at home to help them or they don't. That is a data problem. That is not an analytic problem.

So we need to think about it. If

we want to think about effective risk adjustment, are we using the data that we have accordingly?

Do we get different results if we use different models, in which case we need to worry about which model in terms of the rankings, but also what data do we want to see for making appropriate judgments about what level of resource use is appropriate for a given patient, and do we have that right now or do we need to start collecting it? Those are the issues with risk adjustment.

CO-CHAIR STEINWALD: Tom and then Paul and then Barbara.

CO-CHAIR ROSENTHAL: Well, two comments. One is the fundamental question is good enough for what? I think that has been stated, but we should say it again, and I would submit that, if it gets to shifting major dollars around, to follow up on Joe's point, clearly, what we got now ain't good enough.

The companion comment to that is that -- and it is sort of obvious, but it is a tradeoff between feasibility and specificity. But if you look at a couple of areas where provider entities have taken this on -- and the two that I can speak to pretty straight up are the transplant world and cardiac surgery/cardiology.

What those worlds would consider adequate risk adjusting goes well beyond administrative claims data -- well beyond administrative claims data. But that is expensive, and it is questionably feasible on any large scale, but I would submit, it really -- and this becomes a political statement, not a -- because I am sure people paying the bills would say, hey, we got plenty enough information today to switch the money around, but I don't think the provider world is that accepting of it, and the basis of that statement is looking at least a couple of the condition specific areas where the risk

adjusting is substantially more powerful.

I know the transplant one really well, because I was a transplant surgeon, but the tradeoff is that I don't believe there is -- There is hardly anybody in the transplant world who is a provider level who challenges the accountability they are held to against that risk adjusting. They look at that and go, yep, it is what it is, and I am not going to debate that extensively. But those are the tradeoffs.

DR. BARNETT: Just we were talking about adequacy of risk adjustment, and I supported what Carlos said about the extremes, and only mention the top of the extremes. I think the other place that risk adjustment is very problematic, and it is largely a data problem when you rely on claims data, is at the bottom about people's engagement of care.

So the models -- usually, there are very few risk categories of people who are not engaged in care, and those are real

deficiencies in the risk adjustment models. I think that is a big problem, because that is probably where we can make big gains in efficiency, people who get very little care now, and it is especially worrisome when you have underserved populations or people with limited access, and we know very little about what engages them in care, who is at risk and doesn't get care, those sorts of questions in places where we could make efficiency gains.

So I think one practical thing to think about how you could improve that modeling is if we had multiple year data and data that crosses plans or providers.

So we throw out the people who switch plans. Right? Because we don't have enough data on them. So they are looked as -- So those are some of the people that are at risk, and they are people who we don't know much about their care. So that is a practical thing about thinking about modeling the risk at the low end.

DR. RUDOLPH: Yes. I want to respond to Jack's quest for more data. There's a number of different pilots going on about enhancing administrative data, whether it is with pharmacy data or lab values, then also some efforts that Nado and actually the CEC are making to enhance the data elements that are actually collected, which includes living with another person.

So we are doing a lot of work on those kinds of things to make the data better. However, we really need support in doing that, because providers don't necessarily want to provide that extra detail, because it is, you know, a burden on them.

So -- and the whole issue with race and ethnicity and administrative data is another area where there are fields for it for electronic transactions, but providers are not particularly thrilled about collecting that information, just because of some of the issues related with the type of data it is.

So at any rate, there's a lot of efforts going on to enhance the data. So I think the new models, risk models, are going to have better data to actually use in the future, but it takes a long time to get it around the country.

CO-CHAIR STEINWALD: So do you think that you've got enough content? I would say this. If you are going to write about this in a sort of a forward looking way, you might say that the expectations of -- not this committee, because we will be replaced by another one, but we expect that those expectations would be elevated somewhat compared to what we saw in the submissions in this round.

MR. AMIN: Right. I mean, to just highlight a little bit of what I heard here during our discussion was that the question of -- that the question that was posed, I think, is very clear, that the risk adjustment model should be relevant to the intended population.

I think that was a question where we started, but I think we, clearly, landed somewhere, and additional research potentially, not for a group like this but potential research for the field would think through comparing these risk adjustment models not only on R-Squareds but also how it changes the rating of providers, and looking into the future, additional data elements such as clinically enhanced administrative data, could potentially not only help with the measure scores but actually help in changing the risk adjustment variables -- not variables, actually. The risk adjustment weights, I should say.

The only other additional question that was posed here -- and this was in terms of our large discussion yesterday around costing -- is that in what context should cost measures be used compared to resource use measures?

So there was this large debate that occurred. Now the question is: As we

look forward into informing the field, I think as we have debated the merits of each of the cost and resource use measures, but in which context should each of them be used, I think, is also a question that still remains, which arose from the costing methodology that was submitted between actual prices and standardized prices.

So before we move on from adjustments to comparability, this question still is outstanding in some sense.

CO-CHAIR STEINWALD: You are asking us to discuss that issue here? Does it fit here?

MR. AMIN; Yes.

CO-CHAIR STEINWALD: You know, I thought that from time to time we have been a little bit careless about the use of the word prices versus costs versus -- so what we really mean by standardization, for example.

It also bears on the issue of what is the measure developer's responsibility, and

what is the customer's or user's responsibility, because I think our discussion came to the point where we acknowledged that there are some legitimate uses for actual dollars, and typically what we are talking about is paid amounts when we talk about dollars in that context, versus standardized.

Then standardized, to me, means you are adjusting for the underlying cost of inputs and, therefore, you are standardizing with costs, not prices. Now price of labor can be used as a price, but when it is put into a production system, it is a cost.

MR. AMIN; Right. Okay. So let me just clarify the question, I think, in what we are intending to get at here.

So we have talked about different costing approaches in the measure, some that use standardized pricing and some that use actual prices paid, and those actual prices paid we have termed cost of care measures, and those that use standardized pricing approaches

we have termed resource use measures, in some sense.

So while we have gone along this continuum of discussing cost of care measures that use actual prices paid, there has been a large discussion around potential unintended consequences of such a measure in the inability to -- or the lack of comparability potentially between -- We have had the discussion in the first meeting between Minneapolis and Memphis. We have moved to another example during this meeting.

So the question is that should the question of unintended consequence potentially be relegated to the user or is this something that should be discussed as part of the appropriateness of the measure as it is constructed?

We have discussed it in both ways. We have discussed it in that it -- as a measure of cost, actual prices paid, it has the ability for unintended consequences for

the user, which also interacts with the level of analysis. But we haven't clarified whether or not this would be a way that we are evaluating individual measures themselves.

CO-CHAIR STEINWALD: Who has a view? Jack does?

DR. NEEDLEMAN: First of all, even before we get to review, one of the things we need to do is get our language very cleaned up. Prices are ambiguous. Costs are ambiguous. So what we can talk about are charges. We can talk about payments, and we can talk about standardized prices which are something else.

When we start talking about prices, it is never clear whether we are talking about what is being charged or what is being paid, and each of those have problems right now in the current health care system.

We see pricing for folks at levels that are totally unrelated, both to the underlying cost of production, but also

totally unrelated to what they expect to get paid.

When you see hospitals whose charges are now three times what their costs are, and nobody pays that except the poor uninsured patient who wanders in, and even they negotiate it down if they know what they are doing, charges are not particularly useful, but payments also have a problem when you've got payers with very unequal payment levels. You know, what Medicaid pays for a given dentist is very different from what Delta Dental pays the same dentist for the same service.

So we've got problems with payment, and we've got problems with charges, and that is part of the reason for standardization to understand resource use. But we need to make sure the language is clear.

We need to know what is actually being counted when somebody is counting

resources. Is it the payments? Is it the charges or is it some standardized measure which is an attempt to get the underlying cost of production, and is that adjusted for -- see, that standardized cost of production -- across different areas with very different input costs?

All that language needs to be clarified, and what people are presenting and, therefore, what the measure tells us is going to be very different depending upon which of those things are being used as the basis for measuring resources.

So there is no right answer here, but we ought to at least be clear about what, in fact, we are measuring as opposed to the language we are using.

CO-CHAIR STEINWALD: Just a -- I agree with that, and maybe there should be a box in the report entitled "Watch your language," something like that. But on the issue as you raised it, just to throw it out

there for someone to disagree with, I don't think that the -- I think the measure developer needs to be clear about whether the resource measure is measured in terms of payment dollars or counts or whatever. That has to be clear.

As to whether the evaluation of the measure builds in the potential for unintended consequences, that, to me, sounds like it smacks of paternalism and ought to be an issue between the developer and the developer's users.

You would like to believe that, if the users are going to fork over $100,000 or more, that they understand what they are getting and they use it appropriately, acknowledging that there may be instances where that is not going to be the case, but that is a risk I would be willing to live with.

Tom, do you disagree?

CO-CHAIR ROSENTHAL: Well, no.

Well, mine is a little tangential, but I think it would come back, which is I don't have any problem with the articulation just the way you have said it, that at the end of the day some of this is between, quote, "the developer and their users."

I do begin to have a problem when one of these measurements might be developing as a national standard. I was actually going to pose the question to Helen or the staff, because I am not as familiar within the quality world, how some measures became national measures and others remain what got described variously as, well, there is a one-off registry and, you know, if you are a registry user, it is an NQF endorsed measure, but there are now hundreds of NQF endorsed measures or quite a number, which is all good and fine. But a few of them lurk up and become national.

I don't know. I have a problem within the articulation of, well, it is

between that developer and their little user community, and I actually -- Then the potential for misuse for some of these that are what I would call dollar denominated as opposed to standardized pricing -- the potential for misuse there seems to me profound, because we will have provider A being accused of being inefficient because the payments to them are substantially higher than the payment to some provider B through absolutely nothing that is in their control.

I don't have a problem with that, again if it is in this little micro climate, but I also don't have a sense of how certain of these measures -- so we actually have three layers of NQF issues, one of which is national standard NQF, you know, just the hoi polloi NQF, and then this idea of, well, how do we encourage the world NQF.

I hope somebody will answer my question.

CO-CHAIR STEINWALD: Dolores.

DR. YANAGIHARA: So, yes, there is no easy answers on that, because I think you are right. I mean, there is a potential for misuse. I don't think it is NQF's role to be monitoring that, but there could be some sort of -- in the endorsement, sort of these are the intended uses of this kind of measure. I don't know what those would be, but I mean to just sort of clarify this, it would be appropriate for certain situations or wouldn't be. It is hard to define that, because every situation is so different.

I think, coming back to -- Some of the issues are around reliability of the measurements, and that I think we can address and have standards for reliability of a measurement. Some of them more around just the uses, and I think that is harder to manage, but there may be, like I said, some things like these would be the intended uses of this kind of a measurement.

My other point was just around the

standardized pricing. I am wondering. It seems to me that that is something that could be truly standardized, like instead of asking each measure developer what is the standardized pricing methodology you used -- it is something that I don't think is really situation specific. I mean, it just is you choose some sort of a standard price for each thing, and you apply it.

So it seems like that is an opportunity to have a truly national standard for standardized pricing that doesn't have to be developer -- measure developer specific. I don't know if that is something that in the future NQF could work on. It seems like it would be a great role.

CO-CHAIR STEINWALD: Paul and then Barbara.

DR. BARNETT: So I want to make sure we have the conventional wisdom on health care cost determination, and so that the what is the cost depends on your perspective of

your analysis.

So if the perspective is that of the payer, then the payer amount, the amount that the payer pays, is the cost that is important, but if it is the societal perspective, then we want to know the opportunity cost of producing the service, something we almost never know in health care, although some of the costing systems at various hospitals are vested and may approximate that.

What we are usually stuck with in terms of that is, if we are looking for a standard cost and use that as a proxy for some societal costs, we are trying to get rid of payer discount. We are trying to get rid of the geographic variation of costs, and so we are using some other charge schedule other than the payer's charge schedule, one that we pull from the sky.

