The Workshop met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, NW., Washington, D.C., at 9:00 a.m., Patricia Brennan and Joyce Dubow, Co-Chairs, presiding.

PRESENT:

PATRICIA BRENNAN, PhD, University of Wisconsin-Madison, Co-Chair
JOYCE DUBOW, AARP, MUP, Co-Chair
RICHARD BANKOWITZ, M.D., MBA, FACP, Premier Healthcare Alliance
ETHAN BASCH, M.D., MSc, Memorial Sloan-Kettering Cancer Center

JIM BELLOWS, Ph.D., Kaiser Permanente
DAVID CELLA, Ph.D., Northwestern University Feinberg School of Medicine
ANNE DEUTSCH, Ph.D., RN, CRRN, Brookings Institution
STEPHAN FIHN, M.D., MPH, Veterans Health Administration

JACK FOWLER, Ph.D., Informed Medical Decisions Foundation
LORI FRANK, Ph.D., Patient-Centered Outcomes Research Institute
BARBARA GAGE, Ph.D., MPA, Brookings Institution
TED GANIATS, M.D., University of San Diego

Health System
KATE GOODRICH, M.D., MHS, Centers for Medicare & Medicaid Services
JUDITH HIBBARD, DrPH, University of Oregon
DENNIS KALDENBERG, Ph.D., Press Ganey Associates
IRENE KATZAN, M.D., MS, Cleveland Clinic
LEWIS KAZIS, ScD, Boston University School of Health
UMA KOTAGAL, M.D., Cincinnati Children's Hospital Medical Center
KEVIN LARSEN, M.D., Office of the National Coordinator for HIT
KATHY LOHR, Ph.D., RTI
ELIZABETH MORT, M.D., Massachusetts General Hospital
CHARLES MOSELEY, Ed.D, National Association of State Directors of Developmental Disability Services
GENE NELSON, DSc, MPH, The Dartmouth Institute
KENNETH OTTENBACHER, Ph.D., OTR, The University of Texas Medical Branch at Galveston
GREG PAWLSON, M.D., MPH, FACP, BlueCross BlueShield Association
ELEANOR PERFETTO, Ph.D., Pfizer
COLLETTE PITZEN, RN, BSN, Minnesota Community Measurement
CHERYL POWELL, Centers for Medicare & Medicaid Services (via telephone)
DAVID RADLEY, Ph.D., MPH, Institute for Healthcare Improvement
TED ROONEY, RN, MPH, Maine Quality Counts
DEBRA SALIBA, M.D., MPH, UCLA Borun Center/VA/RAND
MARCEL SALIVE, M.D., MPH, National Institutes of Health
LAURA SMITH, Ph.D., Brookings Institution
BARBARA SUMMERS, Ph.D., RN, University of Texas-MD Anderson Cancer Center (via telephone)
KALAHN TAYLOR-CLARK, Ph.D., MPH, National Partnership for Women & Families
MARY TINETTI, M.D., Yale New Haven Health System
PHYLLIS TORDA, MA, National Committee for Quality Assurance

Neal R. Gross & Co., Inc.
202-234-4433
JOHN WASSON, M.D., Dartmouth Medical School

ROB WEECH-MALDONADO, Ph.D., MBA, University of Alabama-Birmingham

LINDA WILKINSON, MBA, Dartmouth Hitchcock Medical Center

ALBERT WU, M.D., MPH, Johns Hopkins Health System

NQF STAFF:

KAREN ADAMS, Ph.D., MT

HEIDI BOSSLEY, MSN, MBA

HELEN BURSTIN, M.D., MPH

SHEILA CRAWFORD

EUGENE CUNNINGHAM, MS

KAREN PACE, Ph.D.

JESSICA WEBER

EVAN WILLIAMSON
TABLE OF CONTENTS

Welcome and Setting the Stage
  Patricia Brennan, University of Wisconsin-Madison .............. 6
  Joyce Dubow, AARP ............................................ 8

Lessons from the Field - Using PRO-PMS for Accountability (Public Reporting, Payment)
  Greg Pawlson, BlueCross BlueShield Association .................... 36
  David Nuttall, Department of Health, UK ............................ 38
  Liz Goldstein
  Stefan Larsson .................................................... 80
  Discussion .......................................................... 104

Recap of Key Characteristics for Selecting Individual-level PROs for Use in Performance Measurement
  Overview of Related NQF Endorsement Criteria ..................... 140
  Elizabeth Mort, Massachusetts General Hospital ................... 147
  Patti Brennan, University of Wisconsin-Madison, Project HealthDesign .................................................. 155
  Laurie Burke, FDA .................................................. 168
  Discussion .......................................................... 178

Methods that Contribute to Trust - Demonstrating Reliability of PRO-PMs
  Overview of NQF Endorsement Criteria ................................ 202
  Key Issues from Commissioned Paper
  Lewis Kazis, Boston University School of Public Health ........... 221
  Lori Frank, Patient-Centered Outcomes Research Institute ........ 235
  Jack Fowler, Informed Medical Decisions Foundation................ 246
  Discussion .......................................................... 257
Page 5

Methods that Contribute to Trust Demonstrating Validity of PRO-PMs as
Indicators of Quality

Overview of NQF Endorsement

Criteria. . . . . . . . . . . . .229
Key Issues from Commissioned Paper. . .283
Steve Fihn, Veterans Health
Administration. . . . . . . . . .306
Albert Wu, Johns Hopkins. . . . . . . .315
Discussion. . . . . . . . . . . . . . .333
Neal R. Gross & Co., Inc.
202-234-4433

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

9:00 a.m.

DR. BRENNAN: Good morning.

Welcome to the National Quality Forum's Patient-Reported Outcomes. This is our second workshop.

I'm Patti Brennan. I'm from the University of Wisconsin-Madison. I'm very happy to see familiar faces in the audience today. Welcome back to those of you who were at our first workshop in August. I also want to extend a special welcome to the number of people who are connected to us via the phone and the internet. We'll be doing our best to monitor to make sure you have the participation in the meeting over the next 2 days that can help us grow and accomplish our task.

However, before we get onto the tasks today I want to take a moment to remember that this is a very special day in the history of our country and just pause for
a moment to those who might need to have us remember with them what they lost and perhaps learned on this day.

(Moment of silence)

DR. BRENNAN: Thank you. This morning we have a lot to get going with on understanding the difference between PROMs and PROs and PRO-PMs. And we'll have an opening session here that will take you through some of the foundational concepts. We'll have several different panels today on workshops on -- working panels on validity and reliability.

And we have some tasks ahead of us to come to some consensus about the process that we'll be recommending to the National Quality Forum of how to endorse the PRO-PM and by the end of the day you will know what that means.

I'm going to turn over to my co-chair Joyce Dubow from the AARP who's going to take us through our introductory remarks.

Thank you, Joyce.
MS. DUBOW: Thanks, Patti. Good morning, everybody. Yes, if we can all master the vocabulary we'll be in great shape because it's going to be quite challenging I think.

All right.

All right, well the first, can we go back to the first slide, please? Starting with the meeting objectives because here's what we want to accomplish this morning -- actually, during the entire meeting.

First, I hope -- you can't hear? Okay. Okay, better? All right.

We're going to -- one meeting objective is to discuss the methodological issues related to reliability and validity. The paper is very specific. This is not a paper that's supposed to answer all the questions about patient-reported outcomes. It was specifically commissioned to identify and to discuss these issues around reliability and validity when aggregating PROM data into a performance measure. We're going to come back
and talk about those terms in a minute.

We also need to remember that ultimately we're going to want measures that NQF can endorse. And so what we need to do is to think about patient-reported outcome measures in the context of the NQF criteria, the evaluation criteria.

And what we need to think about is whether these particular types of measures present unique circumstances, are there specific or unique attributes about these types of measures that need to be taken into account within the evaluation criteria? So we'll need to be thinking about that.

And finally we're going to want a pathway, a critical pathway from the PROM to the PRO-PM endorsed by NQF for use in accountability and quality improvement. Okay, next slide, please.

So, a word about terminology. And I personally have a lot of trouble tripping on these terms but we're going to be using the
terminology very, very specifically and I think throughout the day we'll try to remind you. But you may want to refer back to this particular slide.

The patient-reported outcome or the PRO is the concept of any report of the status of a patient's health condition that comes directly from the patient without interpretation of the patient's response by a clinician or anybody else. So this is what the patient says.

The PRO measure or the PROM is the scale or the instrument or a single item measure -- I use that word softly -- to assess the PRO concept as perceived by the patient. So an example is the PHQ-9. So it's the tool, it's the instrument. It's not that which we will be endorsing. Next slide, please.

So just a reminder about what a performance measure is. It's a numeric quantification of healthcare quality for a designated accountable healthcare entity like
The PRO-PM is the PRO-based performance measure that is based on the PROM data aggregated for an accountable entity. So the illustration here is the percentage of patients in an accountable care organization whose depression score as measured by the PHQ-9 improve. So it's the PRO-PM that we are seeking to endorse, okay? All right.

Now, in addition to these terms around PROs, PRO-PMs, we also have to remember and had discussion at the last workshop you may recall about the proper way to refer to a person, an individual, a patient, a consumer. And this is subject to very wide debate and discussion a lot at NQF and elsewhere. And I think we need to be very clear up front in the report about what we mean. These terms are circumstantial and they really depend on the context of what we're talking about.
Admittedly we all use shorthand so sometimes people like to say "person," sometimes people will say "consumer." I think there is broad recognition that these terms while not interchangeable are sometimes used that way because we are speaking shorthand.

The disability community in particular is sensitive to the use of terminology and we heard this last time from Chas and from others about their perception of the words. So I think we need to remember to be very sensitive to the language. Vocabulary matters.

We will understand if each of us lapses into the vernacular that we're most comfortable with but I think, at least I hope we'll be understanding, but I think we have to recognize that these terms do mean something particular to different groups and different people and different individuals, and we need to be sensitive to that.

We discussed this, Patti and I
discussed this with Karen and Karen and what's her name over there, Helen.

(Laughter)

MS. DUBOW: Our leader. Our leader, Helen Burstin. And we hope that we will be able to identify a way perhaps, a way of being sure that we meet everybody's needs and understanding without doing violation to sensibilities and sensitivities. So, we are aware of it and we will try to address it.

Okay, next slide, please.

So here we have our bubble diagram that NQF has used for a long time that shows basically an episode of care. And this tries to illustrate the concept -- I'm not going to go through this in detail -- how patient-reported outcomes would be taken into account.

And if you look at the far right, lifestyle and health behaviors, and look at the illustrations on the right-hand side you'll see the types of areas that we will be discussing today that need to be taken into account.
account as we consider these episodes and how
a patient person experiences these episodes.
But we are interested in functional status,
health-related behaviors, symptoms, symptom
burden, et cetera. So if we could move to the
next slide, please.

Just a reminder about what NQF
does. It endorses performance measures, not
the tools. So it endorses PRO-PMs, not PROMs.
NQF endorses PRO-PMs for use in accountability
such as public reporting and quality
improvement. So remember, we have two
purposes. It's we need to be sure that these
measures that ultimately are endorsed are
suitable both for quality improvement and
accountability, not one or the other but both
of them.

NQF has criteria as we mentioned
earlier to evaluate measures that come through
the endorsement process. And these are NQF
board-approved. They have been widely vetted
and as I said earlier we're going to have to
think about whether these criteria need to be
tweaked in some way if at all for purposes of
considering the patient-reported outcome
measures.

So, just to remind you and I think
we'll go through these in a little bit more
detail later. Karen Pace will take us through
that. Importance to measure and report is one
criterion and it's a must-pass criterion. If
a measure doesn't -- is not -- cannot be
determined to be important it doesn't go
through the rest of the criteria.

In addition, the measures under
consideration by NQF committees are evaluated
for their scientific acceptability,
feasibility, usability and use. And this
slide simply gives you a little bit more
information about what the criteria look like.
I think we'll probably be paying a lot of
attention to the second criterion, scientific
acceptability of measure properties because
we're going to be talking about reliability
and validity. Okay, next slide, please.

So, finally -- I did this very fast. We're going to have a conversation.

Finally, what we want out of the workshop.

And you see in your materials a straw man pathway that is probably too big to show on the screen. But you have it in your diagram, at least it's too hard to read here. You have it in your diagram. And this -- we will go through this, you know, at another time.

Everybody says it will be excruciating detail.

But I mean, it's important to understand how this stuff is all going to work and so we have developed a pathway, a straw man pathway to think about how this will happen.

But ultimately we would like to reach consensus by the work group at this workshop on what the pathway should look like.

So with that I think we're ready to have questions and discussion and -- oh, did we want to talk about the -- there's a time line.

Yes, okay.
So, and we are -- you can see where we are here. We're right in the middle of the process. There's going to be a review of the paper which I think you've all received, public comment. The expert panel will review those comments and ultimately this report is going to the CSAC and the board for their approval.

So, is there conversation, questions, observations? How are you, Ethan. Yes. Speak up. And oh, could you -- Ethan, could you reintroduce yourself so the people on the phone can hear you, please?


A couple of quick comments. I'm very glad to see these red boxes at the top which I know wasn't -- we, I think our initial conversation was starting with the measure and now it looks like we're going back and starting with the population of interest and
finding outcomes that are meaningful and important to the population. So I think that's a terrific addition and is consistent with the thinking in other quarters about how to develop tools that are meaningful to patients.

One thing, one question or comment I have is we spend a lot of time talking about the methods for assuring the measurement properties of the tool once we have figured out what we want to use, the validity, the reliability, all this. But we don't really specify how to identify outcomes that are important and meaningful to the target population, number one.

And I wonder the extent to which we want to be a little bit prescriptive about how somebody who's proposing a measure can actually assure to us that the outcomes, right, are actually meaningful in the population for the context of interest.

MS. DUBOW: We are going to have
more discussion about meaningfulness. And in fact Patti just a little while ago volunteered to fill in on a panel where she will be specifically addressing that and we'll have a chance to talk about that in greater detail. I think it's a really important question.

Greg? Greg, you have to tell us who you are, please.

DR. PACE: So let me just orient us to these microphones just to remind because it's a little bit different. The ones on the table, the red light means that it's activated and then you press the button till you get a green light. That means the microphone is actually on. Thanks.

DR. PAWLSON: Okay, and that was a good instruction. We've got the green light. I'm Greg Pawlson.

Sort of an observation. I think that this is going to be, in the continued sort of evolution of NQF endorsement this is going to be a very interesting test case if we
really are saying that you're going to endorse PRO-PMs.

And why I say that is that if you are really talking about accountability measures for a specific population I think you're going to be into what we're going to actually talk about in the first panel is measurement reliability and parameters in a much perhaps deeper way than before.

What I'm talking about is measures that will be endorsed for a specific purpose for a specific population perhaps with specifics around how big the sample size has to be and that kind of stuff. Not necessarily an endorsement of the measure for general use if you will.

And I think this is going to be an interesting dialogue that will have to be played out in the review panels and so on. How far down that road do you really need to go to say that we are endorsing a pro-

performance measure, almost measurement
approach, rather than the PROMs. And I think that that's going to be a very interesting challenge.

DR. BURSTIN: For Joyce's sake I'm Helen Burstin. I'm the senior vice president of performance measures at NQF.

I just want to respond briefly to Greg's point. I think it's a really interesting one. I think the closest analogy we have currently are the CAHPS surveys which actually do have a very similar approach to giving much more details on some of those key nuances around measurement that we've had before.

But you're absolutely right, this is going to be more complex, and we've already seen that for example with the few PRO-PMs we have brought forward and endorsed, including the one on improvement in visual function or the one on depression where I think, you know, all these issues come forward in a much more significant way than I think some of our more
classic process or outcome measures.

The issue about whether it needs to be assigned to a specific use I think is in question. I think the goal would be that you would want to select performance measures based on PROMs that in fact can drive improvement as well as accountability. But I think that's a good question for us to talk about today.

MS. DUBOW: Other questions?

Observations? Kathy?

DR. LOHR: One question about your time line.

MS. DUBOW: -- who you are?

DR. LOHR: Oh, sorry. I'm Kathy Lohr from RTI in North Carolina.

