NATIONAL QUALITY FORUM
+ + + + +
EFFICIENCY RESOURCE USE

+ + + + +

STEERING COMMITTEE MEETING
OPEN SESSION
+ + + + +

MONDAY, JULY 12, 2010

+ + + + +

The Steering Committee convened at 9:00 a.m. in Suite 600 South of the Homer Building, located at 601 13th Street, N.W., Washington, D.C., Doris Lotz and Bruce Steinwald, Co-Chairs, presiding.

PRESENT:
DORIS H. LOTZ, MD, MPH, CO-CHAIR
BRUCE STEINWALD, MBA, CO-CHAIR

PAUL BARNETT, PhD, VA Palo Alto Healthcare System
JACK BOWHAN, Wisconsin Collaborative for Healthcare Quality
JEPTHA CURTIS, MD, FAAC, Yale University School of Medicine
KURTIS ELWARD, MD, MPH, FAAFP, Family Medicine of Albemarle
WILLIAM GOLDEN, MD, MACP, Arkansas Medicaid
LISA M. GRABERT, MPH, American Hospital Association
ETHAN A. HALM, MD, MPH, University of Texas Southwestern Medical Center
ANN HENDRICH, RN, MSN, FAAN, Ascension Health
THOMAS H. LEE, MD, Partners HealthCare System, Inc.

RENEE MARKUS-HODIN, JD, Community Catalyst
JACK NEEDLEMAN, PhD, FAAN, UCLA School of Public Health
MARY KAY O'NEILL, MD, MBA, CIGNA Healthcare
DAVID PENSON, MD, MPH, Vanderbilt University Medical Center
STEVE PHILLIPS, MPA, Johnson & Johnson Health Care Systems, Inc.
DAVID REDFEARN, PhD, WellPoint
JEFFREY B. RICH, MD, Mid-Atlantic Cardiothoracic Surgeons, Ltd.
WILLIAM RICH, MD, Northern Virginia Ophthalmology Associates
TOM ROSENTHAL, MD, UCLA School of Medicine
BARBARA A. RUDOLPH, PhD, MSSW, The Leapfrog Group
JOSEPH STEPHANSKY, PhD, Michigan Health and Hospital Association
JAMES N. WEINSTEIN, DO, MS, Dartmouth Hitchcock Medical Center
DOLORES YANAGIHARA, MPH, Integrated Healthcare Association

NQF STAFF PRESENT:
HELEN BURSTIN, MD, MPH
JANET CORRIGAN, PhD

MAISHA MIMS, MPH
JENNIFER PODULKA, MPAff (Phase 1)
SALLY TURBYVILLE, MA, MS
ASHLIE WILBON, RN, MPH

ALSO PRESENT:

NIALL BRENNAN, CMS
ADRIAN HUSSAIN, HHS
RITA MUNLEY GALLAGHER, PhD, RN

C-O-N-T-E-N-T-S

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

(9:02 a.m.)

OPEN SESSION

MS. TURBYVILLE: Good morning. I want to welcome everyone here today and on the telephone to the NQF Resource Use Steering Committee meeting. We'll be meeting today, Monday, and tomorrow, Tuesday.

At this time we'll first introduce the staff, NQF staff, that are here today.

My name is Sally Turbyville. I'm a senior director in the performance measurement department.

We have Ashlie Wilbon who is the project manager in performance measurements; Maisha Mims, a research analyst in performance measurement; Jennifer Podulka, who is a senior director in strategic partnership.

DR. BURSTIN: Good morning, Helen Burstin, Senior Vice President for Performance Measures at NQF.

MS. TURBYVILLE: So at this time

I would request that we go around the room and have the steering committee members introduce themselves and the organization that they are coming from.

CO-CHAIR LOTZ: I'm Doris Lotz. I'm the New Hampshire Medicaid medical director.

CO-CHAIR STEINWALD: I'm Bruce Steinwald, I'm an independent consultant living right here in Northwest Washington. Until recently I was in the senior executive service in the Government Accountability Office.

DR. STEPHANSKY: Joe Stephansky. I'm with the Michigan Health and Hospital Association.

DR. ROSENTHAL: Tom Rosenthal, I'm from UCLA.

DR. RUDOLPH: Barbara Rudolph. I represent the Leapfrog Group.

DR. O'NEIL: Mary Kay O'Neil, chief medical office for Cigna in the Pacific

Northwest.

MR. BOWHAN: I'm Jack Bowhan, Wisconsin Collaborative for Health Care Quality.

DR. BARNETT: Paul Barnett with the VA Health Economics Resource Center.

DR. GOLDEN: Bill Gold, Arkansas Medicaid.

DR. NEEDLEMAN: Jack Needleman, UCLA School of Public Health.

MR. CURTIS: Jeptha Curtis, Yale University Hospital.

MS. GRABERT: Lisa Grabert, American Hospital Association.

MS. MARKUS-HODIN: Renee Markus-Hodin, Community Catalyst.

MR. WEINSTEIN: Jim Weinstein, I'm director of the Dartmouth Institute for Health Policy and Clinical Practice and I'm recently the president at Dartmouth Hitchcock.

MS. HENDRICH: Ann Hendrich, Ascension Health vice president for quality

and patient safety.

MR. PHILLIPS: Steve Phillips, director of health policy with Johnson & Johnson.

DR. HALM: Ethan Halm, University of Texas Southwestern in Dallas.

DR. REDFEARN: David Redfearn, WellPoint.

DR. RICH: Jeff Rich, I'm a practicing cardiac surgeon now, but in the immediate past a director of the Medicare fee-for-service program.

MR. YANAGIHARA: Hi, I'm Dolores Yanagihara with the Integrated Health Care Association in California.

DR. RICH: Bill Rich, practicing ophthalmologist in Northern Virginia, and director of health policy for the American Academy of Ophthalmology.

DR. LEE: Tom Lee from Partners Health Care System and Harvard Medical School in Boston.

MS. TURBYVILLE: Thank you.

So briefly to talk about some of the project goals and the status for today I'm going to hand it over to Ashlie Wilbon, who is the project manager.

REVIEW OF PROJECT SCOPE, STEERING COMMITTEE CHANGE, PROJECT ACTIVITIES & PROJECT TIMELINE

MS. WILBON: Good morning, everyone. Bear with us today. We are limited on space so we are going to be playing a little bit of musical chairs throughout the day based on who is going to be leading the conversation from the staff end. So thank you everyone for being here.

A lot of this stuff you've seen before. At this point we just want to kind of set the tone and make sure everyone is on the same page. Today is I think the first time that everyone has been in the same place at the same time, and we had a few absentees on the calls sometimes, so it's just nice to regroup and make sure everyone is aware of

what's going on.

So we are just going to review the project scope, the steering committee charged activities and timeline, what we've done so far.

So the goal of this project again is just to provide guidance, and Helen and Sally have been through this already so I won't dwell on it, but to provide guidance for NQF committees, NQF members and measure developers to make the process for reviewing resource use measures transparent for when we do a call for measures later on this fall.

Also for the resource use measurement white paper to provide input for that as well as the resource use evaluation criteria, and then beginning with phase two you will actually be evaluating the submitted resource use measures for recommendation for endorsement.

So again this project is divided into two phases. The first phase that we are

in now is focused on the white paper and actually deciding on what the evaluation criteria will be for when the measures actually come in.

Phase two we anticipate to begin in the fall. We are looking at doing the call for measures in October, and that will begin the consensus development process. We will talk a little bit more about that on day two towards the end so you have an idea of what that looks like.

We've got one steering committee which is all of you for both of the phases, and beginning in phase two we will have five, at this point, five technical advisory panels, based around the clinical areas of the measures that we are expecting. Got 18 conditions, and then we've divided those 18 conditions into five lumps of clinical areas that we'll have the TAPS focused on.

Each of the TAPS will be chaired by a steering committee member, and most of

you guys already know who you are; again, we can talk about that again on day two when we get into phase two a little bit more.

So the roles of the steering committee, to act as a proxy for the NQF multi-stakeholder membership. Each of you guys come from a different - bring a different perspective and represent each of our various stakeholders from different areas to work with NQF staff to help us achieve the goals of the project: evaluate the candidate measures against the evaluation criteria; respond to comments during the review period; and respond to any directions from CSAC.

And at this point we're anticipating that the evaluation criteria once it's decided on will go to CSAC for review and approval to be implemented.

The CSAC is the Consensus Standards Approval Committee, and Helen may want to give a little more detail, but essentially they are the oversight body that

manages or kind of oversees a lot of the - all the processes that - and procedures that NQF implements. I'll leave it at that.

So all of the steering committee members will be expected to review all the measures, and each measure will be assigned primary and secondary reviewers for more in-depth review. And again we'll talk a little bit more about how the measure review process occurs when we get a little bit closer to phase two. Right now this is just kind of to give you an idea of how that will occur.

So each of the measures will be evaluated against each of the evaluation criterions, and you will review the measure evaluations prepared by the staff, and indicate the extent to which the criterion is met for the rationale, and at that point, once you've reviewed it you will be making recommendations to staff whether or not they should be moved forward for endorsement.

So the steering committee charge

for phase one is again to provide input on the white paper, provide NQF operational guidance for the evaluation criteria, and provide guidance on the call for measures which we will talk about tomorrow as well.

So this is just a high level timeline for the project. We had our call on June 18th where we talked about the evaluation criteria. Today and tomorrow we'll be meeting here for the in person meeting. We expect that the white paper will go up for public and member comments towards probably mid to late August. That's a 30-day comment period. Once we received those comments we will reconvene you guys via conference call to discuss the comments that come in, see if you'd like to respond to them, how they should be incorporated and so forth. That call we've already scheduled based on majority availability with the survey that we sent out via email, and that is scheduled for October 5th.

We will begin the call for measures again as I mentioned in October and expect that the white paper will be finalized in November of 2010.

And we will be simultaneously with the call for measures we'll also be doing a call for the remainder of the TAP members. We do have the TAP chairs within the steering committee, but the remainder of the TAP members have not been selected yet so at the time when we do the call for measures in October we'll be doing the call for TAP members. So between October and December amongst all the holidays and stuff we'll be empanelling those TAP measures, and then we'll begin measure review in January of 2011.

DR. HALM: I'm just trying to understand the role of the white paper beyond just sort of getting people together on the same page. You are doing a lot of public disclosure and commenting on the white paper as if it's sort of the final measure, or it's

sort of going to be the bible of resource use; is that correct?

DR. BURSTIN: Basically because this is such a totally new area for NQF, we have never endorsed resource use measures before, this is a pretty evolving field with lots of people unsure what the terrain and the definitions. We thought it was really important to start with and actually it was specifically requested as well by HHS as part of this work that we do the white paper first as sort of almost a grounding exercise, get everybody on the same page and then do the call for measures.

And again, because we're NQF, everything we do is very transparent. We learn a lot from public comment, routinely get hundreds of comments for everything we put out, and we get it from a very wide array of stakeholders.

So we really value getting that. We often learn a lot about things that

sometimes steering committees don't necessarily think about. So that is the purpose of it. We really see it as something that may be also very useful for the measure developers as they begin thinking about what they want to put forward, how to bring that forward.

MS. WILBON: And I think in a broader sense to just kind of help lay the ground and help people better understand how the steering committee was led to coming up with the evaluation criteria. So it's kind of that background and the path leading into the evaluation criteria and the measures. So hopefully that helps.

DR. RUDOLPH: In terms of resource use measures, I know we're going to have probably a long discussion about what these are, but NQF has endorsed measures on length of stay, readmission, and probably others that in my mind are resource use measures as well as sort of quality metrics.

So just a clarification on that.

DR. BURSTIN: We'll have a lot of discussion about that to follow. Those are sort of classic efficiency measures. We've had those before. But we haven't had sort of the building blocks of ones that truly represent cost resource use alone, which I think is more to follow.

But you are absolutely right: we really see this as building that overall portfolio for efficiency measures to follow.

MS. WILBON: Does anyone else have any questions?

MR. WEINSTEIN: There has been a tendency for a lot of the national committees to avoid the issues of cost. Whether it's comparative effectiveness research through the Institute of Medicine or others. Is this - it creates controversy. People use different kinds of language around that.

And I realize the intent here is to get cost into the mix with measures. Do we

anticipate any problems with that publicly?

(Laughter)

DR. BURSTIN: Yes. I've now been at NQF for three years and I actually now say it, and Barb know this, the easy stuff is kind of done. The hard stuff is to follow. And we know cost is sort of the third rail, but I don't think we can proceed without sort of touching it. So that's why all you smart people are here to help us do that wisely, and avoid touching the rail if we can but doing what is needed.

CO-CHAIR LOTZ: I think there is something to be said too with just creating some metrics. What people do with the numbers once they get them is to some extent their own affair, and what I've come to realize with a fair amount of work, not so much with NQF, a little bit of work with NQF, a lot more work with AHRQ, is that while there is a desire to talk about cost, a lot of costing is local, and if we can just get some measures out there

that we can all agree to that start to incorporate some concepts of cost, then the folks local or at whatever level you play at can figure out what they want to do with it.

MR. WEINSTEIN: I was just trying to create the discussion, because I imagine there will be a lot of discussion, and that the committee probably realizes that and is sensitive to that, was the reason for bringing it up, publicly.

DR. LEE: You know, if I could just interject, I mean I do think that though I have questions and concerns about the traditional NQF framework and public reporting of these measures, I do think that some clarity and standardization about how resource use is measured would be very valuable right now. I mean we and a lot of folks I know are trying to develop value report cards with outcomes over cost, and on the same reporting framework, and having clarity and standardization of what we are actually

putting in that bottom half of the page would be very helpful for efforts to improve.

I'm not sure if public comparison will be that straightforward. But for improvement purposes this will be a very useful exercise.

DR. GOLDEN: I don't know when this would be appropriate in the framework, but one of the questions I had as I read through this is that we are kind of - we are still lumping things that may not be lumpable, in a sense for example when we go through this we have Prometheus and the ABMS materials together. And Prometheus is a payment mechanism, and ABMS is a measurement mechanism, and they may use common principles, but I'm not sure that they are comparable in terms of putting them into the same category. And I'm not sure the white paper has done that yet, and I don't know if that is going to be discussed today. But that may help clarity as we put this material forward.

MS. WILBON: We will actually today be dedicated solely to the white paper. We'll actually be going through section by section. There are definitely sections of the white paper that we have pulled out that we want specific steering committee feedback and decisions on so we will be calling your attention to those and obviously everyone will be bringing their own comments as well. So there will definitely be opportunity to discuss that and revisit that as well.

DR. GOLDEN: Just a quick followup, it'd be nice when we start to have a broad discussion before we get into the details, because what I just brought up is really kind of not in the paper but more a conceptual framework.

MS. WILBON: Right. There will be time.

DR. BURSTIN: Just one brief response, Bill, it's a very good question. We've just been through this and brought in

for example some of the Prometheus complication measures as part of our outcomes project. Again, what we would be bringing in would not be the payment model; it would be the measures they use as part of the payment model. So I think we are trying to stay consistent that we would only be looking at the actual metrics themselves rather than the payment model. But it sounds like for the white paper and a lot of the discussions some really useful principles that might be useful to come out of this as well.

DR. NEEDLEMAN: I'd like to go back briefly to Jim's comment and Tom's comments and some of the implications for the work. We are clearly entering another era of where aggressive cost containment is going to be one of the key drivers of what's going on in health care. Bending the cost curve is going to be the mantra for the next decade. But it needs to be done in a way that quality is actually not only protected but enhanced,

given the current levels of quality of care.

So in the long run the kinds of issues that Tom mentioned, the effort to look in an integrated way at how much we are spending and how much quality we're getting for that is critical. And I think in the rubric we've been using in this committee, that's where the efficiency measure is.

What's - what this paper is about, and with the measures we are talking about seem to me to be an important intermediate step. We are going to be talking about resource use at this stage of the game, not so much efficiency measurement. We need to set the stage for being able to measure efficiency with the resource measures that are used, but I've seen in the phone calls we've had, we keep going back and forth between, are we measuring efficiency or are we measuring resources. We are going to need to be clear about how far - the committee name is resource use, and I think that's where we are, that's

where we have to begin. We need to be thinking about how ultimately it may be used in efficiency, but we should not be evaluating whether these measures are adequate to measure the efficiency of health care. They are not going to be for all the reasons that we have talked about on the phone.

MS. WILBON: Thank you, and again, we'll have lots of opportunities to discuss this later on today.

I am going to quickly move through the rest of these slides so you guys can get to your more important or more exciting conversations.

This is just a summary of what we have done so far. As most of you know we started out working with a subcommittee of the entire steering committee to help us get the white paper to a point that we felt was ready to go to the entire steering committee for review. So the subcommittee reviewed the paper in late May, and we had been working on

incorporating their edits and suggestions into the paper prior to distributing it to the entire steering committee.

We had the conference call on the 18th and the meeting today, obviously. And I will actually hand it over to Sally. Thank you.

WHITE PAPER DISCUSSION

MS. TURBYVILLE: Thank you, Ashlie.

So as Ashlie mentioned and as is indicated on the agenda, we are going to spend much of today talking about the white paper. We do have some key areas that we are hoping to get some agreement on, but we realize that there is going to need to be time for broad discussions as well as discussions that may stem from even some of the items that we want. So we realize that there needs to be flexibility. The slides are really just meant as an outline to help guide us through the day. But if they become no longer useful we

are prepared to continue to move forward.

So again we do want an open discussion. We do want input on how to improve the approach that we might take to endorse resource use measures as well as on the paper itself, and any recommendations on key items that the committee may have.

The paper is currently in this structure where we have our introduction, the reason for the white paper. We talk about what's going on in the real world, getting to one of the comments below, perhaps needs some expansion.

The type of resource use measures on a continuum model, and then what the NQF evaluation principles which we will have time set aside for tomorrow to go through with you, but we can bring them in today if we need to as well as the evaluation criteria itself.

For some of you to make sure we're all on the same page we did our best that we could to get the various current measure

development approaches methods in the appendix. We weren't able to get all of them in there, 3M for example. But to the extent that we could we put that in there.

The main reason was to demonstrate how different fundamentally some of them really are, and our sensitivity to that.

So I'm going to hand it over to Bruce and Doris to lead this conversation. We are here for questions or to respond to any comments that you have that you need NQF staff. But we really want this to be a conversation amongst all of you to help guide us through any improvements in the white paper, things that are missing as well as the evaluation approach itself.

CO-CHAIR LOTZ: Niall Brennan is back in the room with us, and there was a question that was asked when he was out, and we said, oh we'll ask that when Niall is in, so I did give him a heads up. But Niall, if you could speak briefly before we begin

reviewing the paper as to how this project bumps up against CMS' RFP requesting the development of episode-based measures.

MR. BRENNAN: Sure.

DR. BURSTIN: Niall, could you come up to a microphone.

MR. BRENNAN: Good morning, everybody.

I suppose the first thing to say is when we conceived of the NQF process the notion of a requirement of a public domain grouper wasn't exactly law yet. So the short answer is, stuff happens. And we need to work around it.

But I do think that the efforts are very complementary. I mean it may seem a little counterintuitive, but to tie into Tom's point and some others specifically, the whole purpose of this project is to try and establish standards and consensus around how resource use measures should be evaluated, what's right, what's wrong, what's good or

bad. And I think a lot of the findings of this process would hopefully feed into the public domain grouper process.

There is an RFP on the streets. We'll be evaluating proposals over the next couple of months and hopefully making an award or awards sometime in the early fall. So I don't know if that answers the question to everybody's satisfaction. So it's a little difficult for me to state with certainty. But fall of this year.

CO-CHAIR STEINWALD: Can I ask a followup?

MR. BRENNAN: Yes.

CO-CHAIR STEINWALD: As you know CMS has a huge task in front of it to implement the provisions of what we are now calling PPACA - just PACA?

MR. BRENNAN: PACA.

CO-CHAIR STEINWALD: PACA? Oh, my gosh. I'm behind the times.

MR. BRENNAN: Affordable Care

Act, with the emphasis on affordable.

(Laughter)

CO-CHAIR STEINWALD: So you heard it and I heard it here first, PACA, and man of those provisions are oriented to moving away from atomistic fee-for-service medicine into different forms of health care delivery. And some of us have wondered whether this project could be complementary to these innovations that you are going to need to identify and implement in any case, and what most of us are hoping that our health system will move in the direction away from solely fee-for-service medicine to more integrated delivery systems.

Do you see a complementarity between that movement and this engagement?

MR. BRENNAN: Absolutely. You know the move to value-based purchasing is one of the key drivers of many many elements of the bill. Accountable care organizations are going to be evaluated against a whole range of quality and cost of care metrics. You have

the expanded physician resource use report program which will ultimately lead to the value modifier provision where physicians will be again evaluated on the basis of cost and quality, and then a value modifier will be calculated for their services.

And even if we do move away from the old style fee for service, first of all, ACOs are really sort of being lowered on top of the fee-for-service structure, and over time we may move to partial or full cap. And there are certainly other models. And even if we move away from payment per widget, we still want to very closely track overall performance and what patients are getting both in terms of quality and quantity of care. So I view them as very complementary. I don't know if that answers your question or not, Bruce.

CO-CHAIR STEINWALD: While we have Niall is here, anyone else like to ask a question on this?

CO-CHAIR LOTZ: One more followup

before we open it up.

So you are seeing the development of these resource use measures that will come after the call up measures as being a component of the assessment of that value and would feed into the application of the value modifiers. You have your quality measures coming in from another pathway so to speak and then these resource measures coming in from this pathway and then leading into the application of that value added modifier that CMS is still discussing at this point.

Is that sort of where we fit in on a bigger picture?

MR. BRENNAN: Yes, the only clarification would be, and maybe it's splitting hairs a little bit, I mean we don't necessarily know and I can't always remember all the dates, the value modifier is like quite a way out there, and this process is right now. So measures that are approved as part of this process, they are the first

generation if you will of resource use measures. So by the time we get around to a value modifier or more aggressively evaluating whatever entity based on quality and cost, they may be different measures, but they will still be based on a foundation of what's happening now. We need to be addressing these issues and developing prototypical standards now and not waiting two more years or three more years. Let's try and evaluate these, get them out into the marketplace, even though many of them are in the marketplace already, and see how people react.

DR. RICH: Niall you saw our timeline and workload. Is that going to be adequate for you to help evaluate the RFP proposals? When I reviewed it it looks like the timelines are - your timeline is a lot shorter than ours.

MR. BRENNAN: Our timelines are somewhat shorter and some of our other internal criteria may be different, but what

we're doing is we're approving work to begin on development of a public domain group or we would imagine that then those measures that come out of that process, we could then look at the evaluation criteria, evaluate them against them.

DR. RICH: I just want to reinforce what you are saying. There is a sense of immediacy at CMS that we have to appreciate, and I think some of the measures, the resource use measures that we will come up with will provide us with a picture of a holy grail seven of eight years from now that we could implement into the payment systems as they evolve. But we have a really complicated payment system now that's based on fee for service, and I know this project is not all about CMS, but the CMS fee-for-service program is in dire need of resource use measurement, and it needs to evolve in its payment mechanisms, and resource use measurement needs to be complementary along the way.

So I would say as we deliberate we should remember that we might not pick the perfect measure or have the perfect goal, but have some intermediate steps and goals that could be utilized by CMS as the payment systems evolve.

DR. GOLDEN: To follow up on that, just so I understand, in this process we are looking at different mechanisms of creating groupers and developing measures. You get general principles which may or may not be what the CMS goes into, and then you have multiple different mechanisms of performing tasks which can then create different kinds of measures like a tangerine or an orange. And I don't know quite what we'll end up with. Will the system accommodate six different silos of measures? Are we looking at principles for different systems, given that they have certain assumptions going in. Even a creative grouper system with certain assumptions and from there

you can create efficiency measures. But we may not match up. So how are we going to put all that together down the road.

MR. BRENNAN: Without wanting to appear to not really answer the question, I think only time will tell. We're both at the beginning of two very complex processes, and it's just difficult for me to state with any certainty what may happen.

I do think one of the unique challenges of this process in particular - not that any of this is easy - is because so many of the measures are different in slightly different ways, and because so many of the post-processing approaches that can be taken can impact what a measure means and how it's interpreted and accepted by people. I think it's a very very tricky issue and area.

CO-CHAIR STEINWALD: Tom.

DR. ROSENTHAL: We can't separate what we're trying to do in terms of identifying some individual measures from what

we think is the long term goal, and the part that I'm having trouble putting my head around is this idea of sort of an individual value for individual physicians, which sort of relates to the grouper methodology, ranking individual physicians with the perhaps more ultimate goal of eliminating fee for service, and driving towards accountable health care organizations. I can't quite figure out how one drives the other. And do you have some thoughts on that based on how CMS is thinking about it? Does the question make sense?

MR. BRENNAN: Could you try me one more time, sorry?

DR. ROSENTHAL: Well, we have this idea of individual doctors, and we are going to rank individual doctors who are practicing fee-for-service medicine and either rank them by some value-based measure or what CMS is developing, yet the real goal is in fact to drive them into accountable health care organizations, I think, and the

elimination of fee for service, and it's not entirely clear to me how the ranking of individual doctors or modifying their fee-for-service payments based on a value purchasing proposition drives the broader goal. A couple of people are nodding their heads.

CO-CHAIR STEINWALD: Jim?

MR. WEINSTEIN: I guess -- just to throw out some ideas on how I frame this, and I don't know if it's helpful for CMS or not. But as Tom said, lots of us are thinking of ways to work within systems that are actually is practical or pragmatic. So you could imagine a grouper for something like knee replacement, which is fairly simple, versus diabetes or something, that people like Ingenix and others have already come up with grouper strategies that we have all read about or are familiar with. And imagine testing that. And I guess my question is, before we jump into policy issues or changing the payment structure, let me ask, is it CMS' goal

to actually test some of these methodologies and measures to see if they actually deliver what's proposed in some RFP or strategy, rather than implementing a broad strategy that ends up failing like the current strategy.

So is that where we're trying to get?

MR. BRENNAN: I think it's a very good point, Jim. And again just to, perhaps, underline the earlier comments, when we referred to ACOs, value modifiers, etcetera. Those are, particularly the value modifier, a very long way in the future. I don't really want this conversation to be about payment reform or payment policy in the short term. I think it's about establishing standards and criteria around getting to useful measures that can eventually inform staff such as that -- as long as we are confident that the measures are giving us accurate information. So I really appreciate that point.

CO-CHAIR STEINWALD: Jack.

DR. NEEDLEMAN: If I was listening carefully enough, one of the things I heard you say was one or several contracts being let. And it ties in, I think, to some of the conversation we've been having here -- which is, the decisions that are made today are going to constrain the decisions that can be made in the future. So if one is picking a specific grouping methodology six months from now, that will have important consequences for the way data is organized, data is collected, data is reported, that will constrain what kinds of measures people use in the future. So where in the -- but we've also talked about, and the white paper makes very clear, we've got a lot of uncertainty about what conceptually is the right way to do this. So where in terms of the development process is CMS trying to hit with the call for development of a public grouper? And is it possible that you're in fact going to have several contracts out there with very

different models being pursued so we have more flexibility in the future to identify the direction we want to go in, based on what we've learned from the initial grouping work?

MR. BRENNAN: I think CMS faces some short-term realities, and has long-term goals. There needs to be a public domain grouper, I believe, and I should be better prepared, by January 1st, 2012. In order for that to happen, CMS has employed a strategy -- as you are probably aware from the RFP -- of pursuing both a short-term goal and a long term goal, a short-term goal that looks possibly at adapting some existing methodologies to better account for measuring these types of things in the Medicare population; a longer term much more expansive goal of: okay, if you had more time and more resources to build a better resource use measure, how would you build it.

CO-CHAIR STEINWALD: Jeff, go ahead.

DR. RICH: Sure, I just want to reflect on what Tom was saying. So just remember that in the Medicare fee-for-service program, there is a Part A and Part B. Part A, the hospital side, already does a lot of bundling, and has a prospective payment system. So resource utilization is felt already a little bit in that system. But on the Part B side, the physician side, it still -- you're paying for widgets, and it's outside surgeons who get paid a 90-day global. Everyone else in health care delivery on the physicians' side is getting paid per diem for what they do on a daily basis.

So in order to modify behavior, you need to measure physician resource use and then apply some payment mechanisms or incentives to drive physicians from single widget producing physicians to a more integrated health care delivery system with hospitals, and that's part of what the value modifier will do. And part of the tiered

structure -- if you read Title III of this program and the value-based purchasing parts -- will be to provide the highest incentives for physicians who are integrated into ACOs, and the lowest to those who continue to remain in the fee-for-service system, which will have to remain. It's a part of the health care delivery in the United States, and it will take a long time to eliminate that. So there has to be sort of multimodal ways of providing incentives to create those organizational and cultural changes in medicine.