Actually, there is essentially a national standard from the RBRVS which is

probably not a very good measure of the opportunity cost, and there is an amazing amount of politics that went into determining that fee schedule based on the leverage that the various specialty organizations had at the time. So that is a concern. It is not very objective, actually, and rewards training and risk and all these, stress and all these other things it is intended to reflect. That was the theory.

So there is one practical thing, if we look toward a standard cost vector, as it were, or charge schedule, a practical matter of implementing it.

So we do this with a VA dataset. So we throw all 600,000 hospital stays and 80 million outpatient visits. We apply the RBRVS. So we know there is a lot of gaps, and so we buy a commercial charge schedule to fill the gaps for things that Medicare doesn't reimburse for, a lot of important stuff that have either HCPCS codes or CPT codes, but are

not reimbursable by Medicare.

Then the very important assumption has to be made about the facility component. So this is -- There are some services that can only be delivered in a facility, and the facility gets reimbursed. So the provider gets reimbursed. The facility gets reimbursed.

So ambulatory surgery, half the payment goes to the provider; half to the surgical center. There's a lot of services that don't need to be provided in a facility, but are. So if you use a standard cost and you say, okay, that specialty outpatient visit occurred in a hospital, so we give the facility, the hospital, this cost that is basically this payment that is equal to what we give to the provider.

Had that same specialty service been delivered off-site in a freestanding specialty clinic, the cost would have been -- oh, I don't know, 5/8th as much. So we you

apply that standard across schedule, you may be regarded as paying too much. It is an interesting question.

So that has been a practical issue for us. It is a non-trivial issue, especially hospital based services, whether you include that facility payment. So it is not so easy to build that standard cost schedule.

That is kind of one of the crucial -- You kind of have to accept RBRVS and the gap schedules that are out there, because there is not really any good substitute before them, but for the facility payment we struggle with that all the time.

So you could offer perverse incentives if you consider someone -- I don't know. I guess you would consider them inefficient, if it were facility based. I think there is an unintended consequence there in applying that schedule.

DR. RUDOLPH: I was going to suggest something, but now that I have

listened to Paul, I am not sure it is a good idea. There's just a lot of different issues in this area that can really change what gets put out and what doesn't.

I as thinking about perhaps some type of geographic indicator. Is this a measure that allows comparability across the nation, across a state, across the region, and not necessarily -- Again, it would be more of a guideline, I think, than an actual requirement, but that it would be useful for the end user to know what the issues would be if you were to try to do this nationally, if you were doing it statewide or in a region.

So it could be something that they should include sort of a statement of applicability across the country.

MS. PACE: I just was going to respond to your question about NQF endorsement. Basically, we say we endorse measures, and they become national voluntary consensus standards, and we endorse measures

that are intended to be used for accountability. Up to now, it has been primarily focused on public reporting, and performance improvement.

So, basically, when we say we endorse a measure, it is considered appropriate to use in a national accountability program.

CO-CHAIR ROSENTHAL: Yes, I am aware of that, and that is part of why I had trouble with some of the measures that were dollar denominated, because the imprimaturship is there right from the get-go.

I would say, though, I resonate a little bit with the suggestion Barbara makes about some notion of guidance about the thing. The one analogy that I am aware of is the various AHRQ measures that got developed all had sort of guidance about use at the bottom. Now they have violated their own guidance recently, but that is a different question. Nonetheless, they did say, you know, this one

is appropriate for, you know, cross-regional comparisons, this one is not ready for public reporting, this one is useful and appropriate for quality improvement. But again, I don't know whether this --

DR. BURSTIN: That is interesting, because AHRQ only submitted a subset of measures to NQF that had already gone through and were validated as part of the reliability and validity testing. So a good number of the measures never came to NQF that they didn't think met that threshold.

CO-CHAIR ROSENTHAL: No, I appreciate that, but the fact is, if you look at their entire set of things on their websites, they would have some guidance around what they felt was appropriateness for use or limitations around use of the various measures that they developed. But again, it may not be consistent with the NQF way.

CO-CHAIR STEINWALD: Yes. So just to follow up on Barbara's question or hope

that there would be some way to deal with the geographic variation in health care costs. So each September, CMS issues a regulation that includes a geographic index for hospital wages, which is a very powerful predictor of regional variation in health care costs.

It has some political issues of its own, how the districts get drawn and when certain hospitals get put into a higher cost area so their reimbursement will be greater. There is also a component for the RBRVS that is a geographic factor, and I think that is determined separately as a separate issue.

So that could be used to help people understand the geographic effects. You know, if someone is shown as a high cost outlier, but you could control for the geography using one of those sources, the effect of geographic wage pressures.

MS. PACE: I was just going to say, if a measure -- if the measure developer thinks that a measure is not valid at this

point in time for comparison, making valid comparisons, then it is probably not ready for NQF endorsement and, as Helen said that AHRQ only brought those measures that they thought would be suitable for those kinds of comparisons.

So it is an interesting -- I mean, we have never -- I don't think we have measures that we have endorsed that say, you know, they are limited to a particular geographic area or geographic comparisons.

CO-CHAIR STEINWALD: Dolores.

DR. BURSTIN: Nothing geographic, although interestingly, we have had this debate recently about whether we are going to start bringing in measures that, in fact, are only ready for EHRs. So I mean, there are -- If we are trying to satisfy the needs of the nation, there may be more advanced users, and I guess that is the question, is that over time you may have a capacity and others won't, but maybe we need to move toward where the pop

will be.

CO-CHAIR STEINWALD: Oh, Wayne Gretzky had a lot of -- Go ahead.

CO-CHAIR ROSENTHAL: And again, I think some of us have been cognizant of that we are operating in a little different space than we have been operating in all of the quality measures.

There are some, I think -- no pun intended -- quantitatively different aspects to the resource use issues than there ever have been in any of the -- I mean, a pressure officer is a pressure officer, and once you have adjusted for it, blah, blah, blah. But here we have this issue that is sort of -- There are, I think, clearly, two views of this which are both valid, i.e., the one that knowing the denominated or the dollar cost is of value, but where, if applied across geographies where we know there are wage and other differences even though there may be politics in the scales that got created to

try to account for it, at least some accounting was attempted.

Otherwise, you are likely -- and I know the differences in the wage price indexes are 20-30 percent apart. They can dwarf the utilization differences. But we have endorsed measures today, or yesterday, that don't take any of that into account, and we did it. So we have, in fact, said, those are ready for prime time. We were the judge and jury on the thing.

CO-CHAIR STEINWALD: You know, I wanted to respond a little bit to Dolores, talking about standards for standardization, I guess. And it is true that the national assistance that exists are largely Medicares, and Medicares are highly politicized. There is 441 areas for adjustments for hospital, but there are ceilings, there are floors, there are special payments for frontier states, whatever in hell they are.

CO-CHAIR ROSENTHAL: Nevada.

CO-CHAIR STEINWALD: Well -- And then on the physician side, there is only 79 areas, even though -- and some of them are statewide, even though there is huge variations in the cost of doing business within states.

So there is that level of problem. I call that political, but then there are some important technical problems, too. If you standardize, for example, by geography, how do you draw the geographic unit? You would like to -- If prevailing wages are higher in Boston than they are in Memphis, then a hospitalization that costs $5,000 in Boston and $4,000 in Memphis may be equivalent when you adjust for those wages. But you can't really adjust for the difference between North Boston and South Boston or, if you try to, then you run into all sorts of problems.

So the standardization -- As a concept, it makes perfect sense in many contexts. How to do it is subject to both

political and technical problems.

CO-CHAIR ROSENTHAL: Does that need to be commented on at the very least in the report, as a compromise, since we didn't exactly reach agreement here, but to elaborate on it in the way that you just did would seem to me sort of the bare minimum that we ought to be doing in relationship to this question.

CO-CHAIR STEINWALD: I think it makes sense to, but I am still -- You know, I think back to some of our discussion before where, let's say, you have a firm that has multi-site locations, and does it want standardization across those multiple sites?

Well, in some cases we are told, they don't. They want to know what the actual paid amount costs are in different locations, even understanding that there may be different costs of doing business. They still want to know those dollar denominator amounts, and I can't argue that they are wrong to want to know that.

CO-CHAIR ROSENTHAL: I am way off the insistence that that never be an approved measure of a dollar denominated one. It depends again upon what the use is, and I don't dispute n the slightest that using your metaphor of when an entity or when a health plan or when the Federal government or when somebody wants to know the dollar denominated, that there is value in that in some regard, even though they may not be able to control the wages in one of their sites versus another.

It is different than holding the provider end accountable for what we are calling efficiency. Again, as I recall, the charge to this group was around efficiency, and efficiency being a component of value, meaning cost and quality, and holding providers accountable.

I personally fail to see how you can hold providers accountable for the dollar component of this unless you have made some

attempt, however imperfect, to account for the differences in uncontrolled inputs, and at least to the extent that this is a problem that we didn't resolve in this go-round, that we articulate the challenge and the problem around this and the potential unintended consequences, if one of these dollar denominated measurements is attempted to be used to rank order providers around their efficiency. I think it is the least that we should be doing in relationship to this.

I am not suggesting to go back and undo what we did the last few days.

MR. AMIN: I have a clarifying question based on those remarks. So as we set up the discussion today, I wanted to focus a little bit on the interactive nature of some of the way the measures are specified in the different components.

So a question I would pose to the group is -- Let's make a little controversial. At an individual provider level, a level of

measurement, which is separate from the level -- or the attribution approach, the level of measurement -- If a measure is reliable and valid in producing reliable estimates at an individual provider level, do we still believe a measure that uses actual prices is still not appropriate?

So is this question of actual prices versus standardized prices changed when we are dealing with a level of measurement that is at the individual provider level?

DR. BARNETT: I was worried that he was going to change the subject and that what I was going to say was going to become irrelevant or past, but it was exactly on this point.

I think that each are actually -- I wouldn't use those words, prices, though, in the way you phrase the question, but is the cost from the payer perspective has an appropriate use and that the cost, the standardized cost has an appropriate use.

Really, one could go through and articulate what the appropriate uses of each are.

So if a plan negotiates a good discount and sends all its elective hip surgeries to a low cost hospital, and you are an employee wanting to evaluate that plan, you would want to include that efficiency that they have achieved by being a clever negotiator and finding the best hospital to send those surgeries to. You would want to include when you are evaluating their efficiency.

Now there would be other situations where you want to know about the mix of services that someone has ordered and the propensity to get patients in the hospital or keep them out of the hospital, and where the standard costing approach would be -- I almost said prices -- would be more appropriate.

So I think, you know, one could go through and sharpen your pencil and really

think of all the different situations where each is appropriate, and I think that would be a good thing to put in the report. I am not sure I could do it all here on the fly. Maybe the group could.

DR. ELWARD: Yes. I mean, it just raised a couple of questions. What if you have just as efficient provider use where the physicians are really doing well, but they are just horrible negotiators? You know, you have got an efficient process, but you don't have a great contracting team.

Either way, it is still -- On the one hand, yes, the people who are better negotiators can provide the less costly care, and then you could be given credit for that. I am not sure how you tease that out, but I think you could build in some -- looking at each process as you go through and build that in. I would make an argument against using overall prices as, I think, Jack said, even at the individual provider level, because I know

our fee schedules are really designed to get paid for the people who will, for some reason, pay us a lot for a given procedure, even if we know we are going to write off something. That is just the way you do it.

If somebody out there is paying --

DR. BARNETT: You mean charges, right?

DR. ELWARD: Charges, yes. Yes. But I don't know whether there is that much difference between -- Yes, I don't know how you factor in prices either. Prices, I think, are almost as much of a problem, because --

DR. BARNETT: If I might -- So, Jack -- or do you want to say it, Jack? So we don't see prices in health care, really, because there is not an open market. What we see is payments, and we see charge schedules.

So the standard price is -- What people use as the standard price is the Medicare reimbursement level. That somehow seems to make people feel comfortable with

that, because it is this national plan and covers most stuff.