And I know you're on a fast track and you have to get moving and so forth and I know that the panel is supposed to be giving you some feedback in October, November, whenever it is, but after we kind of see a second round of the second paper and so forth.
But I was wondering whether you've given any thought to sort of in some sense reconvening 6 months or 9 months or somewhere down the road not only to see whether we've done a good job with sort of giving you all some advice and guidance and our best thinking on all this which is I think a point that Ted might make as well.

But then also to see whether you're making the progress you want, whether we can give you any feedback on mid-course corrections and that sort of thing. So it was really a question of is that the end of the time line kind of thing or can we be of help down the road.

MS. DUBOW: That's a question for staff.

DR. LOHR: I don't necessarily need an answer. You all may need to sleep on it but I wanted to put it forward.

DR. BURSTIN: I think it's a great suggestion, Kathy, and I think the question
for us as well is if we move forward in fact
and do a consensus development project on PRO-
PMs.

But I think one question might be
we'd love to sort of pepper that committee
with some of you. But I think it might be
nice to have a couple of test cases to bring
back for reaction to say did we get it right
as part of that process. That's a great
suggestion.

MS. DUBOW: It does speak to the
broader question that NQF faces with other
measures and that's to get feedback. I mean,
feedback is really very important for all
measures, not just these. I mean these are
obviously of interest at the moment but NQF
needs to get feedback to understand how these
measures are working, whether they are
accomplishing what we seek through the
endorsement process. So it's a really good,
good marker for doing it. I think it's a good
idea.
Other observations or comments

that somebody wants to make? Is there anybody on the phone who would like to weigh in? Do we have to ask the operator to do that?

OPERATOR: At this time in order to ask a question press * then the number 1 on your telephone keypad. At this time there are no questions.

MS. DUBOW: Okay, thank you. Is there somebody in the audience who wants to make an observation or a comment? No? Okay. Karen?

DR. PACE: We have a few minutes before we can begin the next panel so if anyone wants to make any other observations about the pathway. As we said we're going to come back to that in great detail tomorrow. But you know, if you want to make any comments now we can address that or if you have any questions that we can address we can take those now. Otherwise we'll check in with our other speakers and see if they're ready yet.
Kathy?

DR. LOHR: This is Kathy Lohr again and I did have one question. I would second what Greg said about the, you know, red boxes across the top. But I will confess, maybe it's just where I grew up, that I wasn't certain about the process performance measure versus outcomes.

And so in the green boxes like with six and all, is that supposed to be an example of what you would do now with what classically we would think of as process of care measures rather than outcome measures or is it something else?

But I also wasn't certain why you'd have PROM in there, you know, for a patient-reported outcome measure attached to process performance measurement. And it may just be me and age and you know, cohort or something, but the distinction between process and outcome has been around a long time and I wasn't certain why I was seeing outcomes.
mushed up there with processes.

DR. PACE: Right, good question.

And this comes from our discussion last week and I'll just give you a real clear, hopefully clear distinction of a process measure would be just the process of using a PROM in your clinical practice. So the process measure might be percentage of your patients or percentage of your depressed patients that were administered a PHQ-9 versus using that actual PROM data value on the PHQ-9 to say the percentage of your patients who were depressed who are now in remission.

So it's a distinction but it's very important for us to keep in mind. And one of the discussions that we had last week, and that's why it'll be definitely open for discussion, is that the pathway we should take. I mean, ultimately we're interested in outcomes. And so there's some thinking that, well, the first step is to get people using these PROMs before we can get to the step of
having an actual outcome measure.

Maybe there's some concepts that we need to do process measures first. Maybe there's some that we can go directly to outcome measures. And so the pathway kind of shows two ways to get there and that's something that we'll definitely want to have some discussions about.

MS. PITZEN: Hi, this is Collette from Minnesota Community Measurement. I just wanted to make a couple of observations.

One in terms of identifying a population or having something where you have a gap in care that you want to start with I think is really important. And then if a functional status or a healthcare quality of life is key in improving that quality of care then one goes and selects the appropriate instrument to collect that information.

The second point I wanted to make is we've found that in implementing some of these measures in our state we really do need
to do process and outcome measures together
because as you're putting this in place in
clinical practice you want to make sure that
you're capturing enough of your patient
population and how successful are you at
measuring these patients.

For example, we have some
orthopedic measures that we're collecting at
1 year post-op. So at the same time we are
giving groups that feedback of what percentage
of your patients are you actually capturing at
1 year before we can determine if we have a
valuable measure that we can use for public
reporting. Thanks.

MS. TORDA: Hi, I'm Phyllis Torda
from NCQA. I think a number of us, I'm going
to add myself, are making the same point and
that is that the context in which the PROMs
are used is very, very important for a number
of reasons.

The point that Collette just made
is it's important because it's going to give
you the sufficient sample size to actually measure outcomes but it's also important because these processes are much less immature than many other processes that we measure through clinical quality measures. And we need to really recognize that, use measurement to promote adoption of the processes before you can get to outcomes.

MS. DUBOW: I'd just make an observation before I call on you. You know, think that in the endorsement process this is going to be a challenging idea because of where NQF, where the CSAC has been going, and that is really to emphasize outcomes.

And the notion that these processes are immature and that we might need to think about processes on the way to having the outcomes I think needs to be well understood. At least, that's one former CSAC member's opinion.

Would you like to make a comment?

DR. FIHN: Yes I would, thanks.
And I was just going to say exactly what you said, Joyce.

So, we discussed this in the taxicab on the way over here. I work in a system where we have mandates for lots of these to measure with process measures.

And just as a cautionary note we just finished a large survey of all our primary care providers throughout our system. Morale is actually quite low and the greatest barrier to care that they identified was clinical alerts and reminders which include, you know, the mandated screening for depression and so on and so on.

So, you know, I think the problem is they, unless these are really linked to systems and to identifiable mechanisms for care improvement then that could actually be a self-defeating process to put in place mandates for data collection prior to understanding, a good understanding and implementation of the systems to which they
need to be connected. So, and I think -- so
I would really argue more for Joyce.

I think part of that is part of the validation process. And on one hand I
know you're under pressure to get things out and time lines, et cetera. But you know, we're reminded again and again what happens when measures get hurried to market. And we just -- we're also discussing this morning one in which CMS probably hurried to market in terms of the catheter-associated UTIs. And we now learned this week that it's probably a seriously flawed measure despite the fact now it's tied to the payment system.

So you know, I just, I'm not arguing against doing that but I think we ought to be cautious. All too often I've heard people defend this as the -- not letting good be the enemy of -- perfect being the enemy of good here and you know, sometimes we want, you know, we won't get perfect but we may want better than good.
MS. DUBOW: That was Steve Fihn.

John's going to be the last question.

I just want to be very clear. I think this is a necessary step in the pathway but ultimately we know where we want to go. And as a consumer representative you can be sure that there is going to be a sense of urgency to get where we want to go. But recognizing that this is very hard, this is very challenging, just to be clear.

John, you get the last word before we adjourn this session.

DR. WASSON: I hope it's not the last word. But in any case the other point on the diagram and also amplifying what Steve said is when we start talking about outcome measures we are talking about re-contacting patients which really does require consent. And that's a killer in the real world in terms of people even willing to measure before if they know they're going to be contacted later as an after. A lot of people will stop right
there. So that's not in your diagram and
you're going to have to put consent in there
somewhere, and that is a very important
practical step.

DR. BRENNAN: This is Patti
Brennan. I want to thank John for that
comment because it's a segue way to my message
to you.

This afternoon at 4:30 we're going
to be breaking up into small groups as we did
in the past. This is in part response to the
comments from our expert panel and from the
last session there was a great need for more
discussion.

We will have a half an hour to be
talking specifically about what are the
aspects of the NQF evaluation process that
need to be tailored for patient-reported
outcome primary -- performance measures and
aspects of what John identified just now will
be important.

So as we go through the day today
please jot notes on your sheet and we'll have
time in small group discussion with a recorder
from the NQF at each table to get those
thoughts down.

And I'll now turn back to Karen to
continue the program.

DR. PACE: Okay, we'll ask Greg
Pawlson to come up. He's our moderator for
our next session. And I understand that all
three of our panelists are online so I'll let
Greg get started and then we'll go from there.

DR. PAWLSON: Good morning,
everybody. Well, we hope the technology is
going to work on this. It's a great test of
both doing things locally and across the pond
shall we say.

You know, as a very veteran as you
can see by my white hair, my granddaughter
points out my hair is now white. So, I don't
get excited about meetings very often but this
one I did and especially this session.

I think that where we are with
this although we still have clearly some
challenges in terms of the importance of these
measures and how we can illustrate the
importance of these measures. And in
assessing the scientific kind of
characteristics of the measures themselves I
think we're actually farther along in that
than appears to be the case.

But in terms of the feasibility
and usability of these measures and really
going even past the PRO-PM stage and moving
from the concepts to the measures to
performance measures and then to measurement
and the actual application of these measures
in practice I think is our greatest challenge
in this area.

And I think the panel that we've
assembled this morning is really an exciting
one to give us some insights on that because
it represents efforts in three different
countries including the United States to use
PROMs in a very direct and structured way, and
in some cases to actually have results of these in terms of application to specific populations either of disease-specific or condition-specific kinds of measures, or in the case of the Health Outcome Survey that CMS uses the health status of a general population, and what some of the contrasts are between those two approaches.

To guide us along this pathway this morning we're very fortunate to have first of all David Nuttall who's the deputy director of the PROMs Programme with an "e" at the end which I always find interesting. And he is the deputy director for patient-reported outcome measures at the Department of Health of England and he's actually been with that program really from its outset.

Secondly, we have Elizabeth or Liz Goldstein whom I think a lot of you know who is director of the Division of Consumer Assessment and Plan Performance at CMS. And she's going to talk about the Health Outcome
Survey. Liz has always been a very strong proponent of patient-reported outcomes and has also been a very major player in the CAHPS survey process.

And finally, from Sweden we have Stefan Larsson who is the senior and managing director of the Stockholm office of Boston Consulting and has been deeply involved in the program that has been developed for the use of patient-reported outcome performance measures in Sweden.

So without further ado I hope we have David to start off the process. David?

MR. NUTTALL: Hi, good morning.

DR. PAWLSON: And by the way, David, congratulations on your Scotsman's victory last night.

MR. NUTTALL: Thank you very much.

Well, good morning. Can you hear me okay on the line?

DR. PAWLSON: Yes, we hear you quite clearly.
MR. NUTTALL: Can you hear me okay?

DR. PAWLSON: Yes, we can.

MR. NUTTALL: Okay, fantastic.

Well firstly, thank you for inviting me to this Quality Forum meeting. I've put some slides together which are intended to give an overview really of the program of work that we are doing here.

And I think by its nature in 15 minutes it will be a relatively quick overview of some of the issues but I think maybe the important thing is to sort of set out what we're doing completely and then obviously at the end if there's any specific questions I'm happy to go into those in a bit more detail.

So, if I could just go onto the first slide, fantastic. And I don't know if this is going to work with testimony but there's this little animation that will bring in four pieces, a jigsaw puzzle. And I use this slide really to try and demonstrate why
we're interested in patient-reported outcome measures at the Department of Health and what we're trying to achieve.

And in a nutshell I think that the National Health Service here has got a good history of collecting information about care but to some extent it's a bit partial. So I think historically we've had a good set of information about patient's experience from the viewpoint of patients through national survey programs which have been in existence for some time. And I think we've got a wealth of data about healthcare professional's view of patient experience from other routine data sources.

In terms of the effectiveness of care I think again we've got a long track record in measuring from a professional's point of view how effective care is through things like clinical audits and clinical indicators which we derive from routinely collected administrative data.
What I'd argue though is that the effectiveness data is relatively partial insofar as clinical audits are not carried out continuously. They tend to be carried out at point in time looking at particular issues and clinical indicators that we have are generally focused on where things go wrong. So they would be looking at things like mortality, complication rates, transfers to high-dependency care, that sort of thing, less focused on the quality of care for let's say the majority where there isn't an adverse outcome as part of their care. So that problems -- patient-reported outcome measures data is important not as a replacement for the other sorts of information that we collect but really to complement and complete the quality picture that we have about care in the round. So that's kind of the overarching reason and rationale for collecting this kind of information.

If I just go onto the next slide
then. Great. So, in terms of the history of
the program we actually carried out or
commissioned rather I should say a piece of
research in about 2004 which was to look at
all of the available patient-reported outcome
measures that were available sort of off-the-
shelf and to assess their relative merits and
performance, psychometric property and come up
with recommendations for what would work best
in a small number of acute elective
interventions.

So the four on the list there, hip
replacement, knee replacement, varicose vein
surgery and groin hernia repair, and there was
a fifth originally around cataract surgery as
well.

And what that research concluded
was that there was merit in putting together
a generic patient-reported outcome measure
alongside a condition-specific for each of the
five areas we were looking at and made various
recommendations as to what each measure ought
to be.

We then commissioned some further work which ran between 2005-2007 to pilot the collection of those measures in a range of different sorts of units from large hospitals through to much more smaller, elective ambulatory care units and to see how the administration methodology would work, to have a look at collection, to have a look at the acceptability of these sorts of measures from the point of view of both the patients that were completing them and the staff that were administering them.

A couple of points to make from that. I think the first was that we concluded that this was a scheme that had merit, that was cost-effective and ought to be rolled out. But equally that there were some really tricky methodological problems around the collection of this kind of information for cataract surgery with the instruments we had at the time so that the size became flawed over a
In terms of the final point there where I say it evolved over time I think really that's just to say that at the outset we were looking at the collection of this kind of information because we were concerned about the relative performance of different types of units in the NHS where we had good information about cost models and clinical approaches but less good information about outcomes data. And I think the aims and objectives have changed over time as we've become much more interested in facilitating patient choice of provider, of providing much more comprehensive information. So it's evolved over a period of time until now in terms of the sorts of uses of the data, but it continues to attain strong support across the board as a program with merit. If I can just go onto the next slide. The questionnaires that we administer are administered to patients at two time
points, once preoperatively and a second
questionnaire post-operatively at either 3
months or 6 months depending on the
intervention that's in question.

And we have questionnaires which
are effectively batteries of measures. They
comprise a standard EQ-5D profile and the EQ-5
currently. And we paired those up with a
condition-specific for three of the conditions
apart from hernia where we didn't identify a
condition-specific measure which was of
sufficiently good performance.

And around that core battery we
then have on the preoperative one demographic
information. We collect information about
patient comorbidity and so on. And on the
post-operative one we don't replicate the
patient demographic information but we would
also collect some information about
complications, infections, allergy, whether
they'd been back to hospital, and so on. And
that pair of health data measures then give us
a sense of outcome when we look at the
difference between them for any individual
patient.

I won't get into detail about
precisely how we do this but we have contracts
in place with a range of organizations which
help on the logistics side distributing,
producing those questionnaires, scanning them
in and turning them into electronic records.

But during this point in the process is that
it comes back into a body, organization known
as the Health and Social Care Information
Centre which is not part of the Department of
Health but a related organization who have
responsibility for publishing end results as
what we call official statistics.

If I could just go onto the next
page. Just very quickly in terms of the
program we think it's large and significant.
Each -- well, the four interventions comprised
about 250,000 patients per annum. We've been
collecting the data since 2009 and the table
shows the sorts of volumes of information that we've had returned to date.

A large and growing data set I think and I think one of the -- perhaps the most comprehensive data set of its nature as we have a census approach and although voluntary we'd approach everybody that's eligible for interventions from as well as NHS to complete questionnaires.

And we had quite a lot of interest in the program and have spoken to representatives across a whole range of countries about what we've been up to and how some of this can be reproduced. Next slide.

This is just to demonstrate the return rate that we had got. And what the graph is showing that for orthopedics we have historically enjoyed a very high rate of patient participation at around 80 to 90 percent. And we've done less well over time at collecting information for general surgery.

There's a bunch of reasons why
this might be the case from complexity of the
tasknaire in the case of varicose veins to
how much time individuals spend in hospital.
It tends to be admitted whereas the general
surgery procedures will tend to be day cases
and people are in hospital for less time. So
there’s a bunch of reasons.

But I think the key message that
comes from this slide is that actually it can
be done at very high rates. And actually even
within varicose veins and groin hernia which
have slightly lower -- rates. There are
hospitals that will be doing this at nearly
100 percent and have been doing so for a long
period of time. Next slide.