CO-CHAIR STEINWALD: Okay, go ahead.

DR. BURSTIN: Just one comment, just to recognize that even as a measure is endorsed at NQF, it is only endorsed for three years, quite intentionally, because we recognize that time changes, and knowledge and experience moves on. So we really view this as being sort of the initial set. I think there will likely be a lot of evolution. If

you looked at the quality measures, for example, even five years ago, compared to a lot of the outcomes we have now, there's already been that evolutionary change, and I think we need to expect this will happen on the research use side as well. We've got to start somewhere. So I think that was our expectation.

CO-CHAIR STEINWALD: All right, then, one last one.

DR. GOLDEN: Just to understand the NQF-CMS relationship, technically isn't CMS supposed to use NQF endorsed activity, unless proven otherwise or something? So how does that relate here?

DR. BURSTIN: Sure. So NQF is a voluntary consensus based organization; it's a standard-setting organization. So we are the measures that CMS needs to look to first to use as consensus standards. So again, they will look among the many measures we have available and potentially select from that if

they have a program for which they need to use it. It doesn't -- what we do is also not just for CMS though. I mean, this is really a diverse set of measures anyone can use, systems, whoever chooses to, purchasers in Bob's world. I mean, I just think having a place to start is important, and it may be that CMS won't even necessarily use these in the short term, while they wait for the longer experience to follow. We just don't know.

But I think from where you sit, it's just really important to say, how do we get this moving, how do we get this sort of set of principles in place, get this Version 1.0 out there, and gain experience with it.

CO-CHAIR STEINWALD: Okay, we will continue the discussion. I think we should excuse Niall.

(Laughter)

CO-CHAIR STEINWALD: I want to speak for the entire steering committee, and thank you for taking the time that you have,

and also any continued guidance for us as we go through our process from you or your colleagues would be much appreciated.

MS. TURBYVILLE: Time to dive into the white paper.

So certainly within Section 1, a key area that we identified that we wanted agreement from this steering committee and thought -- we tried to do it over the telephone and in person I think it's just best -- there are many definitions of efficiency, but to make sure we get agreement on what the definition of resource use is in the context of this project.

Now clearly I heard a call also for a more broad discussion about the white paper, so I'm going to hand it over to the Chairs, and you can see what fits best. But you'll notice in the slides that are in front of you in the manila folder, we have at NQF staff within most of the sections identified key areas that we really hope to get agreement

on.

That aside, the broader discussion is also certainly welcome and helpful.

CO-CHAIR STEINWALD: I'm going to make a suggestion now. Since there are a lot of people around this table, use this convention of putting your tag card up on its edge if you want to say something or ask a question, and then the chairs can do a better job of recognizing people.

DR. BARNETT: Isn't there something on our microphone that says request.

DR. LEE: It's not hooked up.

DR. REDFEARN: It's a placebo.

(Laughter)

DR. BARNETT: It makes me feel much better.

CO-CHAIR STEINWALD: Jack, earlier you started to address this question about what our focus is on, a resource measure is different from an efficiency measure. It seems that we're headed in the direction of

developing resource measures, but we need to acknowledge what the difference is between resource measures and efficiency, and the staff have tried to capture that difference in this slide, or these two slides.

Do you think they have captured it reasonably well?

DR. GOLDEN: Yes. But as -- and I think the world of the staff here. But as Sally introduced this, she started a single sentence which had efficiency and then slid into resource. And I've got no -- I think this definition is good. I think some of the language that's in the paper is good. But we need to be crystal clear in our thinking, if not our presentation, about what the long-term goal is in terms of measuring efficiency, figuring out the value equation in health care, and what this project is capable of doing, which is focusing on how we measure the resources devoted to caring for patients.

And we shouldn't be slipping and

sliding between language that talks about efficiency and language that talks about resource use. They are separate measures at this point. And that's my only concern.

CO-CHAIR STEINWALD: Mary Kay.

DR. O'NEIL: So two things, first on the efficiency point, and that is that looking at the equation I think that we are trying inform by this process is the number of inputs or resources utilized per some defined population or group of patients leading to a given outcome, which would be the efficiency I would say, is the only way that the resource utilization data will have value in evaluating things.

If we can't compare what this entity putting in these resources here versus this entity putting in this different mix of resources there, then counting resources won't have value. So the resources have to be counted with an eye to efficiency, even though it is a separate process.

And then the other thing I wanted I guess from my industry's perspective, for me resource utilization measures are measures of counting. So you have to count office visits, prescriptions, procedures -- whatever you're counting is what the resource issues are. To put some kind of standard dollar value on them I think is dangerous, because costs vary tremendously for all kinds of reasons. Now at some point in time that exercise needs to be done, but if we're counting resources and looking at what resources it takes to properly and efficiently take care of a condition or group of patients, then putting what the costs are has a lot of, like, geographic-specific issues; it has contractually specific issues; it has site-of-service specific issues, and all of that kind of stuff.

So if we start putting dollars to things early in the process, we're going to get really, really muddy. Those are my feelings.

CO-CHAIR STEINWALD: Jim, and then Tom Lee, and then down --

CO-CHAIR LOTZ: Barbara.

CO-CHAIR STEINWALD: Barbara was next? All right, go ahead, Barbara.

DR. RUDOLPH: I had two comments, one being the resource use also. I agree with the idea that it's really a count of resources used, and then later on more -- as you get to the value equation or the efficiency component -- that you might assign dollars or some type of standardized unit of money to it.

Second, somewhere in the white paper it talks about efficiency, and right below it talks about value being weights assigned, the weight preferences assigned to the various components of the efficiency score.

And I guess I have a hard time thinking about efficiency being without a value, in the sense that no matter what you are doing, depending on who's doing it, it's

going to turn out a little bit different because of their own value preferences. I don't know if you can isolate it. Maybe theoretically you can isolate it -- that efficiency is without any type of value assignment by the parties who are creating the efficiency score. But I just have trouble thinking that that is possible. If somebody has an example, I'd be happy to hear it.

CO-CHAIR STEINWALD: So Jim, Tom and Tom.

MR. WEINSTEIN: Yes. I think this is going to keep coming up, this issue of efficiency and cost, because they are not always related. And I guess whether you want to use a golf analogy or baseball analogy, somebody can swing efficiently, hit the ball 350 years, but it's not straight, and it's going to lead to another problem, and eventually they won't get to the right score on the hole or the ball game.

And yet we all understand the

issues of cost. All this is being driven by cost in the sense that people are suggesting we can't afford the health care system we have. So we have sort of agreed to that being on the table. Yet we haven't really demonstrated what efficient care is for almost any condition and then applied the cost scenario to that. And sometimes I worry that we are putting the cart before the horse a little bit, and wonder if we need to do this in phases, to be explicit, to develop measures of efficiency or resource utilization around sort of best practice, with measures of what that looks like, and then applying costs, as they are a differential across systems for all kinds of reasons -- may be a better methodological strategy.

And I'm not sure whether others agree with that or not, but at least I'd like to discuss it.

DR. LEE: I'm actually finding that by forcing us to read this white paper

over and over again, it's been very helpful, because it's made us sharpen our thinking about what we are trying to accomplish. And so what I'm hearing is that we are all seeing that we are not talking about efficiency and value at this stage. We are talking about resource use. We should make it clear that it's a building block that's necessary to get to the phase where we can look at efficiency and value.

And the other point I would make -- which I think is resonant with yours, Barbara -- is that we should say that as we put together that building block we have to understand that there are different perspectives on value and efficiency. And there is the purchaser perspective; there is the delivery system that is trying to improve perspective; and that there are going to be different tools in different settings, like there are going to be individual peer pressure tools that folks like me need that are

probably never going to be useful to a purchaser, because you can't measure things on an individual doctor basis that are valid and useful in that way for a lot of topics.

So I would say, make clear this is a building block and that there are going to be different perspectives, and that we need to have a number of different types of building blocks to prepare for that phase when we get there.

CO-CHAIR STEINWALD: So Tom, Paul and Jeff.

DR. ROSENTHAL: To weigh in on this theme of sort of cost-versus-efficiency, I'd like to suggest that a criteria for picking a cost metric would be that there is a corresponding outcome measurement that already exists. That would be ideal and preferable. But that we may in fact have opportunities to measure cost in areas that we really don't know anything at all or very insubstantial developed quality metrics, and

we should not rule that out. Because we are so nascent in our understanding of how to measure cost and how we should count things that, if we tie ourselves to having a corresponding outcome metric, we may then end up picking cost metrics that are the wrong ones to experiment on.

So I think it should be a criteria in the best of all possible worlds to have a corresponding outcome measurement, but the absence of a corresponding outcome measurement should not dissuade us from picking cost measurements that would head us down a path of learning.

DR. BARNETT: Yes, just getting back to the slide about resource use measures, what I wanted to address is this issue of how we turn our vector of services into dollars. And the white paper, it seems to me, is a little bit -- it talks about prices, and I think that is careless language. Because -- are prices charges? I think it means

reimbursement. And I think it would be a mistake for us to take the existing resource base relative value scale as a given, and that that actually does represent cost from the view of the payer. But we think that maybe that it wasn't based on it. It's based on tradition. It's based on politics. It's not really a measure of the resources or the dollars that are being used. What we'd really like to look at is what is the opportunity cost, what is the long range marginal cost, for each of these services that's used in the health care system.

And there are vendors that are selling these cost allocation systems used by maybe 5 percent of the health care system to try to actively get their costs attributed to the services, and that we shouldn't just take the existing current fee-for-service schedule as a given. I think that's actually part of the problem of why we are inefficient.

CO-CHAIR STEINWALD: Jeff, next.

DR. RICH: Yes. I think generally I like the way the conversation is going right now. And in particular I like what Jim said and what Tom said, and I want to reinforce that I completely agree that the charge of the steering committee is to create a building block for resource use measurement which will lead to us understanding what efficiency is to avoid cost. But to personalize it, having sat and run the Medicare fee-for-service program, I could see and understand geographic variation in cost very easily. I know that if DRG-1 is $300,000, heart failure with L-VADs in New York, and it's $130,000 in other geographic areas, and I understand those differences in the market, and that there are Medicare basket of market indicators that we use to adjust that.

What I didn't understand that is that $300,000 I was paying for that L-VAD to be implanted in New York, were they using the

same basket of resources that somebody did in Idaho, and were they efficient in the way that they delivered care? Because we were adjusting their prices based on their Medicare cost reports and our historical look-back on what it costs them to implant this.

What I would like to see come out of the committee is to avoid cost but to give payers, and particular CMS, a tool where they can look at the care, the cost that they are paying for it, and know that those particular providers are efficient based on resource use measurement and the quality that they are delivering for their care.

CO-CHAIR STEINWALD: Steve and then Bill.

MR. PHILLIPS: Yes. Thanks. I just wanted to follow up, actually, on Tom's point about -- and maybe I didn't understand it completely -- but where you were saying we shouldn't be constrained by whether there are outcomes measures. And I guess I'm trying to

understand and need a little more elaboration in terms of how you can evaluate costs and resource use without some point that you are trying to get to as far as an outcome.

Was I misunderstanding?

DR. ROSENTHAL: No, and I think the example might in fact be the one that was just alluded to about VADs. They are quite good in the NQF domain outcome measures for a variety of clinical areas, but not for 100 percent of them. We were chatting earlier about glaucoma care. There aren't in the public domain perfectly accepted outcome measures. And I don't think we should be constrained -- I think it should be a principle that if in fact a cost measure and an outcome measure are already coexisting, that would be a criteria for picking the cost measure, but I don't think we could learn an awful lot about resource utilization by picking something like VADs or glaucoma care.

And obviously somebody is then

going to have to go back and say, well, know we know how to measure the cost realm of this thing; somebody has now got to do the work to get the outcome measure better understood and in the public domain. That was I think the point I was trying to make.

MR. PHILLIPS: Yes. I guess the kind of missing link is, you can't really come to a point where you say this is a deficiency, because you can measure the cost but you don't know exactly where you are trying to get to until you have the outcomes piece.

DR. ROSENTHAL: I'm assuming that eventually on anything we would pick to understand the cost framework that somebody then will be able to figure out how to measure the outcomes well. I just don't think we should be constrained on the cost side by saying, well, but there are no good outcome measures for VADs -- which I don't believe there really are. I don't think we really know how to say how the outcomes are different

in Idaho versus New York yet. But certainly we could if we decided to use that as the basis for understanding the cost and resource input side of the thing. So that's all I'm saying.

CO-CHAIR STEINWALD: Bill, and then Ethan and then Jack.

DR. RICH: I want to go back to Paul's comment; I'd like to reinforce Dr, O'Neil's point about -- we should try to use some relative scale as often as possible because the variation of costs, even within regions and among payers. And I think we are going to have to have a hybrid system. If you look at Jeff's analogy of the wide variation of costs for VADs versus macular degeneration, since 2002 the office-based inputs have been standardized, okay, and there is no variation in those relative values. You can do -- there is a $100 device in the office or a five-dollar, you get paid the same. So there is no difference to society.

The problem comes in the facility or the use of device and drugs, and that is where we are going to have to have some standardized adopted Medicare payment for the different drugs and devices. Because certainly in the surgical realm, that's where we see massive variation. And yet some things like the physician side of things, in the office or the facility, are very standardized and have been since 2002. They've been updated. But we are going to have to think of a way of looking at cost variation in drugs and devices, and we are going to have to look at a dollar figure somehow.

CO-CHAIR STEINWALD: Ethan and then Jack. Ethan?

DR. HALM: Yes. I'm still trying to get my head around -- and I think I'm hearing from the committee, that the scope of this steering committee -- because I worry at the margins that we are trying to blow the oceans, and we are trying to be the efficiency

steering committee. And I wonder if it would help -- it's hard enough to get the quality metrics right in the poly steering committees, and then there is this resource steering committee. And I wonder from a process standpoint if NQF is thinking that down the road there would be sort of an efficiency use steering committee, and we just need to worry more narrowly about trying to get the resource use piece right rather than -- right now we are so wrapped up into all the problems with resource use, but resource use in a vacuum doesn't make sense, of course, to other people. It only makes sense if you are a payer. But I wonder if we can sort of constrain -- help us think this through. If there is going to be an efficiency steering committee, then we don't have to -- that's their problem.

CO-CHAIR STEINWALD: Jack?

MS. TURBYVILLE: So to respond to the question, the scope of this project is to

focus on the resource use. We do often have other projects that go on. We focus our projects in on things that are manageable for one steering committee. Our hope would be that we would eventually have an efficiency steering committee that is able to examine and evaluate measures that incorporate outcomes, etcetera, in a sophisticated manner. But we think that right now where we are -- and we are using you to make sure that our assumption is correct -- but where the measurement world is right now, where stakeholders are, that the best kind of value that the steering committee and NQF process could be, would be to focus in on the resource use measures as a building block, as you have heard some of the members comment.

CO-CHAIR STEINWALD: Jack.

DR. NEEDLEMAN: Two things -- one on this issue of: do we reduce things in some ways to costs. Whether we call it costs or something else, ultimately I think we have to.

There is no way to understand the tradeoff between using drugs versus psychotherapy if we are talking about depression, unless we put them into some kind of common metric. We've got lots of different resources that are being used, each of which have their own natural measurement: hours for people's time, dollars or prescriptions for drugs, physical units for different kinds of supplies, but if people are making tradeoffs in the use of supplies versus drugs versus people we will not see that unless we produce some common metric of the resource use.

So I think while it's going to be important to look to keep in mind the natural units, we also have to think about whether the resource measures are effectively aggregating them up so we can understand total resource use; so that is one element.

The other element is, as we think about the use of the resource measures, we've got two different directions I think we are

going to wind up going with, both of which are complementary to each other. One is the issue of: are we getting value for money, the whole issue of what is the outcome, how is the quality, how does that compare with the amount of resources that are being devoted. The other -- which is related to that but is its own analytic effort -- is understanding the sources of variation in resource use. And those are two separate uses of the resource measures that we are going to wind up seeing people use, and whatever we are going to wind up endorsing is going to have to enable people to pursue both research agendas down the road.

CO-CHAIR STEINWALD: I want to agree -- it's always a good idea to agree with Jack, just as advice to the rest of the steering committee. I think we use dollars because it's a way of aggregating across different kinds of services. Now how we do that and what adjustments we make to the dollars for differences in what you pay a

nurse in Boston versus Iowa City, those are important details. But we can't get away, I think, from dollars as the common metric for aggregating across different kinds of resources.

And resource use measures also have a denominator, and that's where all the different kinds of resource measures that we've talked about -- specific services or episodes per capita can be a population of patients or a population at large. So we are talking about measures that have both a numerator and a denominator. Where they fall short of being efficiency measures -- and I think this is kind of what Jack was suggesting too -- is that it's not until you bring health or some other outcome measure into the measure that you've really got a measure of efficiency.

So if we are content with concentrating on resource measures, we don't need to worry too much, or at least it's not

within our charge to say to the measure developers, well, you have to tell us how this resource measure contributes to the improvement, maintenance or restoration of health -- for example.

Tom.

DR. ROSENTHAL: And did we agree that cost is not that same as what's paid?

CO-CHAIR STEINWALD: Yes.

DR. ROSENTHAL: Okay.

CO-CHAIR LOTZ: So then if we go back to the specific charge in front of us now, to look at this definition and create a definition that both guides us and the next several pages of the white paper and in the call for measures, what needs to be amended in this definition to reflect our conversation? Is resource use, for instance, a measure then of the costs for various services? Or does it need some structure -- as Bruce suggested, the cost per something -- which becomes the denominator? It seems to me that a lot of

our conversations up to this point have struggled with what the task is in front of us. And having allowed us to voice some of that struggle, how do we bring it altogether now in a definition that is going to guide the rest of our day's conversation?

CO-CHAIR STEINWALD: Another question, I guess, is -- well, I don't know how close we are to a consensus. The staff needs to tell us that. Do you have enough input on this issue to frame what you believe is the steering committee consensus, or do we need more conversation?

MS. TURBYVILLE: I would say no. (Laughter) Because I see some people throwing out the word cost, others thinking to resource use as just the most obvious example. I've heard some say no to dollars, just utilization; others that we understand the weights. So I think it warrants spending the time now to make sure there is some kind of agreement across the steering committee as we

move forward.

CO-CHAIR STEINWALD: This is pretty fundamental, isn't it.

All right. Those who are putting their cards up, you need to have your name facing this direction, otherwise you don't get called on.

All right, Mary Kay, Barbara, Bill Golden and Tom, and Jim.

DR. O'NEIL: I was not meaning to say we not use dollars. What I was saying is, for us to have a broadly valuable process that we can use in a number of different settings, markets, we need to start with counting inputs, resource utilization, and then have a mechanism by which we monetize it in a given setting.

And obviously, to Jack's point, we need to be able to do relative value of different levels of input -- we need to do that. But I think fundamentally our product needs to be able to accurately count inputs if

we're really going to do resource measurement.

CO-CHAIR STEINWALD: But -- but summarize them using dollars as a common metric, once we've got them.

DR. O'NEIL: Well, I mean there's maybe different purposes for that, so we should be able to do both. But I think the fundamental measure is that we can accurately count what resources have been used, I guess, by a delivery system, whether that's a single doc or something else, on a given population.

CO-CHAIR STEINWALD: Okay. Barbara?

DR. RUDOLPH: I guess I was just planning on speaking to the need to say: according to some sort of population or sample of a population. Otherwise it's just kind of hanging out there, you don't know what you are applying it to.

CO-CHAIR STEINWALD: Okay. I think -- Bill Golden.

DR. GOLDEN: Yes, multiple

thoughts. But getting back to Jack's comment, has the staff -- and I have a followup -- has the staff looked at the thinking that went into the whole RBRVs, because that broke resource use down into different components. And I only have two with that kind of thought. Have you looked at that just from the documents, to look at that kind of elements of what goes into the price of a charge?

MS. TURBYVILLE: Yes, we thought about how different resource use units are defined or weighted. But I think our thinking was that some of that would come through, depending on the measure and the perspective. And Jennifer, maybe you want to add to that, but it certainly was part of the thinking or consideration.

MS. PODULKA: We wanted to make sure that the measures -- remember, again, they are not just going to be used by CMS, so many other payers rely on a similar RBRVs system. But we want to ensure flexibility.

We need to potentially endorse measures, but different payers, different purchasers, can use those for different purposes, and it possibly is entirely valid for a payer who negotiates very different payment rates or prices with different providers in their network to continue to carry along that information. So you might provide fewer resources, considering RBRVs, but have them negotiated at a much higher price point, and thereby overall cost more than comparative physicians.

DR. GOLDEN: And the other thought I had was as we go through this, early in the meeting we talked about the world is maybe going away from solo fee-for-service into more aggregate type of mechanisms. And that has sort of a bearing, because as we consolidate or we get into more global kind of payments, you end up not having to deal with as much complexity, because then the local administration has to deal with how to do its

own metrics, and then you can deal with all the uniqueness of Omaha versus Sacramento. And that is not our problem; it's their problem. That's I guess how the DRGs worked in many ways. So as we go through this, do we focus on a disease -- which gets very micro -- or do we focus on patients? So, patients with diseases, and then you can figure out how a system can manage patients with multiple diseases, and you just get costs. So it seems we want to work toward a simplified system rather than getting real detailed, because I think we can get ourselves lost in the weeds very quickly.

CO-CHAIR STEINWALD: Tom, Steve, Jim and then Jack.

Tom.

DR. ROSENTHAL: I actually think there is consensus. And I think Jack described it pretty well. We can go around the table one more time. But it seems to me the consensus is: count first, monetize

second. And if one were trying to sort of tweak the statement that you've got there, it says resource use measures are measures of input, usually in terms of dollars. And I think if you in fact just rewrote that to talk about this concept of count first, monetize second, and that the monetization is a sort of standardized thing that is separate and apart from the markets, you will have captured what I believe is the consensus of the group, unless I'm missing something.

CO-CHAIR STEINWALD: Steve, and Jim and then Jack.

MR. PHILLIPS: Yes. I was just going to put out there that I think this does capture, as has been said, what generally is out there that you can look at as far as measuring resources. I think the value of NQF going through this is to draw upon creativity and innovative thinking beyond these. But I guess the principle I would raise to maybe add to this is just that the challenge, I think,

is going to be that anything that is maybe more creative in terms of measuring resources, can it be consistently applied across provider groups or physician offices, let's say. You may have some of the more sophisticated groups that can do more sophisticated tracking of their costs. But is whatever measure that is being applied something that can be universally or at least widely used? CO-CHAIR STEINWALD: I think that is one of your global principles. Yes.

Okay, I have Jim and then Jack.

MR. WEINSTEIN: I was just thinking to keep this in the context of the value equation so that we don't get off on too many tangents, but quality will be important, at some point those measures will fit into this resource, which I'm going to call the denominator here. And I like the notion of counts and then monetize, as was suggested, when you do diagnoses, because people have specific -- multiple -- diagnoses more often

than not, especially in the Medicare population. Or do you do it in a capitated sort of population base? But I think the context as suggested is important. People who have to manage populations tend to do that much more efficiently when they have fixed dollars than people who don't. So I think that matters.

But I think that if we as a committee could stay within the value equation and just assume that we are going to deal with resources in the denominator around counts and then monetize, then you start to put the pieces together. Because as the subgroups come together around the different domains -- of cardiac or whatever -- they are going to want to understand value in terms of quality. Because they are going to argue, we are doing something really good as measured by X. And so we want to be able to work with them to fit these pieces together at the end of the day, so that is the comment.

CO-CHAIR LOTZ: So just to jump in to the list here, adding some comment about they should be able to use two, linked with quality in some respect, which is missing still from this definition; not something we acknowledge we are going to work on over the next day and a half, over the next couple of months, but clearly should be reflected in the definition as this guide is moving forward.

You're shaking your head?

MR. WEINSTEIN: Yes, I mean you have to leave quality as the numerator however we're going to measure it, just to stay consistent with everybody else's work, IOM, the Congressional language, etcetera, etcetera. To leave quality out would, I think, make the subgroups and the various subspecialties very uneasy if it's all about eventually cost.

CO-CHAIR STEINWALD: Well, I think it depends on what you mean by leaving quality out. My own view is that, when one

looks at performance measures, one -- historically, we've looked at quality measures more than cost measures, and now maybe they're on an equal footing. Many payers evaluate providers on a cost dimension and a quality dimension separately and have separate thresholds that have to be met for each, as opposed to having a composite measure. And I could imagine that our work says something about how our resource measures should be used in combination or extended to incorporate quality. But in and of themselves, not do that.

MR. WEINSTEIN: That's the notion, just that people realize we haven't forgotten about that.

CO-CHAIR STEINWALD: Okay. Jack.

DR. NEEDLEMAN: Some of what we're doing here is inherently complex, and I think we are going to have to carry some of that complexity through our thinking for a while. I'm a little concerned about premature

closure to get clear definitions when in fact we are not dealing with clear concepts.

So we've talked about the prices that are charged for the final service may not be a good measure of resource use; I've got no problem with that. But we are also going to be talking about input prices for the things that are purchased in order to deliver that care, and there we are going to be dealing with the concept of prices.

We've talked about payments not necessarily as a good measure of resource use. But we often standardize across different places by using the same payment rates regardless of -- even as we make adjustments. So there is a standard DRG rate, even though the actual DRG rate paid to hospitals in Manhattan, New York, are different than the amounts we pay to hospitals in Manhattan, Kansas, because of the wage differentials, which are input price differences.

Sometimes when we are trying to

compare resources we use standardized costs to try to get out some of the regional differences. But when we do that, we ignore some of the reasons for differences of resource use. A community in which advanced practice nurses, nurses that are doing basically primary care, are 90 percent of the cost of primary care docs, and where nurse diabetes educators or other diabetes educators are 40 percent of the cost of primary care docs, is going to have a very different mix of primary care physicians, advance practice nurses, and nurse educators than a community where you've got the advanced practice nurses being 70 percent of the cost of primary care docs, and other nurse educators or diabetes educators being 60 percent of the cost of primary care docs.

Those differences in the relative cost of inputs are going to affect the decisions about what the mix of inputs are going to be. And when you apply standardized

pricing to each of those inputs, you lose that.

So I think we're going to wind up, as we go through this process, having to be thinking about what the advantages are of using standardized input prices or standardized costs for comparisons, versus the actual costs for different communities in comparison. And that, as I said -- I think some of the complexity we have in thinking about this, we just are going to have to live with and carry through, rather than trying to reach premature closure on which group to go in.

CO-CHAIR STEINWALD: Fair enough. I've been told it's time for a break. Dolores, why don't you take the last comment, and then we'll come back and be very consensus minded.

MS. YANAGIHARA: Actually, Jack just touched on what I was going to talk about. But I feel uncomfortable saying that

we should take actual payments made off the table, because I think there is definitely a place for that. And so I'm fine with saying count first, monetize second. Two ways to monetize: standardized pricing, actual pricing. But I think there is a place for each of those, and to say we are not going to go into actual pricing at all just doesn't make sense, because there are times when that is really important.

So to acknowledge there are two different ways, but to not completely take it off the table.

CO-CHAIR STEINWALD: Okay, how much time do we have? Fifteen minutes. All right, this will be an 11 - 15 minute break.

(Whereupon the above entitled matter went off the record at 10:34 a.m. and resumed at 10:50 a.m.)

MS. TURBYVILLE: We'll get started again. At 12:10 we do open this call up for public comment and questions. So I

want to get us started up again so we can finish what we hope to get done this morning, if at all possible. I think we've made good progress so far. So I'm going to go ahead and hand it back over to Doris and Bruce.

RESOURCE USE MEASURE EVALUATION CRITERIA

CO-CHAIR STEINWALD: Well, the consensus out during the break is that we are close to a consensus.

(Laughter)

So Sally has put something up for us to look at, so why don't we take a look. Then if we want to comment on it positively or negatively, please do.

MS. TURBYVILLE: Just to note, I didn't try and change the actual definition too much in that short amount of time. What I did try to do is capture the approaches that we will then want to make sure the definition clearly expresses. But I think we have a few minutes; if there are some wordsmithing suggestions, that's completely welcome.

MR. MARKUS-HODIN: Could you enlarge this slide a little bit?

MS. TURBYVILLE: Oh, sure.

MR. MARKUS-HODIN: Thank you.

CO-CHAIR LOTZ: And we can, as was suggested, move forward in our conversation and then revisit this later on, revisit it over the next couple of days through email.

CO-CHAIR STEINWALD: Mary Kay.