DR. NEEDLEMAN: Yes. Just to reinforce this issue of charges and some of the irrationality -- By the way, it is not the matter of the uninsured patient. There are a lot of folks whose insurance companies have not contracted with individual providers, and when patients from those providers get called in by ambulance, for example, the issue is what level of payment should be there, and the providers frequently say these are our charges. We don't have a contract with you or your insurer to pay anything other than that; this is what you need to pay, and we see that a lot. It has been a major issue in California with some legislation. It is a major issue of litigation in some states, including Florida.

So it just reinforces that the standardization is not obvious. Well, what the level of standardization is, I think we

have tended to default to the Medicare fee schedules -- sorry, the Medicare payment schedules as the default for thinking about standardization.

Clearly, the geographic unit is relevant for thinking about whether we need standardization or not, but it also has to do with the decision making. Medical tourism creates an interest in actual payment levels as opposed to standardized efficiency measures.

So if we are comparing UCLA and the Mayo Clinic or UCLA and the Cleveland Clinic or someplace else, you know, the differences in costs in those places that influence what are reasonable payments may make sense, and standardizing for that in some way may make sense to understand resource use, but if you are a health plan that is thinking about do I want to send somebody to Cleveland or UCLA or Delhi, and the only differences there are about the wage levels, you want to

be looking at the actual amount that you are going to be paying, not simply the amount that is a standardized measure of control that adjusts away the differences in wages across those places.

So the geographic unit plays a critical role in thinking about whether we need to standardize or not, but only in some cases and not others. If we are trying to understand resources and efficiency, then standardization makes sense. If we are trying to understand decision making about where do I send my patients, then the actual payment levels are what are going to be relevant; and both of those are what we are trying to understand with these measures and, therefore, both of them, in some sense, should be incorporated into the analysis that we are getting out of the measure.

CO-CHAIR ROSENTHAL: I just want to add one thing. I was just realizing that I probably overstated the case in saying that

the Medicare reimbursement is widely accepted as the standardized payment amount, because I don't think that is true for inpatient services.

If you look at the DRG reimbursement, you are squeezing out a lot of the variants there, and we really want something else that reflects high cost outliers or length of stay, that sort of thing, because if you are just taking the average -- take the DRG payment, then you are ignoring a lot of the variants, and what drives a lot of health care costs.

So I am not sure that that is very well developed, what the standardized price should be or standardized cost -- excuse me -- should be for inpatient services, whether that is widely agreed exactly how that should be done. I can tell you how we do it, but it involves regressions with Medicare data, but it is not the DRG amount, because you are throwing out so much information.

DR. BARNETT: Jack, I would quibble only with one little aspect of your articulation of the thing, which is if Saudi Arabia or Blue Cross needs to decide whether or not to send a patient to Mayo Clinic or to, say, UCSF, they've got plenty of mechanisms to do that and to ascertain what that cost is going to be, without relying on some NQF publicly reported data element.

I think it is entirely unnecessary, and again I don't know why we would necessarily be doing that. I think there are plenty of mechanisms for doing that. But the non-quibble, or the monkey wrench, actually, though, Paul, from the idea of again Blue Cross making that decision and basing it on sort of, in effect, what are prices -- what are prices, because there is a notion of price; it may not be the charge, but it is what somebody is willing to do the thing for -- there is the whole factor of who can cost shift and in what settings.

So what, in fact, Blue Cross can suck out of a contracted relationship with provider X in community Y may be entirely dependent upon what percentage of that entity or that physician indigent care is and a whole variety of other factors, have nothing whatsoever to do with how efficient they are, and efficient meaning, again, their resource utilization for either provision of care or the avoidance of doing the unnecessary things that you have described are driving the costs.

CO-CHAIR ROSENTHAL: But if I am an employer, and I am choosing between Plan A and Plan B, the fact that Plan B has figured something out about negotiating low payments is important to me, and I regard Plan B as more efficient.

If I am trying to gauge whether a Plan B provider practices in a style that makes the most efficient use of health care resources by minimizing hospitalization, minimizing laboratory tests, that sort of

thing, I don't really care about all those negotiated discounts. I want to use a standardized price.

DR. BARNETT: Well, exactly, and my argument is that, in point of fact, there are 50 ways for the health plan to understand what price they are being asked to do, and I think we are confusing the public realm now with the private realm.

The private realm has every ability in the world to understand and to know which provider -- I am absolutely certain that the payers in our community know what prices, know what costs there are being extracted from which hospitals and which systems are viewed as more expensive and which ones they can get better negotiated deals with, and all of that stuff. Again, they don't need an NQF validated price listing to make those judgments. Maybe I am wrong about that.

CO-CHAIR ROSENTHAL: I would think

that would be true if the products were homogeneous, and they knew that they were actually just buying this stay or that procedure, but the problem is it is a bundle of stuff, and it has got to be case mix adjusted, and I think it is everything about what we are talking about.

I don't know if there is somebody that could represent that perspective that is in the room right now, but that employer perspective -- they really do want to know case mix adjusted for the population, which is the low cost, and that it is not just simply a matter of knowing what the negotiated rates are, but it is more complicated, has to do with how much services are being used, given the patient characteristics.

DR. BARNETT: And if that were true, you would want standardized prices.

CO-CHAIR ROSENTHAL: I think there is a place for the actual payments. I think the actual payments have a ton of other

factors associated with them that are not going to be teased out by the measures that we are in the process of approving.

DR. NEEDLEMAN: I think the conversation we are having -- First of al, I don't think we are going to resolve the debate, but I think it is important to get it up there. But it also underscores the issue of how do we expect these measures to be used, who is going to be using them, for what purpose.

I wanted to run an analogy to the quality measures and then come back to the resource use measures. In the quality measures, we've got two models of how quality is going to get improved.

We have got the J.D. Powers, Consumers Report -- I assume Doris is still on the line -- model of we report the differences in quality and then consumers choose, and that works for some kinds of things, and it doesn't work for others. I don't want to be, frankly,

choosing which hospital I am going to based upon the perception of the relative quality of the nursing.

So we have got a second model, which is what I would characterize as the Underwriters Lab model, which says you buy this toaster oven. If it has got the Underwriters Lab certification, we are pretty sure it is not going to electrocute you. Right? You buy the toaster. It is not going to electrocute you.

To some extent, what we are doing with the quality measurement is saying that ought to be the way the health care system functions. Shouldn't matter which hospital we go to. The care you get should be safe and reliable, and we expect the quality measurement and the differences in the quality measurement not to drive consumer behavior per se, but to drive a professional commitment to improving quality where it is shown to be not as good as other places. That is the

Underwriters Lab model.

I think we have got the same issue as we think about these resource use measures. We have been talking about it right now in the context of consumer purchasing, but an awful lot of the places where we are going to see these resource use measures used are the internal efforts to improve efficiency, improve resource use, while maintaining or improving quality.

So do we have some providers within our community of providers at the Mayo Clinic, at the Cleveland Clinic, at Banner Health, that seem to do a better job of effectively using resources while producing high quality outcomes than others? Can we learn from one another?

Can I learn from looking at the experience of the other providers of the other health plans or the other physician groups in my community to create a standard which says I can do better, because I am seeing others can

do better? But where the fundamental work in improving the efficiency is going to be internal to the group driven by internal commitments, not by consumers deciding where to go buy.

Part of the argument about how important it is to do standardized pricing versus the raw how much does it cost consumers is about whether we expect these decisions to be driven by consumer behavior, consumers choosing where they can get the cheapest care that is of high quality, or whether we think it is going to be driven by the internal decisions of can I look at my experience compared to others in terms of how much resources we use and see opportunities for improvement, even as we try to improve the quality of care.

That, I would argue, argues for more use of standardized pricing to understand what the actual resource use differences are, rather than how much I am paying or how much I

am getting paid for the care I am providing.

CO-CHAIR ROSENTHAL: And If that was a good summary, then it must mean it is lunchtime, because we are at about that time. Is that enough?

MR. AMIN: Yes. It definitely is. There is a lot of complexity here, and we will try to boil it down and get it in the report in the way we discussed it.

CO-CHAIR STEINWALD: The agenda says working lunch. How should we interpret that? How hard working should we be?

MS. WILBON: What we were thinking was that we would take like a 15 or 20 minute break and time for people to get food and then just come back and bring your food back to the table, and then kind of talk and eat for the rest of the afternoon.

I know some people have to leave early. So it would be nice to kind of have as much discussion while we have the majority of the people here for the afternoon.

CO-CHAIR ROSENTHAL: Okay. We will do that.

MS. WILBON: So we will reconvene at about 12:20, 12:25. Thank you.

(Whereupon, the foregoing matter went off the record at 12:05 p.m. and resumed at 12:32 p.m.)

1

2

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

12:32 p.m.

MS. WILBON: We are going to go ahead and get started again. I think we have a plan for moving forward, and I think what we are going to do, assuming all minds are settled about the previous modules that we have discussed up to this point -- we are going to move on to the reporting module, which encompasses the attribution approach, peer grouping, benchmarking, sample size, and defining outliers and thresholds.

So with that, I will turn it over to Taroon to kind of talk through some of these issues.

MR. AMIN: We are going to take reporting and reliability and validity testing together, and we will try to aim for a 45-minute session here.

Some of the overarching themes and considerations from the Steering Committee and TAP is to keep in mind that this section could

be submitted as specifications or guidelines. Again, that is a question that we are posing on whether or not that continues to be appropriate.

Additionally, a question that arose a number of times was around the sample size and whether or not there was an appropriate sample size for a reliable and valid measure, and on the attribution approach.

One of the examples that I will use for the attribution approach was along the lines of the level of measurement, which I am sure we will go into in much more detail, whether it was attributed to an individual provider, was guidance on some temporal logic potentially with the attribution approach.

For example, when we were reviewing the HealthPartners measure, there was an attribution approach that allowed the resource use to be attributed to a primary care provider.

It allowed the potential for a primary care provider to be attributed a patient post-hospitalization, so even before they have actually had their first visit to a PCP, to that PCP, they could have been attributed the actual cost of that resource use, the cost of the provider.

Okay, so reliability testing: So the question there is really around the appropriateness of specification and guidelines, and then a general issue around sample size.

The reliability: I just want to go over a little bit the definition that is used here. It demonstrates that the measure results are repeatable, producing the same results in a high proportion of the time, in the same population, in the same time period, and that the measure score is precise.

So there is really a broad question here of whether or not this construction of reliability is appropriate for

resource use measures, and how well we felt that this was evaluated through the evaluation process.

As part of Carlos' evaluation, he suggested potentially other additional reliability approaches, reliability testing approaches, that might be considered, including the stability of the O to E ratio and the accountable entity over time, and potentially other approaches, including signal-to-noise ratios using ANOVA or intra-class correlation coefficient. But many of the developers used -- They used a parallel development of the episode software and SAS software as their measure of reliability. So it is a question of whether that is sufficiently adequate to our definition of reliability.

Finally, the validity testing: The question of validity looks at the NQF definition as demonstrates the measure data elements are correct, and the measure score

correctly reflects the cost of care and resources provided, adequately distinguishing between high and low resource use, with face validity being the minimum threshold.

So one of the questions here is: Is this adequate for resource use measures, and what considerations should be made by developers when selecting a testing database?

A lot of times -- specifically, I will use the example of Ingenix -- some of the TAPs had difficulty, because Ingenix was testing on a very large dataset that could represent more than one health plan.

So the question of whether or not you really needed to have multiple datasets to adequately assess validity; and also how to assess the data element validity in the context of resource use measures, defining the data element and also the issue that Tom brought up of whether or not there is enough literature to actually demonstrate the appropriateness of each of the data elements.

So I will sort of leave it there. Again, there is a lot that I asked, so we will sort of leave the discussion open to areas that you all felt were the most important and resonate most with your thinking as we are reviewing these measures.

CO-CHAIR ROSENTHAL: Well, I will start. I think the reliability questions are interesting. If I reflect back, though, on our decision making, I think we largely accepted the reliability, but when you posed the question the way you did, I don't think we applied a very high standard either, because most of them had not been tested, really, in real life and multiple settings.