And once the preoperative
administration is within gift of the providers
they physically distribute the questionnaires
to people in their clinic. The post-operative
questionnaire which is sent out through the
post mail 6 months later, the response rate is
much more of a parameter. And the parameter
is about 85 percent for orthopedics of those that complete a pre-op will go on to complete a post-operative questionnaire, about 75 percent for groin hernia and slightly lower again for varicose vein. And that's being consistent -- if we ignore the kind of the wobbly line on the very right-hand side of the data, it's a bit more recent in this graph, then you'll see that they're effectively constant over a period of time. Next slide.

Now, in this slide I'm kind of summarizing an awful lot of information quite quickly but this is really to say that the end product of that data collection, there is a large number of steps which go on in -- period but the data that's collected will form electronic records.

It's linked together with routinely collected administrative data on the sort of nature of the episode. So what interventions took place, how long are they in hospital, what comorbidities were recorded in
the patient record, et cetera, et cetera. All that data is stitched together into an electronic record for each patient. Outcomes are calculated in terms of distance between the pre-op and post-op scores. Measures are available for -- just like the index, the bands in a condition-specific measure.

And we apply a case-mix adjustment to take out of differing case loads of each hospital in this case before constructing average outcomes per unit. This will be displayed on a graph like this. And the funnel plot has become a standard way of us reporting this information back to the providers themselves. And I'm sure many of you are very familiar with it. I think that the volume of records across the x axis, the adjusted health gain changes to the pre- or post-op score accounts for patient case experiences plus unit control limits at 99.8 percent, 95 percent limit.

And then we would publish this
data on a quarterly basis back to providers indicating where their -- statistically different from the national average. They -- where expect the providers to take action as a result. So, as I say -- unfortunately there's an awful lot of work that goes on and needs to be included. This is one of the sorts of output that we would generate for the program. Next slide.

Effectively I started off by saying that one of the rationale for, one of the main reasons for collecting information was to round out and complete the quality picture by giving us a complementary source of information about outcomes and quality. But actually being a bit more specific about that there's a whole range of particular applications that we can use this data for.

And I think the general point is that we don't see that being as primary or necessarily predominantly application of data.

We see the outcome information as being a
resource which could be applied to a whole
range. So, I won't get into this in detail
but it ranges from things like using the
outcomes measure at a local hospital level to
look at the relative performance of clinical
meetings, to have a look at whether the care
that's being offered has particular
consequences on specific domains of the
outcome measures, patient pain or what have
you.

We have an initiative, the name is
Quality Counts where all hospitals have to
provide reports on their performance, provide
a descriptive narrative of why their metrics
look a particular way. And PROMs is being
made a mandatory component of that. So each
provider would have to explain why their data
shows what it shows. Through patient choice
right up to the national level where we can
use this data in aggregate to tell us
something about the relative efficacy of
different interventions from the patient
effectiveness point of view. Next slide.

I just noticed the time so I'll just move through these. But I think over the 3 years since we started collecting this information routinely the data has become part and parcel of the information landscape that we have. I think when we started this initiative the data was seen as very much something which collected in a one-off program and you know, you have to be convinced that there's merit. And I think 3 years on it's just seen as part of the fabric, the set of information that we collect has become terminology that people are comfortable with and are familiar. And it is embedding itself into a whole range of things. We have something, maybe the outcomes framework which is a method for holding the NHS to account and appears -- comprise patient-reported outcome measures. Sort out quality account.

And I think most interesting from my point of view, it's taken a few years but
we're now starting to see the academic
research using this data to flow with some
interesting papers coming out. Next slide.

And just picking up on that point,
I mean particularly looking at case studies
too. This is an example of the sort of
academic research which is coming out. We've
just seen a peer-reviewed paper appear which
is looking at the relative effectiveness of
unicondylar or unicompartmental knee
replacement relative to total knee
replacement. And although I think the
unicondylar are of increasing popularity I
think the research has shown that from a
patient outcome point of view they were very
similar in terms of their effectiveness and
yet the unicondylar has a high revision rate.
So the paper sought to question doing that.
From my point of view I think that's a really
interesting piece of research because it's
going -- leading it away from being just
about patient outcomes and you know, the
softer side of things, to driving some really
meaty clinical -- next slide.

And this is just coming to the end. So, looking into the future we've
covered four elective interventions. We have
quite a lot of work in the pipeline which is
looking to extend the scope of the program.
And this includes a trial which is currently
underway for coronary revascularization, CABGs
and angioplasties where we are piloting some
competent measures with 11 providers. We have
done a cancer survivorship study which was
using elements of the FACS questionnaire and
was sent out much like a general population
survey out to patients which got a very high
participation rate. And looking at other
areas like mental health, care and treatment
of depression, lots there as well.

A few things that we're doing at the moment. One is the development of a, what
we're calling a shorter, sharper generic PROM
questionnaire the point of which will be to
use a much wider range of intervention. And then very briefly we have made some quite significant changes to the way we collect and report the data which allows much greater access to the patient level identifiable data for clinical teams which is something that clinical teams have asked us for. And we are starting to introduce new methods of collecting the data including iPad touch screens, electronic data capture as well as to try and make this the best around.

Then the final slide. To summarize I think our PROMs collection at a national scale with comprehensive coverage gives us a pretty unique insight into the effectiveness of care from the patient's point of view. And we feel confident that the volume and response rate that we've got make the kind of findings that we're coming to quite robust.

You know, a huge amount of work has gone into devising and developing the
methods for getting this data. I think when we started the program there was relatively limited evidence-based literature about how we use this kind of data in the context of routine performance management assessment of this as opposed to the use of it for the appraisal of let's say drugs and the like. I think there's a huge variety of uses we can put the data to and we're now starting to see the evidence build up about the kind of conclusions we can make from this data. I think we'd view it successful and that's why we're starting to roll it out into the other clinical areas. I think that's easily my 15 minutes so I should probably stop there.

DR. PAWLSON: Okay, thank you very much for I think a very good overview of a very complex and somewhat longstanding program that I think is very exciting.

I just, as we start some comments and discussion of this presentation. It's
interesting to me that this is one of the few ways, and from a measurement wonk's perspective very exciting way of starting to get at clinical appropriateness.

We have I think been very stymied in trying to measure the quality of procedures in this realm because of the lack of data of appropriateness. If the procedure wasn't done for appropriate reasons the quality of it I think is very much sort of perhaps even an insubstantial kind of question.

Do you want to comment on how this is beginning to be used in the determination of the appropriateness of some of the procedures that you're looking at?

MR. NUTTALL: Sure. I mean, I think we -- I think our general position is that we don't oppose using the data as a sort of preoperative screening tool. So our position is that it's always at clinical discretion as to whether a procedure is needed or not. And so we wouldn't oppose the use of
PROMs as any sort of screening or rationing mechanism, however you want to think of it.

Instead, the way we see it is actually the value is added by looking post-op and having an assessment of whether actually in different parts of the country we could adopt better scores, had clinical referral, specialty different. So, it's not that the data in itself will allow us to conclude who should have an intervention, that's left to clinical discretion. But we would use the data to assess after the event whether there's something in there which can inform the clinical decision-making if that makes sense.

DR. PAWLSON: Okay. Follow-up comment on that back there?

DR. BASCH: Yes, hi. Hi David, it's Ethan Basch at the University of North Carolina. Nice presentation as always. A quick question, actually a two-part question about missing data which has come up a bit in our conversations here. The
first is a question about response bias, that
certain of the hospital trusts may have higher
or lower response rates. And in general when
they are lower response rates there may be a
lower response from patients who are sickest
who have the worst outcomes. And therefore,
those institutions with the better response
rates may actually be at a disadvantage
because higher response is associated with a
higher number of people reporting worse
outcomes. And how you adjust for that.

And the second is what has been
your approach to missing data not at random,
particularly from hard-to-reach patient
populations who may systematically be missing,
particularly in some of your lower response
rate groups like the varicose vein cohort.

MR. NUTTALL: Sure. Yes, I think
good questions. And we have a piece of advice
which we've put out to providers which is that
we're looking for 80 percent preoperative
participation to ensure that the data we're
collecting is being representative. And the work we've commissioned suggested that over 80 percent participation then the data is generally going to be reliable in and of itself. So that's kind of a first point.

And obviously some providers don't meet that, particularly if you look at varicose veins with a lower average participation rate. So there is an issue to deal with around missing data and potential response bias.

And I think -- two ways of dealing with that. One is that we have done work which has looked at response bias and allowed us to get a handle on what the impact of missing data from particular patient populations is so that we can get, you know, understand what the data is actually telling us when we look at it.

I think the second thing that we -- perhaps most important when we're looking at provider comparisons is through our case-mix
and risk adjustment process then we can effectively give a provider a sort of national average basket of patients and look at what their scores would have been for them. And that's kind of the underpinnings of how our case-mix adjustment works. So, we can take into account the patient mix, we can have a look at what we would have expected to see from them, what we actually got, and correct the data in some sense.

So I think two bits of response. One is that we do the work which actually tells us what the consequence and response bias is and allows us to get a handle on that in the first instance. And secondly, in reporting the data adjust it to take account of that to some extent.

DR. PAWLSON: Thank you. I think in the interest -- I mean this question that you raised and I suspect some others are going to apply to all three data sets. And I think we've just begun to scratch the surface of the
richness of the information they've collected I think and how it impacts.

So what I'd like to do is go onto Liz Goldstein, get that perspective, and then Stefan and then come back and loop in some of the questions. So, after Liz's presentation we'll take questions that are sort of specifically about the issues or methodologies of that set alone. Okay? So Liz, do you want to go ahead?

DR. GOLDSTEIN: Yes, thank you.

So today I'm going to be talking about our Medicare Health Outcome Survey, or often it's referred to the HOS survey. So for the next slide.

The goal in implementing this survey in the Medicare program was to gather valid, reliable and meaningful health status information for Medicare Advantage enrollees. So we use this information in quality improvement, plan accountability, public reporting, to focus on improving the health of
our Medicare beneficiaries. And I'll be going into some of these activities in more detail in this presentation.

The intended uses for the HOS data, as I said, public reporting. And I'm going to be talking a little bit about our public reporting program and how HOS is integrated into this public reporting program. Most recently we're using it for our pay-for-performance program and I'll quickly review that today. It's used for quality improvement activities of the plan, program oversight and in general to advance the science of health outcomes research. There's a wealth of studies that have been done using the HOS survey.

In terms of the HOS overview the survey was implemented by CMS in 1998 so it's been going on for many years. All Medicare Advantage organizations or health plans with at least 500 members are required to participate in this survey.
The survey is done annually with a 2-year follow-up period and I'll explain that in a moment a little more. The focus of the survey is to measure whether the health plan has been able to maintain or improve the physical and health of its members.

In terms of the survey design and questionnaire the sampling unit is a Medicare contract. And as I said before if a contract serves less than 500 enrollees, and we in the Medicare program have a lot of small contracts so this does exclude a number of our contracts that are just very small. The number of small contracts has been decreasing over time so hopefully eventually most contracts will be doing the survey.

If a contract serves anywhere between 500 and 1,200 enrollees all the enrollees are included in the survey. And if they have more than 1,200 enrollees a random sample is taken.

So a beneficiary is surveyed in a
baseline period and then 2 years later we go
back to those same enrollees with the same
cohort to do the follow-up survey. So a
contract at any given year is doing both a
baseline survey as well as a follow-up survey.

Just to give a little information
about how the survey is done, our contracts
have to contract with HOS survey vendors that
are certified or approved by the National
Committee for Quality Assurance. So they
can't use any survey vendor. The survey
vendor has to be approved or certified to do
the survey. It's very important to us for all
of our survey activities that contracts use
approved vendors because we provide oversight
of these vendors, ensure that they're
following standard protocol.

In terms of survey administration
it starts out with a pre-notification letter,
then goes to a first mail survey. Then the
sample member gets a reminder postcard and
then a second mail survey if they haven't
responded yet. If a beneficiary does not respond to the first two mailings then telephone follow-up is used.

And we have found in a lot of our survey activities using this mix mode methodology gets the highest response rate and also gets people more likely to respond by mail or by telephone. Just to note, the survey is done in English, Spanish and Chinese.

The HOS 2.0 survey has 64 questions. And I'll be giving a little bit more information about the questions in a moment. The one thing I want to emphasize here, that this survey is population-based, it's not condition- or disease-specific.

This slide provides just some sample questions on the HOS survey. As I said before the current version of the survey includes 64 questions. It includes the Veterans RAND 12 Item Health Survey. It includes questions about activities of daily
living, chronic conditions. It includes some measures that we collect for our HEDIS survey which includes measures such as monitoring physical activity. It includes information about height and weight, clinical symptoms such as depression and pain items, and a series of sociodemographic questions that are used for our case-mix adjustment model.

The next slide. The HOS survey focuses on two outcome measures of physical and mental health changes for Medicare beneficiaries as I said over a 2-year period from baseline to follow-up. There are -- I just want to give you a little bit more detail on that.

There are eight scales that form the basis of these two summary measures, the physical and mental health status changes. I'm just going to quickly go over these eight scales so you can get some picture of the types of things included in this measure.

They are two questions related to
physical functioning such as the extent to which a respondent's health limits their physical activities. There are a couple of questions related to whether the respondent's physical health limits them in the kind of work or other usual activities they perform in terms of time and performance.

There's one question that determines the extent to which pain interferes with a respondent's normal activities. There's one question that asks respondents to rate their current overall health status. There's one question that asks respondents to rate their well-being by indicating how frequently they experience energy. One question asks respondents to indicate limitations in social functioning specifically because of their health.

There are a couple of questions assessing whether emotional problems have caused respondents to accomplish less in their work or other activities in terms of time and
performance. And there are a couple of questions that focus specifically on how frequently they felt calm and peaceful and felt downhearted and blue. So these eight scales as I said before provide the basis for the two summary measures.

On the next slide the HOS survey was developed under guidance with a technical expert panel. A lot of industry experts have provided input into the initial development of the survey.

We continue each year to look at the survey, look at new methodologies as the state of the art changes. For example, next week we have a technical expert panel that's going to be looking at the survey and seeing if there are additional items to add related to patient-reported outcomes as well as are there any revisions to that current instrument.

Just to provide a little bit more detail, the HOS outcomes are determined by
comparing observed to expected changes in the physical and mental health for individuals in the sample.

One thing that we continue to evaluate each year and it's something that we're going to be paying close attention to in the coming year to see if we want to make some revisions is that the case-mix adjustment methodology is very critical to producing valid plan-to-plan comparisons. The current adjustment for HOS includes variables such as age, gender, education, socioeconomic status, chronic conditions and functional limitations.

I'm going to spend a couple of minutes talking about how we use the HOS data. So HOS for public reporting, and you can go to the next slide. We have a five star plan rating system and HOS is included in this system. So for our health plans that offer drug coverage they're rated on approximately 50 different measures. And so HOS is part of that measurement. So we produce, for every
health plan and drug plan in the country we produce a five star rating and that's our overall rating.

For health plans that offer drug coverage we have nine domains or topic areas. So the topic areas cover things such as staying healthy, managing chronic conditions, experiences of the health plan members, patient safety. So for public reporting purposes we roll it up to these nine domains.

If someone who's using one of our websites wants to look at the individual measures they can go down and look at the individual measures that make up each domain. And for each individual measure we provide a five star rating as well as a numeric number that goes with the measure.

So for HOS it's included in this plan rating system. Last year we made some changes to our public reporting system. Prior to last year all measures were treated equally in the plan rating system so our process
measure and outcome measure were weighted equally, suggesting that they had equal importance.

Last year we moved away from that and right now outcomes and intermediate outcomes receive a weight of 3, patient experience and access measures receive a weight of 1.5, and process measures receive a weight of 1. So the health outcome, those two measures receive a large weight in our system, receiving a weight of 3.

The next slide just is one screenshot from our website. And this shows you when you pull up a plan. Some of the basic information you get up front is our overall plan rating. And then you can drill down to more detailed information as I was talking about.

I'm going to switch just for a couple of minutes and talk about pay-for-performance. Next slide. As part of the Affordable Care Act CMS implemented a quality
bonus payment system for health plans or
Medicare Advantage contracts.