DR. O'NEIL: Well, I like this, and I think the resource inputs as the numerator, the population -- whether it be a geographic population or a single physician's population -- as the denominator, makes perfect sense, but the product has to be the value of the health outcome. Right? The input to the population leads to a result, and that result is better or worse health for value, right? So that's how we're relating -- this is a building block of the equation for value.

CO-CHAIR STEINWALD: A building block?

DR. O'NEIL: Yes.

CO-CHAIR STEINWALD: Dolores, you still have your card up? No?

Okay, anyone else? Go ahead.

DR. RICH: I think I was just reflecting on what Jack said earlier, that some accounts of resources I would see as inputs, and some inputs are already monetized, for instance, wages. So there may be a gray zone between one and two, where some resources are already monetized. And we'll have to deal with how you demonetize something. Do you take the cost of labor and turn it into an FDE? And is that the resource that we're going to look at, number one. And then monetize it in order to compare across providers.

CO-CHAIR STEINWALD: Tom.

DR. LEE: The truth is, I think we need them all. I think that -- in my

organization, which is trying to improve but also dealing with the marketplace, being criticized in the marketplace, we basically are using all of the measures on the spectrum of measures that are described later on in the white paper, and we use them monetized, non-monetized. So I think there are different customers for all of the different frameworks that we might consider here. And I think actually one of the best contributions we can make is help people understand which structures are most useful and appropriate for which settings.

CO-CHAIR STEINWALD: Based on that, would you want to modify the number two, and monetize as appropriate or -- ?

DR. LEE: Well, that is a good point, because there are different approaches to monetize, what you adjust for -- what you adjust for like, do you adjust for GME and teaching hospital status? People at teaching hospitals say, yes, of course you should.

Other people say, we don't care. And both perspectives are understandable.

CO-CHAIR STEINWALD: Right, we don't care because you cost too much.

Anyone else? Oh yes, Ann, sorry, and then Paul.

MS. HENDRICH: I just have a question about number one, which is the count of resources. If it assumed that that is across the environments of where care is provided? Because currently I think that's one of our greatest challenges -- that it's in a very defined portion or episode of illness within that continuum, and we really can't get to number two and the outcomes, as we've been talking about. So that counter resources definition, how are we viewing that?

CO-CHAIR STEINWALD: Barbara, you were an advocate for that. Go ahead.

DR. RUDOLPH: Yes. I wasn't sure in the paper whether or not it was about -- all of this was about within an episode, or

whether it was more broadly aimed. And if it's within an episode, then he could say resource use within an episode. Otherwise, then somewhere along the line you define who the population is and the time period covered. But I'm not sure how to fix that.

CO-CHAIR STEINWALD: Well, I'm for broad-based, personally. And I think the way we've got it there, it seems to me, allows for a lot of different applications of the concept of resource measurement. And the hell part of it that you are referring to, Ann, I think is down the road a bit, or up to the measure developers: if you want to use the measure for this purpose, here's how. At this stage, I think we're at a higher level than the hell.

Jim.

MR. WEINSTEIN: The only thing I would add is, I think we all would say the people are going to do the measures, et cetera, what are they going to adjust for?

That will be a big question. And so just "with appropriate adjustments" somewhere in this will be important.

CO-CHAIR STEINWALD: Well, the monetizing -- the adjustment can be, it seems to me, at any level. But why don't others contribute? Jack and then Tom.

MR. BOWHAN: The way it's written now -- with "count and monetize resources" -- I don't know if there is supposed to be an "or" in there. And to Tom Lee's point, this is supposed to be kind of high level, right? And we haven't figured out what the best way to go is, whether it's counting units or it's dollars. So can there be an "or" in there -- that the resource measures are inputs of dollars or units of service? So we cover the options that people want to go down with building measures, rather than saying it has to be one way or the other.

CO-CHAIR STEINWALD: Tom.

DR. ROSENTHAL: I'm assuming also

that there are going to be a variety of other principles added to this, like attribution and risk adjustment, and a variety of things that we do not have to capture in this definition.

And back to your point, Bruce -- it seems at this stage we could continue probably to sit here all day and wordsmith this thing, but maybe to no particular -- the marginal utility of improvement here is going to decline pretty rapidly.

CO-CHAIR STEINWALD: Steve.

MR. PHILLIPS: I hope this isn't a decline. (Laughter) Or a wordsmith. But maybe more of a conceptual thought. I guess when I think about what we're trying to get at here, there are situations where maybe it's not so much of an input count or cost as, say, you've got a piece of expensive equipment that is sitting there, otherwise would be unused, maybe to test the use of it is not that necessary. But the real resource use is in the payment that's made, for what may

otherwise be a rather low marginal cost. So I don't really see that fitting in there, but I just throw it out there -- that it's not necessarily an input in cost or count, but it's a system cost, as far as the payment is made.

DR. HALM: Sort of a capital expenditure.

MR. PHILLIPS: Well, the capital expense may actually be minimal. So your input cost is minimal, but there is a payment that's made for a test that maybe doesn't have significant marginal benefit.

CO-CHAIR STEINWALD: David. Do you want to introduce yourself?

DR. PENSON: Sure, my name is David Penson. I'm a urologist by training. I am at Vanderbilt and head our Center for Surgical Quality and Outcomes Research. And I'm sorry I was late.

My comment is just, I don't see counting the resources and monetizing the

resources - I think we are building consensus here - as mutually exclusive. And since we are trying to be at this sort of very high level overview the question is when would you use one or when would you use the other. And it strikes me that what we are really trying to do here is come up with measures that are broadly applicable across populations. And maybe that's the verbiage we need to add. I know we are wordsmithing, but I think that people would prefer to have counts, because costs are different between say Nashville and New York. But the fact of the matter is you want to be able to make comparisons. So maybe the idea here is to wordsmith it in such a way that it basically says to make it broadly comparable across populations, if that resonates with people.

CO-CHAIR STEINWALD: Any resonating in the room here? Since there does seem to be a consensus that you don't always need to monetize then maybe you'd add a little

bit of language that says monetize it if necessary or desirable; something like that. And then add the broad usage concept, and we might be done. What do you think?

I don't know that we need all of our refinement right at this moment.

MS. TURBYVILLE: No, but I'm just trying to make sure I capture some of this final silence of resonation.

CO-CHAIR STEINWALD: All right.

Tom has something.

DR. ROSENTHAL: I just want to go back to the note for a second. Because I agree as a principle the resource use is a building block to bring us closer to value or efficiency, but that we may select measures that don't yet have a significant quality component that we currently understand to be developed later. So I don't want this thing to be again implying that we can only select cost measures for which there are existing well developed quality measures.

MS. TURBYVILLE: And we will look at all the principles together. So this is just to note that there is this principle, an opportunity to have the outcome portion, and then also as the group discusses it whether or not to add this other idea to those principles.

CO-CHAIR STEINWALD: Are we good to go? Okay, let's move on.

MS. TURBYVILLE: Let me hit save.

(Laughter)

Okay, so I think the next portion of the discussion we were hoping to have the steering committee focus on was what is section two, I believe, or three, which is looking at the real world implications. We also talk about other NQF efforts that are ongoing to help drive home this idea that we realize that we are just focusing on one portion of some very important things that need to be grappled with.

And then kind of this conceptual

model that we put forth about the approaches to resource use measurement and what they span, and I think it makes sense to talk about this in the context of what we've heard from the definition discussion here today as well. So clearly all of section two and any tie-backs et cetera, are open to comment. But this is an area that we really wanted to make sure the steering committee felt comfortable with our final model, what do we need to add to it, delete, et cetera.

CO-CHAIR STEINWALD: Comments. Ethan? I'm sorry, are you going to make your first?

CO-CHAIR LOTZ: Well, you had wanted me to drive that section, so -

CO-CHAIR STEINWALD: Oh, yes, yes.

CO-CHAIR LOTZ: But you can drive it.

CO-CHAIR STEINWALD: No, no.

CO-CHAIR LOTZ: You're on a roll.

CO-CHAIR STEINWALD: No, I'll

listen this time.

CO-CHAIR LOTZ: All right. I'm shorter though, so that might be a little bit of a challenge.

DR. HALM: So in thinking about the continuum, in the document there I'm wondering if there is a piece of the continuum between per capita and per episode that we may want to articulate in the document? People talk about diabetes care for a year or health care for diabetics for a year. And I often think per capita is the big picture smoothie of Medicare spending per beneficiary in a year. And so I'm wondering if other people sort of think where that kind of construct of diabetes care for a year, or care of a diabetic, for a year, fits in that continuum?

CO-CHAIR LOTZ: Some of the background reading looked at it from an acute and chronic point of view. But I'll just throw that out there. Bill, you're next. Bill, you didn't turn your table tent, but

fortunately I know you.

MS. TURBYVILLE: And just quickly, apparently some people on the phone are having a hard time hearing us, so if you can remember to speak as close to the microphone or closer to the microphone that would be helpful. Sorry for the interruption.

DR. GOLDEN: It just helps me as we go through these technical documents, and your eyes start crossing, I'd like to hear a very succinct distinction between per patient and per capita, a little example.

CO-CHAIR STEINWALD: I'll offer one. I think we're going to run into how the terminology in the field has developed. Per capita as I understand it in the Medicare context was a term that was used to contrast with episode where a lot of the investment had already been made, and for the most part it refers to a population of patients, whereas we're also interested in per capita being a population at large or a population served by

a health plan, or the population of Medicare beneficiaries. So there are two kinds of per capita I think that are of interest. And whether we need to distinguish them with different terms or not I'm not sure.

DR. GOLDEN: I think that the document will be strengthened for other readers by having a little box distinguishing the two. Because I think it's not clear, it's still not clear, when you go over that. Because you start saying, per capita, you start talking populations and it gets real blurry for me. I'm thinking one is a disease state and the other is more global in terms of multiple disease states. Is that correct, or am I wrong there?

CO-CHAIR STEINWALD: Line 277.

DR. GOLDEN: Two seventy seven? For example per patient measures may be the best choice. Physicians performance for the patients. Again, I'm having a hard time distinguishing whether it's a disease state

versus a population within a practice. I'm not sure.

DR. BARNETT: So what I read that to mean was per capita was per covered life, and per patient was per covered life that actually came to see the doctor that year, and that was the jargon that was in the document.

CO-CHAIR LOTZ: Barbara.

DR. RUDOLPH: I just want to remind everyone that there are going to be other users. For example a number of the states are now beginning to collect all payer data systems that cross plans, cross Medicare, cross Medicaid, et cetera. And that we have to really think about not just thinking about specific plan populations or specific payer populations, but across the population as well.

CO-CHAIR STEINWALD: I think Bill wants a return.

CO-CHAIR LOTZ: Bill, are you up again?

DR. GOLDEN: The other question I have is on lines 253 to 255. You have an interesting statement that says, relying on either measure alone could mask differences between physicians and even allow gaming by affecting the physician practice pattern, such that it generates more discrete episodes to appear more efficient on a per episode basis.

Help me out with that. I would think that if you built the concept of time into the concept of an episode you would avoid that problem. So I'm a little confused as to why episodes would necessarily make somebody look more efficient if he has multiple episodes, if you have time built into it.

CO-CHAIR STEINWALD: The back story here, as I understand it, and CMS, you can contribute if you like, this derives from some work that MedPAC did on resource measurement in episodes where they were contrasting different metropolitan areas. And their initial measures showed Miami to be

significantly more efficient if that is the right word than Minneapolis. And the second level of analysis was, well, that's not possible, so what's wrong with our analysis. And it turned out there were many, many more trivial episodes in Miami than there were in Minneapolis. So it was partially a recordkeeping and coding phenomenon that led MedPAC to decide that putting all of one's eggs in the episode basket was not a good idea because it could lead to misleading results.

And that's another reason I think that CMS, while the Medicare improvements, MIPPA, required CMS to develop physician feedback measures based both on episode measurement and what we are calling per capita measurement, in part because of the growing belief that either approach could contribute information that would supplement the other, and provide - and in fact -- I know I'm jumping around, but the document that the staff sent us on CMS' report on their early

experience with physician feedback has both per capita and episode measures constructed in a way that it looks like both contribute to the profile of the physician in a way that neither by itself would accomplish.

DR. GOLDEN: But a followup, was that a problem though of the notion of using episodes or in the definition of an episode, because I was on a committee with Jim that we talked about what is an episode. And unless you expand the timeframe of the episode, you might conclude that, I don't know, unless you have multiple different diseases.

DR. LEE: From the cardiology examples I know well is that in Miami, where they have a low threshold for catheterizing people, that leads to creation of a tremendous number of episodes of unstable ischemic heart disease, of fairly straightforward patients who are basically in and out very quickly. So it's low cost per episode. But they just have many, many more episodes than anywhere else in

the country, or many other places in the country.

CO-CHAIR LOTZ: Jeff.

DR. JEFFREY RICH: I agree, and another example is heart failure. So if you are looking at the cost for heart failure as we did, it's huge. So to become more efficient you discharge your patients earlier but admit them more often. So if you look at the total cost of heart failure over a year - so the point here is you need to have a time element. And this is just an example of racing too fast and you get beyond where you want to be.

CO-CHAIR LOTZ: Ethan.

DR. HALM: So between per capita and per episode can we add per condition. Because you have per procedure at the end, but per condition would capture some of the chronic disease care that people are talking about now.

DR. LEE: And it's not just

disease, but it's also like the high risk population, like the dual eligible population which are in PMPM arrangements in many populations. So there is something in between.

CO-CHAIR LOTZ: Mary Kay.

DR. O'NEIL: I think some of the problems - maybe I'm not understanding this correctly - are if you are counting by how many procedures are done per patient with a condition, versus how many times a given physician or health care system does the procedure to people with that diagnosis, we're going to come up with somewhat different numbers, right?

CO-CHAIR STEINWALD: But part of the problem there is that the frequency of the diagnosis is not independent of the propensity to treat.

Well, guidance to staff, it sounds like some elaboration of this continuum would be helpful. Do we need the per

hospitalization there as a separate element on that. It seems to me that is covered by per procedure - in my mind it is at least.

DR. O'NEIL: If you define a hospitalization as a procedure, but for medical conditions that's the part that gets muddy and has the highest degree of variability.

CO-CHAIR LOTZ: I think what you are looking for is per some sort of an event that is smaller than an episode, or some building block that perhaps leads into a episode.

Tom.

DR. ROSENTHAL: Well, these also get into to whom you are attributing these things. So there is this cross intersection with attribution that it may be at the end of the day that there will be hospital-oriented kinds of things that will want attribution, and then the per hospitalization thing will make sense. In the absence of that it may not

make sense.

CO-CHAIR LOTZ: Jeff.

DR. JEFFREY RICH: That's where I was confused was, what's the difference between per episode and per hospitalization? I always think of episodes of care at least in cardiovascular disease as hospitalizations. And so is hospitalization part of an episode, or is an episode part of a hospitalization?

CO-CHAIR LOTZ: Bill.

DR. GOLDEN: I'm a little concerned, and again, I'm going to flip it back at Jim also, because the committee - there was a previous efficiency steering committee. We got it, and there's a white paper that NQF published that broadens the definition of episode considerably from what Jeff just said. And it talks about timeframes, and getting away from acute events. And I think we're in conflict with what was developed in the previous workgroup.

MR. WEINSTEIN: I also think the

Institute of Medicine did the same thing with some other language around the value equation, quality over cost over time, and to save that longitudinality. And I think the notion is there are a lot of procedures that are hospitalized today, more and more even in cardiac. So I don't know if that captures it. But this issue of time is important, and I think in the text here at least we put diabetes as a one year for the episode potentially. I mean in that time period there are going to be a lot of measures and utilization of resources. And is that an episode? Is that per episode? Is that per procedure? Is it per capita? I mean all of those things are sort of a continuum, right? That's what I thought this line was, was potentially a continuum of services. I didn't think of it as all different. Because you are looking at per capita expenditures. For somebody who may have had one procedure, or one hospitalization, or multiple

hospitalizations, or multiple episodes for different diseases.

And then there is a per capita spending on all of that for that person over time. Is that not what you mean by this?

MS. TURBYVILLE: I think that is right. We are trying to show the breadth of how resource use measures may present themselves for the endorsement process, but not to say that a particular patient isn't going to be within a per capita, or that you are not going to look at episodes. I think we were actually trying to be more broad than narrow in how we are bucketing them.

MR. WEINSTEIN: And the efficiency group that David and I were on, you actually sent that paper out as well, which sort of captured some of this language. So I agree it should be similar.

CO-CHAIR STEINWALD: I think if the message here is that the - that resources measures can be broad based, they can be very

discrete and narrow. They can at the other end be oriented to the care of entire populations, then what we put on the bar doesn't matter that much. And whatever we put there, actually it should be consistent with what NQF has developed in the past.

CO-CHAIR LOTZ: Now is this intended to have the resource use measure developers assign themselves into one of these buckets? Or is this purely for conceptualization for the reader of this white paper?

MS. TURBYVILLE: I think in the end we are going to need to know what kind of resource use measure they intend for us to evaluate. But potentially - and Jennifer, do you want to add something to that?

MS. PODULKA: Well, I think it's going to be helpful to give a signal to the measure developers as to what we might be looking for. It's going to be helpful to the steering committee as well, once you have

those measures submitted and you are considering issues like best in class. A per capita measure at one end of the continuum wouldn't necessarily compete with or crowd out a per procedure measure at the other end of the continuum. So you might want to sort the measures you get into buckets, according to the continuum, and then compare them just within the buckets. Because you might want measures from different buckets.

CO-CHAIR LOTZ: Mary Kay. Some of these table tents are still up from before. Jeff, is yours up new? Or you just want to continually be up, is that what you're -

DR. JEFFREY RICH: No, I put it down, and I put it back up. So all the way to the right is per procedure, also per encounter? I mean procedure sounds to me like a visit to a physician, or a surgeon. But an encounter? Is that where a visit to your family doctor would be, at the very far right? I don't see where on a continuum that would

go.

MS. TURBYVILLE: In my experience it's often part of the definition of a measure that they submit, like a patient must have had an encounter. But if there were a measure like that it would fit somewhere on that continuum. And I think you're right, but I'm having a hard time thinking of an example. You mean just a count of encounters?

(Simultaneous speakers.)

DR. JEFFREY RICH: Just going to your gynecologist for your annual. That's an encounter, isn't it?

CO-CHAIR LOTZ: Paul.

DR. BARNETT: Just trying to understand the context of this, we have in the white paper this thing about applying the unitive measurement and unitive measurement. And it was a little bit hard to understand. In the interest of transparency might be worth thinking about, talking about services and - it was previously the numerator, previously

the denominator. And I found that a little - the numerator and denominator actually I understood. (Laughter) But it is sort of like services - and these are sort of what we're talking about, are they outputs? I'm just thinking that we need to have a little bit better jargon to make it easy for people to understand. So I don't know if these are approaches for measuring resource use as much as they are the buckets that we put the services into, the denominator.

CO-CHAIR LOTZ: Bill.

DR. GOLDEN: Instead of asking a question, I was going to say, I would hope that our steering committee could make comments on preferred measures that we would like to receive. And I think as we go toward the right side of that slide, we are getting into the weeds. And frankly, I don't think, given the past work the NQF has done, I don't think we really want to encourage cost per cath, or cost per ER visit. I think we really

want to get more measures looking at broader measurement and management of conditions. And I would be reluctant to endorse multiple buckets that get into a lot of these very small, very narrow events. I think we'd end up spending a lot of time with measures that are not that interesting and potentially not that useful.

CO-CHAIR LOTZ: Tom.

DR. ROSENTHAL: Well, I think generally I would agree that there is value being more to the left side broadly. But I think we miss real opportunities if we don't leave the whole spectrum on there. Because I don't think the state of the art is such that we really know. And I actually think if we had some measurement that said the cost of cath here and the cost of cath there, at some really prescribed level that we understood well, I mean we have all kinds of innovation projects going on around the country on bundled pricing. I don't think we have a clue

yet really how that is going to shake out. So I don't think we should sort of chop off the right-hand side of that thing, even though I take the point about the broader virtue of per capita measures and global measures and the value. The problem with the per capita measures is that the risk-adjusting methodology becomes so much more challenging and difficult to get your hands around, and will be divisive, and we'll be fighting over it. Whereas on the right-hand side you really can get down to a very prescribed level and it's harder to quibble over what you get at the end of the day.

So I think they are complementary. I don't think that we should sort of say, let's carve off one side or the other.

CO-CHAIR LOTZ: Bill Rich.

DR. WILLIAM RICH: I agree with Tom. Because if large - ideally we should be on the left, I agree with that Bill, but the reality is some of the more clearcut measures

that will address variation can be procedure specific. And there is some suggestion that these can be extrapolated across that physician's provision of more complex services.

So I agree with that, and some of the work that has been done shows huge variation, 100 percent attribution, very clearcut things on very powerful measures that have a great deal of financial impact. So I would not like to see us get rid of those, since we avoid a lot of the problems of attribution.

CO-CHAIR LOTZ: Thomas Lee.

DR. LEE: It's great watching the eye surgeon keep knocking over his water. (Laughter) But that is in the numerator part. But I think to make a comment that directly goes against the interest of my organization, because we do get paid better than our competitors, and we get criticized a lot for it, the truth is most of the world is not

organized enough for the left side as yet. And on the right side, it's actually very complicated to get into prices. We've been trying in Massachusetts, part of our health care reform effort, and of course every payer pays different for a CT scan, for an office visit, and so on. So what in Massachusetts we've done, we've actually organized baskets of service for what you would expect a diabetic to get in the course of a year. And we've said, based upon the median of the fees paid to Partners Health Care and Beth Israel and so on, this is what that basket would cost you for a diabetic; for a routine hypertensive; and so on.

We are not even looking at the question of how efficient we actually are taking care of hypertensives and diabetics, but it's just a basket service, it's two visits, eye exam, that kind of that thing. What would it cost if you got your care within the system? And the truth is, it does put

some healthy pressure on providers to have that out there. Really it's just price. It doesn't reflect what is actually happening to the patients. It's just given the prices in the contracts, what would it cost for a typical basket of services. Because it's not that helpful to say, here are the prices for CT, echo, and that kind of thing.

CO-CHAIR LOTZ: Jack.

DR. NEEDLEMAN: Two things. One, Tom's discussion of looking at the median prices and then applying them to a bundle of goods is a classic example of what we were talking about earlier about the issue of standardized prices versus nonstandardized prices. And there is value for some comparisons in having standardization, and other value for other reasons in not having the standardization. So it's not either/or, it's and/both. And that is going to be an important component of all this work.

As I look at the continuum here,

the most ambiguous part is the per episode. And I read the original, the earlier NQF report on efficiency measures, and I've read what's here, and I've looked at the definitions of episodes. And they are remarkably ambiguous. And they have, as has been noted, very important consequences for estimating who's high, who's low, what's being included or not.

So I think - I don't think the answer to what the right way to do this is necessarily in the documents yet that we have received. And I think one of the issues for looking at what measures are more effective than others is in fact going to turn on how thoughtfully the episode definitions are in fact defined and implemented in the measures to capture the gaming, to capture the aggregation, but also to capture the challenges we have in both attribution and in doing comparisons that make sense.

CO-CHAIR LOTZ: Jeff.

DR. JEFFREY RICH: I just wanted to reinforce what Tom and Bill had said. From a very practical and functional standpoint the payment systems live to the right, per hospitalization per episode. And if this committee is going to have any value to society and to payers and to others, then we are going to create measures that begin there, that are living and dynamic and can move with the payment systems as they move to the left. And that will be our challenge, to pick the right measures that don't stay to the right, but can move to the left.

CO-CHAIR LOTZ: I think as well is where CMS was the sponsor of this steering committee and this effort, that they were looking for specific measures in all three areas - per procedure, per service, per capita, and also per episode. So we can never get too far away from who's paying the bill and what do they expect for that payment from this project. Not to limit ourselves to that,

but I don't think we can get too far away from that. If you want to take something out of the equation as not being doable right now it certainly has to float back to CMS to say, listen, we don't think there is any value in doing this; here's the reason why. And allow them to comment further.

So please bear that in mind as you are thinking about this.

Final comments before we move on? Yes, go ahead, Tom.

DR. ROSENTHAL: Just to the last comment you made. I'm all for leaving these all on here, even though I think we'll find interesting questions about each one of those things. But based on your last comment, we are not obligated to take a measurement in each one of those realms, I assume, despite CMS' imprimatur. Because I share with Jack some of the concerns about the episode basis. I'm not sure the science is where it needs to be right now to pick one of those. But we're

okay with this being broadly permissive as Sally said, but we are not obligated to take one from each area necessarily.

CO-CHAIR LOTZ: And we may not receive one from each area either. So Jeff.

DR. RICH: I'm sorry, I just wanted to clarify my comments. I didn't meant that we should stay to the right. We should be across the spectrum. But we need to be attentive to the right as well, because that's where a lot of the payment systems live and breathe right now.

CO-CHAIR LOTZ: Yes, and there is another directive in our vast amounts of background reading, most of which I have here in front of me in case anyone wants to refresh, is this idea to be actionable as well, and accountable. And I think that it's hard sometimes, depending on who you are to think about accountability at a per capita level, you really do have to get down to something that is a little bit more smaller

unit driven.

Jim.

MR. WEINSTEIN: I guess the question is, to me, one of the other committees, I think it was nebulous for good reasons. I don't think they tried to get into the weeds on every diagnosis. But the notion that there was a population at risk that had a particular diagnosis and then had some followup. But the issue of appropriateness which is where we all sort of fall off, and lots of people have studied this issue, but measures of resource use for things that patients don't want if they were well informed are a problem. And I guess the first committee discussed that a lot, that assuming we are going to use measures that when patients were well informed would actually want what's being offered to them, I don't know where that comes into this definition independent of per capita or procedure. It involves both. Because I think we know from

our own work that over 30-plus percent of people when well informed wouldn't choose what they've gotten in health care. So I don't care how you measure it, but I'd like to see that as part of the measurement of this strategy which goes along with the previous committee.

CO-CHAIR STEINWALD: Let me comment. In a previous conversation I think this was Jack Needleman raised the distinction between technical and allocative efficiency, and not wanting to go into it too deeply, but if you are talking technical measures you can measure the resource use or even the efficiency of performing a procedure whether it's needed or not, right, which is kind of what you're getting at. Whereas if your broader based measure needs to bring in the contribution to health that the procedure once performed makes, and that gets at the problem that you are raising.

So we talked earlier about the

building block concept that we are confining ourselves to resource measures but with the understanding that those measures are intended to be a building block that gets you further down the road at gauging both quality and outcome.

MR. WEINSTEIN: Well, is not a resource measure whether you engage the patient in the decision making? Right now that's paid for very differently, depending on the payers. But it's a resource use that's not traditionally there that affects the utilization a lot.

CO-CHAIR STEINWALD: I think the patient engagement really gets at the concept of health. To me it's, the patient would not undergo 30 percent of procedures if they fully understood the value of the procedure in contributing to their own health. To me the patient preference is indistinguishable from the contribution of the procedure to human health.

MR. WEINSTEIN: I'm just arguing that the resource - it's an important point to me personally, obviously. I don't have any stock or anything, but I think the notion of if we are going to create resource use measures, I'd like to know that the patient involvement as a resource in some active engagement was a resource, part of the measurement. If they're not engaged, that's interesting. In fact we end up spending more money when they're not engaged.

CO-CHAIR LOTZ: Helen is going to trump Ethan and Paul, so hang in there.

DR. BURSTIN: Oh, I didn't mean to trump anybody. I just pulled up the efficiency report, because I think it's actually really useful. And if nothing else probably should just hand out a hard copy of the executive summary; it's just a few pages. Just wanted to make the point that the committee very explicitly said there were domains for performance measurement to get at

the patient-focused episode, the broader view of episodes. And they clearly laid them out as health outcomes important to patients, which includes health status, healthy quality of life, patient experience, cost and resource use as well as the more classic quality measures.

So I think what really still gets back to is this issue of building blocks. I still think we are seeing, we still need the domain of the cost and resource use measures that would then get packaged together into a broad-based view of an episode.

But I think we are still in the case of today at least trying to focus in on again trying not to boil the ocean as somebody said earlier, the piece about cost and resource use, understanding that when you look at the patient-focused episode, our hope is you will pull in the patient experience of care, you will pull in patient preferences, you will pull in outcomes. But what we are

talking about today is the cost and resource use piece really in a more narrow way.

DR. BARNETT: So just to follow up on what Jim just said, so there are important things that need to be done in the health care system that are either inadequately reimbursed or not reimbursed at all. And I'm just still stuck on that last point. It says, if we use the reimbursement schedule to say what things cost, then we're missing the boat.