We just kind of accepted that the computer cranked the thing the same way, it was all computer based and, therefore -- and I don't think, actually, most of our decisions around yes or no on scientific acceptability were driven by the validity side of the equation, but I suspect -- and Carlos might be

our guide on this -- we could be more rigorous -- could have been more rigorous or should be more rigorous in the future on reliability.

I would say, with regard to validity, my own -- The one observation I had from this of why I think I had some struggle with the thing is I am used to looking at data less in theory than in practice. What I found problematic was -- I don't remember, except with one exception and it was, I think, the HealthPartners where they actually showed a chart of how they actually arrayed the data. I don't recall seeing a data element arrayed for any of them.

At least what I would normally do is, if somebody says, well, here is measurement X and it is purporting to measure something or other, and here is the condition, and here is the -- I would expect to see -- I would want to see the data arrayed. I would want to see the confidence intervals. I would want to see how many outliers there were on

the upside and on the downside.

Then in my own mind I would array that against what I knew about that disease state, about how much variation I at least intuitively thought existed in the world, and that would be the basis of my ability to even discern face validity, and I don't think we had that on any of these or with the one exception, and the one exception had, I think, three health medical groups, primary care groups from the HealthPartners. I don't think we ever saw it on Ingenix, etcetera.

So I think that asking to see a sample array of the data in actuality as it was applied would have been extremely helpful for me, and I don't think that is -- again, back to the question of, well, are we just piling on the onerousness of this thing with these developers. I don't think that is a ridiculous kind of request. They surely must have that.

Maybe I was the only one who

suffered from that, but I have a feeling that others were similarly impacted.

DR. STEPHANSKY: Actually, that was some of my concerns about what was in the paper, though. It was describing how we were dealing with the reliability and validity issues. We were kind of missing this piece.

MR. BOWHAN: I guess I will make a comment and then ask Carlos to respond to that.

What struck me particularly in the Ingenix descriptions -- and I think they came mostly the same from what Carlos thought about the validity -- is that, basically, it seemed like the -- what I was getting out of it, that there really wasn't enough information to make an evaluation or a judgment of the validity, but that didn't stop us from going ahead and voting on it anyway, even though we didn't seem to have enough information.

So either we shouldn't ask for it or expect it, or we just go back, the way Dave

suggested on the previous thing. If they didn't have what we need to make the evaluation, we don't move forward until we get it.

CO-CHAIR STEINWALD: Now isn't much of the discussion we had this morning about risk adjustment and some of the related -- aren't those validity issues? Right. And some of that, you have got notes on and are prepared.

I think the way of dealing with the now versus the future, though, is acknowledging that we were the first out of the chute, at least on measure, and we did what we could with what we were submitted. The developers did what they could. Maybe they could have done better, but looking forward, I think we should make it clear that the expectation -- the bar would be raised, and try to be somewhat specific about in which ways we expected it to be raised.

DR. REDFEARN: In terms of

reliability, to go back to that, I am a little concerned about the issue of trying to do sort of test/retest or repeating across time. That is a problem we struggled with when we did provider profiling, because the question we always get is, well, you did this analysis last year; is the ranking of the providers or your evaluation of providers the same when you run it again this year.

The reason that is a little problematic is that the mix of patients that the doctors see changes across time, and the practice of medicine changes across time. So you have built into it some variability that makes it harder to do.

What I would suggest and what I think developers could do is they could do sort of split half-tests, which divide their sample in half and see whether the efficiency scores come up the same when they cut their data, or bootstrap it or something. There's a whole bunch of things they can do with a

single sample.

Then at least you know that these other things are not changing underneath you. It is the same time. It is the same group of providers, but you could certainly split the sample and do that kind of stuff. I think that would be useful. That would tell you something about how stable the scores are.

MR. ALZOLA: Okay. Let's stay with the reliability thing. All we really asked for was -- in terms of reliability was to show that the measure is repeatable. It was the minimum standard that we could ask for.

It kind of -- There are two ways you can look at this. You can look at the repeatability of the algorithms. Is anybody doing -- that has the same information, a different program, going to be able to replicate these datasets in these measures, and the answer is, in principle, yes, because the real variability there is how the data are

input at the source.

If two coders look at the same patient and they assign a different diagnosis, okay, that is where the variability can be. Once you have an algorithm that is going to put them in a database and to use this code, everything will be the same.

Having said that, that is very true for Ingenix, but in measures such as the AVMS, they only gave instructions on how to do it. They didn't give a real algorithm and say this is how you have to do it.

So, yes, it is repeatable, but the devil is in the details. So it may not -- I think we probably have to be very more strict in saying you have to provide enough detail in the algorithm that any program can reproduce -- will obtain the same results, sort of like Joint Commission does.

DR. BARNETT: I think all of these comments are important. I was struck -- Two things: In terms of the message for assessing

reliability and validity, I was thinking in reviewing them that they wouldn't be sufficient to get past muster in a peer review journal, that it just wasn't that high a quality, and that is a bit disturbing, really, when you think about what importance could be attached to the endorsement.

The other thing to reflect upon is the evaluation is being done by oftentimes the same commercial interest that is proposing the measure, and so they have an inherent conflict of interest, which is always a problem in evaluating something, and it is too bad we don't have unlimited resources where we could commission independent assessments. But so given that we don't, then it seems like we have to have a pretty high bar regarding things, meeting these tests of reliability and validity.

I know that, during the meeting, there was some tension about just how high that bar should be. There was some sentiment

that we didn't want to make the perfect the enemy of the good, if I said that right, but I do think that it is important that we get it right, because the concerns about efficiency and health care costs are such a political lightning rod that we don't want to endorse a measure that is not well thought out and then be accused of rationing or convening a death panel or something like that.

CO-CHAIR STEINWALD: Karen, you would like to say? I'm sorry, rationing or?

DR. BARNETT: OR convening a death panel.

CO-CHAIR STEINWALD: Oh, the death panel, of course.

MS. PACE: Yes. I just wanted to provide some clarification about the NQF criteria on reliability and validity, and we had two task forces, a task force last year that spent quite a bit of time looking at reliability and validity and developing some recommendations.

Currently, our criteria -- and that task force was very clear that they want to see empirical data on reliability and validity. Our criteria allow for analysis at either the data element level, which is the repeatability or reproducibility, or at the measure score level, which is more about precision and how much error there is in that computed score.

Regarding electronic sources, they specifically talked about electronic health records, but I think the issue of reliability or repeatability when you are doing a computer programming with claims is applicable. You are going to get the same result. It is going to be repeatable, and we are not really so interested in reliability of a computer program.

What the task force recommended is that -- They acknowledge that that is going to be repeatable and that you wouldn't have to do reliability testing at the data element level,

but you would need to do validity testing at the data element level, which is about the accuracy.

So if you are relying on claims for a particular diagnosis, is that something that claims data is really a valid source of data for that? I am sure you all are more aware of it than me, but the advice we have gotten from the task force and other committees is that it depends on the particular diagnosis, how valid claims are for identifying patients with particular diagnoses.

We have provided a lot of flexibility, so that, if there have been, for example published studies about the validity of claims data for particular diagnoses and that is what is being used in a measure, that can be cited.

So at the data element level, you know, reliability is about repeatability; validity is about the accuracy. Is it right

data? At the measure score level, which is what we are ultimately interested in -- As you know, we are endorsing these measures for accountability purposes and being able to make valid conclusions about differences among providers.

At the measure score level for reliability, we are really interested -- or one type of analysis is signal-to-noise. How much of the difference among those measured entities is actually true difference versus error and noise in the measurement?

Validity at the measure score level -- We do at this point in time allow face validity, which we ask that that be systematically assessed, but if there are other kinds of conceptual relationships that can be tested, if there are other measures of cost or resource use that it can be correlated with, those are certainly things that the measure developer should consider in terms of submitting evidence of reliability and

validity.

CO-CHAIR STEINWALD: You know, on three different occasions that I can think of, I asked a question of the developer: Is this measure in widespread use? And the answer was always, oh, yes, and for years, numbers of clients and so forth.

I would ask, has the measure ever been used in a study that has been published in a peer review journal? The answer was either no or not that we know of. In a way, that is kind of surprising, given the fairly substantial organizations.

I know the argument is, well, who has the time or resources to do that. On the other hand, I wonder if there is a way that we could encourage the developers to get -- I think Dolores said, these are systems; they are not just individual measures. They are systems -- get some public exposure to their systems, either through generating publications or some other posters at meetings

or something like that, more than apparently what they have been able to do so far.

MS. PACE: I think that is not too different from the quality measures either. The developers are not -- unless they have come initially from some academic background or academic setting. We find the same thing on the quality measures.

CO-CHAIR STEINWALD: I would think that -- You know, there is a local university that has graduate students. They are looking for data and topics to publish on. It seems like it would be a natural.

CO-CHAIR ROSENTHAL: I guess it doesn't fit their business model, sadly. Sadly, because I made the comment earlier about the lack of peer review stuff, and that is typically where many of us would go to kind of get an imprimaturship.

One of the observations that I had in my head about the validity question is -- and a little of this was partly the way we set

up the topics, but we had attribution over here, and we talked about attribution in one place. We had sort of what is the disease state and the logic around it kind of over here, and we had the risk adjusting thing all in kind of a separate conversation. But it really is the intersection of those three elements that enable you to draw a picture, again if you are a visual learner, and decide whether that combo of features together passes the validity hurdle.

The way we set up our own review logic didn't tee that up quite so well, and there were times -- and the pieces didn't feel like they fit together. So I think that is something we could do better the next go-round.

MS. PACE: And the current -- The rating scale that the Measure Testing Task Force set up, we really have grouped together validity -- not just the overall validity testing, but the issues that are threats to

validity. So risk adjustment for outcome or resource use measures factors into their validity rating.

Exclusions: Because how you deal with exclusions can affect validity. Missing data, those kinds of things affect ultimately the validity of the conclusion you can make based on that score. So we have moved to grouping those things under the validity category.

I would be interested if you all have seen -- I know you mentioned that these particular measures that you reviewed weren't necessarily published in the literature, but have you seen things in the literature that we could provide as examples of what would be reliability -- a good reliability and validity testing or point us to some publication that we may take a look at? If you don't know today, certainly, send it on to us.

DR. BARNETT: I was just doing some PubMed work here, but there are three

papers by J.L. Adams and McGlynn. One of them was provided to us at the outset, and they have something new in one of the electronic journals about the statistical methods for assessing the groupers.

MS. PACE: Right, and actually, I don't know if it is the same paper or a different paper, but they also did something that we have been looking at for just quality performance measures. They have done some work on the signal-to-noise analysis for precision of measurement, which applies to quality measures as well as resource use measures.

DR. BARNETT: And then I know, and I wish I had read, but I haven't yet -- I have it in the stack -- of the stuff by McCurdy and Thomas, et al. Some of that you gave us, and some of that I got independently. Yes, but they have a contract with CMMS to do evaluations. I know that I have read the abstract and thought this is important stuff,

I need to read this; but I don't know anything more about it than that.

CO-CHAIR ROSENTHAL: But the irony on one of those McGlynn articles was that, in fact, I think their critique of the grouper was that, in fact, it re-rank ordered the physicians from one period to the next, just on my recollection of that.

DR. BARNETT: But I think that not only to read that, but also there are five or six letters that were in the New England Journal in response to that particular one that are also very insightful, I think, about the issues.

MS. PACE: One other thing about the comment about looking at scores over time. Again, we have seen that submitted for quality measures, and it is really not what we would expect for reliability testing, especially in the context of performance improvement. Someone already mentioned it is a different time, that it could be different

patients in the calculation of that score, but also if you are thinking of performance improvement, you know, what is the basis for assuming it should be the same?

So anyway, it is something that we are working through gradually with our committees and measure developers, of really looking at what are some of the real things we are interested in.

DR. BARNETT: If you apply that same method to the split sample, like David suggested, then you don't have that problem. Right? And you should get somewhat similar classification. If your dataset is large enough, then you really could get at the issues. Right?