As part of the Affordable Care Act
it said that in implementing this system it
should be based off of a five star rating
system. And so CMS decided that the quality
bonuses would be based off the Medicare
Advantage plan rating. So it's the same
system that we use for public reporting.

CMS is conducting a demonstration
for the first 3 years of the implementation.
And through this demonstration CMS adjusted
the amount of money or the percentages that
contracts would get for each of their star
ratings, trying to really generate more
quality improvement, more rapid and larger
year-to-year quality improvement.

The slide that's up right now
shows the different amounts for the different
star ratings. So the current -- under current
law contracts would only get a quality bonus
if they had four or more stars. During the
demonstration they do get quality bonuses if
they have three or more stars. Once the
demonstration ends we'll go back to current
law and it will be four or more stars.

So, we've seen since the quality
bonus payments were announced to plans there's
a lot more emphasis on quality improvement by
the plans. A few years ago when we put out
the plan ratings each year some plans paid
attention but they weren't paying a lot of
attention. Right now since they do get paid
based off of HOS and our other data collection
activities they do pay a lot of attention. We
get a lot of questions about data and using it
for quality improvement. So there's clearly
a lot of emphasis right now among Medicare
Advantage plans and seeing how they can
improve performance across all of their
quality performance measures.

The last slide is really just a
summary of what I've gone through. CMS uses
the HOS data to determine performance of
Medicare Advantage plans and to reward high-performing plans. It's really used for quality improvement activities and to just monitor how plans are doing. Medicare beneficiaries can use this information to make a decision about which plan to go into. And researchers have used a lot of this information to advance the state of science and patient-reported and functional health outcomes measurement. So, I think I'll turn it back to Greg to open it up for questions.

DR. PAWLSON: Thanks very much. This is sort of a little hidden gem of the Medicare program that a lot of people don't know about. So I hope this gets it to a wider audience than in the past in terms of both its use for performance improvement but also in terms of health services research.

Questions specifically? Judy? And just identify yourself so Liz knows from whence the questions are coming.

DR. HIBBARD: Hi, this is Judy
Hibbard from the University of Oregon. Hi, Liz. My question for you is about the sensitivity of these measures to change. And I was wondering when you developed the measures was that a criteria and what has your experience been over these years of use.

DR. GOLDSTEIN: So when it was developed that was something they really looked at, how sensitive it was to change. I think the thing that we're -- the struggle that we have right now and it's something we're spending a lot of time looking at is that there's some variation across plans but there's not a lot of variation.

So we're spending some time right now in the coming year really looking at that and seeing how the measurement could be maybe improved or tweaked to, you know, increase that variation across plans. But looking at the sensitivity to change, that was incorporated in the initial development activities.
DR. PAWLSON: I think, Judy, you brought up a very key issue with a lot of patient-reported outcomes in contrast to patient experience measures where we've got some pretty good indicators that a lot of the variance is provider-specific. In this case I think and actually David brought -- one of his PowerPoints where it has the outcomes data published by provider organizations adjusting for differences in case-mix. If you look at that it shows very clearly the 99 percent confidence limits around an estimate of the adjusted average health gain. And they're pretty broad.

It reminds me of when we did the resource use measures and the costs where you see pretty wide variation. And that means the sample sizes for these two have adequate reliability and adequate amount of variance due to provider-specific factors is up in the two, three, four hundred range. And so I think that's a real challenge to a lot of the
measures we're going to see in this field.

Yes.

DR. KAZIS: Hi, this is Lewis Kazis from Boston University. Hi, Liz. I enjoyed your presentation.

I've been involved with -- as a consultant to the CMS project for HOS for many years and what Liz says is absolutely correct. There are some weaknesses in the variability that we're seeing across the plans using the VR 12 broken out into physical and mental.

There is new work that we are currently conducting where we combined both the physical and mental into a utility metric called the VR 16. And in very preliminary work we are seeing more signal across the plans as a result of that. So I just wanted to mention that.

DR. PAWLSON: I think that's another good point in terms of how one aggregates different measures into a composite. Starts to build and I know there's
some methodological approaches that can start
to filter out some of the noise and boost the
signal which is I think a real challenge here.

Thank you very much. Can we go
on? And I understand, Stefan, you're going to
have a little bit of an extra challenge
because you've got some internet problems and
you aren't going to be able to see the slides.
Is that right?

DR. LARSSON: Yes, that's true. I
have them in front of me. So as long as we're
looking at the same slides we should be fine.

DR. PAWLSON: Well, if you can
just as you address a slide perhaps just use
a brief introduction of the title of the slide
to make sure we're on the same one, okay?

DR. LARSSON: Yes and I'll ask you
to switch slides for me. Okay, thank you very
much for having invited me to the meeting. So
I'm Stefan Larsson. I'm a partner with BCG in
Stockholm. I'm a medical doctor by training
but I am a management consultant nowadays
although scientifically trained. I've worked as an advisor as BCG does to the healthcare industry broadly, to governments and to corporates and I have worked very broadly in the industry.

One of the things that struck us was very, you know, it's been striking and I'm sure it has been to all of you is the very, very strong focus on budgets and the cost of healthcare over the past years. And we've found an anti-innovative climate that led us to start looking at these registries.

And being a Swede and quite close to the scientific community here I started looking at disease registries, or we did locally in Sweden, and discovered that there is in fact a means to combine the analysis of cost of care with disease-based high-quality data on the outcomes of care.

And we have since -- well, the past 3 years invested heavily in what we and others call value-based care which is on a
disease-by-disease basis to thoroughly make sure that we have data on outcomes, both traditional outcome measures as well as PROMs, and then use that as the numerator to then compare to the money spent on that particular disease along the entire care chain.

Registries in Sweden were started, the first ones in the seventies. If we switch to the next slide today there are about 100 registries covering -- and they're all disease-based with maybe -- there are a few exceptions. There are one or two which are registries monitoring outcomes for elderly care patients, so multi-disorder patients. But 95 percent of these registries are disease-based.

The focus is on outcomes. These have been research registries from the beginning where clinicians, the specialist societies and not government and not payers took the initiative to set up these registries to discover within the clinician community
what determined good outcomes in terms of which clinical procedures, which medical devices were better than others and so on and so forth.

This has evolved to become increasingly a tool to drive continuous improvement. That's where we as management consultants and interested in transformation of healthcare, the improvement of healthcare found this to be an extraordinarily exciting tool.

The majority of these 100 registries have over 75 percent of the patient population covered which means that from a reliability point of view at least in terms of coverage these are quite unique repositories of rather complete populations.

Furthermore, as these have been used primarily for research purposes validity and so on and so forth has been a lot of time invested into making sure that the data quality is high. And that's also why the
results, the analyses of these registries have had a lot of impact on the medical community.

The measures chosen, the way the data has been captured through genuine engagement by the clinicians and then the validity done by highly qualified peers has led to the analysis resulting to have been interpreted as very important to determine best practice and change clinical procedures and clinical guidelines.

A majority of registries have PROMs as part of the tool set that they use. And there is a lot of interest in PROMs which has grown over the past couple of years. Some of the registries were fairly early to look at PROMs and I'll come back to that. And the data they have shown and the impact that has had on clinical practice has led many others to say we also need to find good PROMs that will let us, would allow us to be smart about how we treat patients. And as was said earlier, to make sure that we have appropriate
treatment and not overuse of surgery and so
on.

The next slide has four points to
it. These are the official arguments why
there is a strong surge of pushing for PROMs
across the Swedish registry. You know, as we
understand the diagnosis and treatment of
disease we realize the complexity of patient
segments and we need to gather more phenotypic
data than before. There's an increasing view
that personalized medicine will not be
answered through genomics alone, but we need -
- we understand by obviously combine genetics
and genomics with phenotypic data, that's
where the observational registries play a very
important role.

Secondly, there is the continuous
growth of the treatment options for patients.
And we need to find ways of understanding who
should have what. If we were trying to
discover that through classical prospective
double-blind studies that would take forever
and therefore these registries turn out to be a very important tool by which we were able to match treatment alternatives to patient segments, patient cohorts.

Thirdly, the healthcare system is changing very rapidly. There are new players coming in. The Nordic in Sweden is privatization happening. There's concern that there will be slippage of indications for elective surgery for instance. Therefore PROMs have been seen as a very important tool to make sure that we don't treat in ways that don't benefit the patients.

And finally, and that's the point I often made above, the individualized medicine need, this is seen as a very important contribution or tool as well.

On the slide after that I've just summarized a set of points to illustrate that there is a national organization for payers in Sweden. That organization has defined as one of its top strategic priorities to promote the
use of PROMs in Swedish healthcare.

So they brought together a group of highly qualified experts who are advisers to the registries who are seeking to develop PROMs to make sure that that's done the proper way through a dialogue with other colleagues who have experience from it, and that the validation models, et cetera, are thought through properly so that the data can be interpreted correctly.

I'm not either an epidemiologist nor a statistician so I will not be able to answer questions about the methodologies used for validation here, but as I understand it this is a well-equipped team that's supporting the registries to do the right thing.

The next page, page 4, illustrates the types of PROMs which are being used to the right of the graph with the dark green color. We've grouped the PROM categories into four. And as you see roughly somewhere between 40-60 percent of the registries use PROM. And more
than half of the registries use at least one of these categories of PROM measures in the surveys they have to patients. So, both activities of daily life, the patient perception of the symptoms they have, general satisfaction measures as well as quality of life measures are being used. Often some of the standardized tools, international tools, in many cases but some cases these are unique for the particular disease and thus specific for the registry question.

On the next page, page 5, I just wanted to illustrate something that we've seen also internationally. I'll come back to that when we look at PROMs in registries across the world. That is differences between segments of medicine.

The orthopedic registries were some of the first ones in Sweden. A hip arthroplasty registry was founded in 1979. Cancer registries were also quite early. But it is interesting to note that, it shouldn't
be surprising maybe, but the orthopedic registries have quite early on invested quite heavily in PROMs and more so than in most of the other disciplines that we have represented on this and the next slide.

It is interesting that in the cancer field even for cancers such as breast cancer where a very large number of the patients treat, you know, nowadays survive for a long time PROMs are very uncommon. And we've seen that globally when we've looked through cancer registries in many countries.

Here the survival data completely dominates and even the -- if the patient survives and many do in some cases, you know, these quite heavy treatments we're not really following up how that influences the activity of daily life short- or long-term. And I do think that if we want to have -- our view as advisers to the drug industry for cancer drugs is many of the innovations do not necessarily extend the length of life but may have
radically better outcomes in terms of other consequences for the patients. And we're not really looking at that in most of the registries. And here we think adding some of the PROMs for some of the conditions will be important in order for the pharma community for instance to drive innovation in a way that also benefits the patients beyond survival.

We can skip the next page. It just shows a set of therapeutic areas that we've -- in Sweden and the degree to which they use PROMs, and move over to page 7. We have over the past year initiated an effort together with Michael Porter and the Karolinska Institute formed something we call the International Consortium for Health Outcomes Measurements, ICHOM.

This is a not-for-profit effort that will be launched first of November this year. And what we've done there is we have gathered the measures from 55 registries across 16 conditions around the world, and in
a very systematic way organize those measures
to make it easy for clinicians around the
world to be able to search and compare what is
being measured across different diseases and
then across geographies.

The intent of it is by having
chosen registries that are perceived to be
leading for, you know, this limited set of
conditions we want to make it easy for
clinicians who want to start measuring to
choose measures that others have chosen and
validated and where there are large amounts of
data so they could easily start comparing
themselves to others.

So the intent of this effort is to
contribute to a standardization of what we
measure disease by disease, and therefore
contribute to a general standardization of the
way we look at outcomes and thus allow the
medical community across borders to an
increasing extent compare -- identify best
practice, compare results, identify best
practice and share that and thus drive the
development of clinical practice forward much faster than will happen if we don't have these measures and are enabled to compare apples with apples.

If I could -- I won't dwell longer on that but will be happy to -- could be a topic for a separate discussion. But we, this is a very exciting effort and we can talk more about that if somebody has any questions.

But this particular slide illustrates some of the PROMs that we have pulled together and which ones are used, which instruments are used for some of the different registries we've looked at. You see for instance the Swedish spine registry, Swespine. We have the Singapore General Hospital low back pain registry. There's a Norwegian arthroplasty registry, et cetera. They're all using EQ-5D for some of the PROM measurements they do.

You see that the same registries
also use other instruments and there is a fair amount of variation. We hope that transparency on this will allow for comparison discussions and maybe ultimately a larger degree of standardization making comparison across diseases and geographies easier.

The final two slides I wanted to show you are the question of so how is this being used, how are these PROMs being used. Do they really influence clinical practice or not? And one of the registries -- there are two registries I'll talk about. The first one is for hip arthroplasty, the other one is for rheumatoid arthritis.

The hip arthroplasty registry started using PROM protocols in 2002 and it came about because of the fact that they had registered for many, many years the duration of the hip arthroplasty. So, and found surgical techniques and medical devices which led to much longer survival rate of the implant and thus higher degree of satisfaction.
of the patient.

But it took maybe 10 years to optimize that and then they've had difficulties in improving further. So they're now turning to say we've taken innovation and development of the surgical technique as far as we can. How can we improve the health of the patient further?

And of course the indication of hip arthroplasty is that you have, you know, pain or you are unable to function normally. You are impaired quality of life and yet the measurements, quality measurements have often been the more mechanical one, whether the implant has a long survival time or not.

So in 2002 they started looking at PROMs. And without going through the details there was a THA thesis published 2 years ago about the PROMs. I spoke Friday met with the leader of the registry and discussed a bit of this with him.

One of the observations they've
made which is very interesting is that there is a strong correlation between the mental status of the patient preoperatively and the degree to which they find that the surgery has become better or has led to significant improvement of their health. We've of course seen that in other conditions in medicine but here numerically it's become quite clear.

They've also found that if a patient has been taken off antidepressive antianxiolytic drugs prior to surgery well then the outcome post surgery is also worse.

So what they're doing now is they're translating some of the data into decision support to try to help the surgeons customize the treatment depending on the profile, pre-surgical profile the patient has, and to make sure that the outcome in terms of quality of life and functionality is significantly better. So the PROMs turns into a decision support tool for the clinicians in order to optimize the outcome beyond the more
classical outcomes looked at.

The next page and the final one is from the rheumatoid arthritis registry that quite early also started using PROMs in 1996. They have taken this further than any other registry in Sweden and earlier. It fully integrated the PROMs into the registry.

Every patient fills in a set of questions prior to visiting the doctor. In the doctor's office the results from the registry are visibly seen graphically and the patient gets immediate feedback as well and can see how the functionality has changed over time. So it becomes a disease management, you could say a support tool for the patient. It increases their disease awareness at the same time as it allows the physician to be more effective during the patient visit by having seen what the patient has reported in the balance of the interview, the meeting.

The PROMs in this registry, you know, has led to generally a much more
holistic view on the disease. It has raised
the patient's knowledge and influence over the
treatment and enabled them to influence the
choices to be made.

It's actually also, because it's
been done in a very rigorous, scientifically
sound manner actually increased the status of
the subjective dimensions of rheumatoid
arthritis discussion or diagnostics within the
physician community. So it's now seen now as
a very beneficial way of contributing to
better clinical outcomes.

And finally, it's -- the results,
the outcomes, the PROM-based outcomes is also
leading to changes in resource allocation so
that more resources are placed in some of the
subcategories of patients. And they're
classified partially through the PROMs now
which contributes to the choice of
pharmaceutical intervention which has been
shown to lead to a better outcome and thus
more productive use of some of the expensive
drugs that you have as treatment alternative
for this patient group.

So I'll stop here. I wanted to
provide you with an overview specifically from
Sweden where there are more registries than in
any other country in terms of high coverage
registries where PROMs have been used for
quite awhile.

And I think we're seeing some very
exciting examples of how it leads us into not
only continuously improving clinical practice
but also serving as an online decision support
tool to the clinicians as well as a tool to
help patients manage their disease. And thus
contributing to more efficient care while at
the same time improving the outcomes and the
quality of care.

So I'll stop there and open for
questions.

DR. PAWLSON: Thank you. You
know, two quick observations. One is isn't it
striking that we're just getting around to
including the patient perspective on diseases that are defined by patient symptoms and patient functional status in terms of outcomes.