CO-CHAIR LOTZ: Ethan.

DR. HALM: I don't mean to playing like the 12th angry man role here, but I think, I like the picture and the continuum I think will be extremely helpful because it will help ground and anchor sort of things that are out there, and then things we hope to see happen. And especially moving towards the left. I think we have a little more work to do to make that visual more representative of the discussions. So to me semantically per

episode suggests something happening. So I'm still arguing that we need to come up with something visually to put between per capita and per episode. Because the words say per capita measures which are total health care spending per person. It doesn't say total diabetes cost per year, or think about mental health carve out. So you have the costs of depression for a year, and then you have an acute episode of acute depression involving hospitalization or ED visits and other stuff. And just the semantics of a per episode, we are not going to want to use per episode to mean, well, an episode could be a year of acute and chronic care altogether. Because all the technicalities, and how you define the episode as looking for sort of quiet periods before and after, and this and that, and I think maybe we can do this more offline, but that is going to be - we've already heard several people talking about, people are actively thinking about measures that are in

between that per capita and per episode part of the spectrum, and coming up with some language to describe that on the diagram I think would be really helpful.

CO-CHAIR LOTZ: Lisa.

MS. GRABERT: I like the spectrum of different measures, and I appreciate that CMS may want to have different measures from each of those buckets.

I have a question in terms of the call for measures that goes out, and whether or not there is going to be an intent to focus on just one of those buckets, or if the intent would be to determine best in class from a per episode measure versus a per procedure measure.

MS. TURBYVILLE: So when we think about best in class we think about selecting measures that are measuring the same thing. So theoretically I would assume that a per capita measure and a per procedure measure are measuring different things as measures;

not that they are not potentially looking at similar types of costs. But it is something to think about.

And then further the steering committee will help us frame and inform that call for measures. So if as we think about how broad to be during a first call for measures, perhaps we go across the whole spectrum and see what we get, and then maybe future efforts would be more narrow. I don't know. So this meeting here today and tomorrow will help us frame that call for measures more concretely. So that is right on question.

MS. GRABERT: I think that is helpful. Then I'll express my strong preference for a per episode scope, because I think that that is where we need the most help right now overall for measures. We certainly have per procedure measures. We certainly have handfuls of per hospital plus days and readmissions measures. I do think that we need a lot of help with per episode now.

CO-CHAIR LOTZ: Mary Kay.

DR. O'NEIL: I was just thinking about the - and this may be semantic as much as anything - but we're kind of looking for an annualized care per individual type measure, whether somebody has a chronic condition. I from a kind of business perspective consider what should be the run rate for a diabetic, what should be a run rate for an asthmatic, what kind of care is the proper level of care, and even the proper level of care for somebody who is well.

But when per capita is utilized I think most people think of measuring populations. They don't think of looking at what is the proper allocation of care for an individual through a timeframe. And that seems maybe the one that is missing between per capita and per episode. I understand per episode can be defined to include that. But I think it's not easily understood to mean that.

CO-CHAIR LOTZ: Jeptha.

MR. CURTIS: So I just want to follow up on Lisa's point. Where I think this is useful is as a construct for understanding the full spectrum of different ways that you can measure resource use. But what is missing I think, what particularly measure developers are going to need is sort of a de facto judgment over which are preferred and which are not preferred. And that's missing at least from my reading of the text currently. And so you may end up with, and I know we'll craft a call for measures later on, but it's easy to say that I think a heart failure should be measured as an episode of care over not just the acute hospitalization; and that is in fact superior to a pure procedural or per hospital measure.

And I think we can make those judgments, and to the extent we can now it will be better in terms of setting it up for actually judgments, measures, as they come

forward.

CO-CHAIR LOTZ: Bill.

DR. GOLDEN: I think this is a very important section, and it's not an easy section to write. So I have my sympathies. It is a little wonky. And I think if we have other readers - if we are having trouble teasing it apart. I think that a lot of readers - this will be read a lot I'm sure by many people - I think it'd be really useful to have a box of just examples of what we are talking about, concrete examples. Otherwise we're going to I think lose some folks. I think it's a good start, but we need a Readers Digest version and some examples.

CO-CHAIR LOTZ: Tom, you don't have your table tent, Thomas Lee, up, but you just drew a box that you shared with Helen, and Helen is nodding and there is a conversation going on down here. So why don't you share your conceptualization in a box?

DR. LEE: This is like my sort of

line about we need all these things. So my table is like those four columns - per capita, per episode, per hospitalization, per procedure. Although I agree with Ethan that there are things in between those categories, but then the rows I have are unadjusted dollars, like what purchasers value. The second row is adjusted dollars, adjusting for pair missed wage inputs, and people will disagree what you adjust for, but it's for policymakers, and then the third row is nonmonetized utilization data which is what providers need for actual improvement.

And the truth is, I actually think we need things in all 12 of those cells.

CO-CHAIR LOTZ: Could you add a row to deal with the patient perspective in some way or other, and then maybe we've got everyone's focus.

DR. LEE: My assumption is that patients care most about the first row which is unadjusted dollars. They may be shielded

right now from the unadjusted dollars. But ultimately I think they are bearing more and more of it.

CO-CHAIR LOTZ: Additional comments before we close out this section? Jim.

MR. WEINSTEIN: I guess I'm probably not being very clear about what I'm trying to suggest as a resource. There is a unit of service that requires patients to engage like a laboratory around decision making that is currently not paid for that is not in the system. And because it's not there we are spending 30 percent more than we need to potentially. That is just a potential.

How do we account for those kinds of resources, independent of the one I'm suggesting, patient decision tools?

CO-CHAIR LOTZ: Jack.

DR. NEEDLEMAN: If I understand what Jim is saying, let me try to add some concreteness to it. So we've got patients

with prostate cancer who are trying to decide what treatment to take. And some practices have hired people - some practices rely on the physician to provide that counseling, and there is more time required of the physician visit, and maybe that gets reflected in the way the physician has - the CPT code the physicians use to bill that, so it looks like they are using higher resources, or maybe it isn't. Some practices have hired individuals with expertise in talking to patients about counseling them about their options. And that will not be a billable service; it will not show up in our resource use.

Other practices have put together online programs to enable patients to look at the literature, hear discussions that other patients have had about the different procedures, and the cost of developing that and of making it available to the patients are also not billable directly and may or not be showing up in our resource measures. Each of

those represent three - and then some practices because it's not billed well will not be doing very much of that in any organized way.

So we've got four different models about how patients get educated about their choices about prostate cancer. Each of us have very different resource profiles, only some of which are being shown up in our billing, and if we are using billing methods to capture resource use, only some of which will be showing up in our measure of how much resources are being devoted to the treatment of prostate cancer.

So I think Jim's issue is, down the road we want to sort out whether - which of these are better methods, which of them are better methods for which patients. But in - even if we are in the realm of resource use how do we capture the differences in the way practices have organized themselves to provide information to patients around decisions the

patients have got to make.

Have I got your problem and a concrete enough example?

MR. WEINSTEIN: You're hired as a resource.

CO-CHAIR LOTZ: Inasmuch as I put out there for folks to recall, you know CMS has these buckets. They are not precluding the idea of being somewhat aspirational and creating new buckets. So that may be part of the framework that is put out there, and it may be that when the calls for measure go out, no one populates that with any measures. But putting it out there I think brings in an important point. And I'm speaking personally now, I think it is an important point.

MR. WEINSTEIN: There are measures that IOM and others have been looking around, patient values, knowledge-based measures about the decisions. People are working on those kinds of measures. The notion of it being a resource that currently

is not in our normal thought process to me doesn't exclude it as a resource. And I think we need to think out of the box about that if we are actually going to have some benefit for the future going forward.

CO-CHAIR LOTZ: We'll have to be a little concrete in the use of our resources. I know from the perspective of where I sit, we create all sorts of programs. And I not infrequently ask myself, does anyone want these programs?

All right, next I have Jack, but you're down. Mary Kay.

DR. O'NEIL: Well, speaking of out of the box, in my company we have purchased intellectual property to decision support processes for 12 different conditions, and we deploy them on our membership as the insurer. And so in part because we have a patchwork of capabilities across the country in terms of different delivery systems and physicians. So in fact that is being invested

in and done as a resource. And if we look at the whole big global per capita spending in the U.S. on health care, I mean the stuff we're spending money on is an expense and an investment. And we can argue whether we should be doing that or not, and whether we should be paying people to do that. But that is currently being done for the 12 million people that have their coverage through us.

DR. NEEDLEMAN: Immediate follow on to that, and once again it goes to, those costs will be showing up in the ASO cost, the administrative cost of the insurance, that wasted portion rather than in direct patient care expenditures in terms of direct billed services. So again we've got to understand what we are trying to measure, and we've got to appreciate that some of the things we're trying to measure right now are not being billed for explicitly, and the question is, how well will any measures of resource use that people are putting forward take into

account these kinds of variations in the way care is delivered that are important in terms of patient decision making.

CO-CHAIR LOTZ: Ann.

MS. HENDRICH: This was the reason for my question earlier about counting of resources, because I think the field is so trained to not give credit for the types of innovations that we just heard that I think a question that we're all seeing perhaps in a different way is, is part of the role of this paper to help discover the innovations that are out there, and frankly stir up the measures and have them come forward in a way that can get connected to outcomes eventually.

So I think these two dots, as I think we're all saying in different ways, have a very important real connection of getting at quality and cost.

CO-CHAIR LOTZ: Paul.

DR. BARNETT: So I think the question about, for example, the unreimbursed

prostate cancer counselor, is whether we want to put in the call for measures that we are seeking a measure that would involve actually figuring out what things cost as opposed to what we actually reimburse for them. And there are products, activity-based cost allocation systems, I'm sure, that people would be interested in submitting them as potential measures.

On the other hand I could see this as falling into one of those categories of things that is not immediately feasible, and something that we ought to kick forward to the next committee. But somewhere we need to at least acknowledge that limitation if we don't address it and acknowledge the limitation and say this is to be accomplished.

CO-CHAIR LOTZ: Ethan.

DR. HALM: Yes, and I wonder if one way to solve this issue of what to do about things that we know take real inputs but aren't currently reimbursed for, if we ought

to have a section that directly addresses that. So as a primary care doctor, a big chunk of what I do is not currently reimbursed for. So a lot of the coordination, counseling and other services, the extent to which it doesn't show up in E&M coding and email and phone stuff, whereas in some of the new pay reforms they are trying to acknowledge that as inputs that are real, that need some other way to sort of recognize that. And so the shared decision making is one area, but there are other examples as well. And I wonder if in the document we may want to comment on the fact that there are real inputs out there right now that we just don't have very good ways of measuring or counting, and that we need to do so so that people understand that it is not just - if it ain't being charged for now it's not important or doesn't exist.

CO-CHAIR LOTZ: Renee.

MS. MARKUS-HODIN: So I just wanted to as probably the only consumer rep on

the committee, I wanted to just echo - I feel like we are reaching consensus around that. I agree that we should be using this process to kind of drive the - to attempt to drive the kind of behavior we'd like to see in the system.

And so paying for the kinds of counseling that we have talked about, the use of shared decision making tools and the like I would absolutely support. I'd add another one, and I think it absolutely should be in this paper, which would make this I think for consumers a much less inscrutable paper to read.

So I would add that we might want to include other things, more proactive things, like evidence-based programs such as chronic disease self management programs, things like that, that those could also be things that we would count as resources.

CO-CHAIR LOTZ: I just want to do a real quick time check. We've got 15 minutes

before we'll open up the lines for any of the public that are listening in on this to comment. So right now I have teed up David, Tom, Joseph and Bill. That might be all we can get through in about 15 minutes. So bear that in mind.

David.

DR. REDFEARN: I'll be quick. I just kind of agree with some of the comments, that just the practical issue is, a lot of these services, there are no codes for them. The example we ran into in California is telemedicine, which got to be supported by the Medical program in California, and there were some codes developed but nonstandard codes, and you are always struggling with them. And from a prospective payer point of view, if you don't have a code you can't pay it, so it doesn't exist; that's part of it.

The other thing that just occurred to me is, the rules that are being pushed out now about loss ratios for carriers is going to

push the - we are going to start developing codes for these kinds of services because we want them to be in the admin costs when we calculate our loss ratios. So that is actually going to push in the right direction I think. Right now there is a huge gap but it's going to move in the right direction fairly quickly, I would suspect.

CO-CHAIR LOTZ: Tom.

DR. ROSENTHAL: Well, we've identified kind of a generalizable problem it seems, that the way we count things is by billed services or, in a capitated system, encounter data. And how do you account for things that aren't billed? And so this is certainly one that we've seen about these kind of patient teaching kinds of services. But the other one is there are entire health systems that don't generate bills, and wouldn't it be nice if in this thing you'd be actually able to know how well is Kaiser doing in relationship to the commercial world, or

how well is the VA doing in relationship to the commercial world. And this may be a bridge too far as I think it was Paul alluded to. But it certainly is a problem that we might want to think about, because we've made an assumption that the only way we're going to count things is by, whether it was paid or not that actually somebody has generated a billing slip which drives the accounting.

CO-CHAIR LOTZ: Joe.

DR. STEPHANSKY: I think we need to be aware that there are actually quite a few payers, insurance companies, that are experimenting with sets of codes for these kinds of services. On Blue Cross Blue Shield in Michigan for example has developed some codes for social worker contact; for group educational sessions; for different kinds of nursing assistant contacts that other insurances may not be recognizing at this point, but there is going to be a lot of data there, potential measures that we can learn

from.

CO-CHAIR LOTZ: Bill Golden.

DR. GOLDEN: Again, very briefly, again, talking about the left side versus the right side of the curve, as we start going toward the right, we start really focusing the system on paying for widgets. You pay for it now; therefore we can continue to measure it. And we really want to move away from paying for widgets, and moving in more broad kind of service packages, which is just a whole other dynamic as we move forward.

CO-CHAIR LOTZ: Bill Rich.

DR. RICH: Well, actually a lot of these services that are not quote counted now, the codes exist and they are valued, they are just not paid for, including consultation, email consultations, coordination of care. And I think what we are going to find is you are going to have successful economic models of delivering quality of care, and if you do a look back then you will see what comprises

those things. That's what happened with the medical home. The Medicare demonstration care project for chronic care, there were four or five sites that had really, they did quite well. So you go back and you look retrospectively what were the components of Michigan, Dartmouth, and Geisinger, and you put those inputs together, and that's how you can actually value and define what is an effective medical home. And I think that's what we are going to find here. I don't think up front we're going to be able to count these things. Because there is now way of accounting - even though there are codes for them, they are not valued.

DR. STEPHANSKY: They are being paid for as part of the developing medical homes.

DR. RICH: But that is just in Michigan. But there are many parts of the country where there is no way of capturing that work being done by primary care now.

CO-CHAIR LOTZ: All right, I'm going to try to summarize a little bit, and the point of doing it is just tell me where I'm wrong.

So what I've heard was is that there is a desire to take this continuum that is up there and relate it to prior work that NQF has done, in particular the model that came out of the evaluation of episodes very recently. And Helen said we can get the executive summary. So hold on to that thought if you want to look at that, and then comment on it again tomorrow. But clearly the list needs to be a little bit broader, so we need to create some new buckets. Some of these new buckets may be somewhat aspirational, but should include the patient decision making as a resource that we should capture. Care coordination was mentioned, and maybe there are even some other things that we haven't quite articulated them in a nice sound bite, but what else is happening in medicine that

should be incorporated here.

So thinking about it also in terms of a linear continuum or possibly a box as Tom Lee put out there. And I think that summarizes what I wrote along the margins of my paper. What do you have, Sally? What else is missing? Or from the group, what else wasn't captured in that summary?

MS. TURBYVILLE: I would just add that we would, as we in particular I heard the requests to add something in between per capita and per episode. And that's mainly to address the way people think of the two as - that there is something in the middle. But I think we do want to continue to present this as a continuum and not just a series of buckets, so that we don't back ourselves into a corner. I see people nodding their head, okay. And then making sure whether it shows up in the continuum or the language within this section that there are very important resources, whether it's through innovation or

which have been ongoing for awhile that are not currently reimbursed that can affect resource use. And we need to make sure that we are putting that in this section, because that is real world, and we need to make sure that we are addressing that, along with all the other comments that I heard including taking a look at how to further add information for the users and readers of this document so it's not this continuum at a very high level; that they can actually think about how it applies to the measures that they have been exposed to.

CO-CHAIR LOTZ: Well, you guys are models of efficiency, because we are actually 10 minutes before we open up the lines. I don't know what the NQF dynamic is about that. Can we open up the lines early? Do we have to talk about our children for 10 minutes until we open up the lines? Or where do we go from here?

MS. TURBYVILLE: Well, I'll

defer to Ashley and Helen, but we do need to make sure it's open at 12:10, because there is an agenda that is posted to our website that said at 12:10 we'll be open to public comments. So and without knowing how many questions will come in, I don't know what your experience has been.

So we'll just go ahead and open them up for public questions of the steering committee.

Could the operator please open the line for public comment or questions?

OPERATOR: If you do have a question or comment from the phone line you may press star zero and then a voice will prompt you on your phone line to indicate when your line is open to make your comment or a question. We ask that you state your name before you pose your question or comment. And once again if you want to ask a question or make a comment, you may press star one.

DR. MUNLEY GALLAGHER: This is

Rita Munley Gallagher for the American Nurses Association. May I get in the queue please?

MS. TURBYVILLE: You're in the queue. We can hear you loud and clear. Please go ahead with your question or comment.

DR. MUNLEY GALLAGHER: Thank you for the opportunity to comment. I would respectfully request that the steering committee utilize more inclusive language as appropriate throughout their process and also within the white paper which would be reflective of clinicians, practitioners, and health care as opposed to referencing physicians and doctors and medical care.

In addition, while I've not had the benefit of the white paper to review, I believe that I heard reference to gaming during the steering committee's discussion. And I would call the committee's attention to earlier comments sent to NQF by the then presidents of the American Medical Association and the American Nurses Association opposing

the reference to gaming by clinicians within NQF documents.

Thank you.

MS. TURBYVILLE: Thank you. We have taken note of your comment.

Any other questions or comments?

OPERATOR: We have no one else on the phone lines at this time.

MS. TURBYVILLE: So at 12:10 we'll open it back up just to make sure that there isn't a public comment that was going to call in specifically at 12:10 based on the materials they've seen. Otherwise we'll just keep it on hold for now.

DR. GOLDEN: Doris, I have a question. There are some comments about the text - I only got the text in the last couple of days, and there were some comments about some of the trials that I didn't want to spend during the meeting. Who do I send comments to?

CO-CHAIR LOTZ: Sally. If

they're handwritten.

DR. GOLDEN: You probably want me to type something to you?

CO-CHAIR LOTZ: It would be easier for us to manage and track. We do want to share all the comments across the board. And I don't have the whole timeline in front of me right now but I think it's mid-August that the white paper gets done. So you have a couple of weeks after this to continue to write.

MS. TURBYVILLE: So for the white paper the next two weeks will be devoted to comments preceding this meeting and after this meeting incorporated those, and then we do have to send it to our publication department who does an excellent formatting, editing, making sure we are using the appropriate NQF language that they prefer. And then it gets posted for public comment . Then we have another opportunity with all of you to review that public comment, and again, make the paper

better and better. In the end the final paper will be in November. But this next two weeks really is critical to get it to a point where the steering committee feels it's ready for public comment. So in particular for red flags, please let us know if there is something you think we must address or we can't go forward. And then all edits of course and suggestions we will be paying very close attention to.

CO-CHAIR LOTZ: Bill - I'm sorry, I've lost the order. Jack, you're up next and then we'll go to Steve.

DR. NEEDLEMAN: Two things. First, in terms of the comment we did receive about using more inclusive language, particular as we began to think about this work, is supporting efforts to find more appropriate, more integrated ways of delivering care. I think the comment about using more inclusive language about who is delivering the care is absolutely appropriate,

and we should be looking, making sure that we are being rather aggressive in broadening the language as we do the revision.

The second thing is in the discussion of all the comments I think Sally mentioned this but I don't want it to get lost, we've got a whole variety of individual practices and integrated systems, sometimes working with the insurers that are in fact delivering many of these services right now; the total amount that they are compensation from whatever services they bill for are in fact being used to pay for a variety of these services.

So I think as we write the paper and as we think about editing it, we need to draw the distinction between the resources that are used to deliver care, some of which are paid for explicitly and some aren't, versus the specific services that people are being billed for and reimbursed for. And to draw the distinction as we measure resource

use between measuring on the basis of billed services and measuring on the basis of the way care is organized and delivered, and some of the tensions between those two things is appropriate measure of resource use. And we should not - again I think it was implicit in what Sally added to the things that were on her list, but we need to make very sure that that is clear and it's communicated in the paper.

CO-CHAIR LOTZ: Steve.

MR. PHILLIPS: Thanks. Maybe it was at the end of the previous section, but it's just a specific comment on the language. I didn't have a lot of written comments, so I'll just offer this up now. At line 229 it talks about some of the specific objectives of these measures. And I wanted to insert the note, because it doesn't mention there - it was mentioned I think earlier in the discussion this morning that they should be useful to the health care providers in terms

of the information that comes out of these measures. And I thought that that should be mentioned in that section as well.

CO-CHAIR LOTZ: Tom Lee.

DR. LEE: You know the more you think about this stuff, the more complicated it gets. One thing which hadn't occurred to me was, to what extent do we imagine adjusting for the type of insurance products that make up the patient population of whoever is measured, you know, PPO versus HMO.

CO-CHAIR LOTZ: So you're thinking about benefit design or the composition of physicians?

DR. LEE: Well, I think in benefit design and in high deductible. I mean it takes you down a road that frankly probably just gets you feeling lost. But these are variables I think that I know that when we start showing this kind of data to other people out there they are going to go, well, if it's just for this and that. We might as

well at least anticipate those questions and prepare a response.

CO-CHAIR LOTZ: Again, I'm thinking that as we think of our building blocks strategy, that that is something when you are building you have to consider what resources a patient may have by way of their benefit design.

Mary Kay.

DR. O'NEIL: It's not even benefit design as much as like the consumer-driven model of health care, and how that drives individual decision making. It really changes things that we know in our industry from an actuarial standpoint, we have actuarial data on the impact of those different types of products. So it's really product type and benefit design.

The other issue that could get you really crazy about benefit design is in the commercial insurance world the trend over the last number of years, I don't know when it

really started for increasing percentage of our business being self insured. That means somebody comes to me and says, does CIGNA cover this, I can't hardly answer the question. I mean I can say what we do on our fully insured book of business, what we do on a consumer driven, and what we do on our own policy for our own employees. But given the array of plants that we have, we have a lot of different plants, so there is a lot of complexity out there.

DR. LEE: I just want to say that contemplating it I think that we might want to decide up front it's hopeless to adjust for these things, and just make it clear that it is hopeless. You'd never get the information on their deductibles and that, and that the data might end up being biased, but I think we have to accept that there are going to be issues that we can't adjust for.

CO-CHAIR LOTZ: But at least acknowledge it.

DR. RUDOLPH: Well, just a couple of different things on that. One might be a stratification process like we do with race, ethnicity, those kinds of things. Because I think we don't want to adjust away differences when consumers are able to make choices; they should be able to see what those differences are, and thereby make a rational choice.

So if we are going to do something, I would propose stratification as opposed to adjustment.

CO-CHAIR LOTZ: All right, we're going to pause here again to see if there are any public questions.

MS. TURBYVILLE: Operator, at this time we would request that you open the line again for any public comments or questions.

OPERATOR: Again, as a reminder, if you would like to pose a question or comment, you may press star one on your telephone keypad at this time.

(Pause)

It appears that we have no questions or comments at this time from the phone lines.

MS. TURBYVILLE: Thank you.

All right, so we'll put it to the group. We've got a few table tents up. We can take the three that are up right now - oops, two that are up right now, or move to lunch.

David, and then Kurt.

DR. REDFEARN: I'll be fast in the interests of the food.

We have an interesting experience with ACO in California with some pilots that are starting up right now that reflect on these issues of benefits. First, the issues that came up -- I think stratification makes sense in terms of looking at this - I don't think you can adjust it away, stratification.

There's been some push back from the medical groups in terms of looking at

individual enrolled business, because of high deductibles, policies that are sold commonly in that industry, and if you are going to take responsibility for managing the whole care, and you have this huge gap in the benefit structure, it makes it harder for the medical groups.

The other comment about ASO which I thought was very appropriate, essentially we've taken essentially all the ASO business off the table for the very practical reason is that you have to go back to the groups and get permission to do that. That's not insured business which we sort of control.

So again it's like a stratification. It's what populations go into your analysis. And I think it makes sense to think of it in that context.

CO-CHAIR LOTZ: Kurt.

DR. ELWARD: Yes, I'd just echo those, and also mentioned that to your point about the number of self-insured business on

one hand, those self-insured policies can provide a lot of innovation, because the self-insured groups will do some things that you just don't get done otherwise.

At the same time it really does complicate things, and we're - I think benefit design is one thing, this person does the surgery for this policy is really really cheap, but I guess we do want to include anesthesia, if you don't have that. Or it may go to diabetes. I'd think it'd be very important for an employer or a purchaser to know that when they are buying diabetes care from X, it includes a package of services that they really want their employees to have, and that extra resource input may be worthwhile. So I think it'd be really helpful to people.

CO-CHAIR LOTZ: Well, we are actually on schedule. So Sally, do you want to see if there are any last public comments before we break for lunch? Just go ahead and break for lunch? When do you want us back?

MS. TURBYVILLE: So we will break for lunch now. And for those on the phone we will reconvene at about 12:50. So thank you.

(Whereupon the proceeding in the above-entitled matter went off the record at 12:16 p.m. and resumed at 12:57 p.m.)

CO-CHAIR LOTZ: Steering committee, lunch is officially over.

You may now sit down and digest, but please do sit down so we can begin. Our afternoon is very, very ambitious, so the sooner begun, the sooner done.

MS. TURBYVILLE: And to the operator, if you could please make sure that the line is open, though not open for comments or questions at this time, just in case you disconnected during lunch.

CO-CHAIR STEINWALD: All right, we're on to Section 3, types of resource measures. We've already had some discussion of this right before lunch. But before we get onto the next topic, would anyone like to say

anything more about types of resource measures?

One thing I guess I'd like to emphasize, what I said before is that when the per capita terminology has been applied as I understand it at CMS and elsewhere to mean a population of patients. But we I believe intended to incorporate more than just a population of patients. It could include populations at large, populations served by a health plan, the population of the state of New Hampshire.

And I'm not sure whether we need to have new terminology or just make it clear in the text of the paper that per capita means more than just one thing.

Anyone else? We talked about it enough? Okay, that we can go on.

MS. TURBYVILLE: So this is still within Section 3, and what we want to have the steering committee discuss and then come to some agreement, is, is there a way to classify

what we called phases of resource use measures. And it's steps that a measure developer and an implementer of those measures would need to take in order to successfully then roll out the resource use measure.

And the way we classified three phases, and we could call them modules as well, include data preparation, which would be steps that are taken in order to make sure that the data are in a form or robust or sufficient enough in order to support the resource measure. A creation of the unit of measurement which in essence could be the creation of the denominator for the resource use measure the way we've defined it. We shied away from the term, denominator, because we felt that there was some disagreement on whether or not you can call them denominators. But we're very open again to what these are called. And then how you apply those, and that includes once you've defined your unit that you want to measure, whether that is a

population, or is it an episode, et cetera, you have your clinical logic that is in that second step; you have your creation of an episode if it's an episode measure, or your creation of the population, whether it's those patients with diabetes; and then how you apply that unit of measurement which would include which resource units you want to measure. So those would potentially include the reimbursable and as we have stated acknowledging that there are unreimbursible units. So your ED stays. Your evaluation and management, et cetera. And then also whether or not and how you would monetize those. And all those steps are outlined in the paper.

But we would like some general agreement on this attempt to put resource use measurement into these three phases, and then that will help us discuss with all of you which phases and which steps within phases are subject to evaluation by this steering committee.

So it's an attempt to make sure that as you think about evaluating these measures, recovering all the various steps which are quite numerous, what is true is that these are not all mutually exclusive steps that would happen only for resource measurement. You might find them in quality measurements. But we think it's really important to send the signal to the measure developers and others what exactly we will be evaluating, and how we are bucketing those particular steps.

So with that I'm going to hand it back over to the chairs to either further add to that or just kick off the discussion.

CO-CHAIR STEINWALD: The chair is looking for names being turned up to the vertical position.