CO-CHAIR ROSENTHAL: Well, and just not to quibble on the thing, but there is some element of which it is a reflection of how good the risk adjusting methodology is, and the signal-to-noise ratio, because I think, if you looked at this the same way we

would look at control charts within an organization, there is a certain amount of this that is just within two standard deviations, and what you are seeing year over year is, in effect, the normal variation that you would expect to see, and yet we are attributing in one snapshot in time as a difference A and B as being some profound statistically significant difference.

So it goes back to the notion of even something that is less than p less than .05 has a one in 20 chance of being no different, and that is what we may be seeing.

MR. ALZOLA: I think the developers were rather confused about what we were expecting of them. As I look at the submissions and the answers they gave to the reliability questions, in most cases they didn't answer the questions. I had to look hard for evidences of reproducibility.

So I think it would help if we could provide the concrete examples. In that

respect, what HealthPartners did was the best example I saw for signal-to-noise ratio.

CO-CHAIR ROSENTHAL: Carlos, were we unclear in the way we asked or were they unclear in interpreting it, or both?

MR. ALZOLA: It is probably an issue of both. To me, the question was clear enough, but they didn't get it.

DR. PETER: Do they get -- Do measure developers get a sample measure that has like a model for them to look at?

MS. TURBYVILLE: My recollection was that we did point them, since we knew the developers so well, to the Testing Task Force report.

MS. WILBON: But we didn't have any sample measures, because we just didn't have measures before us. So we did our best to try to give them examples from the -- pulling from what we knew from quality already to draw attention.

DR. PETER: Right. Maybe now that

we have -- I don't know if you can de-identify them. I guess they are in the public, but maybe you can provide them with a sample. It is appropriate to point them to one of the best ones that you have received.

MS. WILBON: Yes, absolutely.

MR. AMIN: One other quick observation from what I am hearing from this conversation is: Some of the elements, specifically, that were in what is termed right now Module 5 in the reporting is clearly part of the reliability and validity of the measure. So this "adequately demonstrating with certain sample size and the level of analysis that is articulated in the measure" needs to be more precisely defined.

So allowing this level of flexibility through submitting guidelines may not be appropriate in the next phase of work. So we will take that. I just want to make sure I have heard that right. Okay. So we will take that, Dolores.

MS. YANAGIHARA: I was just going to comment on that. I think -- I wonder if there is any thought in terms of reporting of having -- I mean, there are industry standards for reliability of measurement. So I am just wondering if it would make sense to have demonstration of that kind of reliability.

That sort of gets to the sample size issue and getting down to individual provider level and things like that. As long as the measurement is reliable, I mean there still may be reasons why you would or wouldn't use it, but at least technically it is ready to be -- or meets the standard, so to speak.

So I don't know if there is some way to work that kind of reliability requirement or industry standard. A .7 reliability, I think, is often used. There are different ways to measure reliability, but it might be something to think about.

Sample size is part of it, but it is only part of the equation, I think.

DR. BURSTIN: Just one thought, and it is spawned by several of these comments. It also seems like it would be useful, especially as Barb has reminded us repeatedly of measure developer burden on this as well, is as we take maybe a retrospective look at what we ask for on the form, what we ask for that we actually didn't use perhaps, that maybe they jumped through a lot of hoops to provide, and maybe what we asked for that they didn't answer, maybe we could be a bit more focused, because we asked for a whole lot of stuff.

It would be really helpful to go back and look to see what really added value into our decision making and what didn't. So your input and thoughts, particularly you, Carlos, as we look at that section, could be very useful.

CO-CHAIR STEINWALD: I think there will be bad news and good news for developers. The bad news is that some of the bars are

going to be set higher. The good news should be of two kinds. One is there is certain information that we decide we don't need and we are not going to ask for, and then the second is along the lines of being specific about what we want and, if you give us what we want, then the chances of you getting what you want are improved compared to the first time around.

MS. YANAGIHARA: I actually also think that it would be good -- and maybe it was in the instructions, but brevity, I think, is really good, and not putting in the whole PR spiel. I mean, like some of this stuff was just so -- I was just like cut pages out of this stuff. There is just like the same stuff over and over that wasn't even relevant. I was like, oh -- So just focusing on just provide the information asked for. It can be brief. It doesn't have to be volumes and volumes, and just really emphasizing that, because some of the stuff was just --

MS. WILBON: I think some of that was a product of different developers and being new to the process, and there was an effort of the staff to try to -- What you guys got, believe it or not, staff actually had done some back and forth with the developers before you got it.

At some point, we just have to be like, okay, we have to move on with the process, and we just have to move the measures forward, but I think you could probably see a different level of how the questions were approached by the different -- and I think that is just a matter of experience. Anyway, there is an effort to help. No, it's a great point, and we completely agree.

DR. REDFEARN: Perhaps when you have a measure developer submitting a number of different measures, all of which rest on the same foundational logic, they could present it with one description of the underlying logic and then the specifics of

each measure separately. That would bring the volume down, and we only have to read the garbage once.

That actually -- I was thinking in terms of Ingenix, because this is all fundamental, but it applies to NCQA, too. It is like when we went to COPD, it's like, oh, same as asthma except different definition. So all of the fundamentals of the data cleaning and stuff like that can be presented one time, and then the specifics of each measure, and that would be a lot easier for us, and probably easier for the developers, too, because then they could develop it a little bit better.

MS. WILBON: Right. So we try to do that. I am not sure it came across that way. So that general methods document or that item in the submission that we allowed them to submit -- we asked them two or three pages; for some of them, it was actually much longer than that -- a just broad overview of how they

-- what their approach is, and then for the measure submission form to be specific to that measure.

There was an attempt to do that. If you have suggestions on ways that we could maybe communicate that better. IF you remember, in the beginning we did that one webinar in the beginning that was supposed to be like a general methods webinar where each developer kind of presented their general -- So there was an attempt to try to go with that approach, but obviously, it being our first time, it wasn't completely successful, obviously.

So any suggestions you have on how we might be able to narrow that or make it --

DR. REDFEARN: Make the methodology -- Make a methodology section that forces them to do that and then the specifics of the measures, and just label it separately. That might help.

CO-CHAIR STEINWALD: Jack.

DR. NEEDLEMAN: Along that line, if they are told to create an appendix with the details of their general methodology and it is attached to the specifics, but we know it is general, then we know we have seen it before. The first time you are reading it, you know you have to read the appendix. After that, you are looking for the exceptions, the tailoring of the method, and that may work as a vehicle for both allowing them to present the stuff, but also allowing them to quickly attach it to a whole bunch of measures simultaneously without saying there is a whole separate thing we have given you which has this. You need it there, if you need to refer to it, but it doesn't have to be 10 pages of text in the middle of every application.

So that is a possible strategy for just helping them organize what is generic and what is unique.

DR. BARNETT: Just to observe that we -- So the areas that the measures were

chosen were pre-defined -- right? -- as diabetes and COPD, etcetera, because it was thought they were high value. So we almost never use that value section.

Sometimes interesting to read their take on it, but it almost never had anything to do with the measure that we were reviewing. It was something about the larger literature.

So I think that whole section, actually, could be dropped, since we already asked them to submit something in an area that we knew was going to have high value.

MR. AMIN: Okay. Just in the interest of time, because I know a lot of people have to leave, I am just going to go over usability quickly, and then we can move into some additional conversations around efficiency measures.

So the current usability criteria are whether the measure performance results are reported to the public at large in

national or community reporting programs, and has the measure demonstrated results that are meaningful, understandable, and useful for information for public accountability and process of performance improvement, and can the measure be deconstructed to facilitate transparency in understanding.

So one of the questions here is that how do we assess the usefulness versus whether the measure is in use, as we have discussed this many times where we fell back on the principle of, well, it is in use, maybe not necessarily addressing whether it is useful in the way it is currently expressed.

Also, this larger conceptual issue of balancing the need for transparency with the potentially inherent complexity of the measure. So can there really be transparency when the measures are this complex for the intended use, and who is the intended audience, which we started to address?

MS. WILBON: So, Karen, I don't

want to put you on the spot. We started to try to bring in some of the ideas from the NQF Task Force. So NQF has a task force that is in place right now that is looking at this particular criteria and trying to see how it can be reframed and kind of revamped to address exactly this issue.

So would you mind kind of summarizing a little bit? Sorry to put you on the spot.

MS. PACE: No, that is okay. We actually have a conference call next week, but I will just say that the -- So we have trouble really evaluating usability, not just for resource use, and really wanted to have a group take a look at it, also because NQF to date has focused a lot on public reporting, but given the current environment and other accountability applications, we wanted to make sure that that was also encompassed, though the ultimate transparency is for public reporting, and we want to still encourage

measures to be publicly reported.

Basically, I can just tell you this is in process, but the way the task force is going is that for measures on initial endorsement to really look at usability as kind of a hypothetical construct, and that asking for a rationale of how it could be used for both accountability and performance improvement, to give a rational for that, and also because ultimately these measures won't have any impact on performance unless they are actually used, to again in a hypothetical way talk about -- or probably even more concrete, what is the plan for getting these measures to be used in an accountability application.

You know, what is the plan? do they have any commitments, and what is the timeline, because ultimately then, by the time of endorsement maintenance, what we are going to be -- What the task force is talking about is then actually asking about, is the measure in use? Is the measure in use for what

accountability application? Is that national, regional, less so? Is it public reporting or is it some other less transparent use in an accountability function? Then also what impact is it having on quality or, in this case, efficiency?

So it is really kind of two-tiered: At the initial endorsement, to look at the potential use for performance, impact on performance improvement, and a plan to actually get this measure into use; and then on endorsement maintenance, is it in use, and what impact is it having?

So that is kind of the basics of what that task force is recommending, but it is still under discussion.

CO-CHAIR ROSENTHAL: I am going to re-raise this concept of the cul de sac that I mentioned this morning. So we saw a number of measures, and they are all in use, as I understand it. They are in use for private purposes, but the users are willing to fork

over substantial amounts of money, and presumably they are using the information, but they are using it in a different way than these criteria suggest, and I see you are well aware of that.

Switch over to the Medicare program for a second. It is our nation's largest insurer, in the world, I guess. The U.S. Congress requires the Medicare program to have what in essence is a physician profiling purpose prohibiting any public reporting.

So we have our largest national program using resource measures for the purpose of providing private feedback to physicians -- a lot of states, no public. So I am thinking we are a little bit at odds here with how these measures are actually being used.

MS. PACE: So just a couple of things. In terms of use, we are defining that as use in an accountability application, not jus use in a private application, because

again, if it is just going to be used in a private application, there is really no need for NQF endorsement. I mean in terms of what NQF's mission is.

So -- but in terms of the public reporting, you are right. Not all accountability applications result in public reporting, but one principle of NQF is as much transparency and openness as possible. So we would see -- you know, again if it is being used in an accountability application such as required reporting to CMS or perhaps for a payment incentive program as moving along that line of transparency and, hopefully, eventually measures will be publicly reported or at least publicly available, or perhaps in some cases there will be evidence or data suggesting that it is not useful for public reporting applications.

So we are still working on that, but I think the idea is that NQF's mission is to improve health care for the American

public, and quality also includes efficiency. If they are not going to be used in that context, then it is going to be a specific question of why do we need to continue endorsement of a measure that is not being used out to actually facilitate that goal.

CO-CHAIR ROSENTHAL: I think, had we had to apply that standard, we might not have approved any of the measures, because --

MS. PACE: They are not required to be in use at the time of initial endorsement.

CO-CHAIR ROSENTHAL: No, I understand, but it is also not necessarily clear how any of them really will be available for national public reporting either. But that is okay. I mean, again I am not second guessing our decisions, because I think we did some things to move the ball down the field.

The observation I was going to make, though, is that we did, I think, assess the usability or how they had been used, and

from my own reflection on that, the NCQA one -- you know, 300 members have used this. They go out of their way to provide the information. They pay for the privilege of doing so. They accept the fact that these reports are there. There was quite a lot of detail around that.