And the second is I have this vision dancing in my head of orthopedic mental health teams collaborating.

(Laughter)

DR. PAWLSON: That may be a little farfetched. Questions specifically for Stefan? And then we'll open it up for general questions. In the back, all the way in the back on the right then I'll move this way.

DR. SALIBA: Hello, this is Deb Saliba from Los Angeles, UCLA VA and RAND. I think it's phenomenal what you guys have done to use your data and to take a look at it and ask questions about improvement and what's been driving improvement and why you've seen a flat line in some areas of improvement.

One thing you may want to think about when you're looking at arthroplasty, I'm
not seeing a whole lot about post-acute care or rehabilitation issues. And I completely agree that mental health is really important but I think you also want to look at what's happening in that recovery period.

DR. LARSSON: Yes, that's absolutely right. Today reimbursement is increasingly looking at the 2-year period and the outcomes achieved in that 2-year period is what some of the prior providers are being paid for. That of course includes the whole recovery phase and it provides a strong incentive to do not only the surgery well and avoid surgical infections but actually be very early on in mobilizing the patient.

I think that what the registry allowed to do quite early was to, you know, to at the same time as you put the RG system in place you have the registries in place. So the length of stay for hip arthroplasty patients is extraordinarily short as I think it is in many places in the U.S. But it's
been demonstrated through the registries that early mobilization has many advantages and not the opposite. So absolutely, the patient part is absolutely an element of what they measure that didn't show that properly.

DR. WU: Yes. Hey, Stefan.

Albert Wu.

(Comment in Swedish)

(Laughter)

DR. LARSSON: Nobody has understood that, but that was Swedish.

(Laughter)

DR. WU: Oh sorry, I forgot where I was.

(Laughter)

DR. WU: I think this is fantastic and I wonder if you could give us some examples or if there have been examples where you have used PROMs to compare one institution to another since you have such good coverage, particularly for some conditions.

DR. LARSSON: Let me see. I think
in the rheumatoid arthritis registry they've seen big differences between centers and how they have -- what they have achieved on PROMs. And the registries take a fairly active role in not only making the data public. You know, they each -- there are roughly I think for many of the registries somewhere around 70 hospitals involved and much of the data is made publicly available today. So not only can you see where you are yourself but you see where all your peers are.

So, the outliers typically will be contacting those and have -- those with poor results contact those who have very good. And for some of the PROMs in the rheumatoid arthritis registry there have been big differences between centers. And that typically leads to a dialogue where they then figure out what to do to improve.

I can't give you an example of specifically what those, you know, examples of those measures but I have heard several
examples were cited that the big differences
that are leading to learning across and
changes in clinical practice.

DR. BASCH: Hi Stefan, it's Ethan
Basch at University of North Carolina. Nice
presentation.

To Albert's question because I
think that actually is, it's a really
important question. So the Swedish Rheumatoid
Arthritis Quality Register collects data at 64
clinics and it actually looks at change scores
over time and compares them between those
clinics. It looks at swollen joints, tender
joints, EQ-5D and a couple of other measures
of functional status.

But to my knowledge that's the
only one of the registries that's really
actively been comparing between practices.
And I think it really highlights the
difference between using registries for
effectiveness research versus performance
assessment.
There are lots of registries that use PROs but they haven't explicitly been designed to compare performance either within or between practices. So I think the challenge here is to take some of the techniques that have really been honed in CER and various contexts and bring them into the performance improvement setting.

Nice presentation. Thank you.

DR. PAWLSON: Very good point. I think we'll open up now to all the presentations. So if you can just direct your questions on who and we'll go to the back there and then we'll move forward.

DR. GAGE: Thank you, Greg.

Barbara Gage from Brookings.

A little bit of a follow-up on Ethan's point. In thinking about -- that was a very nice presentation and in thinking about the use of performance measures for payment policies and some of the monitoring that happens in a regulatory nature, Stefan, could
you say something about the discussions that occurred regarding holding a provider or a clinician responsible based on the patient-reported outcomes relative to the clinician's assessment of those outcomes? That's been kind of a critical issue over here in the States.

DR. LARSSON: Well, first of all there's been no reimbursement linked to the outcomes of the registry so far. There are some early events that are happening in Stockholm that have actually been so far no reimbursement, you know, difference linked to the performances. Sorry, I lost part of your question. Could you repeat that?

DR. GAGE: Sorry about that. The question had to do with the discussion even in holding -- even in using the registries in order to make the decisions about the treatment options to --

DR. LARSSON: -- to the physician responsibility. Yes. So, what the view is in
many of the registries that use this is that the team treating the patient has a very essential role to play in ensuring that the patient reports satisfaction or that the functionality for a newly operated hip joint patient, et cetera, is high.

In discussions with the hip arthroplasty registry, they clearly see that as a failure on their part if the patient is not reaching the target levels of satisfaction and physical function that they had set out. So there is no responsibility in legal terms, that's obviously of big importance in the U.S. because we don't have the litigation component very much at all in the healthcare system, but it is more from a professional integrity and peer pressure point of view very important. And you're seen as in charge of making sure that the team, be it rehab team or others who -- or physical therapy teams that will contribute to your reaching the functional target. You as the physician are seen as
responsible but not in a legal sense. I hope that answers your question.

DR. LOHR: I wanted to compliment all three of the speakers, I thought they were terrific presentations. And this question is directed at any of you who might care to answer. And it has to do with, say, feasibility and usability and costs. And I'm curious whether any of you have any information that would tell us something about the administrative burden, the burden on patients, what these systems cost for getting this kind of information.

And sort of moving on from that set of questions to whether anybody has ever really tried to look at the cost-effectiveness of doing this. I mean, we're all here because we believe passionately perhaps in using patient-reported outcomes and so forth but lots of people are more skeptical and I'm curious whether there have been -- any analyses of sort of the cost-effectiveness,
the return on investment of doing this kind of effort. Thanks.

MR. NUTTALL: It's David here. I can have a stab at that if that helps. In terms of from my experience --

DR. PAWLSON: We just --

MR. NUTTALL: Hello?

DR. PAWLSON: We're losing our foreign end. Not so foreign speakers.

MR. CUNNINGHAM: Are you still on the phone?

MR. NUTTALL: Yes.

DR. PAWLSON: Okay. So David or Stefan or Liz, would you address that issue?

DR. LARSSON: David started I think, at least I heard him well.

MR. NUTTALL: I'm not sure how much you got. I'll have another go. I'll keep it short. But I'll just say from my experience that I think the cost question is probably going to be contingent on the nature of the system in which you're trying to
collect this data.

So, in England, in the NHS we make
the collection of this data a mandatory
requirement. It's a contractual term that
anyone that's providing NHS-funded services
must collect this data as, you know, a term
and condition of doing business with the NHS.
So, in terms of the burden around
administering a preoperative questionnaire,
that falls to the provider.

We have done some work to have a
look at what that burden is and to be honest
I think it's correlated highly with let's say
the quality of management of the institution.
So, some providers tell us this is trivial, we
fit it into existing practice, it takes no
time at all. Others tell us you know they had
to appoint a senior manager to oversee the
entire process.

And so the kind of cost of that
side of it go hand in hand with how
complicated they've made the administration
method. I think the message would be it can be done very light touch and easy.

And then the way we have implemented the actual collection of the data and the processing is to do that through an outsourcing contract. And I think it's -- while I can't go into the detail in terms of unit costs of that contract I think it is fair to say that over time as people have understood the "ask" a bit better and understood precisely how to collect and process information, then the unit costs are falling. As people become more familiar with what's required we learn how to iron out some of the bit that we put into specifications which perhaps are less necessary and we've streamlined the collections and so on and so forth.

So I think burden is dependent on how it's implemented. I think costs are dependent on the nature of the system which are implementing it, and I think over time
those costs will come down.

DR. PAWLSON: Stefan or Liz?

DR. LARSSON: Yes. The situation in Sweden is similar to in Great Britain. The majority of healthcare is publicly delivered and publicly funded. But even now that the privatization is growing quite rapidly contributing with data to the registries is compulsory for the private providers as well. It's part of the contract that you have to submit the data. And it becomes part of the quality control also of the private providers.

The costs involved have been you could say hidden in the monthly salaries of the staff. It's been seen as one of the tasks of team members to gather the data and submit the data. And there have been some calculations done as to so how much is that all in all. We've seen some of those. I'm not sure they're accurate.

The specific funding that's come from the government to support these efforts
were very, very small. You know, at the time a couple of years ago when we did our first study on this there was 6 million Euro being paid by the government to the registries which supported some staff across some 65 registries.

We did a business case for the government on this and said you should at least fivefold this allotment to make sure the registries are sufficiently staffed and that IT platforms are being built appropriately, et cetera. So now the -- and the decision was taken by the government to do so. So this year the budget is 20 million Euro and it's growing to 30 million over the next year or two.

But that's still a quite, you know, small amount of money. So much of the time is in fact salary time paid by the health system as such.

When it comes to the return on investment, you know, one of the things that
PROMs would allow you to do is to look at the appropriateness of care. And we have a team currently looking at the U.S. healthcare system from the point of view of if it was managed more on outcomes and value what would be the savings.

I would argue that it would be hard not to have a business case that would be convincing for PROMs when you compare the healthcare cost in the U.S. and the outcomes compared to other OECD countries. I'm absolutely convinced that the business case would be staggering. I don't know if that's the right word for it -- that's the word that comes to mind.

But there is, I mean a challenge is data-gathering is done by clinical staff that are pressed for time and who if they spend time gathering data like this would see one or two fewer patients a day which is where the bread comes in. So the challenge will be not that the business case overall from a
societal payer point of view wouldn't be convincing, it would be to motivate finance to gathering of data and to motivate the clinicians to do so at the expense of potentially seeing fewer patients and reducing revenues. So I think that's one of the key hinges here.

And finally, I think it would be key to make sure that the data-gathering is as simple as possible, that some of this is integrated into the electronic medical record, some of it is web-based, automated as David described in his presentation is happening over in the UK. And in Sweden that is happening but it's been reasonably slow. So it's actually taking a lot of time from clinicians to gather the data so far.

But I think the business case would be absolutely convincing, you just need to find the -- move money from one place to another in order to fund it.

DR. PAWLSON: The point you just
made I think is really important to keep in mind, and that is the extent to which this information is really critical in treating the patient. And I think a reasonable amount of it is, or at least should be. It's nice to know what our outcomes are.

I think we can probably take two more questions. So, here.

DR. FRANK: Hello, this is Lori Frank. I'm with the Patient-Centered Outcomes Research Institute. Thank you all for those presentations.

My question is a follow-up to the last question about optimizing use of collection of these PROMs. In what ways is the full value of the PROMs being recognized in terms of their value for direct-to-consumer communication? So, you know, for David I know there's some efforts with regard to choice and for Liz, I'd be interested in hearing more about the rating scale and how that's communicated and how PROMs play into that.
DR. PAWLSON: Okay, Liz, you want to address it first and then David?

DR. GOLDSTEIN: Sure. So for the Health Outcomes Survey we've made these outcome measures very prominent in our system, or at least prominent in our calculations of our overall rating for a health plan. So, these measures receive our highest rate as an outcome measure. So, we're trying to convey to the public that this is really critical information, actually, one of the most critical pieces of information in our rating system.

We encourage -- we do a lot of testing of our displays for Medicare beneficiaries. And so it's really hard to get them to drill down to this detailed information. Some of them are really information gatherers and go down to the details a lot, just use that overall rating and ignore the details below it.

But something at least CMS is
doing a lot of work on as well as doing more research about how to get people to drill down and really see the value of these measures as well as the other measures included.

DR. PAWLSON: Thanks. David?

MR. NUTTALL: I think certainly with our program we have a challenge in terms of making best use of this data and conveying it to patients. I think when we set up the program our thought was that the data would be used primarily by patients to help them make informed decisions about where to go and we would publish it via website and we would go from there.

And I think actually the key piece of learning has been that the main audience for the data at the moment other than kind of managers of organizations looking at the aggregate stuff is the clinical teams who want to have much greater access to the disaggregated data so they can have a look at which dimensions their patients are doing
particularly bad in, you know, is that pain, is it to do with pain control, and so on and so forth. So, I think that's been a really important learning point, that it's the clinical teams that has the bigger oversight than patients in the first instance.

And I think the other learning point was that we have a lot of work to do to try and convert EQ-5D profile scores, EQ-5D index scores into meaningful information for patients that's kind of digestible. I think aggregate scores saying one is better or worse than the national average is probably fine but turning in a kind of 2.7 percentage point move on the EQ-5D index is kind of meaningless to a lot of people. So that is a challenge I think.

DR. PAWLSON: Thank you. Steve?

DR. LARSSON: If I could just comment --

DR. PAWLSON: Sure.

DR. LARSSON: -- on the way it's
been used in Sweden. You know, the data has been public but oftentimes the data has, as David said, very complex to interpret. And I think an effort to just broadly disseminate this information I think would be completely confusing to patients. And in order for it to be a useful tool we, clinicians need to spend time making sure that we help with the interpretation of it so it becomes meaningful.

So I think the specialist societies play a very important role here in making it available in a way that makes it useful for patients.

We just did a survey across 9 nations for 1,000 consumers randomly picked in each country. And it turns out that clinical outcomes is the determinant that patients primarily want to see in their choice of medicine. It's often been viewed that proximity and a hospital close to home is important but it's clear that patients don't have access to the information but when the
question is asked if you had access to high-quality information what of all these variables would you be using, would be most important to you, and by far clinical outcomes comes out as the measure they want to have access to. So I think we would do medicine as well as the public safer by helping package that in a good way.

In some cases media has done it reasonably well with, you know, scientific journalists who understand enough of medicine to make sense of the data. But there can often be errors there as well.

DR. PAWLSON: Steve.

DR. FIHN: I think this question is directed mostly to David but perhaps Stefan could respond as well. I'm curious about the extent to which you've been able to evaluate whether implementing the PROMs has changed the case mix of patients who are undergoing the procedures. You could think about changes that would be both in a positive and negative
direction. And I'm just curious the extent to which that front end part has been evaluated as well as the sort of change and outcome piece.

DR. PAWLSON: And that was Steve Fihn from our VA.

MR. NUTTALL: It's a good question. To be honest I don't think we've done a huge amount of work to have a look at changes to, you know, as you suggest the case mix which is going through. I think there's probably a couple of points I'd try to make there.

One is I think, you know, although it has been running for some time since 2009 I suspect it may be too early to tell in terms of it's only really it feels that now, kind of 3 years later that we're getting a lot of traction with people looking at this data in a very serious way and using it to kind of think about clinical practice.

So it's almost like the first
couple of years were just embedding the
collection and getting it firmly lodged into
people's minds as a valid data collection
tool. So it may be too early to kind of do
that kind of analysis although it is feasible.

I think the other sort of factor
to bear in mind with all of this is the impact
of the general impact of the economy in terms
of our funding of interventions. There's a
line of argument which is that we will be --
irrespective of the data set we would be
seeing the more severe cases going through now
as commissioners decide to -- or implicitly or
explicitly put more stringent referral
criteria on. And I think that's something
that we could look at in time.

And then the only other kind of
final, final point on that would be I think
what we have seen is that on average the mean
scores have gone up over those 3 years. And
so that would be consistent with either
quality of care going up which would be great
but that's a short period of time for that to happen in, or that we're actually focusing on the cases where there's most potential clinical benefit. So that's not really a full answer, more just a set of things to think about I guess.

DR. PAWLSON: And Stefan, do you have a quick comment on that?

DR. LARSSON: I can't say that it's influenced case mix to my knowledge. I know one of the observations done in the rheumatoid arthritis registry was the observation that smokers turn out to much less responsive to TNF-alpha inhibitors. I think that piece of information came through the PROMs. That clearly led to changes in the prescription pattern for those patients. So I think that might be an example but it's not the typical PROM measure either. So I would have to look into that.

DR. PAWLSON: Well, on behalf of everyone here I'm sure thank you so much for
your really excellent presentations.

(Applause)

MR. NUTTALL: Thank you very much, it's been a pleasure.

DR. LARSSON: Thank you.

DR. PACE: David and Stefan and Liz, thank you for joining us. And you're welcome to stay online and listen as long as you like but we're going to take a short break here. And again, thank you so much for taking time out to share your experiences with us.