Tom

DR. LEE: I think that those steps work, chronologically and logically. I think that the middle one is where there is a

lot of intellectual work to be done. I think that looking over the white paper draft, I think the one section within that number two that I think really probably warrants some more discussion is the risk adjustment section. Because a na‹ve reader might look at it and think that risk adjustment is something you do and it's done like scanning a computer for viruses, which of course isn't that straightforward either. (Laughter) But risk, it's more like a philosophical or religious experience than something you just do, it's not a commodity. So the risk adjustment approach of Prometheus is quite different from other kinds, like the DxCG. Under Prometheus they've taken the approach that if you do more procedures to a patient it must mean the patient needed more, and you move into a different bucket, because they wanted providers to be comfortable that this wasn't something that was going to punish them for taking care of sick patients.

So it's a spectrum. And I think the risk adjustment section should probably include some longer discussion so that readers can be realistic about it.

CO-CHAIR STEINWALD: Whenever I've dealt with risk adjustment issues in the work I've done, mostly at GAO, the question or the issue of the best being the enemy of the good always comes up. And if you can't achieve perfection, how much do you need to achieve in order to go forward, even if risk adjustment is imperfect.

And I don't know if that is worth some discussion among the steering committee, but it might be, because some guidance to the developers about how good is good enough might be helpful to them.

DR. LEE: Just to jump out and say, I agree with you completely. It can never be perfect. There is no completely risk adjusted status. You can't adjust for all the socioeconomic factors either, and so I think

that's like being, having perspective on the data knowing that they are not going to be perfect. There may be biases that we have to live with.

CO-CHAIR STEINWALD: Paul, and then David, and Bill.

DR. BARNETT: So I think that the phases are right; I'm a little worried about the units for measurement. It's just the nomenclature; I think I mentioned that before. It's a little bit hard to understand what is meant. And I appreciate you're trying to make a generic term. But it's not real clear to folks.

Just in terms of, in the document itself, everywhere I saw the word, price, I circled it, and I thought, I wonder what really they mean here. And I think in many cases it meant reimbursement - you know we had that issue this morning about that.

There was also mention in the data cleaning part about removing high cost

observations. And I think Winsorizing is one thing, removing them altogether is quite another. So I thought that might be just a little editorial comment. I don't think you probably meant that.

And then I think vis-…-vis the risk adjustment, I think there is a whole category of people alluded to about pay status, socioeconomic status, that are not really developed much in the document. And at some point we might want to think about the issue of teaching status of the provider, and disproportionate share status of the provider, and how that enters into these efficiency measures, and how do we credit those things, or do we ignore them; important things to think about.

CO-CHAIR STEINWALD: David.

DR. REDFEARN: Per the risk adjustment, I think one thing that might be prudent is to make sure that there is no implication that somehow when you do risk

adjustment you adjusted away risk. The thing I just keep in mind is, and I've had a lot of experience with the models, like the DxCG model and the symmetry ERG model, is that prospectively those models predict about 25 percent of future cost variation. That seems to be - there is no danger of getting something perfect, because we are stuck with something that is actually fairly crappy. (Laughter) And in my social science training when you account for 25 percent of the variation you've failed.

So they certainly help, but they don't solve any problems. You say, well okay you do risk adjustment; you set categories; you apply these models; and risk goes away. No, it does. Some of it goes away, but not much.

CO-CHAIR STEINWALD: Bill.

DR. GOLDEN: As we go through this, I think it'd be useful to have somewhere in there that the risk adjustment should be

based on the total universe of the patients' claims, and not just on the claims of the entity being looked at.

What do I mean by that? Well, for example, if there is a diabetic or a heart failure patient in my practice, I may not code for mental illness issues. In fact if I did code for mental illness issues I'd be penalized because I get paid less because I'm caring for the depression. However everybody knows if they have mental illness comorbidities, their costs go up. So I think it'd be a useful principle that it should not be just based on the narrow diagnoses being cared for but all the diagnoses that the patient carries in the claim set.

CO-CHAIR STEINWALD: Tom.

DR. ROSENTHAL: On the risk adjustment question, my observation would be, it sort of depends. I would say that to the extent possible we should in fact include socioeconomic status. Because as I've learned

in about the last year in a half it's an incredibly important driver of cost of care, and to not include it would be a major mistake.

But to the question of how good is good, and when is perfect the enemy of good, I think it depends on what use the data is going to be put to. If the use is going to be to publish the names of every hospital or every provider in a state, region or country on the front page of the New York Times, I have a feeling that all of those individuals so arrayed are going to be very much interested in as good a risk adjusting thing as we could possibly come up with.

If on the other hand the goal of the exercise, and what we are trying to get, is performance improvement, I suspect that good enough is generally good enough for most organizations to look at and say, it may not be perfect but we'll work on this because it's probably directionally correct.

So I think the answer is, it's going to depend on the uses as to how much risk adjustment is going to be viewed as believable.

Jack Needleman and I were talking the other day, and there are two criteria for this entire exercise it seems to us. It's got to be believable and actionable. Believable and actionable are going to then depend on how the data is collected and to whom it's portrayed for action.

CO-CHAIR STEINWALD: Bill, Rich and then Lisa.

DR. RICH: On the point on risk adjustment that Tom raised, it depends on what part of the spectrum you are, working left to right. If you are all the way on the left where Bill Golden is you need more and more granularity, and it's very important that you do I think look at race and socioeconomic status.

When you get over to the right-

hand side for procedure specific, I think that the risk adjustment should be, the measure should have a very finite endpoint, so there is not - almost have a clear line and a little granularity. Now I don't know how you verbalize that, but I think you can develop some pretty powerful measures on the right-hand side where you don't want granularity. And on the left-hand side you need that granularity if you're ever going to have meaningful risk adjustment over there.

So I think there is a continuum of risk adjustment, depending on what kind of measures we are looking at. I don't know if that makes sense.

CO-CHAIR STEINWALD: It makes sense to me.

Lisa and then Ethan.

MS. GRABERT: I just had a comment about how we're discussing this. The last bullet says, later discuss the extent to which each phase is subject to evaluation. I

actually think we need to have that conversation first, because I don't know what parts of the methodology I should put into which bucket unless I know whether or not they are a sufficient or necessary part of the overall evaluation process.

Previously in the earlier draft that we had, data preparation and creation of the unit for measurement were both necessary and sufficient, Part 3, applying it to measurement was not necessary. And there are certain things that are in that bucket that you might argue are more or less necessary such as attribution, once you know what you plan to do with each of those buckets.

MS. TURBYVILLE: So what was previously put forth was just to get the steering committee talking about it; there was no assumption that three would not be subject to evaluation. It was like here was an example, you wouldn't have to include them all.

Right now, the thinking after we heard the comments based on the webinar is that potentially all three are subject to evaluation, and maybe there are steps within - honestly we don't know. We really want the steering committee to inform this piece - which parts will be subject to the evaluation and which parts will not. And there are many reasons and implications for those decisions.

DR. HALM: Nothing stops providers quicker than the risk adjustment issue. And there have been lots and lots of thoughtful discussions of this, and I wonder, obviously 15 lines on this doesn't capture what's already known and been talked about with risk adjustment. So I wonder since there is an element of the white paper that is to educate people and somewhat aspirational as far as encouraging your more sophisticated measures if that's what's indicated, if either NQF or their other well endorsed sort of policy descriptions of the different

domains, or sort of goodness in the pursuit of the perfect with regard to risk adjustment that are out there that we can articulate so that people can kind of see in the hierarch of the kinds of domains that are sort of optimal, okay, and not so good.

I'll also comment that in the era of electronic medical records there is going to be much more that people are going to be able to do. So even like in our center we are coming up with measures of social chaos and adherenceness using electronic means.

CO-CHAIR STEINWALD: I'm going to pretend this is for the benefit of the guys with earphones. Would you repeat what you just said? Social chaos and?

DR. HALM: Adherences.

CO-CHAIR STEINWALD: Adherences.

DR. HALM: But I wonder if we can sort of amplify or have an appendix of just some - a lot of the thoughtful work that's been done on risk adjustment so that we don't

have to see that as the purpose of the resource use steering committee, but really sort of reference that or educate people about what's already been known and done on that.

CO-CHAIR STEINWALD: Right. So you're saying let's not even try to create the wheel, but let's point to where the wheels exist elsewhere.

I'm sorry?

DR. HALM: We need to have more than 15 lines in the document that reflects that current state on that, and that would help.

CO-CHAIR STEINWALD: Lisa, your card is still up. Did you intend it to be?

All right, I'm not sure we've given enough guidance to staff on this. One question is, are these the three distinct phases that need to be identified as such? Do we agree with the phases and how they are described?

Data preparation I think, it

sounds a little bit more techy that we intended it to be. That is, it's not just the preparation, it's identifying what data will be used for a resource measure, and then putting those data in a form that meets the scientific acceptability criterion, I would guess. Did you mean it to mean more than that?

MS. TURBYVILLE: Kind of diving into what we have within, so the data preparation is to determine any changes that might need to be made to the data, and some would be optional; others would be mandatory. And some of that could depend on whose resource measure you are using. So for example an Ingenix episode-based measure may have suggestions on how you might want to set up your data. Are there certain claim line outliers that you would want to either Winsorize or actually eliminate? And what we are proposing is what steps need to be taken in order for the resource use measure to be

valid based on the measure developer's experience be explicitly included in the specification. So that if there are steps that a user of a measure needs to take in order for the end measure to be valid and reliable, that they make sure that they're explicitly telling the measure developers what those steps are.

DR. LEE: Although this part isn't the sexiest part of the document, it's important for states that have developed all-payer databases, like Maine, Massachusetts - we actually outsource it to Maine, because they actually had experience working through these issues, and wanted our data to be comparable at least to Maine. So there really are some methods - there should be clarify and consistency in the methods used here. So this is important methodological stuff.

CO-CHAIR STEINWALD: Jack, Bill, Barbara.

DR. NEEDLEMAN: The goal of the

work is to produce comparisons that are informative, whatever we're comparing folks to. And the whole purpose of risk adjustment is to eliminate things that say, yes, but you can't compare that. That's apples and oranges. So we've got lots of methods for adjusting our comparisons so that they are informative and I'll use the word believable as a condition. And we've got methods specified in here that risk adjustment typically means some sort of additive or multiplicative adjustment process, or - but we've got those methods. We've got stratification. We've got truncation. We've got exclusion rules. And ultimately when you are looking at a measure you've to look at the combined way that all the things they have done to create comparability in the numbers. And that I think is the most important thing to communicate here.

And then I would agree that some of the specific methods, and some of the

things that NQF has already learnt, looking at other kinds of measures about how those comparisons are made might very fruitfully be in an appendix or a more extended discussion.

But the essence of what we are trying to get at is important, and then the mechanics of the kinds of things we've seen in other measures to do that are also important. And in that regard, just again at the risk of nitpicking on the language, around line 524 it says, unlike quality measures which normally compare performance to an agreed upon standard, and it talks about the lack of agreement here. Well, we've got some quality measures that are process measures, and there we tend not to risk adjust. Because once we've decided that for this population something should happen, it either happened or it didn't; we don't need to risk adjust it.

But we've got other measures, particularly anything that falls into the realm of outcome measures, where risk

adjustment is absolutely essential. So we've got mortality measures for cardiothoracic surgery, and we risk adjust the hell out of those measures precisely to produce this informative comparison basis. So I think that language is a little bit too strong and doesn't capture the full range of measures that we actually are using, some of which are risk adjusted, some of which aren't.

CO-CHAIR STEINWALD: Bill.

DR. GOLDEN: If you want to technically get into what we are doing here, I'm not sure that we are capturing it. Data preparation is really, I think you are looking at really, you are trying to describe an analytic protocol as opposed to just - and I don't think we're capturing it. I mean I've seen the material on 313, 318. I've seen people make substantial errors here. And really what you're trying to say to folks is that they need to make sure that they know what they're looking at, and they have

competent people doing the analysis. And I'm not sure that's truly data preparation. It's really preparing the dataset. It's making sure you have done the right basics before you can get started. And putting in an inclusion or an exclusion as the next step, that actually shouldn't happen I think until after you've defined your measures. So I think you've kind of put some things in a funny order there. But let's - I mean basically you have all the material there, but I'm not sure you have the right steps put together.

And what you are trying to do here I think is put together a roadmap for the analytic team to make sure they don't make fundamental errors.

CO-CHAIR STEINWALD: Barbara.

DR. RUDOLPH: I was doing to talk about some similar things. The data selection really, there should be a basis for why that particular data is being utilized, given this is by other parties. And if it's something

new that has never been done, then I think there needs to be a rationale as to why that might be a viable option, things like that, in addition to then when you start to prepare the data.

But I really like the idea that all the steps must be explicit. And having to do that is so torturous, because you make a million decisions when you are analyzing data. But for the end user unless that is made explicit you aren't going to be able to really know what happened or replicate it.

For risk adjustment I agreed with what Jack was saying, depending on where you are on that sort of measure continuum, if you were doing just a procedure or a clinic visit you may not need to risk adjust it. So if there is risk adjustment necessary or not necessary, it should be at least discussed in the data submission, and a rationale should be given for sort of the extent of that adjustment activity.

And I actually on page 47 of the report, I liked what was in the exhibit, risk bifurcation, and then Prometheus model where they actually came up with some criteria for the predictors that were selected. In this case it was like greater than or equal to 30 episodes per category of positive coefficient; a low variance inflation factor; a high partial r square; and clinically plausible. Because what I see happening oftentimes is, those decision points aren't made explicit. Why was that predictor included? I look through for a little private project that I did, I looked through all the public reports on CABG procedures, and the risk adjustment models for all of them and found that about 80 percent of the risk predictors actually weren't significant, yet they were still being used. And to me that just muddies the water. I mean if you are going to have predictors they ought be actually contributing to the model as opposed to just additional factors

that people think might contribute but actually don't.

CO-CHAIR STEINWALD: Tom, then Kurt then Jim.

DR. ROSENTHAL: I had a clarifying question is, this discussion about data presumes that somebody submitting a measurement will have very carefully identified the populations that are being talked about, and to whom it would be attributed, and all of that kind of stuff, and now we are only talking about the data elements that would be necessary and how you would get them and all that.

Have we assumed that that step is done, or have I missed that?

MS. TURBYVILLE: If I'm getting your question right we would have assumed that they have specified the measure and tested it in some way or another, and so clearly identified which elements a user needs in order to support the intent of the measure

they are putting forward. So if it's an overall resource use, and they say pharmacy data is critical to this metric, you would assume that they have a rationale for that.

DR. ROSENTHAL: So if, in fact, somebody were saying I'm going to measure Blue-Cross-capitated populations in California across organized medical groups, and then this is only relevant to them. Here's the datasets, here's how I get the data; here's how I risk adjust the data, and so forth. The presumption is all the other descriptive stuff would have been done.

MS. TURBYVILLE: Correct.

DR. ROSENTHAL: Okay.

CO-CHAIR STEINWALD: Kurt.

DR. ELWARD: Just following on Bill's comment on what we're trying to do. Unless data preparation is a formal term that we always use, I'm wondering if we might think more about data protocol or development of data design, something like that. When I read

data preparation I almost think of data cleaning, like you've got the data and now you're going to prepare it. And maybe it's just me, but perhaps we could be a little bit clearer about what we actually mean by data preparation.

CO-CHAIR STEINWALD: Data development?

DR. ELWARD: Data development or data protocol.

CO-CHAIR STEINWALD: Jim.

MR. WEINSTEIN: If I could follow up a little bit, so these are resource use measures that are data elements I guess. And then is it implied that the IT people in either the payer world or in the provider world will actually have the ability to do this in their systems?

CO-CHAIR STEINWALD: Is that a rhetorical question?

MR. WEINSTEIN: Well, I guess the question is, we're running up against what

meaningful use is for IT strategies. And is it a recommendation of the committee then that these become meaningful use measures or not? Because it does create a cost to systems to have to do these things, and we should be cognizant of that.

CO-CHAIR STEINWALD: What do others think about that issue?

DR. ROSENTHAL: Doability was a component of what --

CO-CHAIR LOTZ: Feasibility.

DR. ROSENTHAL: I don't know all the lingo yet. Obviously, feasibility has got to be a component of the thing. Of if it is a profoundly important bit of data that is not in fact currently feasible then we'd have to make some judgment of -- this ought to be part of meaningful use even if it's not. But I think we've got to adjudicate that against what comes. I don't think we should lay it out prospectively yea or nay.

DR. RUDOLPH: To add on to that,

I think one of the questions has been over the last couple of years in the CSAC is, if data is available to only one party and to no one else, is that really a feasible measure, and I'll raise that here.

CO-CHAIR STEINWALD: A tree falls in the forest and --

Would anyone like to respond to that? Well, I guess we will discuss the four criteria and what constitutes feasibility. Should we defer that? Go ahead, Tom.

DR. ROSENTHAL: Well, I think that is a very important question, because some of the materials that we've seen hasn't been validated by anybody and isn't particularly transparent. And I think on any of these there has to be at least a certain element of transparency, and ability for independent replication before it can be put into place. So I think your point is well made.

CO-CHAIR STEINWALD: Helen, go ahead and jump in.

DR. BURSTIN: One clarification, though: I think there is a distinction between the measure and all the details of the measure, and how you do the measure being fully transparent which is an absolute paramount thing for NQF. That's clear, you guys will get to evaluate the insides -- there are not black boxes. You guys will get to set the insides of any box submitted, should boxes come in.

The issue I think Barbara is raising is slightly different, which is, what if the measure is fully transparent but the data are not? And this has been a continuing issue we are going to talk about again this week with the CSAC of you know the ultimate goal of NQF-endorsed measures that they are publicly reported to the public at large. But we recognize this as a continuum of public reporting, so if a measure is only reported for example to the health plan or only reported to CMS, that is potentially along

that path on public reporting.

But the point Tom is raising I just want to make very clear, the measure itself must be fully transparent. And we actually even have a carder that we have created for submission of measures by proprietary groups where it is fully transparent to the steering committee completely, all the details. We will then have a limited license to view. If it's endorsed they will have to provide a limited license to view so that anybody who wants to look at using that endorsed measure will have the capacity to again go under the hood, but potentially the issue of costs involved in paying for a proprietary system would get woven in under feasibility.

CO-CHAIR STEINWALD: Lisa and Bill, you both have your cards up. Did you intend them to be? Yes, okay, go ahead.

MS. GRABERT: This may be just splitting hairs, but I think that the NQF

project team may be interested in my specific comment that I think pricing methodology should be included in the data preparation bucket. That is certainly the way that I've always thought of it when I managed the physician resource use reporting program for CMS, that was definitely a data preparation step before we did anything else with the data in terms of putting it into a measure.

CO-CHAIR STEINWALD: Ethan.

DR. HALM: Yes, I think there are a lot of good steps in the analytic process that are listed out here. I'm hearing comments that I don't know that these three phases make the most sense as a way of classifying them or chunking them. And I'm trying to figure out from a process standpoint how to handle that as far as advising people with regard to the white paper. We would see maybe some additional important analytic steps in there, and then we can think conceptually of how we would chunk these things and name

them. Because I think there is a lot of great content here in the steps, but the names themselves, I think, are less informative than the individual pieces. And I hear us struggling with that a little bit. And I don't know that we can solve that in real time.

CO-CHAIR STEINWALD: I don't hear or see disagreement. I guess again, could the best be the enemy of the good? Do these three work for purposes of providing guidance to measure developers, let's say, that there is some value in trying to disaggregate a process into components, and we've got, as Sally said earlier, that calling them phases implies that they are sequential in time and don't actually necessarily mean that, so they could be modules or something else. But if - can anybody suggest any modifications or improvements to - other than what we've already talked about in data preparation?

Do you think you have enough

guidance for the time being?

MS. TURBYVILLE: I do, thank you.

DR. RUDOLPH: Can I make one suggestion?

CO-CHAIR STEINWALD: Sure.

DR. RUDOLPH: I think this is where it kind of threw me off. I think it'd be helpful to have a category above data preparation called "measure specification" so that people would then realize some of the things that we've talked about, that would be already - that you would start out with this package of measure specs, and then you'd do data preparation, and then creation of the units for measurement and explaining it that way. I don't know how others feel, but --

DR. PENSON: What do you mean by that? I'm over here, sorry. I'm not too sure what you're driving at with that so help me out.

DR. RUDOLPH: The measure developers have to provide to NQF the set of

specifications including all the coding, the - et cetera, all the description of the data, blah blah blah. And then, as an end user you get that description. But then there is also - you'll have to do data preparation activities which would include some of the things that - the next steps, the data prep, the creation of the units for measurement, the applying the units.

DR. PENSON: I see where you are driving at. I mean, I think it's all implied there because effectively your phase two and three are your numerator and denominator for your measure specifications. So I'm wrestling with the same thing everyone else is wrestling with which is - I think it's semantics. I don't know if adding a fourth helps. I see what you're getting at. But it strikes me that these are not - they are not sequential, they're concurrent. I mean they are components, but you sort of - maybe the whole thing is measure specifications; maybe that's

what this is, which is each measure has to have certain specifications. The numerator is the unit of measurement, the denominator - and you're going to have to do certain things with the data to get there and you've got to tell us what they are. Maybe that's the right way to do it, maybe.

CO-CHAIR STEINWALD: Bill Golden, your card is up? All right, then Bill Rich.

DR. RICH: You know one of the things we're struggling with and we're talking around, when you look at this white paper and then you read the McGlynn Thomas paper and the acumen things, are any of these steps, they're going to give us transparency, but are we going to have a product that is going to be different and useful and meaningful? Or are we going to have to have a retrospective analysis like those papers do? I think what we'd like to do is try to, in this process, try to identify the things that would make these things actual, reasonable, reliable, and

do a better job than explaining 25 percent of their variation.

And then I struggle with the whole white paper. I'm not sure that any of these steps are going to predict whether these things are going to be reliable measures or not.

CO-CHAIR STEINWALD: An even more fundamental question, I think, is, will any measure developers develop measures that meet all the requirements that we're laying out here? And what can we do to turn that around? How can we have the white paper be a resource that encourages measure developers to innovate a bit? I'm not sure that's very responsive to what you just said. But if the white paper reads as if it's full of requirements and restrictions, then you can't view it as something that's encouraging, and how can it be framed in such a way that we're trying to encourage measure developers to think a little bit outside the box and doing something

different from what they've done before.

Yes, sir.

DR. BARNETT: I am not sure - I saw in the appendix that one of the measure developers had distinguished the inessential care, the care that is not indicated, not appropriate. And I don't think that that, as being a criteria or a part of it, is in the front part of the white paper, that whole issue of appropriateness. And of course we very much care about - we want to see resources used on things that are appropriate and not on things that are inappropriate, and to the extent that that can be incorporated into a measure is certainly desirable.

CO-CHAIR STEINWALD: Well, do you think that the way the white paper is written now it's precluding that?

DR. BARNETT: No, I just think it's an oversight. That could be another attribute that some measures have that others don't and that hasn't been articulated.

CO-CHAIR STEINWALD: Tom.

DR. ROSENTHAL: I think the point that you raised about encouraging development is a really important one. Because it's not entirely clear that any of the things that might really - that are currently phrased in the thing are quite ready for the front page of the Washington Post test that I hold it against. And maybe this is a situation where the NQF's success over the last 10 years might work a little bit against this effort, in the sense that we're at the 10th or 12th year of quality measures and sort of a little bit of maturity to them. And maybe the notion that could be incorporated into the white paper is that the notion of pilot measurements would be solicited in addition to those that would get the full NQF treatment, and I don't think that that concept is in there yet, and - or at all, and maybe that's a way to address the notion that this is going to be a multi-year process to identify things that are really fully ready

for the whole full-scale NQF treatment.

CO-CHAIR STEINWALD: Your Washington Post - I know I heard New York Times earlier.

Ethan.

DR. HALM: Just thinking, I wonder if - I'll just throw it out there as a suggestion. You know perhaps the - if we called things measure specification creation, and application, would that potentially fit?

CO-CHAIR STEINWALD: You're suggestion renaming --

DR. HALM: Or just playing with it a little bit. Because I'm hearing now it's all about the measure so rather than putting the word data or unit - I mean the word data and unit, I get confused on. But I know what you mean by a measure. So maybe playing around with measure specification, creation and application or further kind of brainstorming about making it a little bit more intuitive.

CO-CHAIR STEINWALD: All right. Go ahead.

DR. RICH: Yes, that was a little bit where I was confused here, because when I saw data preparation, when I think about data preparation, I think about application of the measure that's already specified. And I think I'm stuck where Barbara was, you know, trying to figure out, is this a measure specification for developers section? Or is this talking about how you take data that you've collected and apply to an already specified measure?

CO-CHAIR STEINWALD: Go ahead.

MS. TURBYVILLE: And I appreciate this conversation. We did struggle with the language and we still are, which is one reason that I'm very thankful that you are being thoughtful about this. It is intended to parse out the measure specification. And the signal that we are hoping to drive through this process, whether it's painful or not, is to make sure the measure developers understand

that they can't just submit to us some high level specification and think that that is a done deal. They need to be very explicit about what data are needed, what needs to happen to the data protocol. It's not how you collect the data necessarily. But it would be implicit because it would tell you what types of data need to be there. You need to tell them explicitly -them being the users - how to create your denominator or your unit for which you are going to apply the resources.

You need to then say which units of resources are relevant to that denominator. Because there is a huge menu of potential resource units that could be applied to any measure. So we want the measure submitter, the developer, to say, for this measure it is valid to measure ED use, to measure monetized evaluation management. That all should be part of the specification. So, in essence, this is the specification as was mentioned earlier, broken out because we don't want any

one step to be ignored, realizing that a lot of these developers have not going through this type of endorsement process before. So to the extent that we can make it more clear and use better language, and whether or not three -

CO-CHAIR STEINWALD: Mary Kay.

DR. O'NEIL: I'm new to this process as well, so I guess there is the tension between being inviting, to want people to actually submit things, and being clear about what's required to participate. And so perhaps it is in the wording to just say things the way that you've said them.

I did come to this meeting with some local feedback that the NQF measure requirements are too expensive for many entities to participate in, and obviously we don't want just anybody in here doing things that are meaningless and nonreproducible. But maybe there does need to be a little bit of balance if part of the message is to invite

people to participate in that.

CO-CHAIR STEINWALD: The remark that someone made earlier is that NQF's past success may work against it. So the ability to get measure developers to do something they haven't done before may be something that should be confronted right up front, and the notion that a measure doesn't need to be a completely developed measure, let's say. Well, let me rephrase that.

(Laughter.)

I obviously misspoke.

Well, then, rather than get fired again, language that would be more inviting and encouraging, I think would be very helpful if we can figure out a way to do that.

DR. BURSTIN: I'd just make a point, I think in some ways some of this is just the different lexicons that these different fields use. As I read through that section again, this whole data preparation-data cleaning looks like the algorithm that a

company lists of our measure specifications. So it may even be sort of a side by side of this is what we usually think about in terms of a measure, numerator, denominator, exclusions, the algorithm that accompanies it, and then link it to what would happen, because in some ways what they are actually saying in the paper here is that the data piece is actually somewhat distinct and precedes the measure calculation. So there is almost a pre-phase in this that doesn't exist with most measures unless you sort of -- so they just need --

PARTICIPANT: It really should.

DR. BURSTIN: Hmm?

PARTICIPANT: It really should be for all measures, all the time.

DR. BURSTIN: Right, but this degree of specificity, I think, is somewhat unique to this area, and I don't think we want to lose that specificity when these things are submitted to us. I think we can work to try

to make some - help the lexicons connect.

CO-CHAIR STEINWALD: Okay. Anything else on this matter? Let's go forward.

DR. BARNETT: One brief thing is, I think that everything is here and that what people are struggling with is, they want to change the headings -- the names on the headings and the order of the text. And so that suggests there is not really a lot to worry about.

DR. BURSTIN: That was such a simply stated comment. Does she feel like she has enough in terms of what should the headings should be changed to and how they should be reordered? But suggestions on paper are welcome as well.

MS. TURBYVILLE: So we did start out with denominator and numerator. We have been having conversation with many of the measure developers to get their feedback along the way, and quite a few did not like that at

all and felt that it was trying to force them into the quality measure domain, and our measures are really different. This language came from one of their suggestions, but I still think it's really important that we are clear to our readers as well. So it is kind of this interesting development of what is going to be the best nomenclature. And since we are a little bit ahead of schedule it might be worthwhile to spend a little bit of time on it now rather than just through email, since we are not behind schedule yet. Is that okay with you?

CO-CHAIR STEINWALD: Paul and then Barbara.

DR. BARNETT: So Ethan had something he said earlier. He said what were good headings. But I think the one thing we're struggling with is that word "unit," when what you mean is per capita or per episode. And it would be more transparent, isn't it?