Again, I would contrast that, that I felt like there was somewhat less robustness of what we heard back from, say, Ingenix, and of course, from the American Board of Internal Medicine there was none, which I think was one of the things that queered that up. But my observation would be perhaps there is an efficient way to ask that, because I don't think we asked it in an efficient way, and that perhaps is something we could contemplate on how to do. I think we grappled with it, but we didn't ask it in an efficient way.

DR. BARNETT: Yes, just to pile on that comment, I think the things that Karen said were very clear in terms of what the

criteria are, and I think the instructions -- In retrospect, we were not specific enough in what we were seeking, and that would be helpful.

So the first section says current use -- semicolon -- or colon, excuse me. So that gives people a lot of wiggle room, and we didn't really say exactly what we wanted. I think, if we were more specific about what we exactly wanted, we would get more useful information out of the process.

DR. REDFEARN: I would like to resurrect the issue of inviting commercial grouper vendors, expensive software vendors into this, which is a decision that I think NQF made on sort of an experimental basis.

This is directly relevant to usability, I think, because my personal feeling is, when you get into products that are going to cost hundreds of thousands of dollars to implement, it reduces usability dramatically. Frankly, I don't think there is

going to be any commercial carrier that would want to use the Ingenix methods unless they already licensed the ETG software.

All the commercial carriers are doing something. They might be using another tool. If they are paying Thompson a million dollars a year for the Medstat system, why would they want to go out and pay two or three hundred thousand dollars for ETG's?

So I guess this is a question for NQF. Was this experiment successful or is this something you might want to rethink?

DR. BURSTIN: It is a great question, David. This has been an ongoing issue for us for several years now: Should we bring in proprietary systems?

In fact, we got criticism initially saying, but you are leaving out a lot of the innovation and where a lot of this work is happening, and particularly in this area, there were so few developers who didn't live in that private space that we didn't

think we had a choice. However, there are ongoing discussions.

The NQF Board of Directors is going to be dealing with this for the next couple of months as we revise our measure steward agreement, and one of the issues we are going to have to decide is does NQF bring in measures that are associated with charges? To date, we have allowed that corridor. It has been almost never used except for this project and one other project we did on readmissions.

So I would be curious from your perspective. It would be very useful input. Do you think it is something we should continue with or was really kind of the juice not worth the squeeze at the end of the day, given the efforts involved for them in terms of getting the costing data, the information and the number of measures that came out?

I am not sure it would have been terribly different, David -- and I would like

your perspective; it would be interesting -- if the other vendors had come forward. I suspect we probably would have tripped onto some of the same issues over and over again with all of them.

DR. NEEDLEMAN: I think we are going to keep ping-ponging questions back and forth between us and NQF, because I think this has to do as much with the philosophy of what you are about, and what your certification of a measure communicates.

So part of it is a question for David in some sense, which is: Do you need an NQF reviewing measures from these vendors to help you decide or for other payers or other groups to decide whether they should buy it, basically, or is that something that -- If there is a value there for folks who are going to use this for private purposes, never going to get to public reporting, but clearly are going to use it in a whole variety of accountability ways privately or internally,

but it is a large and growing vendor community, and I am new in this field. Is there a value to the potential purchasers of a commercial product of having you vet it or can we assume that the purchasers are sufficiently sophisticated, have enough both financial capacity and technical capacity, to evaluate the value of an Ingenix measure, an NCQA measure or a Prometheus measure if they commercialize that without it going through your process and getting your imprimatur?

That, seems to me, to be part of the issue for the discussion here, and I don't have any clear insight into that.

DR. REDFEARN: Tom and I were arguing a little bit yesterday about whether there was any value to Ingenix from being certified, and I think there certainly is some value to have one of their measures certified. But my comment was, when you are looking at a million dollar product, I doubt if that would make very much difference in terms of

purchase decisions by companies.

I think there is value there, and that is probably one of the reasons that Ingenix is interested in participating in this process, but I wonder how much difference it is going to make to them selling the product. I don't know.

I think companies make decisions about these products based on there are some internal business needs and what they think will work and what they need to do. If there is some external -- powerful external force that tells the large carriers like WellPoint, Aetna, CIGNA, United, that you have to do this if you want to compete in the market, then it will happen, even at these price points, I think, but I don't see that. I don't see any entity having the power to do that.

The only one is CMS. I mean, if CMS certifies an episode grouper, I think that is going to send a lot of -- a huge message to the community, but these individual measures,

I don't see that having that much influence. But that is my guess.

CO-CHAIR ROSENTHAL: The only -- Back to the comment about what the currently stated NQF goals are in having approved measures being nationwide public reporting for driving both improvements and transparency, unfortunately, these groupers don't pass the big test, in my opinion.

They may pass some small tests, and it gets back to a little of the dialogue we had yesterday around are there quality measures that correspond to these things, and the answer was, yeah, there are some where they are more registry based, etcetera, and if you part of the registry, you can reap the value of them. But that doesn't pass the big test either, quite frankly, it doesn't seem to me, and the big test is the one that you posited as the major goal for this, which is transparency, driving improvement, etcetera, etcetera.

So I think the jury is out on the answer to is this pursuable.

MS. PACE: One question that Helen posed is, in this particular space, are there any non-proprietary measures that you all were aware of? Did we? Oh, okay.

MR. AMIN: Yes, the ABMS measures and the Prometheus measures would be two large players.

MS. YANAGIHARA: We have some non-episode based measures that we just didn't have the resources to put through the process, but I was just going to comment on -- I think most of the measures, with the exception, I think, of the HealthPartners measures, are in use for quality improvement, really only, pretty much.

There may be a few users of the Ingenix that are using them for accountability purposes. I don't know of anyone except from HealthPartners that are using for public reporting. So I think that this criteria was

always a struggle for me, because I am like it doesn't meet that public accountability or even any plans for public accountability, as far as I could tell at this point.

I take that back. NCQA, and that is publicly reported, the relative resources measure. But I think that is a struggle, and I think, if it really is only for internal quality improvement, we don't need this whole process. It really is when it is for public accountability that I think that this endorsement process becomes really important.

So I do think it is important to keep that in mind and, when measures are being evaluated, to really take that into consideration, because I feel like we were kind of lax on that, and maybe for good reason, because we are early in this process. But I think it is really important going forward to make sure that it is not just for internal quality improvement, but really for that public accountability/transparency.

DR. BURSTIN: This has been an interesting issue we have been dealing with a lot in the usability task force as well, is that it has been for quality improvement and public reporting, and before that it said for public accountability, and I think we have now moved toward a broader set of accountability functions.

So I think a lot of docs would argue that, if health plans are using these measures to tier docs and pay them differentially, that they should be in the mix. So I think there is a difference of saying you are using internally for QI, and whether there is still some public facing way that -- Barbara is probably going to say something along these lines.

DR. RUDOLPH: Yes. I mean, I am thinking of the court case in New York where the plan was sued about the appropriateness of the tiering and so forth.

So I think these things -- their

getting endorsed is probably going to lead to a path eventually of public reporting, just because there is enough health plans and others who probably already have access to many of these Ingenix groupers, etcetera, and other proprietary groupers that will eventually become publicly reported, and the endorsement helps them in a sense that it legitimizes what it is that they have done and, therefore, they would be doing standard practice, like in the medical terms -- you know, the customary practice.

So I think it provides protection for them.

CO-CHAIR STEINWALD: Your Wayne Gretzky metaphor -- your only sports metaphor? So I would say this. You know, if you viewed this experiment as one whose sole output was a handful of endorsed measures -- maybe not, but if you are going to skate to where the puck is going to be, more public accountability, more public reporting, more system-wide

accountability measures as opposed to small groups and providers, but then that begs the question of what do you do next? Right?

So what does skating to where the puck is going to be mean for NQF? That is not for us to decide. Right? That is up to you.

DR. BURSTIN: Although I think that -- you know, and one of the things we haven't talked about yet is where we really want to go. Right? It is efficiency measures. So we kept stumbling on the fact that, hey, it is pretty hard to do efficiency measures if you don't have anything on cost and resource use.

So we feel like this is our foray, but that is -- The puck, I think, is efficiency and value.

MS. PACE: And I should -- I think that is a good reminder, that I don't think NQF would really advocate that just these resource use measures be publicly reported, because what does it tell you other than cost

is different. It doesn't tell you whether high cost is associated with better outcomes or worse outcomes. So it is really not information that people can rally act on.

Also, as Helen mentioned, although public reporting and maximum transparency is going to continue to be the goal that we are really recognizing and trying to work into the criteria, and how that will be evaluated is progress along that goal, and certainly other accountability functions would count and, for some measures, it may not be appropriate for public reporting.

CO-CHAIR STEINWALD: This is a good segue into our last topic or conversation. Tom, and then Dolores, and then Paul.

CO-CHAIR ROSENTHAL: Well, I think one of the -- To the point of, quote, "where should we," again one of my disappointments is we really spent most of the time here grappling with the groupers, and gosh -- and

coming back to something Paul alluded to at the beginning of this conversation which is, there have got to be 50 or 100 other sort of, kind of metrics that would be efficiency/resource utilization ones, and I jotted a few down that would get at the issue of that 30 percent that is arguably, quote "unnecessary care" that we are not capturing even in this sphere, because with the exception of the population based things, none of these things grapple with the appropriateness part of the equation. But you can get the things that Dartmouth just did on use of cancer drugs in the last week of Life.

You've got bed days per 1,000 in the HMO world. You have got ER visits per 1,000. You got MRI measurements. You got a ton of things, and for whatever set of reasons, we ended up spending virtually all our time grappling with grouper methodologies, but I would hope that in some fashion in the report we might solicit a wider variety of

submissions, and perhaps even offer up the lure that there could be -- I don't want to say a fast track piece, because that implies there is some shortcut to it, but that it would not be necessarily some hideous, onerous process to get one of those kinds of things approved, and it clearly would expand the world in which we are grappling, and there has got to be dozens of those things out there.

MS. TURBYVILLE: Could I quickly just jump in? A lot of the measures that came through, including the NCQA measures, included ER visits, discharges, and I think the committees and the TAPs did focus mainly on the costing part, but including the Ingenix measure, though they were built into the grouper. They did include these utilization metrics within the -- not to disagree with your statement that we can't cast a broader net, but within the conditions we did see those coming through.

CO-CHAIR ROSENTHAL: Well, yes,

there is no doubt within the conditions, it is all costs of those things. So it would include -- It clearly includes the drugs, but there are population based metrics around those things that are independent of the specific condition that again casts a narrower net of the thing being measured, and that again are widely accepted measures around the country.

Again, there are other kinds of things related to the end of life use, etcetera. I won't belabor the point.

DR. ELWARD: This is Kurt.

CO-CHAIR STEINWALD: Kurt? Okay, go ahead.

DR. ELWARD: Yes, this is Kurt. Is it okay if I make a comment?

CO-CHAIR ROSENTHAL: Sure, go ahead.

DR. ELWARD: I think Tom's point is very well taken. It strikes me, though, that some type of grouping episodes of care is

going to become increasingly important as things like ACOs, patient centered home development -- you know, the public is going to be given a lot of information, true or not true, about how well this care is provided and at what cost, particularly since, I think, no matter where we go with reform or not, we are going to be dealing with significant cost challenges.

So I think that there is -- As more specialty care gets done as outpatients and you start centering again moving toward patient care models, I think that thinking about some type of episodes of care model is going to be, actually, very appropriate, and to keep the playing field level, I think the next things you have to focus on -- you know, the things that are out there will really prevent a lot of -- I think, will serve the public safety a lot better. Thanks.

MS. YANAGIHARA: Two points. One is I think the HealthPartners total cost of

care and resource use index actually does, at a population level, break down into all those different kind of cost categories. So that is, I think, a kind of format measure.

Then my other point is on, really, the kind of discussion that was going on a little earlier. The physician and hospital quality certification program that NCQA has actually has a number of standards that I think are applicable and may be useful in terms of guides.