We'll take a break now and try to limit it to 10 minutes. We'll reconvene the next panel at 11:25.

(Whereupon, the foregoing matter went off the record at 11:13 a.m. and went back on the record at 11:28 a.m.)

DR. ADAMS: Okay, so our next panel is going to be on a recap of the key characteristics for selecting PROMs for use in performance measurement. First I'm going to introduce our panel to you but if you recall
from our first workshop we spent quite a bit
of time on these characteristics and also
everyone was sent out a survey because staff
took an attempt to distill that information
which was very rich.

So we're going to have a chance to
do a bit of recap not so much to spend time on
the first workshop but to really use this as
a springboard for our discussions as we take
deeper dives into discussions around
reliability and validity.

So I did want to introduce our
panel for this session. My colleague, Karen
Pace, who is our evaluation methodologist
expert here at NQF. And Karen's going to for
each panel you'll see she's going to go
through in a bit more detail the NQF
endorsement criteria and how this relates.

And of course what we're going to
be asking from you and including our audience
and those listening on the line is what are
some considerations we might need to take in
regards to PRO-based performance measurement.

I'm also pleased that Liz Mort is joining us from Massachusetts General Hospital and Laurie Burke from the Food and Drug Administration. Regrettably, Jennifer Eames who was also going to be on the panel could not join us today. She does send her regrets but Patti Brennan is going to fill in for her. So, I want to thank our panelists for the prep that really helped shape this and for their contributions during the session.

So I'm just going to get started. If we could have the next slide, please. So, we're always going to touch back to this terminology. As Joyce said earlier, the terms we use of course are important. And here we're going to be talking about the characteristics that you had identified for us last time from a PROM which is an instrument or scale to a PRO-based performance measure, a PRO-PM.

And as we think about the entire
day we're building a pathway. So we have a schematic representation that was sent out in your handouts. We have a color version at your seats. But what we're trying to do very pragmatically is build this tool or these building blocks as we think of how we would go from a PROM to a PRO-PM. So we're keeping that in mind as we go. So if I may have the next slide, please.

So I'm going to do a little bit of history. At our last workshop, the expert panel, we discussed what I would consider the highest leverage characteristics for identifying PROMs which are most ready for prime time as we start to think about performance measures.

And I think that there was general consensus that the psychometric properties that were very elegantly detailed in the paper with David Cella and colleagues was really a baseline. And so David, allow me to thank you and your team for the terrific work you did on
the paper but also from the last workshop, some really great discussions.

So we felt that what we're calling affectionately it was Table 4 in your handout really was a very good baseline for us. However, I think that the discussions last time really offered some very helpful guideposts. You know, what do we hold true particularly as we think about moving this into accountability type of programs.

And so from that rich discussion the NQF staff -- I say this humbly because we got such great feedback -- we tried to distill that into some additional statements which you provided input on. And those were sent out to you and you completed a survey saying, you know, did you agree, did you agree with some modification or did you disagree.

From a survey perspective we're very pleased with the response rate, over 70 percent from our expert panel so thank you very much. And I think particularly since
this was done during holiday time at end of August I thank many of you for doing that from your vacations.

So today we're going to really spend time on refinement of these characteristics, and I think in particular looking at this in relationship to the NQF endorsement criteria. And I think from my perspective having done some of the synthesis here with our team these are very mutually reinforcing. So I'm going to ask for the next slide.

So, if you'd like to look along in your handouts that you received we provided an attempt to do some redline edits to the statements. And we had statements around actionability and meaningfulness and facilitating shared decision-making and implementableness.

As opposed to going through the redlines at this time we really wanted to focus on the high-level concepts and of course
our reactor panel here will be helping us. And as we said, terms and words are important. So as we go throughout the day we will further refine these statements. But I thought it would be important to share back some of the comments that were received and some of the themes that converged around this survey data. And I hope that we've captured your voice adequately but I know that you'll be able to help us out here. But I think in this theme around actionability it's this notion that key end users, and importantly patients and persons and providers and systems should be motivated by the PROM to lead improvement.

And this thing of amenable to change or would this spur, but I think this was an important theme, particularly in the written comments that came through. Also, that the evidence should indicate that care can be improved in a relatively short time period for the patient respondents. So we're
always taking our patient- and person-centric view and that there should be value with this. And many felt that this was very much linked to our meaningfulness criteria. So many of these characteristics aren't mutually exclusive but I would say mutually reinforcing.

Also, it was pointed out that certainly randomized control-level information or evidence is critically important, but that we should, and particularly when we're looking at this now and in our early stages take into consideration a range of evidence. And certainly what we're doing now with expert-based opinion, face validity, other things that those would also be important things to consider. I think we had this discussion last time, that certainly evidence is along a continuum and it needs to be applied appropriately to what's being examined. So, but this was reinforced with several of your comments so I did include it here.
And then this was a very enlightening comment for me. I think that we touched on this a little bit but I welcome additional insight here. It's that some outcomes are worth measuring that might not be amenable to change by providers but patients need to make informed decisions and it could be quite useful to them. And so some of the examples were given were pain after intervention and functional status and treatment.

And because I spoke with Jack about this earlier I'm going to tee him up for that. And I thank you for raising that for us because I think that is -- it's beyond a nuance and I'd love for us to have a bit more further discussion on that.

And so these are the key themes from our first area around actionability. We're going to go to our next area in the next slide and meaningfulness.

And when I looked at the
qualitative comments that were provided we saw lots of intersections with meaningfulness and burden, and in particular burden to respondents and administrators, and the implications for that.

In our chart that David Cella and team very succinctly distilled for us burden was called out as a characteristic in and of itself. It raised to that level. But I think Patti, you as well as others remind us but also saw very strong connections to the meaningfulness.

And when we think about the PRO it's the concept that the PRO is capturing, that it's -- we have to think about what the PRO is capturing, not just the PRO itself. And that it's important to include the patient's perspective on the impact of the condition or the treatment and this impact on their life. And I think when Karen speaks to the NQF criteria this certainly is something that we'll touch on a bit more.
I think many in our disability community emphasize for us and we tried to capture that in the redlines but we welcome further refinement here is that we often think of things like health-related quality of life, symptoms, et cetera, but that also we would want to think of certain long-term care service and supports that you would want to capture that move us beyond an acute episode of care and into our community support. So keeping in mind that we have been pushing beyond the hospital walls.

And importantly that when we think about perspectives that caregiver's perspectives would be important to include as well.

So I'm going to go to the next category. And this is the facilitate shared decision-making characteristic that we would like to apply to PROMs. Based on the comments and naturally we welcome discussion here I don't think that the respondents disagreed at
all. They supported that certainly shared decision-making is a critical concept and a critical process that needs to take place.

What there was some indecision and so I put this forth simply to capture some of the insights that were shared is that not all performance measures need to facilitate shared decision-making. This was a common theme amongst a couple of the respondents. And so if this is going to be a characteristic that we apply universally is this something we should consider. I put that forth. Obviously this is up for discussion but since it was raised.

And also there was a comment that shared decision-making, not all patients want this or it may not be evenly distributed, et cetera. So are these things we need to take into consideration.

And a couple of respondents thought that this might actually have some redundancy or duplication with actionability.
Once again, not to say that the respondents firmly supported shared decision-making and that this part of engagement is critical, but just some considerations in that regard.

If you recall we, in some of the discussions from last time we talked about how there might be a -- how we might sufficiently standardize some of these outputs and roll them up to a population or accountable entity. And some felt that saddling that only onto shared decision-making wasn't quite fair because really these aggregation issues apply across all our PROMs.

But I think here it was, you know, how do we -- since we want to look at patient's preferences and things like that how might we standardize that. Because you're customizing but you need a level of standardization. So, whether we should in that statement not hang the aggregation issue specifically on that because it is crosscutting across all of these.
And then you know, there was a comment around is shared decision-making too broad, should we say that this is patient engagement, that certainly patient engagement and using PROs as a step. We can discuss that more. I think the important thing is that we wanted to bring out patient engagement and shared decision-making. And there could be certainly additional wordsmithing to the actual statement but where we stand on these other issues is part of our dialogue today.

And then last but not least we had the criteria which I have difficulty pronouncing but implementableness or is it -- you know, when you're typing this stuff it's like how do you -- is it -ability, is it -ness. But anyhow, implementable. And so -- thank you. That was a tongue-twister for me.

And you know, in this characteristic I think we were trying to capture a lot of things. We talked about IP issues. We talked about many things. And so
several respondents did say you're covering a lot here and it may be hard to map to all these requirements. And what measure, not only a PROM, could possibly meet all these requirements.

But the important message is that implementation issues and as we learned from our prior panel are very critical and they should inform our decision-making. And as we look on our action pathway when we go from PROM to PRO we have number 5 which talks about how do we implement this in clinical practice.

However, maybe this criteria we might want to look at how we can streamline this. Some felt it was already covered under the other topical areas of actionability and meaningfulness. And then we kept wanting to insert disparities because that was very important to this group but some felt that disparities were not indicators -- disparity-sensitive measures were not indicators of implementability.
And then importantly it was raised that we need to think about ease of fielding. I think a lot of this gets into missing data elements and things like that, but ease of fielding this and testing it could have impact on implementability.

So this is a quick synopsis of your feedback. And then if we can go to the next slide. Okay, great. Just wanted to make sure we were at the last.

So, with that discussion it was just a bit of a primer. We wanted to feed back to you what we heard from the survey, what some of the key themes are. We'll continue to refine those statements over time. We're going to discuss in this order. Liz is going to speak with us about actionability, Patti about meaningfulness and then Laurie certainly last but not least but certainly around the implementableness. And certainly you have experienced that. But Karen's just going to give us a brief overview of the
intersections here with the NQF endorsement criteria. Thanks, Karen.

DR. PACE: Good morning. So, what I'll be doing at each session is just giving a brief overview of some of the criteria. Next slide, please.

So we'll be always -- next one. There we go. So again keep in mind that NQF does endorse the performance measure which we're referring to as PRO-PM versus the instrument. However, our criteria, there's a lot of overlap of how these things apply to the data that are going to go into the performance measure. So I'm going to just highlight some of the criteria that although we're talking about performance measures also have an intersection with the PROM, the instrument or scale. Next slide.

So, I'm just going to quickly mention psychometric properties. We're going to have three panels talking about that. But basically our criterion about scientific
acceptability of measure properties specifically focuses on reliability and validity.

And we do allow for testing at the data electronic or the performance measure score. And we'll get into some of those distinctions in the specific panels. But just wanted to mention that we do have reliability and validity of the data which would be from the PROM instrument or scale. And then we're very interested in the reliability and validity of that actual performance measure or the score that a hospital or a healthcare facility would receive. Next slide.

So in terms of actionability we would see this falling under our major criterion of importance to measure and report. And there's a couple of related sub-criteria. One is a performance gap or opportunity for improvement. And generally when we endorse a performance measure we like to endorse something where, you know, everyone's not
doing well because we want to put our
resources for data collection, reporting,
analysis on those areas that are really going
to push us forward in improving healthcare and
health.

But the key one here is our
criterion about evidence. We do under
importance to measure and report have a
criterion about evidence. We say under this
criterion that the measure focus is a health
outcome or it's evidence-based meaning that
there's evidence that links the structure,
process or intermediate outcome to a desired
health outcome. And when we're looking at
this evidence we're looking at the quantity,
quality and consistency of the body of
evidence.

And I want to emphasize that our
criteria do not require that this be
randomized controlled trials. Depending on
the focus of measurement and healthcare in
general there's lots of different kinds of
evidence but we are talking about empirical
evidence, not just expert opinion. Next
slide.

The reason we have "health outcome
or" and Helen mentioned this at our last
workshop, that we really have a preference for
health outcomes. And we think that health
outcomes by their nature have a special place
in terms of performance measurement because
that's the reason for providing healthcare
service, it's the reason for seeking
healthcare service. And for these -- and also
there's multiple processes that influence any
particular health outcome.

So we really for health outcomes
ask for a rationale that supports the
relationship of that health outcome to
processes or structures of care. So some of
the PROMs that we're talking about would be
considered health outcomes. Some might be
considered intermediate clinical outcomes.
And I think that's an area where we'll have
some discussions.

Certainly for intermediate clinical outcome and an example of this in a clinical sense would be something like blood pressure or a particular lab value or a process or structure is that we really want to see a systematic assessment and grading of the quantity, quality and consistency of the body of evidence, that that particular aspect of healthcare that's being measured actually is linked to desirable health outcomes.

And then certainly experience with care, what our guidance states is that the evidence -- there should be evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information, or there may be evidence that experience with care is correlated with desired outcomes.

So, all of these things that are already in our criteria we think relate to the idea of actionability but we probably will
need some more discussion about where we have
our emphasis on health outcomes. Next slide.

So, meaningfulness. We think this
again relates to our NQF criterion of
importance to measure and report, certainly
evidence that the measure is a health outcome
or is evidence-based and what I just mentioned
about experience with care relate to
meaningfulness. We also have a criterion, a
sub-criteria of high impact, that it's related
to a national health goal or priority. And
we've heard that patient engagement, patient
experience are certainly important aspects of
our national health goals and priorities.

And certainly we look at the data
on numbers of persons affected, high-resource
use, severity of illnesses or consequence of
poor quality. And then our criterion about
usability and use also relates to
meaningfulness. Next slide.

And this is a criterion. In your
handout it's an addendum. We're in the
process of revising this particular criterion so this is the new information here. But this really focuses on accountability and transparency and improvement.

And we've already mentioned that NQF-endorsed measures are intended for use in an accountability application. And we really want to see that they end up being in use or that there's a credible plan to get them in use. But certainly all of the measures should be available for improvement. And again, we want to look at how these measures are really helping with progress in achieving high-quality and efficient healthcare. Next slide.

And implementable really relates to our criterion about feasibility. And next slide.

And the elements that we look at under feasibility are that the data -- if it's clinical data generated and used during care process we think that this certainly makes it more feasible. And I think this again is an
area that we'll talk about. What we're
talking about here is if it's something that's
done outside of the care process it becomes
more burdensome.

And I think that's part of our
discussions about PROMs actually being used in
clinical practice. If they're relevant for
clinical practice then the data will be there.
Certainly electronic sources make things more
feasible. And we want to see information that
the data collection strategy can be
implemented.

Okay, I think that's the end of my
slides. And we'll turn it over to Liz.

DR. MORT: Thank you, Karen. Can
everybody hear me okay? No. The mike is --
can you hear me now? Okay.

So, I've been asked to speak a
little bit about accountability. I think some
of the comments that Karen and Karen made were
very clear and I don't want to be redundant
but I do want to put some focus on this from
the provider's perspective.

I think the actionability issue is critical and that actionability defined as responsiveness to healthcare interventions, if we don't have a healthcare intervention that's going to advance or improve a patient-reported outcome that measure really has no business being in an accountability framework. That measure may have a very important role in a measure to help stratify patients or understand what basis you're dealing with from a patient's perspective, and that measure may have a very important role in research in quality improvement but I think it's just not a good thing to be taking measures, as compelling as they may be, as interesting as they may be, as important to patients as they may be, I don't think unless there is an intervention that those should be in an accountability framework.

And then the reason I say that is that we have enough to do in the course of
taking care of patients where there are interventions that are known that we should be spending our time working on the things that we know we can improve. And I know myself and my clinical colleagues feel bothered when they're asked to do things that they feel detract from the clinical value that you can bring to a conversation. Again, not that many, many of these things that we want to do aren't important in the research environment, but until you have a defined and either proven through RCT or other forms of evidence they really don't belong in the accountability framework.

So, what about those actionable measures? Now, there are actionable measures and there are very actionable measures and there are barely actionable measures. And that's one of the reasons why I think the ones that are clearly not should just not be on the table. And I think it's important when we look at these measures and we evaluate them
for endorsement that we consider the spectrum of actionability because it's very, very important.

So let me start. And this really comes ripped from the headlines. We're trying to implement these. We are implementing PROMs at Partners Healthcare. I gave those of you who were here last time a brief overview.

Well, it's marching ahead apace and many, many of our clinicians are extremely enthusiastic and anxious to get onboard this data collection effort. And not surprisingly the clinicians who are interested in participating are those that have measures that are highly actionable.