MS. TURBYVILLE: Right, that is the challenge, because at that point it's not - and maybe it's a different way to think about it, but that's prior to applying the resource unit, so there is no "per" at that point. It's just defining either the clinical logic, the episode construction would be - which would be like a trigger, whether it's an event or a procedure or a diagnosis, and then the end data for a population, it would be whether based on demographic descriptions or belonging to a certain health plan or whether it's this other in between, which is that per patient, like someone with diabetes, just identifying what in essence is the denominator, and then the third piece was intended to not apply as an implementation, but now that you have that unit or that denominator, which was the way I was thinking about it, and I got pushback on the concept of a denominator, to that numerator, then you say which resources, whether or not they need to

be - have a cost method applied to them, whether it's standard or allowable charges, or some other - and then from there you get your -- exactly as you described, Paul, is it at that point per patient per this, is it a per episode? Is it average, episode costs, et cetera. There are all those steps that that third place, that third bucket has the most steps of all, at least the way we presented it to all of you.

DR. BARNETT: So what is the definition of unit?

MS. TURBYVILLE: It's the core defining features which include the clinical and temporal logic of the claims that identify a distinct and homogeneous - though they may not be completely homogeneous - units for measurement.

DR. BARNETT: But don't you really mean there episode or case or person, patient?

CO-CHAIR STEINWALD: If you do

doesn't the word, unit, kind of cover all of those things?

DR. BARNETT: But unit could be, you know, dollar, it could be emergency room visited. It's so generic that it doesn't say anything, and that's why people are having trouble understanding it.

CO-CHAIR STEINWALD: Barbara and then Tom.

DR. RUDOLPH: Okay, I was going to talk about something else, but it sort of fits in here too. I think that for these types of measures, many of the people developing them do not come from an epidemiologic framework. They come from a different - either economic or business logic or sociologic framework. So numerators and denominators aren't really part of that way of thinking. So I can see where a lot of them would be resistive to that kind of nomenclature for it.

The other thing is that oftentimes

when you do numerators and denominators, you simplify things that are much more complex than that. Like some of the methodologies used for risk adjustment, you really can't even think about it in terms of numerators and denominators; it doesn't make sense. Based on techniques and other kinds of things.

So I just think it's probably best not to do numerator/denominator. And I don't know what the correct - how you would define it exactly, but this more complete definition helps me a lot.

CO-CHAIR STEINWALD: Tom and then Dolores.

DR. ROSENTHAL: I think the thing we're struggling with a little bit is, again back to the question I posed earlier, the presumption is that somebody has already thought, what am I going to prepare to what? I want to compare the cost of glaucoma surgery where I'm going to measure every ophthalmologist in the country or I'm going to

measure heart failure at the individual doctor level using an episode grouper, these units are, well, what are those costs that you would include in whatever the episode is? But it presumes that you've got the episode categorized in your head, and now you are just getting people to put the data together is the way I see it. And if I were you, I'd push back to the pushback.

I mean, here we're all sitting and saying, when you describe what you think we are trying to get them, anybody who would develop one of these things, to do, it makes perfect sense to us. And the only reason it sounds like you are sticking with those title headings is, well, some of the developers didn't like it or didn't understand it.

I'd just push back, I'd push back the pushback.

CO-CHAIR STEINWALD: Dolores.

MS. YANAGIHARA: I'm just wondering if -- I like the idea of bucketing

it, but it seems like different methodologies might bucket different things in different places. So that gets kind of messy. But I'm just wondering, I mean this is really all part of the specification. So instead of trying to bucket it into different categories, I'm wondering if we just say, here are the requirements of the specification and describe - instead of trying to figure out what terms to use to bucket them. Just say, these are the things that we need, and you have to use some terms obviously, but then the description, I think, makes it more clear. So it's all part of the specification. I don't know if that helps or not, but I think, to try to figure out how to bucket it, and what terms to use for the bucketing, it just may be more complicated in the end than just making a list of what needs to be included in the specifications.

DR. HALM: You've used the word "steps" several times. So even just thinking

about steps for measure specification and creation and application may get away from - may get closer to what people really do.

CO-CHAIR STEINWALD: David.

DR. PENSON: I think, from hearing the comments from across the table before, I don't want to use the term "denominator," Barbara. But on the other hand it goes back to something we were talking about this morning, which was that spectrum on left and right per episode, per capita. The first step is deciding what your "per" is going to be.

PARTICIPANT: The per is not -- they don't do it in terms of denominator

DR. PENSON: But effectively it is, because it's a rate. Whether it's for the population, for the - but however you want to word it, that's the first step. Because that defines - I mean the first step in measure specification is what is your `per', for lack of a better way to put it. Because that then

is going to determine everything that goes on afterwards, whether it's risk adjustment, whether you count, whether you monetize, what it is, it's all what your `per' is. I don't know the right term to use, it's categorization, how you are going to -- unit, I don't know, but that's really what we're trying to say here if I'm reading it right.

CO-CHAIR STEINWALD: Paul and then Jack.

MR. BOWHAN: I hate to disagree with my colleague from Wisconsin, but I think it does come down to keeping it simple, because we are directing this toward people who are going to create performance measures, as I understand it. Let's keep it simple: it's a measure specification, and in that specification you need something that describes a numerator and a denominator, and you need the specifics behind that. And I think we are trying to make it too complicated, and each of these paragraphs or

sections almost always talk about defining a population. Well, you are repeating that same kind of language through all the sections. We are making it too complicated, I think.

CO-CHAIR STEINWALD: How are you feeling about this?

MS. TURBYVILLE: Like I did last week.

CO-CHAIR STEINWALD: Well, I do hear something that sounds like a consensus at work. We're talking broadly about the specification of a measure. And that specification has a number of components to it, whether you call them steps or units or whatever else.

Now I heard someone suggest that rather than have three buckets and try to defend them it might be better just to go on to discuss about them, in some detail as you have done, the elements without trying to create and defend that there are three buckets, and here's how they are defined.

Does that work? In other words -- can you live with that?

MS. TURBYVILLE: Is there anything missing in this step? We've talked about what is there, but is there anything that is not there that ought to be?

CO-CHAIR STEINWALD: Go ahead, Ethan.

DR. HALM: I think Tom earlier brought up the sort of all payers perspective, but in the section on inclusion and exclusion there's some nice subtlety talked about, people moving in and out of different plans or payment systems. But I would just amplify that, because there are plenty of people who are just in more than one and stay in more than one. So we have dual eligible Medicare and Medicaid people. We have people in the VA and Medicare.

So the extent to which we really want developers to have all inputs in relating to a patient I might just add a few sentences

about the reality of acknowledging and handling sort of dual eligible individuals.

CO-CHAIR STEINWALD: That makes sense to me. Do you want to move on?

MR. PHILLIPS: Could I, just on the question of whether there was anything missing, I guess when I was looking through here I was trying to find - oh here it is - just as far as the general NQF measure evaluation criteria. And looking through here and maybe this isn't the place that we would touch on these, but the first one on the importance of measuring this whatever it is you're trying to measure, and is that something that needs to be touched on here, as we talked about the mechanics of developing the measure, but we should also include in here something on is this an important area to measure.

CO-CHAIR STEINWALD: Well, that is one of the four criteria - go ahead.

CO-CHAIR LOTZ: I was just going

to say, I'm looking through some of the notes from our earlier conversation. We brought up the issue of perspective and how perspective changes, is that something that should be spoken to as part of this development. Does that change the steps you would take?

CO-CHAIR STEINWALD: Payer, patient, provider.

CO-CHAIR LOTZ: Policy.

CO-CHAIR STEINWALD: As long as it begins with a P.

CO-CHAIR LOTZ: Yes, we're okay with that.

MR. PHILLIPS: I guess my thought is just as the developer going through the thought process, it might be worthwhile just to include something here in terms of the steps, to assess the importance of the measure.

MS. TURBYVILLE: So our approach has changed a couple of times where we have the evaluation, criterion discussions as part

of the white paper, and then where we've broken it apart and it's become it's own separate deliverable. But we are managing the timeline such that any revisions to the current submission form and the evaluation criteria for which the measures will be evaluated by the steering committee and the technical advisory panels will be available to the measure developers before they submit. So that call for measures that is coming out of here, and we are going to talk a lot more about the evaluation criteria tomorrow, and the evaluation criteria will go out to them at the same time. So, excellent point, thank you, and you will have the opportunity to make sure that we are hitting everything that needs to be on those subcriteria.

CO-CHAIR STEINWALD: Well, do we need a discussion and decision on all of these things? Or have we covered - you've got a lot of the details. I don't know that we need complete discussion of all of those details,

what do you think?

MS. TURBYVILLE: We don't as long as you go through it or if you see today anything that's missing or you think doesn't belong on the list we certainly need to know that and provide the group an opportunity to respond to those comments.

What we would be moving onto next - we had bucketed them but now maybe thinking of all of these as components of a resource use measure specification which pieces will be subject to evaluation. So will the measure developer have to submit, and when way say specifications, that is like following a very clear recipe. It's not like your peer comparison is something very vague. So when we are requesting them to submit a specification on a peer comparison or how you would estimate your comparison benchmarker expected that they would have to be explicit, and that you would evaluate them. So there may be things as you work through that you

might want it to be an explicit specification, or options for the users, or what have you, and we need that guidance so we can make sure we get that guidance to the measure developer. So we are ready to move on to that next step.

CO-CHAIR STEINWALD: Rich.

DR. RICH: So as a measure developer, am I going to develop a measure for data or for patients from any dataset? Or do I have to specify the data source? Because if you create a measure that has sophisticated risk adjustment like the STS mortality CABG measure, that has a lot of chart abstraction associated with it, whereas if you have just a resource measure that is coming from an administrative database, risk adjustment can't be as strong within that measure. So are we assuming that a measure has to apply to all populations, or can people submit measures for a specific population, administrative data set only or for more comprehensive data sets?

MS. TURBYVILLE: So I mean they certainly can submit a measure for a specific population. One of the four criteria that the steering committee will evaluate will be feasibility of a measure. However if you think that the data that they are requiring is so narrow in scope that it is not very usable, or at this time perhaps it's ahead of itself, you would weigh that in your evaluation of the measure. So one might specify a measure for a Medicare population and perhaps it wouldn't work as well in another population. That would be part of their specification, and you would evaluate it as such. But you also may look if they specify a population that is so narrow that you are not really sure how useful it will be. So it's a balance in the context that you have put in between usability and feasibility as well and importance to the industry as a whole.

CO-CHAIR STEINWALD: Bill.

DR. GOLDEN: Just a direct

comment, on page 17, lines 478 through 483 or 482, I don't think that - you've made things way too simple. In the private sector some plans such as HMOs that use gatekeepers, which very few do anymore, assign patients for primary care physician making the attribution of patients' resources relatively straightforward.

I would warn you that is a gross understatement. And especially in PPO plans where there is an assigned primary care doc, attribution is still a nightmare. So I would significantly tone that down.

CO-CHAIR STEINWALD: I think the issue gets to usability, feasibility, I certainly wouldn't want to reject a measure just because it requires chart abstraction. Even if it was based on a database where chart abstraction had already been done on a small subset of patients, if the measure seems to have great promise, then it may convince other users that it's worth the cost of chart

abstraction. Of alternatively as electronic health records become more prevalent, then there is an alternative to actually going to paper charts and obtaining the data from electronic records.

Go ahead.

DR. RICH: So then I think when we do our call for measures we should give some specific guidance and language regarding that. Because if I'm a measure developer, I'd like to know that you are willing to accept measures that have a lot of chart abstraction or measures that aren't only associated with administrative databases, et cetera. You may have a different response. Invitation and participation, I think has been said, so you may not want to participate if you think the invite is too strong or too prohibitive.

CO-CHAIR STEINWALD: Helen would like to respond.

DR. BURSTIN: Just to briefly weigh in, feasibility is one of the four

criteria, so the ability to collect data through the routine byproduct of care or administrative data is one component. So it's already built into our process that you can bring in anything. But I've got to tell you, having sat through enough steering committees in the last year the appetite for true chart-based measures unless they are in a very focused area like a registry or something like that is just plummeting.

CO-CHAIR STEINWALD: Just like the value of my portfolio.

(Laughter)

Kurt.

DR. ELWARD: One question, too, while we are talking about EMRs, it's clear to me that a lot of the EMRs aren't designed at all for this kind of data provision, and how are we thinking about how we can make this set of measures interface with EMR manufacturers to say this is what you need and this is how you get the data out of there? Because in

some ways I don't see it any easier on an EMR than I do a paper chart.

CO-CHAIR STEINWALD: Bill or David, you both have your cards up.

DR. REDFEARN: This is a comment starting on 494. You quote the MedPAC article that said the attribution method did not significantly affect physician's resource. There's an Annals of Internal Medicine paper that just came out that says exactly the opposite. This implies that this is easy. I'm more inclined to say it's really hard. This kind of dismissed it. You can't do that. I think there is a lot of credible evidence that it's really hard to do attributions, and attribution methodology does make a difference in terms of the scores and how the physicians come out on these things as the providers.

CO-CHAIR STEINWALD: Bill.

DR. GOLDEN: Comment on page 20, line 543ish, 541, talks about peer groups and

that kind of material about comparing docs. And I would just be cautious here, it may not be true for all specialties, when you start getting into things like internal medicine, probably family practice, others, you can have somebody with a specialty, but somebody who is a hospitalist versus an outpatient doc, somebody is a nursing home doc or a palliative care doc, and the - they get all put in the same bucket. And I was one of the people who got one of the 300 Medicare profiles. It was one of those things, I'm a pack rat, and I had it in my hand about a week or two ago, and I think I threw it away after having it in a pile somewhere. Otherwise I would Xerox and send it to you all. But it was interesting, the data for internal medicine was remarkably flat. The difference between the 10th percentile and the 85th - 90th percentile was about $100 a year, $200 a year. But then there were some interesting tails, and I think the tail at the far end I think were clearly

misattributed docs who had atypical practices. They may have been internal medicine doc who worked for an oncologist, and there are people like this, and I'm not sure how you get around that, but you have to be careful. Not everybody who is an internist is an internist.

DR. BARNETT: That's the fear, correct, not the attribution.

DR. GOLDEN: Sorry, that's - well, yes okay. That's right, the peer group argument. It was an inappropriate peer.

DR. REDFEARN: We even have a name for them - people practicing outside their specialties. Zebras. We find them all over the place. And I tried to do a little statistical work trying to impute specialties going backwards from episode data, trying to impute the specialty. And there are some statistical tools you can use right now that kind of help you do that a little bit. But then at the end result you get this reattribution, reclassification. You say,

well, what's right? I know in California the Blue Shield plan actually hired 20 temps and had them call every one of the doctors in the list and said, what are you. It's not like, how do you want to be listed in the directory. They said, we are going to compare you to a peer group. What peer group would you like to be part of, and ask them point blank. I tried to get our own people to do it in the company, and they said no way, too expensive, and they didn't do it.

DR. GOLDEN: Let me follow up on that, because I've had experience with my Medicaid program, and every now and then somebody will send me a data analysis of outliers, and you look at these docs, they charge them by level five. They're an outlier.

And there are some docs who are charging level fives for head colds, but half of them or more were people who were running a foster care clinic. Where somebody who was

running a tertiary GI clinic, and you almost have to - before you go out and publish these things, all you're doing is identifying outliers, and very often or a good percent of the time there is a reason that they are an outlier, a perfectly legitimate reason they are outliers. And I think we ought to build that in at some point. Outlier status does not necessarily mean that there is something wrong.

CO-CHAIR LOTZ: I want to ask another part of the section that we're looking at here talks about determining the expected - specifically this page we were on before, 20, talks about the observed to the expected but this concept of benchmarking is something we've talked about on our conference calls. And does the committee have anything that it wants to tell measure developers or incorporate into our evaluation tool about how to come up with an expected or benchmark or comparison group. We haven't brought that out

in conversation yet. Are we just leaving it up to them and whatever they provide? Will it stand? Are there some guiding principles or some fundamental aspects that need to be brought out? Because I didn't see them when I read through the paper, sort of the must-have criteria.

CO-CHAIR STEINWALD: Thank you for that contribution.

(Laughter.)

CO-CHAIR LOTZ: Yes, well, just invite me to any party and that will put a quick end to it. No, I mean, I was very concerned about that --

DR. O'NEIL: But being a kind of new area it would be kind of hard - I'm trying to think of where we would get that benchmark from the get-go. I mean I think that's something that over a couple of years you might be able to start seeing if you have a clear definition of these measures. I mean it just seems like unless there is a bunch of

data out there from a particular system that - which there very well may be that I don't know about, I don't know where you would start today.

CO-CHAIR STEINWALD: I've been sort of thinking, maybe this is along the lines of what you're saying, how can the measure developers be encouraged to be innovative if that's what we would like them to be? And we're in a relatively new area, and a health system that is evolving, a delivery system, and so how can we either - if we're not going to provide them guidance and invite them to be innovative in developing their own benchmarks, or benchmarking in a way that hasn't been done before, or even using foreign data for that matter, how do we do that and at the same time say you have to meet specifications in order to satisfy NQF, but at the same time we'd like you to be forward thinking and think about how measures might be utilized, not only tomorrow but five or 10

years from now when our health care system looks a bit different than it does now.

I think that is a dilemma. I'm not sure how to solve it, but it seems to come up in a lot of the individual comments that people are making.

Yes?

DR. BURSTIN: Just to go back to a comment I made earlier, that's a really important point, the reality is these measures are going to be called for in just a few months in October, so if they haven't developed them yet or if they are not in their sort of ultimate testing, they ain't going to be submitted. But I think it's critically important that part of what this steering committee does is say what should be developed and what should be the next generation of measures, while in the interim we kind of deal with what gets submitted in November. I just think nothing is going to happen between now and November in new de novo measure

development. You might for example spark an interest in a developer to maybe modify the way they were thinking about it for submission, but what you are really talking about is informing the next generation. And that's a critical role for steering committees; we really value that. We'll put that out there and hopefully that will then bring the next generation to measures we really want.

CO-CHAIR STEINWALD: I see some cards up. Jack.

DR. NEEDLEMAN: It's not clear to me that at the moment we know enough to be able to specify what a good risk adjustment model looks like or what the right peer groups are. As Helen has said these are questions that the folks who are developing these measures are not unaware people are asking, and we are going to see measures come in that attempt to deal with all of these issues. And I think I'm expecting to learn a hell of a

lot from looking at how they have tried to do it, and what documentation they provide about how successful they are. To be able to critically assess whether I think they have successfully dealt with them.

So Prometheus we already see in the documents we've got a very clear model of what we include in our risk adjustment model and so forth. So we - I think the level of information to the developers at this point is adequate. We're worried about these issues and we're looking to see how you have solved them is about where we are I think in terms of the guidance we can offer.

CO-CHAIR STEINWALD: Lisa.

MS. GRABERT: I think I really agree with the remarks that Jack just made. I think I would feel most comfortable if a developer went through every single one of these steps and then applied to the measure. The reality is, they're probably not going to go through every single one of them. But I

would like to know what they've thought about applying each of these portions of the methodology behind it to their measure, and where they decided to not do something let us know why.

CO-CHAIR STEINWALD: Bill.

DR. GOLDEN: In the zone of limitations, unresolved questions, et cetera, a potential unintended consequence, what happens, what do we do with folks who work at multiple clinical sites with different profiles? So a private practicing physician works in a teaching clinic a half day, two half days a week or spends two days in a underresourced clinic or a charity clinic. They may have very different kinds of outcomes and resources, and they may have less control over what gets attributed to the patients they're taking care of.

Can people opt out? Or are we going to potentially incentivize people to no longer do those activities?

So I don't know how you play with that, but I'm just throwing it out there as something to put on the table as a potential consequence of this kind of profiling.

CO-CHAIR STEINWALD: Mark Kay, one more and then we'll break.

DR. O'NEIL: Well, in this sort of whole new field of resource utilization measurement, I mean basically when we're talking about benchmarking and whatnot, I'm not sure that we have real baselines. I mean these measures are being developed to see how people are doing, which is essentially a baselining operation. It's not really to see quite yet if this intervention changes resource utilization.

Are we at a baselining, or at we at the changing of the system point with this?

DR. BURSTIN: I'm not really sure we can make that distinction. But I do think we are beyond just baselining. I think we want to put something out that people can

begin to use to compare providers. And the question is what comfort zone is there, depending on the level of attribution and the level of comparisons.

DR. O'NEIL: But working on all these definitions of how we're describing these things, this data doesn't currently exist out there; is that correct? Or it does? So utilization data is out there currently?

DR. BURSTIN: There is a lot of it out there.

DR. O'NEIL: There is a lot, I know, but in this sort of systematic way. I mean I know we have utilization data in our system, but it's not generally applicable to other systems. I mean we can say a lot about what's going on in our population, but we are trying to come up with measures that are more generally applicable, right? I'm just struggling with using these things - you can rank, order, if you are counting or monetizing something, you can say this doc is driving all

these expenses for this category of care, or this system is driving all this utilization for this defined population. I'm just - but in terms of benchmarking, because every subcomponent of the system has different data there isn't a benchmark out there that's generally applicable to these measures; is that correct?

CO-CHAIR STEINWALD: A single benchmark, yes.

MS. TURBYVILLE: I think what we were thinking is the measure developers as part of their submitting their specifications would either have to describe to the user how to estimate or create a benchmark, and if they had some external benchmark available they would have to make it available.

CO-CHAIR STEINWALD: All right, Barbara, and then Jeff, and then we break. No more cards.

DR. RUDOLPH: Just a statement on the history of some of these groups, they've

been doing this a long time. I know I went to presentations back in 2003 by a variety of vendors who were already doing this kind of work. So they are going to be ahead of where we are in terms of their experience with this. That doesn't mean that they have done everything right, because they probably haven't. But I think anybody who will have had to be doing this for awhile in order to meet the submission requirements, for testing, reliability, validity, all those kinds of things. So I think we are going to learn a lot from what they are able to tell us, and many of them have had access to physicians' claim data for a long time.

CO-CHAIR STEINWALD: Jeff.

MR. CURTIS: This may be just restating what people have said already, but I think the important thing is that it's the expectation that the measure developers will specify how it will be implemented, which I think is what this is getting to. How will

outliers be identified? How might this information be conveyed? And I think at least from my work on the outcomes measure, that is something that developers are oftentimes on purpose - or purposely vague about, and this is trying to get them to put their nickel down.

So I think we don't need to set the benchmark for them. We just need them to tell us what their proposed benchmarks are, so we have an idea if there is face validity to it.

CO-CHAIR STEINWALD: Very good.

All right, let's break until what time?

MS. TURBYVILLE: 2:40?

CO-CHAIR STEINWALD: 2:40, okay, see you then.

(Whereupon the proceeding in the above entitled matter went off the record at 2:23 p.m. and resumed at 2:49 p.m.)

MS. TURBYVILLE: So as I've been

saying throughout the day thank you again for all your thoughtful inputs thus far. We think at this time we can move onto the next section which is Section 4. It is still labeled Section 5 in the white paper, I apologize, starting on line 621, unless there are any other absolute must-share items for the sections that we've been talking about, not that there won't be future opportunities to circle back. But I know we kind of abruptly stopped for break. So are we all ready to move on to the next section of the white paper? Please?

MS. YANAGIHARA: Did we determine which of the aspects are going to be part of the purview of this group and which are not? I'm not really clear on it if we did?

MS. TURBYVILLE: No, because we are going to discuss that more again. So we're not getting away from that, and then we are going to open that conversation up later on today, and then as we go through the criteria,

kind of rethink any thinking we've had before that to make sure that we are meeting the needs of this measurement effort.

So I will go ahead and hand it back over to our co-chairs.

CO-CHAIR LOTZ: All right, we're going to discuss Section 5 - it's labeled differently in different places, but it's Section 4 or Section 5, and this is talking about limitations, implications, unresolved questions. Sally has written a few things on the slide here, but open forum. Barbara, did you want to say something already?

So again, analogous to our prior conversation, much has been suggested based on the emails and calls today about what some of the limitations might be. They are baked into the white paper. Did we capture them correctly? Is there something missing? Is there something that needs to be amplified, clarified? What kinds of limitations will we communicate or what kind of considerations of

limitations do our measure developers need to communicate to the steering committee when they submit their measures?

` Jim.

MR. WEINSTEIN: I was just curious, in this section I don't know if this is part of the black box methodology section as well, but I didn't have those specific references that you had in Ingenix and Reuters and didn't know if the measurement developers will have access to those kinds of things or not.

CO-CHAIR LOTZ: Jim, what kinds of things? Could you just elaborate?

MR. WEINSTEIN: You have in here - I'll get the page number - on page 27, anyhow, the reference is 17 to 18.

DR. REDFEARN: Ingenix at least it's a website, once you go into the website all you do is agree not to steal their proprietary information and the entire documentation is available; it's completely

open. I haven't looked at the Med Stat one but I assume it's similar.

MR. WEINSTEIN: So the question is, are we going to require people who work on measures to provide that kind of detail.

MS. TURBYVILLE: So yes, anyone who wants to submit a measure for endorsement, is that what you might mean?

MR. WEINSTEIN: Yes.

CO-CHAIR LOTZ: Yes. The answer is yes.

Paul.

DR. BARNETT: Actually I think David is in front of me.

CO-CHAIR LOTZ: All right, David.

DR. REDFEARN: Just the discussion starting on 636 about standard error of a mean. I might be really confused, but I think there is something missing here. I think the variability of the mean for the physician or the provider that you are comparing is also important to the confidence

interval. It's not just the standard deviation of the norm; it's the distribution or the variability. So I just would - maybe I don't understand the calculation, but I always think about it as terms of the variability of the physician being compared, not the variability of the norm or the peer group.

CO-CHAIR LOTZ: Paul.

DR. BARNETT: So one of the problems of all this, I notice in the description of the Prometheus they talk about adjusting for regional variation. And so there are these regional level variations in practice that are quite profound, and it doesn't seem like something we want to adjust out. We want to ding them if they are in a region that's bad and get everybody in the region to do better. But that is part of the evaluation problem is to detect what's regional and what's not.

CO-CHAIR STEINWALD: My response

to that is, everyone is favored; it depends. A lot of the measures so far in use of resource are confined to metropolitan areas. And part of the reason for that is lack of confidence in being able to compare across metropolitan areas.

And so you might - a lot of the measures are, well, Dr. X compares to his peers in Phoenix; he's 20 percent above. And Dr. Y in Sacramento is 20 percent below the average for the peers. But the median may be very different in those areas.

And I think that's kind of been the evolution of the measure development. If you are taking small steps before big steps you do the comparisons within geographic areas before you then try to compare across. At least that is my observation. Not that we shouldn't eventually want to compare across. But I think again going back to the measure developers and their specification and justification for the measure they need to

tell us whether this is a measure that can be used within geographic areas or it could be used across, and the criteria might be a bit more stringent for comparing across.

CO-CHAIR LOTZ: Barbara.

DR. RUDOLPH: I guess I was wondering if we should talk at all about using different Bayesian methodology to address issues like small cell size, and whether that is going to be acceptable as understandable to the physicians being compared. And if in fact it is going to be acceptable, then what types of information should the measure developer provide that would support that use.

CO-CHAIR STEINWALD: CMS, when it was developing its physician feedback protocol addressed that issue, and I don't know all the details, but there was a tradeoff between providing all the information that you would want, that someone might want to know right up front and making it look very complicated versus keeping it simple and understandable

but not satisfying the people who wanted all the details.

So I think what they did is they used drill downs. They have a report, a feedback report for a physician that gives the basic information on how you compare to a peer group. And then for those who want all the details of the methodology whether it involves Bayes theorem or not, if it's an electronic report, then they can drill down or they can call the telephone number of the expert who developed it and actually have a conversation.

And so I think there are I'm sure many users and maybe even some developers who wouldn't feel comfortable having to explain the methodology to every person or provider who is likely to be compared as a requirement, but as an option, maybe so. Does that make sense?

Others have more experience with these measures, that issue I know gets

addressed all the time, I think. Go ahead.

DR. ROSENTHAL: Well, when CMS measures hospitals and looks at things like congestive heart failure and mortality outcomes, they end up with quite a number of the measured entities not being statistically different. I don't see any problem with that, if you end up with a big chunk of the measured being not statistically different, you have a little bit around the small N problem. The other way to approach it is, let's not pick something that has a small N.

But if you do you are going to end up with a bunch of them being not being statistically different, and why is that not okay? Maybe all we're looking for is to define some very high and low end outliers, and again it gets down to what's our purpose in the measurement.

CO-CHAIR LOTZ: David.