They give you credit for having a majority of your measures that are either NQF endorsed or by a national accreditor or by the government. So that is one of their standards.

You can only use cost in conjunction with quality. You are not supposed to use it for action, -- and I will define action in just a minute -- you know, cost on its own or resources on its own.

Then the three definitions of kind

of using it for action are public reporting, payment or benefit design. So that might be a way to kind of frame that accountability -- you know, the aspect that we are trying to get at. It is not necessarily just public reporting, but for payment purposes or for benefit design purposes. That all would fall into that category.

MS. PACE: And, actually, some prior work has been done in terms of identifying those accountability applications, certainly payment. Different incentives could be accreditation or certification, all of those kinds of things.

DR. BARNETT: Those were NQF?

MS. YANAGIHARA: NCQA physician and hospital quality certification. Actually, it was in response to the whole Attorney General case in New York, and they came up with certification standards for health plans and other organizations, and we actually just went through that on behalf of our health

plans for the stuff that we do for them.

DR. BARNETT: I was just going to follow up on what John said, and appreciate his thoughts and ideas about what some specific appropriateness criteria might be.

So in our health care system, we have had for more than 10 years now a quality enhancement research initiative. There's 10 centers funded to do implementation of guideline concordant care, quality improvement projects, and it has been done pretty much without regard to cost, although we have done some economic evaluation.

So these 10 centers each -- you know, they have at any given time a half-dozen projects that are quality improvement efforts. So our national director of the Quality Enhancement Researcher Query Initiative, David Atkins, has asked each of the queries now to come up with a de-implementation or a disinvestment program.

So if doing inappropriate things

is clearly not high quality care, so doing less of the inappropriate things is a kind of quality improvement and, obviously, it has economic implications.

CO-CHAIR ROSENTHAL: One other thing that occurred to me just in the moment is that one aspect of this is to try to change the whole public dynamic, but the other imperative, it seems to me and to some others, is the idea of encouraging, if not promoting, integration.

In some way, the ETG methodologies that tried to get down to the individual physician level, obviously, have their purposes. It is not clear to me, though, that they promote integration. In fact, you might argue that they don't promote integration at all, because the goal is to get -- you could have a doc out completely on his own and, if you have attributed the patients correctly, he or she doesn't have to be involved in any organized entity whatsoever.

So one of the things that might be where the puck ought to go is to say that measures that specifically encourage integrated behavior would be desirable and, like it or not, one of the integrators in this country are hospitals, ironically, and it has been demonstrated pretty substantially that, if you hold hospitals accountable, somehow they manage to figure out how to get their physicians engaged, even if they don't own the physicians.

So it was surprising, again, not to see any hospital oriented measures, but if there were hospital oriented measures, I believe it would have the additional payoff of creating an imperative toward integrating; whereas, maybe somebody could argue or debate with me on this, but I am not sure most of the ones we saw have that as a specific outcome of the measure.

DR. NEEDLEMAN: I have been thinking about the nature of the measures we

have been looking at, and what we mean by efficiency is going to be -- differs across the different kinds of measures.

We have got some measures that are measuring acute episodes, often surgical episodes, the hip/knee sort of thing, and there the resource use concept, I think, is clearer. You know, patient comes in -- The patient starts into treatment. Something happens. They get done with treatment.

Then we have got -- So the concept of an episode there, I think, makes a lot of sense. When we start looking at some of the care that represents primary care, coordinated care, long term care for patients, we begin looking at these concepts of episodes around diseases, and that just increasingly feels wrong.

You know, Helen has made the point in several ways that all of her patients are coming in with multiple conditions, and that is increasingly what we are seeing. If we are

thinking about Medicare episode groupers, that is extensively what we are going to see. The 20 percent of the patients that are 80 percent of the costs are all very complex patients or they have extremely -- you know, half of them are very complex patients that are in the system for the long period.

So perhaps we need to -- For that set of patients, if we are trying to understand resource use, we need to perhaps encourage the developers to be thinking about patient centered definitions of who these are, so we stop thinking about COPD as a disease that needs treatment, and we think about patients who have COPD.

We saw a little bit of this with the NCQA measures, but really think about the fact that that patient may also be a patient with -- and add to the problem list. We need to think about dealing with complex patients, and maybe some of these things are together sufficiently we can think about what the

resource use is for patients with this cluster of diseases.

Then we can begin tying the resource use to measures of outcome or measures of specific outcomes or levels of maintenance of disease progression that will tie back to the measure of quality. But the core question here is, when this patient comes to you, how many resources are being applied to their care, and how much bang for the buck are we getting for that? Are there ways to get the same bang with fewer resources? But it is about the patient, and it is about the complexity of the patient.

We need to think about how the grouper methodologies or the episode methodologies fully capture the complexity of the patients that are there, and what kinds of resource levels are needed to care for them.

MR. AMIN: Jack, that was very well said. As we are sort of thinking about the last 15 minutes here, one of the things

that we really wanted to focus on was, as we are sort of thinking about the NQF endorsed measuring framework for efficiency, we have really had a large discussion around making sure that the measure -- these types of cost measures and resource use measures are not pursued individually or in isolation, but rather as an essential subcomponent of the larger groups of measures.

I wanted to bring us back to our sort of evolving conceptual model, the patient centered episode of care, which really is getting to what you are speaking about, what you just spent about five minutes talking about. But I just wanted to make sure that we are not in the framework of thinking about episodes in the way that we have gotten measures from.

Let's think about this, really, in the sense of understanding the patient through their trajectory of care over their care continuum, and also in a sense of having

multiple co-occurring conditions, so this issue that we have been discussing around how to deal with multiple comorbidities, and how the clinical hierarchies would work for a patient through a patient centered episode of care framework, and also thinking about even the questions that Paul had raised earlier about what types of costs we are thinking about when we are thinking about it from a patient centered episode of care framework.

As we are sort of thinking through the way forward of where the puck should be, how do we start to give guidance to the field of developers of how to really construct these measures in a way that are truly patient centered, patient centered episode of care that captures a care trajectory that not only groups them according to the underlying conditions, but captures the patients and their inherent complexities?

It also begs the question of how do we actually start to think about the

alignments of these resource use measures with their appropriate outcomes. Do we expect that the way that the measures are paired have aligned denominators or have paired populations that we are measuring, has alignments on risk adjustment?

How does this actually -- How do we actually think through how these efficiency measures are developed? And as I framed the first question, how do we start to integrate this into the patient centered episode of care model in a way that is truly patient centered, as we are thinking through the future of these types of measures?

CO-CHAIR STEINWALD: It seems to me -- Actually, David stepped out, but others can answer this, too -- that most resource measurement, especially of the kind that the developers who sent us some measures have developed, are used in conjunction with quality measures, but they are not combined into composite efficiency measures. Right?

Like if it is an episode of care for congestive heart failure, there are some quality measures, and then they have a resource measure, and they are trying to see separately if they re meeting certain standards of both quality and -- they probably call it efficiency.

I am not personally very familiar myself with any composite measures like, you know, at the conceptual level it is the cost for a given level of quality. I think it is cost provided that certain standards of quality or measures of quality are satisfied.

So I am, in part, raising this to ask the people who are more familiar, have I got that right or has the field gone further than I had supposed?

DR. BARNETT: So there is a Kindig book on purchasing population health, which talks about this. A lot of people talk about this. So when we do cost effectiveness analysis, we think of looking at an

intervention and its impact on cost per quality adjusted life year.

So there is a standard way of valuing health outcomes in terms of morbidity adjusted survival or qualities, and the U.S. Public Health Service Task Force kind of enunciated the standard method of doing cost effectiveness analysis.

There are recommendations they published in 1996. Tufts has a cost effectiveness registry that Peter Newman and the people in his shop have created, which is basically what they call a league table, but a list of all the cost effectiveness findings for every intervention that has been published, and he has, amazingly, been able to try to keep this up. so there's thousands of entries in this.

So peter also published a paper a while ago, another "do not do" list, which were things that were disseminated that have very high cost for quality, in other words,

low value/high cost care.

So far we have nibbled at the margins, thinking of things that we shouldn't do, and interventions that we have evaluated with cost effectiveness studies, and so things that are clearly cost effective, things that are clearly ruled out, some stuff that is kind of in the middle, because they are so close to that threshold, we think that the health care payers -- well, various estimates, started out being $50,000 per quality is as much as we would pay for any intervention at U.S. Health Care.

There have been a lot of publications about what is the appropriate value now, international studies that say it should be about the per capita income for quality. Where they got that from -- it just seems to be what the health plans in various countries will pay.

So the problem of using that approach -- and you know what Kindig's whole

thing about the limitation is -- is we don't have cost effectiveness analysis on everything. Heck, we don't even know whether a lot of the stuff we do is effective at all, whether there is any marginal benefit, let alone what is the size of the marginal benefit. Right? That is the comparative effectiveness gap, research gap. We know it works compared to placebo, but we don't know whether it works compared to the alternative treatment.

So that is like the Holy Grail, is to know, everything we do, exactly what the payoff is going to be, or at least what the probability of the payoff is. So all we can really do is know about the things that are extreme outliers. Gets back to this disinvestment idea.

I think it is interesting to note that Peter Turk said, hey, here's a lot of things that are really not -- we shouldn't be doing, because they are so low value, and

saying this is -- and actually, what his paper says is this could be the basis of designing a low cost health plan. If we just say we are not going to give you these low value -- and our health plan just won't offer these low value, low payoff things, you could save a tremendous amount of money and offer people a health plan that would deliver a lot of value.

MS. YANAGIHARA: I think at a kind of higher level, I think what I have seen most frequently, and what NCQA does with their relative resources measures is looks at kind of a quadrant, and so you look at -- One axis is cost of resources, and the other one is quality, and you see kind of which quadrant different organizations fall into.

So I think that is what I have seen most commonly. We have been grappling with trying to figure out some kind of a value calculation, bringing our total cost of care measure and our quality composite together somehow.

What we have been advised by different people is that, when you start combining it into a composite, you lose sort of -- Everything kind of could wash out. So you kind of lose the high cost, low quality or low cost, high quality. I mean, it kind of all comes out average.

So it is really tricky trying to figure out how to combine them together. I think that is why people end up doing the quadrant, and I am just seeing where people fall on that graph.

You could then just make a judgment and say, okay, we are only going to look at the high quality, low cost group. So they are going to be some kind of differentially paid or something than these other ones, but it is tricky, and I don't know that there is a lot of work around actually combining it into a composite.

I can't remember now, to be honest, who gave us that advice. I think it

was actually Wisconsin, to be honest. I'm trying to think of who it was, but anyway --

MR. BOWHAN: John is definitely on the quadrant. There is no question about that. We played around with it, but just what you said, you don't know what is driving that actual measure, that score, if it is quality or cost. So your quadrant makes it evident.

CO-CHAIR STEINWALD: Just as an observation, it is just remarkable to me how much mistrust there is of the concept of cost effectiveness, going way back to the Office of Technology Assessment and its demise and all the things that have happened since this, and we saddled with this term comparative effect in this, because it is like saying Voldemort, you know, to say cost effective, and the very idea that we would use something like quality adjusted -- cost for quality adjusted life here is sort of like -- Isn't that what those Socialist countries across the Atlantic Ocean do?

So I think it is a shame, but if NQF could kind of advance that ball a little bit, I think that would be extremely useful.

MS. YANAGIHARA: That is kind more to the very granular service level, though. I think, if we are talking about the kind of measures that are a little bit more global -- I mean, maybe for some utilization kind of measures or something, it might -- you might have those kind of direct comparisons, but at the higher level, I am not sure that you would.

MR. AMIN: Just to frame a little bit of, I think, where this question and where this concept is coming from is that, you know, in essence we recognize that efficiency measures really need to -- maybe link is not the right word, but resource use measures and cost measures need to be reported in the context of quality measures in a fairly robust fashion in the way that they are used, recognizing that the resource use or cost

measure needs to be scientifically acceptable in its own right. But as we sort of think through in the future of the guidance on how we are starting to think through measures of efficiency, how do these measures come together in a way that is representative of the efficiency of the care system, recognizing that we still want to keep it in the context of the patient, that it is not just looking at in some sense a disease specific model, but it is actually patient centered in some way.