So let me give you an example. In the urologist world we have lower urinary tract symptoms. So frequency, urgency and the like. I won't get into too much of the clinical details.

But you can scale those, we've been scaling those for decades, and you can
intervene in many, many different ways
including everything from watchful waiting if
that's what the patient's preference is to
lifestyle changes, to medications, to
minimally invasive surgery and to surgical
surgery. And guess what? The symptoms
change. So that is I would say a PROM measure
that is a very highly actionable measure.

And that the other key point I
would say is that what we find in RCTs in
efficacy studies translates reasonably well in
real life or in effectiveness. And so that's
why I think this particular example hits on a
bunch of different issues that are important
when you tease apart actionability. So it has
to be highly actionable and it has to be
demonstrated outside of an RCT that you can
actually do this in practice. So I think that
kind of -- a clinical scenario with those kind
of measures and those kind of interventions is
well suited.

Another one that I'm getting lots
of interest in is from the orthopedic surgeons who are very, very interested in shoulder repair, either rotator cuff or shoulder replacement. And there are good measures, and there are strong interventions, and there's demonstrable improvement. Again, it's important to stratify and so on and so forth but it's highly actionable and there are strong interventions.

So let's move to the next category that I would say is more moderate, depression. So we can measure depression and we can act on depression. You can't treat everybody. You can try and there's lots of different things and there's referrals so I'm saying it's a very important measure to have as an accountability measure but we have to take that into consideration when we set it up.

And how it's used as an accountability measure because the last thing you want to do is put out a measure that incents providers to treat and treat and
treat, to add medications, to try referrals
that may end up for some patients not working
and potentially harm them. So that's what I
would say is kind of a middle ground.

And then I'll give you an example
of what I would call a weakly actionable,
probably weak to not actionable would be
dementia, Alzheimer's dementia. There are two
classes of medications that are used now in
practice to help slow the rate of decline in
patients who have Alzheimer's-type dementia.
But it's very hard to really demonstrate
improvement in clinical trials let alone in
practice. So I would not as a physician at
this point in time want to be held accountable
despite the fact that there are drugs out
there that have been shown to have impact on
some patients. It's a weak intervention,
therefore a weakly actionable, weak to not
actionable measure if you're looking at some
kind of functional status for patients with
dementia.
So I think that gives you the range of types of actionability that must be considered, that should be considered when you're trying to set something through a process to decide whether it's a good measure for accountability.

But I hope I've made the point that if there's no evidence for actionability that it really should be off the table. There's too much disease burden on the table that we can treat. That's what we should focus on first. And then all the cautionary notes about strength of actionability, differences between efficacy and effectiveness.

And the last point I'll leave with because I'm looking right at Jack Fowler and I can't help myself is that and another really important consideration is patients, how they view all these symptoms. So you might have a patient with a symptom that's highly actionable, but if their tolerance or their
preference or their bother with it -- thanks, Jack. Jack's one of my very favorite mentors. If they're not bothered by the symptom and it's perfectly legitimate for them to have the quality of life they want then trying to move that symptom would be harming the patient. So I think you have to take that into consideration as well.

DR. BRENNAN: Hi, I'm Patti Brennan. I'm among other things the director of the National Program Office of Project HealthDesign which has been a multi-years project trying to better understand health in everyday living. So much of the work I'm going to be talking about today and the basis for my comments will come from my experience with Project HealthDesign.

The concept of meaningfulness is well laid out in our early discussions here. And the idea that a patient-reported outcome measure or a person-reported outcome measure should be meaningful largely is linked to the
idea that it's meaningful to the individual
patient, meaningful to the individual or the
persons involved in that individual's care.

So I am going to be focusing significantly on
how meaningfulness intersects with the patient
and the person's experience of their health
services rather than meaningful in the course
of a treatment plan which I recognize is also
important.

Then we've been asked to talk
about two issues, how do we engage patients in
the selection of PROMs and then how do we
document that engagement.

So I wanted to begin discussing
meaningfulness by using three C's, three
words. First of all, meaningfulness on the
conceptual level, secondly, meaningfulness on
the contextual level, and third,
meaningfulness on the consequential level.

On the conceptual level we're
focusing on to what the PROM actually
measures. What is the patient-reported
outcome or the concept underlying this. And often those PROs, the outcome -- sorry, the concept is linked to some kind of clinical or professional definition of what is health and what is healthcare. As a starting point we need to relax that a little bit and we need to listen to the way patients identify their health, their health experience, what they understand.

And while we divide the world up into clinical specializations or locations of care, patients live in one body and they divide the world up inside of the inside and outside of their body, period. So as we're talking about the conceptual basis of a PROM listening and engaging the patients in a dialogue and engaging people in a dialogue around how the language shapes and determines and provides meaning to the concepts that will be measured is a critically important first step.

In our case we use the phrase
"patient-defined and patient-generated" to describe two rather separate areas. Patient-generated are the sensations, the experiences, the ideas that an individual and only that individual can report. But they may be reporting them in response to things that we say, how is your pain, how much can you bend your arm. Those are things that have clinical meaning in our professional practice.

Patient-defined is the individual's experience of that stimulus that leads them to be able to report or generate a response about it in response to their health status. And that may include, for example, can I lift up my child. Can I walk to work holding hands with my son. The idea of not can I flex my arm this much but can I do what I want to do in my life.

Other aspects of patient-defined may become much more abstract. The tenor of a conversation at dinner being tense or not tense is something that's really quite
difficult to map to a specific professionally validated indicator but it may for that individual patient be the concept that tells them that their medication for depression is improving or not improving.

So as we begin to bring patients into the concepts and setting meaningfulness of our patient-reported outcomes we need to recognize that we first have to listen to them. We don't stop at that point. There may be very good reasons to have professionally selected, professionally defined and patient-generated PROs and patient-reported outcome measures, but we can't begin the process without first listening and finding the cross-mapping between what is meaningful in the patient's language and what is meaningful to us about that individual.

The scope of care defines this very much. And I heard discussions in our earlier comments today of the tension between a clinically targeted measure and a general
measure. And in fact I guess I'm asking to think about a third way of defining the measures, neither clinically specialty-focused or general, but actually patient-focused in terms and the experience that individuals have.

Their experience of care is not divided up based on the offices in a hallway. Their experience of care is the experience of the individual. So we may need to think about indicators that transcend different points of care as we find which concepts are most meaningful to the individual.

Second, I want to talk about contextual. Where do PRO and PROMs become meaningful to the individual? They often become meaningful we heard earlier today in being able to learn of a provider's performance and therefore begin to select providers. This provider is responsive to or has interventions for something that is important to them.
But people also find it meaningful to participate in the larger social exercise of health. And there is an altruism experienced by individuals when they believe their input is sought and respected in determining the quality of care for themselves and those around them.

So to make the experience of patient-reported outcome measures and selecting PROMs for individuals meaningful to patients may include thinking about how it allows them to participate in the larger social discourse about care.

Third and importantly, we heard in our previous slides about the usefulness of PROMs for an individual's own care. And I want you to think about that from two perspectives, one in the moment when we're capturing that information, how meaningful is it for a person to know this about him- or herself at that moment, and second in the clinical encounter. How meaningful is it for
that person to share that observation or share
that PROM with their clinician to make a
determination about how their own care is
progressing.

I've heard a lot of discussion
about the second kind of care, that it would
be nice to have measures, PROMs that can be
both useful to tell us about how a system is
functioning and how this particular patient is
improving or not improving. But I want you to
also think about the meaningfulness of the
individual who's capturing that, who may not
be sitting in the clinic waiting for a visit
or filling an after-survey, but may be 2, 6,
8, 12 months after the encounter and being
asked to recall something -- provide something
about their daily life to help us interpret
how good the care that they received was. At
that moment it would be also useful for the
individual to use that marker as a self-
assessment, as a way to meaningfully
understand how they have changed in their
Now I want to move onto consequential. The meaningfulness of PROMs can be established by looking at the consequence of knowing about them. And although we don't think of this as a patient-valued or patient -- an individual patient experience, by ensuring that there's good quality providers around we are providing something for the patient. So the measurement or the use of PROMs is in fact useful on a consequential level.

So it is incumbent upon us as we begin to observe, as we begin to collect patient-reported outcome measures that they feed back into assuring that the practice is of good quality, that the clinicians are fairly compensated, that the practice can financially sustain itself or the institution can sustain itself.

Assessment of patient-reported outcomes through PROMs can in fact have
important consequences on the availability of
health services for individuals, the type and
the responsiveness to the individual's needs.

I want to close my remarks by
making two more points. The first is that our
work today is really not about PROMs. It's
about PROMs -- patient-reported outcome
performance measures. We're here to discuss --
recommend to the National Quality Forum how
to address the performance measurement.

And we don't specifically think
about how patients will be engaged at that
level, what thresholds should be set, or how
much of a gain or loss is tolerable. And yet
it's important that we consider not only
having patients involved in the selection of
the measures but also in the way that they are
used and interpreted and reported.

So moving forward I have four
points for a national agenda related to
meaningfulness and engaging patients in
selecting PROMs.
The first is that we need to have
a national agenda. We need to have the
kindergarten curriculum of patient outcomes so
that it becomes a national everyday experience
for individuals to know that their reaction to
the care services they received is not only
important but meaningful in terms of shaping
their own care as well as shaping the
availability of care in society.

The reports we heard from England
and Stockholm today didn't occur because there
was a 2-year or a 5-year national program.
They occurred because of the context that took
many years to develop. And it's timely for us
to think about how we develop this.

Secondly, the patient voice
matters. The patient voice matters. That
means not only the individual who speaks for
him- or herself at various points in time, but
the ability for an individual to set policies
that are subsequently respected over time, the
inclusion of surrogates where appropriate and
properly denoted, and the inclusion of caregivers who may not speak for the patient but speak about the patient in a PROM-type situation. It may be as useful to know that there's a difference between what a spouse and the care recipient perceive about the quality of care that an individual received.

So recognizing that the patient voice comes in three different forms as I see it, the person him- or herself, the surrogates and the caregivers, and sometimes it's important to keep those separate as opposed to aggregating them together.

Third, I've heard over the process of these two workshops some indication that we will have a national list that institutions can select from -- of PROMs that institutions can select to measure concepts, some indication that there might be a set of explicit measures.

We need to consider at the point - - that the involvement of patients in
selecting PROMs will vary based on whether we're talking about setting a national agenda or we're talking about a clinic in a small town who's trying to pick certain outcome measures. There are different ways to engage patients over those times to do each of those. And we can't simply call patients up and assign them to a task force immediately. We need to cultivate and build the skill sets for that.

And finally, I'm going to call for some perhaps non-traditional ways of understanding the patient's voice and the patient's experience. I'm going to encourage groups that are trying to bring the concept of patients meaningfully into the selection of PROMs to not only think about sitting patients down at a table but also to look to see how the creative literatures can help us understand the patient experiences and help us to target what might be useful or meaningful about an adolescent who's facing a pregnancy,
or an older person living alone. Think about ways to engage not just the person of patient but the concept of patient as we select participation.

We also need to find increasing ways to use social media to get reactions from individuals over time. Not only in the selection of the PROMs, perhaps even in their assessment, but at this point in time I think using social media, using Twitter, using Facebook as a way of making it more widely known that there is an interest in selecting PROMs and an interest in hearing the patient's voice about them might give a way to get a very broad and very diverse set of viewpoints on the selection process.

I thank you very much for your time and for listening to me instead of Jennifer. I hope that I represented this perspective well in the conversation.

(Applause)

MS. BURKE: I made one slide.
Everything I have to say is on one slide. I have five points and I just thought it might be helpful if I put those up in front because I might not be able to read what I scribbled on my paper.

But my -- I'm assigned the topic of implementability but I think that the implementability depends on this topic of the key characteristics that we started out with here and for selecting PRO measures.

First of all, my first point is that we -- 2001 is the year that's the 11-year anniversary of the birth of the term "patient-reported outcome." And I can keep track of that because of the monumental nature of that year 2001.

And it was generated because there was an initiative called the Health-Related Quality of Life Harmonization group that were deliberating on how to define health-related quality of life. And we finally accomplished that but then we realized that everybody was
using health-related quality of life to mean everything else that health-related quality of life does not mean. And so we needed a larger grouping term for everything patient-reported and that's when that term was developed.

And it's really important that we maintain the meaning of that term just for communication purposes. And I really appreciate NQF and their efforts at doing that for this meeting.

And so there's good measurement science that's under development right now and that we're talking about profusely in this last workshop and today. It does not apply only to patient-reported outcomes, however. It applies to measures, all measurement used in healthcare. And I wanted to make sure we keep that point in mind because of course there could be other things besides patient reports of things that may be useful to use as performance measures.

My second point is that the key
characteristics that we're talking about are not characteristics of the PRO measure. These characteristics apply to the use of the PRO measure in a particular context. And this is really important to know about because when you select a PRO measure you have to think about these characteristics in the context that you're going to then apply it to.

I think that we very often lose sight of that. And the red boxes at the top of the diagram were added for this meeting and I really appreciate that. And so if we continue to keep that in mind. This is a very key point for the implementability of these measures in terms of future use.

My third point is that contrary to classic psychometrics teaching it is not efficient to test reliability first. Now, I'm sure I'm going to get some pushback on this topic but that our experience in the regulatory setting where we're talking about good measurement has really borne this out.
And for example, you have to first decide what you're going to measure because when you're going to measure something it's not just grouping a bunch of things together to create a score. You need to somehow characterize a meaningful state in a context of use that you're planning to measure it in, in a disease group, in a patient group.

It's -- the score has to represent a meaningful concept. I just was in a meeting last week about an instrument to measure suicide ideation, for example. And this is a checklist where we have five graduating more serious ideation concepts that the patients check yes or no.

And it's an excellent measure. But when I said, "Well, what does this score represent?" they said, "Oh, we don't have a score." But yet when you collapse the data you try to assess whether or not we have suicide ideation you do actually look at everybody who checked for the fourth or the
fifth one and that's your score of suicide ideation. So, we have to think in terms of scores and what we're trying to make the instrument actually represent. In this case it is suicide ideation.

The items and responses need to be available, need to be presented in a hierarchical order so that we know that it makes sense to a patient as they are assessing and trying to match what they're reading on the page and marking as a response to this question.

There needs to be adequate coverage in the target population. All of these things represent what we would call validity of the item.

Now, if you wait to assess all of that until after reliability is tested you may find out that the PRO measure is not measuring the thing that you have been targeting. And the content may need to change. And if that's the case then repeated reliability testing
needs to take place. And if that -- and if the content is not changed you will have some compromised sensitivity to change with the measure.

And this has been borne out over and over in our experience in reviewing clinical trial data. I do not think these principles for good measurement are any different outside of the clinical trial and in the healthcare setting.

So then my fourth point is that validity testing is iterative as the PRO measure content develops. So that if you spend adequate attention to the validity of the measure to measure the concept that you have intended in the context of use that you have intended then you can very much influence the reliability of this measure and therefore its ability to change in the clinical setting.

For example, there are four major types of variability as far as I can identify. There is true patient heterogeneity. So that
is very much related to your context of use. You need to identify all the ways that your patient population really does validly vary and you want to be able to consider that for the PRO measure. So, that can be considered in the early stages in this iterative approach to generating validity of your measure.

Then there's random variation which is often called error which is also something that can be impacted by careful consideration to the content of the instrument and minimized during that early phases of instrument development.

And then there's systematic non-random variation, also something that can be considered based on your population. What are the characteristics of your population that would cause -- what are the items in the instrument that may be obscure or not understandable to certain parts of your population that are going to introduce this non-random variation.
Then there's the experimental design and the conduct error which is a huge concern when you start administering your instrument in large populations. This is a big reason for a lot of the missing data concerns that we've heard about earlier today and it is the reason why we are focusing more and more on the training and instruction component that goes along with an instrument so that it is adequately administered in the way it's intended so that you can in fact collect the results that you're looking for.

We find that training and instructions are always an afterthought and in the past we rarely even asked to see it or review it. But it's really an important aspect of being able to understand what you're collecting and how you can actually implement it. And if it's being administered in a way that's not intended then you have to think about that validity. So, this is just following on the point that validity really
needs to be thought about first, not last, and some of the aspects of that.