DR. REDFEARN: The only comment I'd make is that I think you need to do

statistical comparisons when you are reporting this kind of data. I think that is kind of a requirement. What we have found is that confidence intervals presents in a way that it's more understandable than using some of the fancier statistical techniques. It seems to work better; it seems to be understandable.

And I'll tell you frankly that in our work we have - I generate three classifications of a physician: efficient, don't know, and inefficient. The "don't know" category is real big. And it's a byproduct of doing a 95 percent confidence interval on the data, and the small sample sizes, and the extreme variability you see in the performance of the doctors because of presumably changes in medical practice, variability in the underlying patient that we have not measured; severity differences that we can't measure; all those kinds of things. But we get a big "don't' know" category.

And in terms of reliability across

time, you see people moving between the efficient and don't know and inefficient and don't know. You very rarely see somebody moving from efficient to inefficient across time. So it's a fairly conservative approach. And I've presented this to a couple of medical groups in California, and they seem to get it. So that's just personal experience.

CO-CHAIR LOTZ: Tom.

DR. LEE: I think that the way this plays out in real life is that you end up with situations where you have a small hospital or small volume practice that seems way off the average, but they are described as statistically within the expected range. And then you will have a bigger hospital or a bigger practice that actually has performance which is better or worse. Because they will be an outlier, because they have - a statistical outlier because they have more volume. And people will go, hey, wait a minute, we're better or worse than them, but

we're statistically significantly different.

But the price of not using statistical analysis is too great. You just get clobbered. So you just have to explain that these little outliers are going to be classified within the statistical norm because of statistics. Because the alternatives are worse. That is our experience in reporting in Massachusetts so far.

MR. CURTIS: Let me just follow up on that because we bump up on this continuously with the outcome measures that we develop with heart failure and pneumonia, and in terms of the choice of presenting it. Obviously there are pros and cons for hierarchical or regular logistic regression in different approaches to the statistical modeling. And I think everyone agrees that you need to have some form of risk adjustment.

The problem is and one that we've run into time and time again now is that there are such extreme opinions on it yet there is

no external truth out there that this is the right way to do it or not. And I know that NQF has continuously struggled with this as well, but I think what we are going to see are a huge range of different approaches, and unless we establish an external gold standard for better or worse, it's going to be very difficult to make good comparisons across the validity of these different measures and different approaches. This is a larger issue than the steering committee here, but I think NQF really needs to develop a format or a form by which we can get to consensus on this, because it's I think tearing outcomes measures apart.

CO-CHAIR LOTZ: Jack.

DR. NEEDLEMAN: On that point I think one of the issues that we see is that people have adopted a specific method, and the impact is somewhat blinded. So CMS when it's doing hospital compare uses a Bayesian shrinkage model which moves people - small

places get their numbers averaged with the mean based on the relative size. So we start out in all of these methods with some raw data. Including standard errors around estimates and standard errors around individual estimates for subunits that have numbers. And then things get done with them. So I think at this point since we don't have a gold standard, and we don't have an agreed upon method, it will be helpful to understand what the raw numbers are that emerge from these systems, and how the statistical adjustments that the methods developers prefer change what is being reported from raw. So if we got Bayesian shrinkage, let's see what the original number looks like and what the shrunk number looks like. And so in the face of this, rather than our having to decide ex ante what the right method is, I'd like to see us get enough raw data from the folks who are submitting that we can understand how their methods change the interpretation of the

analysis that is being presented out of their measure.

CO-CHAIR LOTZ: So a requirement to provide both basically.

Tom.

DR. ROSENTHAL: Well, it seems to me ultimately this is not a trivial philosophical point in regard to how the information is going to be used, and the example I would give relates to some of the AHRQ quality measures that are now getting incorporated into practice, and one of the people in our system has been heavily involved in the validation of those, gave a presentation to our group recently and said, well, this particular AHRQ measurement has been validated to 50 percent reliability. And was pleased that this was going to be put into public practice. And my response was, it seems to me doing something that is potentially significantly wrong is actually quite harmful, and that we should err on the

side of actually knowing something, and knowing it pretty affirmatively before we start using it in pretty substantial ways either for public reporting or for - and now we are talking about money. So for payment differences, and that's where I come back David to your comment of, you know, I understand the Bayesian methodologies and you can manipulate things. But I think we want to err on the side of being affirmatively right and being able to demonstrate that it's affirmatively right than being able to come up with something that by some manipulation we can jigger into something that may be statistically valid.

So I think this is a not unimportant philosophical debate that we will likely see played out as we get these thing in.

DR. RUDOLPH: If that were Bayesian they would have taken out the random error, so it would be on the positive side at

50 percent correct, after they removed the random error you would be in the positive zone, so you would know something more than just a guess.

CO-CHAIR STEINWALD: One criterion that might be applied, I am getting at the sample size issue is, if there is a difference between two groups or between a group's performance and a benchmark, that is economically significant, then the sample size and the power that goes with it ought to be able to detect that difference as being statistically significant. One often uses the same criterion in clinical trials, a clinically significant difference ought to be statistically significant. And in this case since we are developing resource measures, I think the criterion would be economically. So if there is a difference in the resources associated with hospitalization of $10 between two groups, that's not economically significant. But you as the user need to

decide well, what is economically significant? Is it $1,000? If that's what it is, then our power including the sample size ought to be good enough to detect that difference as being statistically significant.

CO-CHAIR LOTZ: Lisa.

MS. GRABERT: I can sort of let you know what methodology we use for the physician resource use measurement program when I was at CMS for a minimum threshold, or to get around small N. We had a minimum threshold for each and every benchmark that we had, and the benchmark was defined as specialty condition and geographic area. And if you changed one of those three factors, you'd have a different minimum threshold. So each time you had to calculate a new threshold. The rate of reliability that we used was point five, and we felt comfortable with that for purposes of confidential feedback reporting.

For other levels such as public

reporting and possibly payment, the rate of reliability we always felt needed to be much higher than that.

CO-CHAIR STEINWALD: How did you determine where the threshold was? If each threshold was different how did you determine the threshold?

MS. GRABERT: As long as you hit a rate of reliability of point five, depending on those three parameters of the benchmark, that would define what your minimum case number had to be. So for example for any given specialty in a given geographic area, for just diabetes, your minimum threshold to hit a reliability of point five from everyone underneath that benchmark might need to be 25 cases. But if you were to change the geographic location to a different part of the country, because there may be more variation in that geographic area, your minimum caseload - minimum threshold might be 55 cases. So it changes depending on what your benchmark looks

like.

CO-CHAIR LOTZ: Jim.

MR. WEINSTEIN: It's a little interesting how you determine what your meaningful clinical difference is. Given that we don't have data that allows us to understand what change we are looking for, to calculate some sample size et cetera. I think it's a good discussion, but I worry about the implications. Who is going to determine what is meaningful, what a meaningful difference is? Is it geographically defined? Is it individually defined? Is it case defined? And then you are talking about dollars, in this last example. Geez, I don't know is $100,000 too much for a transplant difference? Is it $50,000? In the Bayesian models what's cost effective? Is it $100,000 in the U.S. system? Is it the U.S. Euro qual versus the European Euro qual measure? This whole area of measurement, for all of you who are better at it than I am, is not so simple. And

depending on the measure du jour and how you use that you can find an answer. The nice thing about Bayesian is you do some sensitivity analysis, and sort of look at the model, and you do some bootstrapping and things like that to help. But I guess I just - I don't know enough about resource utilizations to say what is right for a given diagnosis.

There is a commonality of resource use that we'd like to look for, but then with some outcome, again, how is that determined what is good or bad, and what's that change score? So we all realize around the table the limitations, and we realize the altruism of what we are trying to do. But I think not stating that is problematic.

CO-CHAIR STEINWALD: Well, would it be sufficient from your point of view just to require that the measure developer address that issue, rather than us trying to say here is what the threshold is?

MR. WEINSTEIN: Yes. Yes.

MS. GRABERT: I think so, because I'll just add, in order to do what we did with CMS, we had $12 million and several FTEs, and I think that that is a level (laughter) to expect from a measure developer that may be too high.

CO-CHAIR LOTZ: That's Helen's next budget for the steering committee.

Ethan.

DR. HALM: I was just going to add, since we see the resource use as a building block to another committee who is going to worry about efficiency that small differences in money, in the denominator, or even the same amount of resources spent, when you start looking at different outcomes, you might in fact be from a patient's or society's perspective become big differences. So I would stay away from trying to specify that, because I think in a vacuum it's not going to be as useful.

CO-CHAIR LOTZ: Mary Kay.

DR. O'NEIL: I was going to make a similar point to look at what the significance of the different amounts being spent without an outcome is a little big dangerous just to rank order people. In our system we will look at characteristics of physician practice not just by how much they're spending at a particular time but the impact on the cost to the system or the individual over a longer time frame. So that if you are seen by a physician with a particular cost and utilization profile, that will predict the trend of cost for somebody with that diagnosis. And it's a much longer timeframe for evaluation.

But I was a little concerned about some of the language in this section about making sure that the measures all have the performance of the individual physician as required for the measures, because I think that the way things are going with health care

into accountable care organizations, and the organization of care, and the variation of what kind of services are available if you go to different organizations, and I think those differentiations are increasing right now, to require that every measure starts with just the individual physician's performance is a little bit limiting, and maybe not exactly really where we want to go. And then of course if you are talking about individual physicians you get the small N problem.

So I think that there are some important characteristics of practices that really change outcome even if somebody is being seen for maybe a limited service, just because certain systems prospectively look for different characteristics of patients and address them, and that is increasingly emphasized in how people develop their practices.

CO-CHAIR LOTZ: Helen

DR. BURSTIN: This is more a

question than an answer, but it sounded like part of what we were also hoping to do as part of this white paper was to also identify what were the unique issues around testing these measures. And it sounds like it might be useful for this committee to come up with a list of the key things you'd want people to report on in terms of the testing of the measures to get at some of these analytic issues.

The other thing I wanted to mention that I think is also an important question is, one of the things that also kind of plagues us at NQF and probably in the real world as well is the issue of when does the measure specifications end and implementation guidance begin. So there is oftentimes a lot of discussion about you may for example put an implementation guidance, sample size, statistical significance, things like that that aren't necessarily uniquely part of the measure specification itself. I think it'd be

really helpful for this group to also help us think through when do some of these issues become - this is core to the measure versus these are sort of ways you can potentially vary it as you use it in practice. Things like level of analysis would be baked into the measure because scientific reliability is so dependent on for example where you would look at that. But things like sample size for example, sample size is sometimes baked in, not always. But there are many issues in terms of the way the data are actually reported out, for example using stars or above or below or things like that, choosing a level of statistical significance that sometimes are actually outside of the measure specifications, and more so in the realm of how the end user uses the measure.

So this would just be helpful, just a series of questions, but things that I think would really be useful for us to try to get a handle on before we even go out with a

call for measures here.

DR. GOLDEN: You wanted ideas about testing. Two ideas here. One would be, I think you want to document that there really is variation; no point in having the measure if there is no variation. Or then the question is, if there is variation, then how much variation? So if you profile 500 people and you find out three person variation, is that meaningful?

I guess the followup to that, something we mentioned earlier, have you looked at who the outliers are? So are the outliers true outliers? Or did you actually analyze the outliers and find out that there was a reason for the outliers?

CO-CHAIR STEINWALD: I'm thinking of, Helen, what you just said, you've got the measure and then we've got the use of the measure. And the question is what is required of the measure developer to provide guidance to the user? And that is where issues of

small N and outliers and other things come in. But I think it's a good question though as to how responsible is the measure developer for a potential user's misuse of the measure. And I think it's a good question, but I don't have a good answer personally.

CO-CHAIR LOTZ: Going back to the way Helen teed up these two questions, issues of testing we've talked earlier today about multiple sites of service. We've talked about limitations of administrative data sets. Do we want to bring any of the aspects of those conversations into this part of our day to include that in an emerging list of guidelines to developers?

Bill, is your card up still? Kurt.

DR. ELWARD: It might help to have them submit known issues that they would anticipate, some of which have just been mentioned, but potential problems with the measure, potential variations in how well it

would work in different settings, ways a measurement might actually vary based on the setting which is being evaluated. I think giving people some idea initially that they have thought about these things and they can anticipate certain problems in various settings I think would be very helpful.

CO-CHAIR LOTZ: Tom Rosenthal. Oh, I'm sorry, Tom Lee.

DR. LEE: I was just going to raise a question: is it too high a hurdle to ask them to actually analyze reliability the way the RAND folks did in the New England Journal paper? I mean I know it's actually a hard thing methodologically to read, but the actual formula itself is not that difficult. That might be actually something that could help people put measures in perspective. They need real data.

CO-CHAIR LOTZ: Sally is thinking; she's just not thinking out loud.

What about that is making you

uncomfortable, Sally?

MS. TURBYVILLE: I guess my question would be, is that the only way, or is that a suggestion that respected colleagues have produced as an approach to examine reliability but that there may be other avenues. So being - it's rather new. We would be going out for a call for measures in October. Is that something that if you have done this this might be a way to demonstrate reliability but not be so narrow? That was my discomfort reaction, kind of thinking how new it is and how much people have actually had a chance to practice it in the real data arena. But that is not to say no.

CO-CHAIR LOTZ: All right, well that inspired a lot of conversation.

David.

DR. REDFEARN: That particular measure of reliability is fairly controversial. Ingenix for example says it's absolutely the wrong measure to use when you

are using this kind of methodology to evaluate performance. So I would say that is controversial. I think what Sally was sort of implying is that that is not a standard; that's a suggestion. Perhaps a suggestion that the developers consider statistical reliability in terms of analyzing their results would be appropriate without being terribly specific.

I personally, I think I agree with Ingenix, I think a statistical test is a better fit for doing this kind of evaluation, so I would - I don't think reliability works in this context, but that is an opinion.

CO-CHAIR LOTZ: Steve.

MR. PHILLIPS: This is probably I think more of a question than a comment necessarily, but the idea of linking with electronic medical records has been mentioned a couple of times and I guess this discussion makes me think of it. If I recall it seems like there is an effort within NQF to try to

look at the issue of measures that are - can be incorporated within electronic medical records, and I was just wondering, does that overlay with what we are talking about here? And maybe it falls within the just the general evaluative criteria which is I understand going to be a separate document? But I was just curious about that question.

DR. BURSTIN: It's an interesting area, just because so much of the data here is actually claims based, at least in terms of, on the cost side. So I guess the real issue is getting at the interoperability issues, how do these data relate to perhaps some of the richer clinical data you could pull out of an EHR for risk adjustment to get a better handle on the patient population. I don't think we know yet, but I would suspect if you think about the meaningful use trajectory, these are likely more like 2015 rather than 2010 or 2013. But I think we will wait and see.

CO-CHAIR LOTZ: Tom Rosenthal.

DR. ROSENTHAL: A couple of quick points. On this business about the acceptance of these things, given that there is almost no science, at least not any peer reviewed science, unlike when we started on the quality measurement realm there was at least some peer reviewed basis that you had to start out, coming back to the idea of encouraging the developers, I'd still put out the idea that maybe there are two levels of acceptance. Level one would be, it's not been validated. There is not statistical proof. It has not been tried in populations, but boy it sure sounds intriguing and likely to get us to something in 2015, and the sort of phase one endorsement by this group would enable those entities potentially to get grant money and other sorts of things to actually do the science to get to the next phase.

So I'd put back the idea that there might be sort of two levels of acceptance where there may be other things

that are perfectly ready for primetime and ready to roll out.

And the other point that I wanted to come back to was the thing that Mary Kay said a minute ago that I wanted to be sure - I didn't hear any validation of, which is that even in this white paper document we are not going to insist upon physician level attribution; that we would in fact solicit a wide swatch of potential attributions, possibly hospitals, possibly individual physicians, possibly physician groups, et cetera. And I didn't hear any validation of Mary Kay's point which I think is really critical.

DR. STEPHANSKY: Hear hear on that.

CO-CHAIR LOTZ: David, your card is up as well.

DR. PENSON: So I wanted to build on Tom's comment and get back to David's comment, the issue of reliability. And we

keep dancing around validity. And if you look at the criteria which NQF has used for quality measures, validity has always been there. There is even a line, and I pulled it up because I've been on these TAPS as everyone else has, if face validity is the only validity addressed, it should be systematically addressed. I think you have to have validity for these resource use measures, and I do think you have to have reliability. It doesn't necessarily have to be what the RAND group used, but there has to be some sort of assessment. We may be early here. But I think Tom's point is very well taken. And I think that actually it is worth explicitly saying that validity has to be assessed and reliability should be there too. I'm not hearing that. People are saying reliability doesn't have to be there.

CO-CHAIR LOTZ: Barbara.

DR. RUDOLPH: I was just going to say that I wouldn't want to prescribe the

exact tests of reliability or validity, but it certainly is part of the endorsement process, and if you didn't have reliability and validity the measure would be unlikely to pass.

CO-CHAIR LOTZ: Bill Rich.

DR. RICH: Again this addresses some of my initial concerns that the way the white paper is constructed gives us no way to evaluate how effective these things are going to be. And I think we may have an out if we have some direction for implementation and put it up front as one of the criteria, some measure or intent to look at reliability and stability. And again I don't feel confident, as David pointed out, in specifying what that should be. But I do think we have to address either up front or in the directions for implementation that there be some measure of the stability and reliability of the measures.

CO-CHAIR LOTZ: Barbara, did you go back up again?

David? Mary Kay?

DR. O'NEIL: We've come around in a circle from this morning. If we are talking about the resource utilization building block of really counting resources that have been used, we do have to have a reliable way of counting those things, but I think we get a lot more nervous about counting it when we start rank ordering docs by how many resources they are using, and then we get nervous by rank ordering them without a outcome of product of quality or medical outcome at the other end.

I guess being in my end of the business, I feel like we can count resource uses pretty reliably, but I guess the other reliability part is the significance of any of this or the attribution of it to a physician implying clinical decision making and quality of practice and all of those kinds of things. So the counting I think we can do, but if it's just that building block and nothing else,

then it's hard to figure out what the significance of that activity is.

CO-CHAIR LOTZ: Bill Rich, are you up again, or are you still up from the first time?

CO-CHAIR STEINWALD: It's fun to do that.

CO-CHAIR LOTZ: I haven't tried it yet.

CO-CHAIR STEINWALD: I've been trying to figure out how we could implement the spirit of what Tom suggested about sort of different level of standard without actually putting it that way. Because again if we want to be inviting and encouraging to measure developers to develop measures that are new, in essence, it seems to me that the standard that one would apply should be somewhat different than the standard that one would apply if we are talking about measures that are incremental improvements in the technology that has existed for years. But

how we present that to measure developers in a way that doesn't make it seem like we have a double standard, because I know NQF doesn't have a double standard, I think is a bit of a challenge. But I would hope we could find a way to do that. And any suggestions for the staff on how to present that I think might be very helpful.

Well, let me add to that. One of the four absolute criteria than NQF has, importance is first, scientific acceptability is second. But scientific acceptability might have some wiggle room to it.

CO-CHAIR LOTZ: David.

DR. REDFEARN: The only thing I would suggest is that you could, without being specific about how the evaluation was actually done, you could ask that the developers do some sort of separate sample validation on their method, or split half validation. I mean to get to this point of reliability, you can think of reliability in a general sense

is: if you take a doctor and you take half his cases and run your methodology and you take the other half and run your methodology and compare the two scores, and you would expect the scores to agree.

So in a general sense of reliability that is absolutely a standard for what we have, and I think it's not unreasonable to ask these developers to document and demonstrate that level of reliability using the data. Any of the developers that I think are going to come to this and are going to propose this are going to be sitting on tons of data, developmental data, huge amounts of data, so it's not ridiculous to ask them to go back and document this with their sample data they used to develop it. I mean that's how you do development: you develop the model, you hold back some samples, you run your model again and you see what happens to your R squared. Does it stay the same or does it go down? And

I mean that is routine in terms of the developmental process. And when they do a developmental process they should be able to document what happens when they do that. I don't think that is unreasonable.

CO-CHAIR LOTZ: Kurt.

DR. ELWARD: I would agree. And as much as I think we'd all like to imagine it wouldn't happen, I think the reality some people if they developed a measure that is particularly to their advantage will use this for commercial advantage. There is a lot of gamesmanship that can be done. I think if they - support them being required to go back and say this has some validity. Or if they don't, but if it's a promising measure, they should at least be able to say, we have this plan of evaluation. So that at least we have an evaluation plan in place. So we know that if this seems to be a very very promising measure that there is some kind of - we will know eventually whether it's valid or not.

CO-CHAIR LOTZ: Dolores.

MS. YANAGIHARA: This is a question for NQF. I seem to recall that for the quality measures that there are two levels of endorsement, is that right? There is like a full endorsement and then there is a time-limited endorsement for those that haven't been kind of tested in the real world? Is that the differentiation? I mean could that sort of differential apply and be what Bruce is looking for in terms of kind of different levels of endorsement or whatever?

DR. BURSTIN: To date we do have a time limited endorsement category. It is really intended for measures that otherwise would completely pass all the NQF evaluation criteria except for the fact that it hasn't been adequately field tested. There is actually a testing task force that just put out a draft report on our website currently that really much more specifically outlines what we really mean by testing reliability and

validity.

But I think that in the last year or so there has been increasing discomfort with bringing in untested measures, especially untested complex measures. So in May the Board of Directors narrowed the scope of what could be brought in under time limited measures. It has to be an area where there is clearly a gap. It has to be where there is a legislative mandate or a clear need to bring in untested measures because of a quick need to get measures in. And the third thing is they can't be complex measures. And I think many would argue these are fairly complex, risk adjusted, things like that, composite. So I think - and again I'd be curious to potentially get CMS' perspectives on this. But I think the idea of an untested resource use measure getting through the process I think at this point are low.

CO-CHAIR STEINWALD: I don't know that anyone, even Tom, or me, was suggesting

untested. The question is what standards of testing might you apply to different methods at different levels.

DR. BURSTIN: What might be useful the testing report is still - actually closes for comment tomorrow. There is a very nice table of their proposed evaluation ratings of high, medium and low for testing for measures, that I think might be useful to share for the committee for our discussion tomorrow. And the question would be, how would you look at this in light of these kinds of measures? Do you have to modify this rubric slightly? But it really gets into the level of sort of testing both at the data element level or the measure score level. And I think these seem like the kind of measures that really gear toward measure score as opposed to testing of the data elements. But we'll have to - we'll share that with you tomorrow as we sit down and think that through.

CO-CHAIR LOTZ: Go ahead, Tom.

DR. ROSENTHAL: Of those criteria that you listed, the one that would seem to offer some possibility here is the notion of gap. Because in a way there is a gap entirely here. And it seems to me part of the gap isn't being able to just do the statistics. It's to have a group of people who have given some deep though to the problem and say, right now I've got this array of numbers. Does this really tell me something about the difference between this physician and that physician, or this group and that group, and that group and the other group. And I think that's where there is a tremendous gap, with the exception of a few of the health plans and maybe a little bit of the Medicare demonstration project, I don't think those - I mean there are value judgments implicit and explicit in that. But it may be that we are going to have to do some of those. But until you do those, you don't want to have the thing out there as

a fully endorsed measure. So I think that is where it maybe even not as much a statistical methodologic problem as it is a value judgment problem in assessing the thing for, is this really likely to play out in real life as being a meaningful difference between these doctors' practices.

But the gap piece would maybe - the complexity piece is what doesn't give us any room in that formulation.

CO-CHAIR LOTZ: Paul.

DR. BARNETT: I'm a little dismayed by the direction we are heading here, because I think that this whole topic of efficiency in health care is a political lightning rod, and that anything we do that is half baked is going to be a real problem. And that what we ought to do is find a few things that are extremely well validated and start with that. That's all.

CO-CHAIR STEINWALD: Well, okay. You know if we truly believe we are on an

unsustainable path and we need to get control of health care spending in this country, and we want to do it in a way that is as efficient as possible, I don't want to ignore the politics, but I also don't want to acquiesce to the knee-jerk reactions that often occur, that we are going to have rationing, we are going to have death panels if we develop comparative effectiveness data. I think we can be above, or if not above, in a different place than that political discussion. And again I wouldn't want to discourage a measure developer from being somewhat innovative because they would worry that their measure would immediately be put in a political context and trashed if it weren't completely validated in an - to the extent that expectations would be unreasonable.

Now and so for example if someone were to develop a measure of - a resource measure that was suitable for let's say medical home, and yet we don't have a whole

lot of medical homes and the ones that we have are different from each other, maybe the measure developer can only test it in one or two sites, and yet we think that the concept is going to grow and develop nationally so that having a measure of resource use for medical homes would be a very useful thing to have now, but especially five years from now, I think we would want to encourage measure developers to go that route even knowing that their ability to completely validate the measure might be very constrained.

I don't know, what do you guys think?

DR. PENSON: You know I think we can learn a few things from the experience with quality measures. You don't want to stop people from doing this; you want to encourage innovation. But let's be honest: a lot of our quality measures, which are process measures, are nonsense. They are not tied to outcomes, and while they have some face

validity, when you look at it in the end when you look at it a lot of them are kind of meaningless.

So Bruce, I think we have to balance it, and I think there is something to be said for saying, you have a responsibility with this measure. We don't want you - you don't have to come here with a full statistical validity reliability assessment, but on the other hand you have some responsibility to show that this is a meaningful measure.

I'm concerned that if we just sort of say, well, it seems like it might be a good idea and it might be a great idea in five years, and we let it in the door, then we'll end up with nonsense. And I do think a lot of the quality measures we have now, the process measures, are problematic. So I think we can learn from our prior errors.

CO-CHAIR STEINWALD: I'm completely with you on the meaningful measure.

And I guess where I am though is not wanting to have unreasonable expectations, especially for the development of innovative measures that are not sort of incremental improvements of what already exists.

CO-CHAIR LOTZ: Jim.

MR. WEINSTEIN: I guess there is lots of data out there, that I know we have done and others have done, predictors of utilization as just sort of what people have been doing for the last 10 years. And if you look at what they are going to do the next 10 years, you probably can project that with pretty good confidence intervals and reliability and Bayesian modeling. I think that - I want to agree with you that I think we need to go out a little bit further on the limb here, but I do think there are some groupers that we can look at that are fairly simple compared to others and do the experiment, whether it's like the pay for performance activities that use mostly process

measures. And we were able to do 2 percent better than others and get some benefit and some cost sharing, or shared savings. I think that is an interesting experiment.

I think the same thing can happen here, and I don't imagine that we actually know the measures. But I know that in large databases you can find predictors of utilization of resources by various provider groups all across the country, and we have seen that over and over and over again. The question is, what are we going to do about it? So now we have the models. Now we know what the resources are. Are we going to change the payment system, which does get into politics? And that's what people are afraid of. They are actually afraid that you are going to use that information to change their payment structure, and therefore not have the same number of resources to spend for the resources that they have been using.

So that context is out there. So

let's not shy away from it, but let's take some pretty good examples where we can get fairly confident measures of use of resources, and yes eventually tie them to outcomes so that we can get to the value equation, and yes eventually change the payment system. Because unless we are going down that pathway, I'm not sure why we are sitting here. The fact of the matter is, we do have to change what we are doing. I think the voice at the table is, let's do that in some rational way with some experimentation that suggests by the providers that are in those experiments that that does make sense, and yes we can do it better and differently.

CO-CHAIR LOTZ: Bill.

DR. GOLDEN: You know as we evaluate these materials that come in, we might even want to, over and above passing a judgment or commentary on a particular measure or measure set or developer, look at their methodology and actually take some lessons

learned from the methodology in general, and since you are from Dartmouth I can beat up on Dartmouth. When you looked at the payment models, one of the Dartmouth model as presented at a meeting I was at talked about using historic costs and taking a percentage of the historic costs. Which made me very nervous, because if you were a traditionally over-utilizing region or you were a traditionally - you were an early adopted of technology, you got locked into that pattern as opposed to somebody who hadn't adopted the technology. And so I was kind of very uncomfortable with that approach.

Now that would mean we could have a discussion about do we like that measure with historic costs as the basis. But we might even want to pull that out and just discuss the notion of historic costs as a method that has its positives and negatives, and perhaps inform the rest of the community about when they design further measures.

CO-CHAIR LOTZ: Dolores.

MS. YANAGIHARA: One comment, and one question.

Comment is, I think we would be better served to have fewer really usable tested well vetted measures than a bunch of ones that are maybe not quite so solid. So I just think we need to make sure that the science is behind it.

The question, though, is, because I also agree that we want to encourage innovation, how - so we've got this initial call for measures. What happens next? So the developers that aren't ready to go out in October with their measures, when is the next time that they would be able to submit? Is there a certain cycle? Is it ongoing? Because if a measure is not ready it is not ready. And so when would the next opportunity be so that we can let people know that process?