Again, it may be just a question for thought, more or less, than an actual answer, but is it that the measures in some way need to be constructed so that -- or not constructed, but evaluated in the way that they systematically have the same denominator populations or appropriate risk adjustment models that span both the quality and the resource use measure, or is that we just sort of evaluate the same construct?

I am not sure. I am just

throwing that out as a question to the group as we are thinking through that.

DR. NEEDLEMAN: I think this -- You come back to how these measures are likely to be used, and where the shoals are in the stream bed as we try to go down the river. God, that's a horrible metaphor. I apologize for that.

I think there is a perception that we are spending a lot of money with no value, that we can find examples of people that are getting better outcomes and using less -- fewer resources to do them, or getting high performance and using fewer resources to achieve it, and we want to move the system to look like that.

The issue of is it worth paying this to get this additional stuff, which is implicit in the quality measurement issue, is one that can be deferred until we get everybody doing at least as well as we now know how to without -- you know, just by

squeezing the waste out.

So if I am looking at where we are in terms of where the public is, we don't want death panels. We don't want to say it is not worth saving your life, if it is going to cost $10 million to do that, even though we make those decisions all the time in reality. But the place we are is, if this provider produces high out, high, good outcomes, and is spending a lot less, how do we get people to look like them, and how do we get to understand what they are doing and how they deploy the resources they use?

That is where I think these measures are going to have their immediate impact and where the immediate use is going to be. So I do think the quadrant is valid, and we are trying to move people to the high value, low cost quadrant, and ultimately the question is how are these measures going to feed into our ability to do that.

That is the place we ought to be

starting, and I think, to some extent, we ought to defer the issue of asking is this care worth it, in the sense of spending more to get more. That is a different debate, and that is a different forum than I think we are going to be seeing these measures used in.

MR. PHILLIPS: Just in thinking back over where we were, I started in thinking, entering this effort, from the perspective of, okay, we are looking to merge resource use with outcomes measures and come up with efficiency for various episodes, to where we have kind of focused more in on resource use and then, I guess, coming into thinking as we are headed to the end, you know, the real tough work around developing appropriate quality measures that will tie to these and how far away we are from that.

Thinking about that, the thought occurs that -- I mean, we have had DRGs and prospective payment systems for other providers, but generally accepted is -- you

know, have their problems that people are tinkering with, but generally function pretty well. They pretty much rely on an accepted case mix measure that pays based on the average within a case mix, without actual outcomes measures to show that the average is good or bad. People have kind of worked with that.

I guess it has brought me to the place of thinking, you know, this effort is a very good first start in trying to identify some of the issues that can at least get us to the place where we can hopefully get to where we can come up with averages that people are comfortable with, based on there is good case mix adjustment.

Hopefully, we need to still develop the outcomes measures so that we were able to be confident that the averages get enough or is good, but I guess I am maybe more confident at this point than I was a little while back as far as this first step moving us

down the road.

CO-CHAIR STEINWALD: Paul.

DR. BARNETT: Yes, just to follow up on what Dolores said about thinking about the world as a two-space of cost on the y axis and quality on the x axis, and we want to be in the right -- the corridor where the costs are lower and the quality is higher.

Actually, there is subtle variations on that, but you can divide the world in half based on your -- that space in half based on your judgment of how much quality is worth. But rather than get lost in explaining that too much, I think the issue is whether you can do that at the level of an organization or whether you do that at the level of a specific intervention.

The problem with doing it at the level of a specific intervention is there's just too many darn interventions to evaluate to provide advice on every possible thing.

The problem with doing it at the

organization level is that the organization can't take your analysis saying that they are, say, high cost and delivering low value, and turn around and have any specific actionable item, because they don't know which interventions they are doing are wrong.

I have also heard a lot of, I got to say, more rhetoric than proof that improving quality saves cost. So there is some thought that, if you avoid the bad events and you don't have the central infections and you do the right stuff that you are going to save cost.

I think it is also -- So I will observe that most of the quality improvement efforts that we have engaged in have been costly and may be cost effective. Maybe they are adding cost at less than $50,000 of quality, but it is hard to change provider behavior. Even when you have a very specific guideline recommended car that everyone agrees needs to be done, it is very hard to get

people to change their patterns, and it takes sustained and expensive effort to achieve that.

So I think the way to move organizations is you figure out some specific measure, and then you manage to a performance measure. I am just saying we need some that are performance measures based on the tradeoff between quality and cost or value and cost, and that you can hold up organizations and say this is how you perform, but I don't think it gives them enough to actually manage to make a change.

CO-CHAIR STEINWALD: We are getting close to the time that we had hoped to adjourn. This is our last face to face meeting. Right? No applause either.

So I wonder if anybody would want to take the opportunity to suggest anything related to pairing cost and quality outcomes or any other thing for the benefit of NQF and, in particular, the preparation of the report

that still lies in front of you.

MS. WILBON: Right. As you guys are thinking about that, we just have this one last slide to kind of reiterate where we are going.

We are anticipating evaluating the public sector episode grouper sometime next year, and in that same space, hopefully, we will be taking another look at the criteria based on your feedback today and throughout the process, seeing how we might refine that a little bit for any future efforts we have on evaluating individual resources measures.

Then to Bruce's suggestion, if you have any other suggestions on things we should consider as we move forward on those and other things you think we should be looking at or process suggestions. The suggestions you made on the submission form are extremely helpful.

So any other thing like that, we are open to that. So thank you all. It has been a really amazing, extremely rewarding

learning experience for me, too, because actually, this was my first foray into resources measurement. It has been really working with you guys.

So thank you for your efforts. We realize it was a tremendous undertaking, and it has been quite an extended project, too. So we appreciate your time.

On that note, there will be two more conference calls. I know you guys thought you were done with this meeting. When the comment period ends for this draft report, we will have a call. Sheila, our administrative coordinator, will be sending out an email to you guys to schedule a call to discuss the comments that come in from the first draft report, and then we will have another call after the second draft report, and then that will conclude the Steering Committee calls at that point, but we would like also to get your input on the comments that come in from the report and see if there

are any ways we need to revise it or if there are things that came up that, for some reason, you guys didn't consider. You did a pretty thorough job, but there are often things that come up from the public and the membership that sometimes either weren't discussed or --

CO-CHAIR STEINWALD: So there will be comments on the report that has already been sent out.

MS. WILBON: Yes.

CO-CHAIR STEINWALD: Then you are going to draft a second report that incorporates a lot of new information, in addition to those public comments, and then you will send that to us in draft.

MS. WILBON: Yes.

CO-CHAIR STEINWALD: Seeking our comments on that. Then following that, a final conference call. Then do we get a little -- We get a letter from -- Yes, Paul?

DR. BARNETT: I just had one question, and perhaps I was napping or

otherwise distracted when it was announced, but what happened with the cancer measures in the cancer TAP?

MS. WILBON: They were all ABMS measures. There were four. It was two breast cancer, two colon cancer measures, and they were all submitted from ABMS, and those were all withdrawn. So the cancer TAP actually did an amazing job, too. It is unfortunate that we didn't get a chance to move those forward, but those kind of dropped out of the process with the ABMS withdrawal.

MR. AMIN: The only thing I also would add is that I truly -- you know, from all of our project team, truly appreciate all of the hard work that each of you have done, and I sincerely appreciate the leadership from Bruce and Tom and all of the TAP chairs that have really taken the time to review all these very extensive measures. Really appreciate your leadership on this, and I hope that our work as we sort of redraft the report is

reflective of the quality thinking and quality effort that you have put into this process. So we really appreciate that.

CO-CHAIR STEINWALD: Sure. Well, Helen, any final?

DR. BURSTIN: Thank you for learning with us. I think we have gotten a little bit further down the ice, my only first run on this. Take it to the end.

I think, while this project will end, it is clear this is not the last time that we will encounter many of you. I just really thank you. We have learned an amazing amount.

I have always said at the outset that I thought this project had a heavy dose of learning and, if we got some measures out of it, that would be nice, too. But I do think we have learned a tremendous amount. You guys have been wonderful on getting us there. So thank you.

CO-CHAIR STEINWALD: Certainly

true for me. Anything final? All right. Yes, Jack?

DR. NEEDLEMAN: A couple or three things. First of all, as one of the more vocal folks complaining about how the hell could you schedule a meeting at the end of August in Washington, D.C., I want to thank the staff for arranging gorgeous weather for us while we were here.

More to the point, I've got to compliment the staff and say how lucky Helen is. It has been an extraordinary group of people, an extraordinary group of materials that you have been able to pull together. So we are -- AS a committee member, I am deeply in your debt for the work you have done that enabled us to do the work you have asked of us.

I want to second Taroon's comment about Bruce and Tom and the work they did, which was also extraordinary, given the complexity and the details.

Thinking about the work ahead, we have spent a lot of time talking about methodology and how the measures are constructed, and all that has assumed the data that whoever is proposing the measure says that should be used. But I think our conversations have repeatedly underscored the challenges with data for doing this work in at least two ways.

The carve-out stuff looks like carve-out stuff, but it is really about integrating data from multiple places that have a piece of how much has been spent on different kinds of care, and when you begin looking at the public grouper work, and particularly thinking about getting resource measures out from Medicare, you are going to encounter the same issues in spades.

We have got Medicare Part D for drug data, but that ain't the only place drug data sits. You have got all the Medicare Advantage plans, some of which have internal

billing systems. So the billing data is sort of there, if you can get it back from them -- that are also doing some drug stuff and some other stuff.

You have got some Medicare Advantage and care plans like Kaiser which do not have good encounter data or where it is just encounter data, because people are capitated, and we know the encounter data has been crap, because there has been no incentive for doing it.

So we have got all kinds. As you look at these public groupers, you really do need to think about whether the data will support what the groupers intend it to support, and what kinds of things are going to be needed to get the data that actually enable you to measure how many resources are being used for different kinds of Medicare clients/beneficiaries. They go beyond the issue of does the grouper get an interesting number out if the data are there.

So I don't see how the data issue is separate from the measure construction issues, and that, I think, is one of the key lessons of the work we have done.

The second thing is the point that I have said and try to say nearly every meeting, which is billing data does not include all the resources that are used in care or all the things that make care effective.

If we want to understand why a Kaiser or a UCLA group have or if somebody else has better performance than others, we need to look at some of the things that aren't showing up in the billing, like do they use nurse educators for the patients? Do they use diabetes educators? Have they got other kinds of specialized staff that are not being billed that they use to make sure care is effective?

If we are going to the health plan area, we have got this issue of what kinds of health education activities and resources are

the plans making available to their beneficiaries, at what cost, and how are those being used with what effect?

So we have got all kinds of services that are in the system that are not being billed for and, therefore, invisible to this enterprise, and one of the long term goals, if we want to understand the resources and how to be effective in delivering care, is figuring out how to make those resources visible and understand which of them are worth doing, and which strategies for doing them are more effective, so they are really worth doing.

That is a second long term challenge and, even as you begin getting into the details of we got all this billing data, how do we organize it, you need to keep that second agenda in mind as you try to move the larger agenda of understanding what resources are worth investing in care to deliver the outcomes we want for patients, as the long

term goal for all this work.

CO-CHAIR STEINWALD: We do have public comments.

MS. WILBON: Operator, is there anyone on the participant line from the public?

OPERATOR: Yes, we do have people on the public line. Would you like me to open their line?

MS. WILBON: Yes. Could we open if there are comments.

OPERATOR: Again, our lines are open.

MR. AMIN: If anyone has a comment, feel free to go ahead and make your comment at this time. Okay, thank you very much. You can go ahead and close the line.

CO-CHAIR STEINWALD: All right. The meeting is adjourned.

(Whereupon, the foregoing matter went off the record at 2:16 p.m.)