MS. TURBYVILLE: There are two

answers to that question. One is we do have a standard three-year maintenance cycle in which the developers would know that in three years there is an opportunity to resubmit and measures that have been currently endorsed are reexamined and compared to the new generation of measures. And that is true for all the measures that we have.

But if another project were to come about that's another opportunity too. We could have another project that comes forward and says, okay, you kind of got the very first little building block. Now it's time to go to the next building block. So there are at least two opportunities. And Helen, did I --

DR. BURSTIN: I think, again, this is such a new area for us, so that what we often do with new areas like this is, we'll have a group help us think it through. What we really want to do is mainstream it. So we had a composite steering committee a couple of years back set up an evaluation framework.

How do we look at composites? What are we going to ask for? What is the measure submission look like?

The reality is, now, any project we do, composite measures are welcome. And I think the idea would be, going forward, once we get past this initial hump, big hump, little, whatever it is, I don't know, we will ultimately as we go through all of our endorsement maintenance projects over the next three years, as a condition comes up or as cost counting error comes up, resource measures would be welcome as would any other kind of - any other sort of parts of quality measures, et cetera.

So I think that is the way we'd ultimately want to say if for example in September we are doing surgery and cardiovascular. That's a little soon obviously right now. But if we did a call for measures on anything related to cardiovascular care, we would within two or three years say,

sure, bring in resource use or efficiency, if we get to the next level we hope by then, going back to Ethan's point. We would say, please bring in your efficiency measures to the topical area as well. We won't have them pigeon holed into these measurement type projects any more.

CO-CHAIR LOTZ: Tom.

DR. ROSENTHAL: I would agree with the proposition that given the full NQF criteria that we will be lucky to come up with two or three that pass the full spectrum of things, and that that ought to be our focus. So I agree with that completely. And the only thing I thought we were discussing was how not to close the door at this stage of the game on people who are thinking about other things, and I would hope that it is not a sort of taste-great-less-filling conversation, but just how do we keep the doors open at this point. I agree that we will be lucky if we come up with two or three, and they need to be

perfecto, they can't be - oh, maybe they were okay, but maybe they hadn't been tested. That would be a catastrophe. But it may also be the case that we could also learn from the quality thing that maybe this is an opportunity to accelerate the field. Because again I think we are starting at a different place than the quality conversation started where again there is not a lot of public science about the thing.

And I saw somebody frown when I said there is no science about this, but there is not a lot of published science. What we saw last week in the New England Journal article was about the first thing I'd seen that anybody - and I've Googled it and I couldn't find much more that really gave me what you would classically get from the science. So I don't think this is a taste-great-less-filling conversation yet.

CO-CHAIR LOTZ: Mary Kay.

DR. O'NEIL: It's a question, if

in the measures that are submitted it's clear that there are significant gaps in issues that are not being addressed by those, is there a process or method to go after getting those gaps filled without waiting for a three-year cycle? If a couple of measures addressing one part of the question come up, and everything else is kind of ignored?

DR. BURSTIN: We are faced within this project measures that are submitted to the project. Part of what we also rely on steering committees to do is to say where are the good pockets of measures out there; find them; bring them in. So if you can work with whoever your stakeholders are, whatever the group is, if there are some good things out there, we would also hope you would identify those for us, and we can get them into the mix early.

The other thing is that there is a fair amount of back and forth between steering committees and developers as well, so it may

be that - and Jeff knows this as well. But as measures come into us, there is often a sort of back and forth saying this measure probably could make it through but there are significant concerns with the following, and these are the conditions. So you will have a little back and forth. We can't wholesale rewrite measures or make them really different than they started out to be. But again I don't think it's a three-year cycle. Because I think again once we get comfortable with this area of measurement we would expect these kinds of measures to flow into every single one of our - whether it's cancer or pulmonary or cardiovascular or care coordination. I man there may be ways to bring these kinds of things in regardless of the topical area.

CO-CHAIR LOTZ: Paul.

DR. BARNETT: The other thing I was just thinking about, and this was said in the morning, but in terms of the criteria, I think these resource measures are going to be

more acceptable if they can somehow be linked to a quality measure. And I heard this great quote, I don't know where it originally came from, but that efficiency without quality is unacceptable and quality without efficiency is unsustainable. And so I think that whole idea of efficiency without quality being unacceptable is something that we have to realize and that we would love it if there were criteria that - you know one of these resource criteria was linkable or really tied in with some quality metric too, and it was a value for money thing. So obviously we would rate that much higher if we had that linkage.

CO-CHAIR LOTZ: Helen.

DR. BURSTIN: This is a really interesting question, and one we've actually spent a lot of time thinking about. The question is, at this point in the game, do we know enough about these resource use measures to say, you should use this resource use measure with this outcome X? Or is that

something that might be evolutionary over the next couple of years. And again if we sort of stick to the idea these are building blocks, the NQF measurement framework couldn't have said more clearly that resource use measures should only be used when coupled with quality measures. We are saying that; that is up front; that is a given. But I think the idea of necessarily saying that, going back to the point somebody was raising earlier, we may not have the right - your point - we may not always have the right outcome measures we want to put in front of it. But certainly Jim would recognize that there are some great issues around joint replacement. We've got really crappy, at least to date, quality measures around knee and hip replacement, yet such an important area, really high cost, high variability. So I think again if we kind of stick to the building block approach, knowing we can frame it in the context of at least from where we sit the measurement framework

has already clearly said, only use when coupled with quality, that is probably the best approach for us at this point. But I'm certainly open to other ideas and suggestions.

One thing Sally and I were talking a little bit about is, it might be useful when we do the measure submission to at least ask the developers to indicate which quality measures they have coupled these with, just so we can begin to learn. But I don't think again making it the requirement, or automatically saying this resource use measure goes with this outcome measure, I'd be curious to know your take on it, but it's an interesting question.

DR. BARNETT: Just to make clear what I was proposing is not that we make it mandatory, but that we include it as another criterion, and if somebody meets it then obviously we are going to rate that measure higher.

CO-CHAIR LOTZ: Kurt.

DR. ELWARD: Yes, I appreciate Helen's comments, and I just want to make sure that that's something that will get past the board, too. You were talking about things that haven't actually been proven yet. So that would pass - your sense is what you were saying would pass the board? I'm trying to make sure that I'm following you. When measures haven't been truly proven yet, earlier I thought you were saying that that would have problems getting past the NQF board.

DR. BURSTIN: I still think these would be tested resource use measures. I'm just trying to make the point that I'm not sure we necessarily want to say we absolutely know at this point for certain that these resource use measures should go with outcome B. In terms of the coupling piece of it. Separate, I think the resource use measures we are still saying we want them to be reliable and valid.

DR. ELWARD: I think from my point of view, also, we almost have to couple them with some kind of quality measure.

DR. BURSTIN: We are absolutely saying they must be coupled with a quality measure. The question is the level of specificity. Do we say for example this resource use measure on hip and knees must go with the Oxford hip and knee functional tool. Or must go with the Ottawa tool. I mean this is where I think it maybe fuzzy. I'd just be curious about the committee's perspective on that, because I think that's a tough issue.

CO-CHAIR LOTZ: Okay, we started this section talking about limitations and implications. We actually drifted a little bit into our next section, which is supposed to be - I'm going to ask Sally to say it the way she said it over the break, because it's really getting to the point where it's got some useable and actual information out of our conversations. So what are must-have

criteria, what are things for us to continue to think about? How can we take the day's conversation and distill it into a ranking or a sense of importance of what we have talked about that will then tomorrow inform how we actually turn those comments into criteria and into our call for measures. Kurt. Oh, I'm sorry, Jim.

MR. WEINSTEIN: Are you into the discussion?

CO-CHAIR LOTZ: Yes.

MR. WEINSTEIN: Because I think as you mentioned like line 806 et cetera gets into the resource use measures as building blocks toward the future around quality and appropriateness performance, and I just - I think quality and appropriateness are different.

And then maybe examples in this text of how these things might be measured. And then as you go through the next several bullets like 816, 818, it might be nice to

have some specific examples as well as 824, resource use scores, et cetera, for the reader. And the reference 20 doesn't really reference anything. But I was just - or if it does I don't know what it is. But those are my comments. I think this could have been helped a lot by some examples of actually in use systems that are doing these measures that would help the people who want to submit measures understanding what kind of thing we'd be looking for in submitting.

CO-CHAIR LOTZ: Do you want to reframe the conversation the way you did over the break?

DAY 1 RECAP

CO-CHAIR LOTZ: Again, just doing a time check, we've got about half an hour, a little more than half an hour before we open up the lines for public comment. We are supposed to have a break in between then, but I'm also told that we should wrap up by about 4:50, so if you all feel you need a break, or

if we - it's 3:55 right now, so we have a little less than an hour total, part of which includes some time for public comment.

Bill, go ahead.

DR. GOLDEN: Since we have a little time, I was just curious, maybe we could hear, and you might not want to say, on page 32, there was a talk about direction for a physician compare website. And with the problems of getting that granular, are there any thoughts as to what that may or may not look like?

CO-CHAIR LOTZ: Or would you rather actually stay with the agenda and just take that afterwards?

DR. RICH: I'll sort of answer that for you, it was going to be based on PQRI, people who successfully participated in PQRI used that measure set. Not a process measure as was said before, but that's what it was going to be based on.

DR. GOLDEN: That helps.

MS. TURBYVILLE: So knowing that we were going to have a very full day, my hope was that prior to adjourning this first day that we at least start to get the steering committee to think a little more concretely in the application of evaluating the resource use measures as they come in based on the conversation that had happened.

I think you guys have actually already started down that path prior to the formal kickoff of this section. So in thinking about the evaluation criteria that exists now, in thinking about the evaluation criteria that we had as a straw man during the webinar where we started playing around with how we might expand language to just reframe - we've done away with what we were calling the phases, but there are certain components that we have identified, and I'm just going to throw this one out there as being one of those hot topics, like attribution, that would or would not be subject to the evaluation

process.

And this is not meant to come to decision today because we will actually go through the principles and evaluations tomorrow. But make sure we've answered questions or kind of really dove deep enough into the various topic areas to inform that conversation tomorrow.

So I don't think we are trying to get to decisions unless it's absolutely not, let's not look at that, and then get the group to respond to that.

CO-CHAIR LOTZ: Jim, are you up again? Not yet? Okay.

CO-CHAIR STEINWALD: Well, just to make sure I understand context, we have four criteria, four main criteria. They are: importance, scientific acceptability, feasibility and usability, right? And so we are not arguing those four. Those four are etched in stone. So what we are discussing is, okay, how much granularity do we need

underneath those four?

MS. TURBYVILLE: Well, my vision which may not have been right a week ago given that we are now in play was that that would be tomorrow. But today thinking about what we've discussed about the components as far as measures, are there any things that are off the table, things that would help us prepare what you need tomorrow to really dive in.

And putting it into context, it's almost you guys help us do a wrap up for today and bringing it down to the more concrete slice in which - affects evaluation. Which I really do think you already had naturally started having that conversation about.

And yes, tomorrow when we go through the actual sub-criteria, I mean a lot of the sub-criteria we saw are still applicable. Validity is still applicable. Reliability is still applicable. Some of the things that you started talking about. How will we guide the measure developers in what

do they need to submit to us that would demonstrate that they are meeting those sub-criteria. We will have those conversations certainly in more depth tomorrow as well. Or we can choose to go down another path. Honestly these slides are just meant to help guide us and make sure that we are getting everything we need, but not meant to be restrictive to the steering committee at all.

CO-CHAIR STEINWALD: What's the next slide? Oh okay.

MS. TURBYVILLE: We could also just kind of - staff could present to you some of the criteria and considerations that we know have to happen. And then maybe that will allow you to think through the conversation today. And then tomorrow we can dive more into the steering committee discussion. We're open to that, because this is kind of canned, already, pre-canned. If everyone is wearing out.

DR. HALM: You know the extent to

which you can highlight from the discussion or the email comments of some suggested additions to kind of the standard canon that people have suggested we might need to include for resource use specifically beyond just the four horsemen.

CO-CHAIR LOTZ: Barbara, go ahead.

DR. RUDOLPH: Well, I was just going to talk about the first criteria is importance. And it might be under the criteria for evaluation of measures. And I think there might be something extra that would be valuable there. Right now you need to meet the sort of one of the important areas that the priority partners have decided upon needs to be high volume or high cost, I'm trying to think what the other ones are. But I have a feeling that this area, particularly pertaining to efficiency -- as it relates to efficiency, might need a little more specification on the importance side. Because

I'm thinking about like how do you select the - say if you were going to do procedure specific, or you were going to do a condition, if you were going to do an episode, where do you start with this? And what would drive your choice about which area to measure? And I guess there as representing purchasers of commercial populations, I might have an interest in for example deliveries, and that might not be on a priority partners' list of important things, yet it's say 20 percent of the consumer spend or the commercial spend is on deliveries. So I mean I'm thinking that there might be some other criteria that may or may not match with the existing criteria for importance to measure.

MS. TURBYVILLE: So just to maybe aid a little bit more in this discussion, I forgot to mention, in your manila envelopes are in a table that is the thickest document in there, on the right-hand side, are the current NQF evaluation criteria, and on the

left-hand side would be what we have proposed. It looks like the memo is the one that is in landscape on the front page.

DR. BURSTIN: Let me just put this in context for a minute, Sally, if that's okay.

So the idea here was that we already have a set of NQF evaluation criteria that we use. A couple of them are being updated as I mentioned. There is a testing task force, and I've got that table for you that I'll share around reliability and validity, as well as a group that is really focusing on sub-criteria 1(c) which is the evidence for the measure focus. But the idea was, how these resource measures still feel like a bit of a round peg in a square hole. And so the idea was, how do we adapt, not create new ones, but adapt the current criteria or perhaps add a subcriteria or two to make it work best to evaluate these measures, since they are kind of a bit of a

different beast.

DR. GOLDEN: A question for you. I was going to put this in writing, but I'll make it a statement now. There are two aspects of the paper that I thought got a little on the editorial side as opposed to a little more analytic. And I thought - and I just think as we go through this there are ways of producing a compelling document without necessarily turning off audiences we want to reach.

So the first paragraph I thought was a little strong in terms of going after the health care system. You can go over things, but I don't think we have to - I think the approach was a little - I think we can make it a little more value neutral in some ways.

But the whole notion of the term - and I think it's in the first page here on the document - the third page - poor performance, the term poor performance, especially when it

deals with resource use, it might be better to use the term, inferior performance. You are implying a value judgment on the performance on the basis of the data without necessarily - it's all comparative. And I'm not sure we can state that it's truly "poor performance." It's definitely a lesser performance; it's an inferior performance. But it's not necessarily a poor performance, depending on the circumstances.

CO-CHAIR LOTZ: Barbara. Dolores?

MS. YANAGIHARA: I was just going to go back to the question I started with this afternoon. So there are these lists of things that are supposed to be part of the specifications, and which ones are we going to be responsible for, and which ones are - it's kind of the question that I think Helen was asking. So which are part of the evaluation criteria, and which are the implementation? And I don't know that we answer that. I don't

know if we should answer it now or tomorrow or what, but it might be helpful to go through that list of all of the things that we talked about and ask, is it in or out as part of the evaluation or not at some point. I don't know if it's now in preparation for tomorrow so we can think about how all these criteria would apply, or if it's tomorrow. But I think that needs to be done still.

CO-CHAIR LOTZ: Ethan.

DR. HALM: I'm not sure if you are asking us to kind of wade into this table now and comment on the different pieces. But in heading into the discussion I think on the importance side, I think the perspective seems different for this resource use steering committee than some of the quality measure ones. So sort of proportion of heart attack patients getting an aspirin, you don't have to worry about what the perspective is there in generally. But here we have heard conversations about, is it the prospective

society, of the payer, the provider, the organization, the patient. And so I think this is going to come up particularly in the importance that people need to talk about things being important, from which perspective. And I'm not sure thinking ahead on these other dimensions what about resource use is inherently different than the more traditional quality measure construction issues are. That seems like one of them to me.

CO-CHAIR LOTZ: Bill, are you up again? Mary Kay.

DR. O'NEIL: I'm thinking back to that spectrum that we discussed earlier from procedures and episodes and things like that. So when you look at those things you can say who does the most efficient job of doing a knee replacement from inputs or costs of inputs. On the other hand if you go further to the other end of the spectrum when you are looking at the application of resources to a

population, the outcome can be looked at as what the costs are, what the medical trend costs are for that population over a longer time horizon, and that can be the surrogate for how healthy they are, in terms of what their medical needs are on a going-forward basis. So in other words what kind of inputs are most efficient in producing health over time. And of course in our industry we are also interested in productivity over time, because we are mostly looking at working age adults. So there are some really different ways of looking at what the outcomes are, what you are measuring and the impact that you are tracking that is pretty different than more narrow very time-restricted interaction of the patient with the health care system.

MS. TURBYVILLE: I wish there was an easier way to pull up the various components because what we had thought was to think about the phases and the steps, but now it's just pretty much the steps. So let me

see if I can - they are all included in here, but just separately. So if we don't think of them as just kind of specification. (Pause) Bear with me for one second. (Pause)

So this may seem obvious, but for example would there be any of these steps of a specification that would not be subject to evaluation? And then we can just, there are a few others. So would there be any reason why there wouldn't be an expectation as we go through the evaluation criteria and we're helping guide this through, that the measure developer could step away from clearly defining the unit of measurement, et cetera.

And so just kind of quickly in the last few minutes, it doesn't mean that it's all in cement or anything, but are there any things - and I'll pull up the rest of the list - are there any things in these steps that you do not consider to be subject to evaluation when we start thinking about the criteria in more detail tomorrow?

CO-CHAIR LOTZ: Bill.

DR. GOLDEN: Do you want me to react to what's up there? Or do you want to have comment on one of the measures, or just see whether it should be there or not?

MS. PODULKA: I'm sorry.

DR. GOLDEN: I'll flip it. The first item, the first bullet, I'm not sure is what you are trying to capture. The issue of acute versus chronic is not as important as the timeframe. So I mean you could have acute episodes that last five days or last - for the same issue - last two months. And I think we need time in there as opposed to adjectives.

MS. TURBYVILLE: Does that work?

DR. GOLDEN: I'm just making the comment.

MS. TURBYVILLE: So would that address your concern a little bit better, if the group agrees?

DR. GOLDEN: It goes to a similar thing we did before.

MS. TURBYVILLE: Right. Okay.

CO-CHAIR LOTZ: Paul.

DR. BARNETT: I'm not sure when we evaluate measures we are sort of - we want to look at something, and I think the top level, the four horsemen as it were, are the right place to start, and that drilling down to this level, if someone has a measure and we are convinced it's important and it's valid and it gives actionable information, I don't really care to second guess how they got to that point, or drill down and say hey, should it be tweaked a little bit. I really want to focus on those top level things. Is it important? Are there are a lot of costs involved? A lot of people involved? Have they done some validation that it works? And can we do something with it once it comes out? I think those are all the things that we should be looking at and not so much how they got there.

CO-CHAIR LOTZ: I don't

understand it as a how we got there, but as a refining or clarification as it's to be used in this particular application. I'm thinking about the scientific acceptability. And someone earlier mentioned the idea about the difference between clinically significant and statistically significant, and the analogy here being economically significant versus statistically significant. And do we - telegraph that in the call to measure, and consider that when actually looking at the measure. So as I have kind of understood this exercise it's to take those four essential criteria that NQF has used and intends to continue to use and refine them or amend them for this particular process. Is there anything else we need? Or do we need to be a little bit clearer on how to apply those to this particular project? Tell me if that is wrong.

DR. BURSTIN: I think that is right, but I guess the question would be,

would it make more sense to start with the criteria and the suggestions for ways to flex them, and then come back?

MS. TURBYVILLE: I think that is a good question. I think the challenge is even - and so maybe this will come out more naturally tomorrow - is when speaking to measure developers about these types of measures and submitting measures, it's not clear to them what the specification would be that they need to submit for the steering committee to evaluate.

And so while it seems like it's getting into the details, that's because it is. So what are they just going to submit the description of their measure and we use a clinical logic that we have tested but you don't need to see it because we'll just let you know that it was tested? These are the kinds of things I'm trying to push the steering committee to think about so that as we develop the call for measures and the

submission form, et cetera, that we are able to clearly tell them what to submit, that we are also not getting volumes and volumes of things that you might not be interested in evaluating.

So it is the details, and that's kind of what the thinking was behind it.

MS. PODULKA: It has an implication not just for the signal to the measure developers about what you want to see, but it has an implication down the road for the users of our measures. And for instance we want to keep abreast of whether users are using actual NQF-endorsed measures or if they are tweaking or somehow changing those in their implementation. But there could be changes or tweaks that you as a steering committee are agnostic or indifferent about. For instance in one of those pre-steps whatever we are calling them now about data cleaning. There is a step for Winsorizing or excluding certain claims. If some measure

developers Winsorize to the 1st and the 99th percentile, and another measure developer Winsorizes to the 2nd and the 98th, you might be agnostic or indifferent, and you might say, well users could choose what they want there. Or you might say, this step is integral to the measure, and needs to be considered in view of our four criteria.

CO-CHAIR LOTZ: Mark Kay? Tom?

DR. ROSENTHAL: I mean I think this is pretty good. I mean this is almost a recapitulation of the discussion we had a little earlier, trying to define what it is, and how much detail are we going to insist on. I have a question and a comment.

The question is, when we have defined the unit episode, let's say for the sake of discussion that that was going to be a bundled payment for joint replacement, are you saying that we do or don't have to ask for any greater degree of specificity about what kinds of things they would put in there? For

example in the working group that we had on this, the question is, do you include rehab services or don't you include rehab services? And it's a fairly important question. Is that the kind of specificity that you are looking for in the definition of the unit? And if so you may want to provide a little more guidance.

And the comment I have is, I think the attribution question is maybe more important in this business than it has been in the quality realm, maybe it didn't matter quite as much, but here I think it matters a lot. And maybe that needs to be its own something up there that guides people as saying we really want to be very certain to whom you are attributing this, because I think our selection of one measurement over another may hinge upon the degree to which the attribution is really clearcut and not ambiguous. And so that may be the only thing that I don't see on there that might be worth

spelling out, unless again that's implicit in your idea of defining unit construction logic, but now we are back into, will everybody understand what you meant exactly by that.

MS. TURBYVILLE: I just haven't cut and pasted it on there yet. So yes that would come from - you are absolutely right, when we go back to the PowerPoint, you were right to say it's getting back to the earlier discussion. So when we started talking about applying, that's when you see this list is quite long, 29 and 30, when we start talking about the attribution, et cetera, absolutely. So the question was if there are any of these steps in here that would not be part of evaluation.

CO-CHAIR STEINWALD: So we are not explicitly excluding anything, but we are giving a lot of discretion to the measure developer about what really needs to be highlighted. In other words, it may not be important - there may be a measure where

attribution is not an important dimension; but some will. And they might need to say why that's true. But that could be a paragraph as opposed to 10 pages.

MS. TURBYVILLE: Just to comment, not to necessarily - to add to your comment - how we handled it on the quality measure, there are times where it's - risk adjustment is a required component, or you must provide your justification of rationale of why it is not applicable to this measure. So you could latch on to that kind of language for some of these where it is a required criteria, and where they don't do it it must be an intentional well thought out rationale that then the steering committee thinks is an acceptable rationale.

CO-CHAIR LOTZ: What I've noticed as kind of a common theme, or what I think has been a common theme throughout our discussion is that there is a lot of these criteria where we say, just explain to us what your thinking

is, but we are not creating any absolute thresholds like you have to do your statistical analysis in this way. What I've heard is, we have to do some kind of statistical analysis. Just tell us what it is, and tell us why you chose it, and tell us what kind of strength comes along with that.

Paul, Barbara and Jack in that order.

DR. BARNETT: So now you have convinced me of the need to get into all of the detail. But I think that the last two comments are exactly right. So if I'm peer reviewing one of these, I'm not going to have any independent information whether a 60-day or a 30-day or a 90-day clear period is the best way to define an episode. But I want to be convinced that they knew why they chose the one that they did.

CO-CHAIR LOTZ: Barbara.

DR. RUDOLPH: I was just looking at one of the comments on the criteria, and it

was discussing like laboratory and I think pharmacy as well. And so I can see situations where either laboratory costs and/or drug/pharmacy costs are included, and in other cases perhaps the measure developer didn't include it because either they didn't have access to the data or whatever. Again I think that might be something that there would need to be a rationale for not including some of those costs. If they were relevant.

And again maybe it's this discussion of relevancy is, because if you got two groups in, and they -- one group included both lab and pharmacy and the other said, well we tested that but it really didn't make any difference, I could live with that. But things like that would certainly be a consideration here.

CO-CHAIR LOTZ: Jack.

DR. NEEDLEMAN: Barbara's comment raises the issue of understanding not only what the unit of measurement is intended to be

but what the scope of the measure is, what costs, what resources, have been taken into account, which have not been taken into account. So I think where that fits in the description of what we want is not completely clear to me, but we certainly need to make clear we want that.

Missing from this list is something we talked about right at the beginning of the day, which is what kinds of cost adjustments or cost standardizations have been introduced into the measure. And that should probably be made explicit.

The other thing that I had a question about is whether the flow on this list is important to our consideration, because the attribution problem is going to be a major issue as we think about the application of this. It's post-attribution that the physician comparisons and the Adams McGlynn paper come up in particular, and reliability issues come up. But right now

it's on the top of that second page, and I actually think the attribution issue, attribution goes down after we have measures of resource costs either at the episode or the patient level. Then we need to figure out who we are assigning these to for the purposes of making any cross-provider, cross-clinician comparisons, and what the appropriate level is.

But if it's relevant about the flow here, I'd want to see that move down in the flow. And if it's not relevant at the moment then it's not relevant at the moment. But we should be thinking about the flow of the analysis and which numbers we get early and which numbers we get late in the course of understanding what the measure is doing.

CO-CHAIR LOTZ: Additional comments? Should we just go ahead and open up the phones?

MS. TURBYVILLE: Operator, at this time we'd like to open the call up to any

public comments or questions for the steering committee.

PUBLIC COMMENT

OPERATOR: Press star one for any questions or comments.

We'll take our first question. Caller, please go ahead.

DR. MUNLEY GALLAGHER: Hi, this is Rita Gallagher. May I comment?

MS. TURBYVILLE: Yes, please.

DR. MUNLEY GALLAGHER: This is Rita Gallagher. May I comment?

MS. TURBYVILLE: Yes, you can. Can you hear us, Rita? Because we can hear you?

DR. MUNLEY GALLAGHER: I can now. Thank you.

Again thank you for the opportunity to comment. The efficiency resource use steering committee has been eloquent in its discussion of what the measure developers are to be doing in specifying and

submitting measures for consideration for NQF endorsement. I would respectfully suggest that there is a parallel need for NQF to ensure that the members of the various technical advisory panels and steering advisory committees have the necessary expertise and time, and are adequately and uniformly prepared to fully engage in the evaluation of those measures once they are submitted.

Thank you.

MS. TURBYVILLE: Thank you. Any other questions or comments?

OPERATOR: No further comments or questions on the phone.

CO-CHAIR LOTZ: All right, there is a desire to wrap it up. So one last time, final comments? Otherwise we will talk about the logistics of reconvening in the morning and how we'll spend our time.

So according to our schedule we are back in this building I guess - why am I

doing logistics, I have no insight into this. Ashley, actually Ashley I think is our logistics person.

MS. TURBYVILLE: We are indeed back here tomorrow. And day two starts at 8:45 for a continental breakfast. We start at 9:00. We dive right in, we do our best to do a recap to make sure that we haven't inadvertently gone down the wrong path. We will talk about some external market implications and give everyone a chance to discuss that, so we can learn from all of you. And then we will dive right into the evaluation criteria table.

Now the slides are set up to go through what we have seen with the subcriteria, and we'll talk a bit more in detail as being different. But I do want to make sure we spend time on what we have, because we may have missed something. So we will think about how best to approach that tomorrow. And that includes a discussion by

the way of the evaluation principles which are reflected in the white paper and will also be maintained in the evaluation criteria documents.

So that I think will be the starting point before we dive into the more detailed discussion.

Please.

DR. BURSTIN: I'd just mention, she said Table 2 is the draft table from our testing task force report. It might just be useful as a starting point for our discussion tomorrow to look at this and see how that would work, not work, need to be modified potentially for the resource use measures.

MS. TURBYVILLE: Thank you everyone.

(Whereupon at 4:33 p.m. the proceeding in the above-entitled matter was adjourned.)

Q

S