Then finally, because of all these things that I've already mentioned and lots of other examples we -- I think it's really important that we quit, in terms of the terminology that we use going forward, that we stop calling -- we stop talking about a measure as being validated. And this is people who are in the inner circle of measurement like all of you are need to be really careful about this because people who don't understand validation are going to take and run with it and say well, they said it's validated so we can use it and plop it into whatever context of use is currently under consideration. And so I think that this would be really useful as we're focusing on terminology and how we think about measurement that we stop using that phrase.

Okay, those are my five points.

DR. ADAMS: Well, first I'd like
to thank all our reactors for really providing us some thought-provoking comments. And I'm going to open up to the room for some discussion questions.

First, I want to tee up the operator because we do want to be able to bring in our people who are participating with us virtually so that they can start queuing up while we're doing a few questions. And I want to make sure that our external audience that's here with us live today, we do have microphones and stands for you too so we welcome your participation.

But I'm going to start the moderating now. And I feel I need to stand up because I can't see in the back.

OPERATOR: At this time in order to ask a question press * then the number 1 on your telephone keypad.

DR. ADAMS: Great. Yes, please,

Ted.

DR. GANIATS: Ted Ganiats in San
Diego. Not in San Diego but from San Diego.

I wanted to thank folks. The morning just, you know, I've always been excited about patient-reported outcomes and the morning's presentation, the beginning just increased my already high enthusiasm. And then these three talks were able to highlight one of my big concerns for performance measurement but that means I can list all three people.

You know, Laurie talked about scores, so important. Not all PROs have a score. Could just do a count of number of joints, for example. But we oftentimes try to collapse them into a score.

And I think that's left -- that means Patti whose presentation was just so great forgot one thing in that she talked about it being patient-defined and patient-generated but I think it should be patient-scored.

And I think we have a very, very
hard time if we're going to try to be actionable if we take a population preference scored or just an average of I'm going to count up the number of yeses and give us a score and have that be actionable because it may not mean anything to the patient. So that's my opinion and I'm wondering if the panelists would react to the idea of how important it would be for an outcome performance measure instead of a process performance measure to be patient-scored.

In my opinion if it's not patient-scored it should be a process measure. If it is patient-scored it might be able to be an outcome measure.

DR. MORT: I'd like to make a comment. I think it's a really important point. I think it's -- there are ways to address it maybe that get at what you're talking about without necessarily having a patient score everything, every aspect of it. But understanding how bothered the patient is
or how satisfied they are with the outcome
might be one way to do it.

I have a story of a patient with a
shoulder problem and the shoulder PROM didn't
really improve but he was happy as a clam
because he could do what he needed to do
because he figured out a way to play curling
was his thing, ice sports, and he figured out
how to use a broom in a different way that
didn't -- so he was pleased with what had
happened but you wouldn't have been given a
good rating because the shoulder functional
measure hadn't changed.

So I think you have to be a clever
in the way you assess patient's preferences,
how much they're bothered by it. And that
also gets complicated because they may not
have known enough to realize that they could
have actually improved more and so on and so
forth. But it's a very important point
because the clinical or the provider-based
assessment may not tell you the whole picture.
MS. BURKE: Well, I can't agree more with the importance of the score because the score is how you're going to make your conclusion. If your score doesn't -- first of all, if it doesn't represent the thing you think it represents that's a big problem. And that's part of the validity of an instrument is making sure at the end of the day that the people who are filling out the -- are actually providing the input that generates the score will agree at the end of the day that the compilation of what they have reported represents the thing that the score represents in a way that they intended.

And this is a very complicated issue, a lot of weighting of items and making sure you have the right things that represent the score. And change in the score represents change in the thing that they're telling you about. And this is critically important, yes.

DR. GANIATS: For example, the EQ-5D which I'm going to understand every single
question that's there and I'm going to respond appropriately for every single question that's there but the score is irrelevant to me because the score is based on population preferences which may or may not reflect mine.

So, that's an example of my seeing an EQ-5D score or worse if a clinician tries to improve a patient's EQ-5D score you may end up going in the wrong direction because those scores just like the shoulder, it may not move the patient in a way that the patient is getting better, it's just that the score is getting better.

DR. ADAMS: Ethan, you have a response to this?

DR. BASCH: Yes, I was just going to comment to this. I mean, I think this is one of the reasons that many have really focused on specific symptoms or very specific dimensions of functionality. Certainly regulatory agencies have focused on this recently and that's because we want to know
what we're measuring. And we also want to make sure that a change in that score is meaningful to a patient.

And there are techniques for demonstrating that a particular score change is clinically or I should say meaningful to patients. And it's much harder for example with the EQ-5D to demonstrate that, a composite EQ-5D score of those five different dimensions is meaningful to an individual patient. I couldn't agree with you more. I'm sure Laurie would follow up.

DR. ADAMS: Yes, and I know, Patti, you wanted to respond as well. Is that a direct response, Laurie? Go ahead.

MS. BURKE: Yes, and I think that it's really important to think about how we present the results of whatever we measure. For example, if you are in a clinical care situation you're measuring an individual patient and that individual patient's change with respect to whatever the thing is you're
measuring, that is -- it's critical that your instrument is adequately assessing that thing on an individual basis.

Now, when we take that instrument and move it into a clinical trial or in an observational setting where we're looking at populations of people then we're looking at change in terms of an average or a proportion of responders or whatever that metric becomes. We -- that same degree of change is not -- does not carry the same meaning in that mean score, for example. So, this is why we're looking at other ways of displaying data so that you can understand the distribution of response across a population, for example. And that has to be -- that's probably another whole conference on how to display and present the results that are being measured in a population of people because it's very difficult to understand the range of response and how a patient can look at that population-based data and interpret it in terms of the
meaning that it brings to them personally.

DR. BRENNAN: That's a nice segue way into the comment I wanted to make which is a call for greater research into the methodologies of aggregation.

We make a lot of presumptions that everything is monotonically related and increasing in an equal fashion and then we add things up. Most of the time they don't.

And the question about the extent to which a score maps to my personal experience versus the population as a whole remains somewhat ignored by the methodologists in the field and it may not be able to be over time. Because if we want an individual to interpret a gain from 7 to 9 is the same degree of change as a gain from 14 to 16. We have to be pretty sure for that individual that experience is in fact trustable and has that kind of an underlying process.

I think the other part is while we need to look at summative measures they might
not all be scores. They might be
classifications, they might be displays, we
might be looking at Euclidian distances. So
think flexibly when you think about scores.
Thank you.

DR. ADAMS: Ted, did you have?
And I see we have someone in the back. Can
you come up to the microphone and Evan, if you
can have our audience member teed up after
Ted. Thank you, Ted.

MR. ROONEY: Ted Rooney from
Maine. This is a terrific panel and terrific
day. So I'm thinking as the discussion goes
as part of my informed decision-making if I
had -- I'm a male so and I'm going to have a
prostate problem someday. And I remember
seeing the videotapes of the informed shared
decision-making, you know, the differences
between the two docs explaining it.
So I'm going to come down this
road. I'm going to want to know although
there's different approaches who gets the best
outcomes and the quality of life and so I want to see that there.

And then I want as part of my informed decision-making I want to know if I have three or four urologists over here who gets the best scores on their treatment of patients, so I want to know that.

And then when I go in and I potentially have my procedure, afterwards I want to know how my score relates to what potentially could have been. Is it worth it for me to go through whatever else I might have to go through to get a better score. So this is an incredibly rich discussion and very complex.

But you know, so I want to make sure that whatever I choose fits me but I want to have enough standardized information that I could choose among different treatment options and among the people who deliver those treatment options because I'm sure there's a difference between the people who do medical
management and the people who do surgical
management. So, this is a great discussion.

DR. ADAMS: So those may be things
to benchmark against. Go ahead, Patti.

DR. BRENNAN: I want to comment on
that. The challenge is in part the
mathematical literacy of a population and
trying to help people understand that the guy
with the best score still might have had a
turkey day one day. And so they're not
necessarily a guaranteed performance but
rather a typical, not even likely-to-happen-to-you performance.

DR. ADAMS: And Liz?

MR. ROONEY: And I think that's
part of what I would want to know so that if
I make my decision that that's -- I play the
odds.

DR. MORT: I love your summary of
the complexity and it makes me think do we
advise NQF to go slowly and go deliberately
rather than just throw a whole bunch of
measures through.

We've learned the hard way through process measures that throwing a whole bunch through may end up doing some harm that we would have avoided had we don't it a little differently.

So it's very complicated to do the patient-reported work. So many issues, it just makes me think boy, let's do a few really well so we understand before we open the floodgates.

DR. ADAMS: Sure, Ted, and then we're going to go to our audience.

MR. ROONEY: So in Maine we're playing around with a lot of these things. And I agree, for the ones that have strict accountability I absolutely agree with you. But how about having some others in the field that people could work with that get us directionally towards where we want to go that may not be a clinician accountability but may be a group or an organization accountability.
So, I think there's a lot of room in there.

DR. ADAMS: Okay, we're going to go the back. Thank you.

DR. KELLER: San Keller from the American Institutes for Research. And this is kind of related to what the last speaker said, that I think we're talking about the context of use. There's always been a tension between meaningfulness in terms of the semantic meaningfulness and precision of measurement and the need to make decisions. So, a single symptom is meaningful to the patient but may not be precise in a statistical sense. And that's why we create composites for the signal-to-noise ratio, but also to allow us to make decisions. So what we've done in the past is take a composite score and then show the relationship between that score and a single symptom or a single ability to do this or that so that you could have the best of both worlds. I'm not sure that that's possible but it's a longstanding tension.
DR. ADAMS: Thank you. Any comment or reply from the reactor panel? And I think we'll get into it a little bit too as we get into our next two panels. Go ahead, Laurie.

MS. BURKE: San, are you talking about a composite of many things that you put together in a score or are you talking about a general concept of measurement that the score represents? Is that the composite that you're talking about or is it -- I'm thinking of composite as multiple scores that are then used in a way to get sort of a profile.

DR. KELLER: I think there are two different things. If you want a composite to make a decision, so you want to accumulate across different kinds of outcomes and you know, have them come up with a single answer even though they're very different that's -- I think that's what people usually call a composite.

But if you want to increase your
signal-to-noise ratio then you want things
that are very similar that just, you know,
enable you to do that. So, thanks for
pointing that out because I was kind of
conflating both of those.

MS. BURKE: Sure. And I just, my
response would be that it doesn't really
matter. If you have a score and your score
represents something that's -- we should be
able to name what that thing is and then the
measurement properties in terms of key
characteristics should relate no matter if
we're talking about a general, a score for a
general concept or something very specific
like a single pain intensity measure. The key
characteristics still apply.

DR. ADAMS: We have another
question. Go ahead.

MS. OKUN: Thank you. I'm Sally
Okun and I'm from PatientsLikeMe.

I really appreciated this panel.

I think there's been a lot of really important
points that have been brought out. I'm wondering whether the -- it will be important for us to consider things like we consider now clinically important differences.

Maybe we need to be really thinking about is there a measure for patient-important differences and really understanding with patients targets that they might want to set that they can reach and that the individual clinician can help them reach in terms of how they're measuring how they're doing, but then also being able to that into consideration when you aggregate that data. Because I think as we start talking about patient-reported outcome information it's going to vary so much by all the different circumstances that patients putting into what they want as their target versus what the target composite score might indicate for performance.

So the incentives that a patient has to feel better, the incentives that the
clinician has to help them feel better and then also to get paid particularly I think are aligned quite differently. And so understanding what the minimal clinical important difference is and then what the minimal patient-important difference would look like I think will be something that we can begin to talk more about.

DR. ADAMS: Okay, thank you so much for that insight. Al, you had a comment? And then I'm going to -- we queued up people on the phone. I don't want to forget them. So after, for those on the phone you're up next. Go ahead, Al.

DR. WU: On the last point, the idea of what is a -- or how much is a patient-important difference is a very good question but it's not so easy to measure because it's one person. And so there is going to be a lot more noise around that measurement. And it's just a difficult question to confront. I think it deserves to be looked at.
DR. ADAMS: Okay, I'm going to check with the operator. Do we have anyone on the line that has a question?

OPERATOR: At this time in order to ask a question press * then the number 1 on your telephone keypad. Please hold for the first question.

DR. ADAMS: Thank you. So we're ready to take the question from the person on the line.

OPERATOR: Your first question comes from Susan Tavernier.

DR. TAVERNIER: Hi. I was wondering if the panel would address the issue of response shift and the fact that that can be defined in a number of ways, whether it's in the scoring of it versus conceptual response shift and how the panel feels that relates to these patient-reported outcomes.

DR. BRENNAN: I think that's a great question if I'm understanding response shift the same way that the caller means it to
be. And I was thinking of this actually a few minutes ago as Sally was speaking.

When we talk about the variability in measures we need to consider the same measure over time with the same person, populations at the same time, variability within that, and then populations over time. So we have all these different dependencies.

And there is a -- I would imagine that there's both a testing effect and as well as historical effects that need to be considered as we get socialized to be more mindful or less mindful of certain things, whether it's prescription drug in the news because of our sports players or the expectation that you should live to be 90 and still play tennis. There's a shifting in what we expect as a baseline.

And measures need at one and the same time to be both sensitive and resilient. And so I -- all I can simply say is yes, I think that's important. No, I don't know the
methods to handle it.

MS. BURKE: Well, we're fortunate in a clinical trial which is most of the data that I review that we -- well, at least we hope we're fortunate in that the randomization between groups takes care of the response shift that no doubt happens in both groups. So, we don't have to worry about it too much in that setting. But in an observational setting that is clearly a concern and that has to be taken into consideration somehow.

DR. ADAMS: Thank you. Any other comments? Oh, Lewis, yes.

DR. KAZIS: I had a fellow a few years ago who we published an article looking at response shift. And what we found is that if you ask a subject how much their health has changed over the past year, over the past 12 months they're basically going to be telling you about their current health and that's correlated much higher with the current health than an actual change score that you've
derived. So that's been published and I think there have been other articles as well.

DR. ADAMS: Thank you. Ethan, you have a response?

DR. BASCH: Yes. I mean response shift happens but it's a small effect overall. We know from many, many, many, many studies done over time that we are able to detect change over time, that people's scores do change over time both in response to interventions and as a part of disease trajectory. So I think response shift is real.

That said, you know, the answer that many people have to this is that what the patient says about how they're feeling is how the patient is feeling. And so if the patient changes the context in which they think about a particular experience because their general outlook has altered then that is actually an accurate subjective portrayal of how they're actually feeling at that point in time, right?
So you know, perhaps -- I'm an oncologist, right, so you know perhaps the patient's view of nausea or fatigue changes after they've had multiple rounds of chemotherapy, but how they feel about those symptoms is actually -- at a later time point is actually an accurate portrayal of how they actually feel.

And so when we compare between groups if you want to -- I don't want to belabor it, but if you want to say to -- if you want to portray to a patient before starting chemotherapy how patients like them will experience chemotherapy-related symptoms later on it is actually a more accurate conveyance, some feel, to explain to them how people at the later time frame who have had a response shift experience those symptoms because that's where a patient might actually be.

DR. ADAMS: Well, we're one minute away from lunch so I'm just going to -- I
think we have one more question. We're going
to break for lunch at 12:45 and resume back at
1:30. But we have time for one more question.
I don't see any hands up. Oh, great.
Phyllis. Give us our great closing remark,
yes?

MS. TORDA: I thought that Patti
made some really important points about the
importance of the process of going through
patient-reported outcome measurement to the
patient. And I just wanted to note that if
that's the case that suggests that process
measures might measure that and we shouldn't
be apologetic about it. If it's a process of
being asked about how you feel and your
symptoms and all of that is important to
patients we can measure that in and of itself.

DR. ADAMS: Thank you, Phyllis.
So with those closing remarks let us thank our
panelists here. Well done, thank you.

(Applause)

DR. ADAMS: And lunch is in the
back and we'll see you back at 1:30. Thank you.

(Whereupon, the foregoing matter went off the record at 12:45 p.m. and went back on the record at 1:29 p.m.)

DR. PACE: Okay, good afternoon and welcome back from lunch. So we're now going to really focus on the performance measure. And this first panel is about reliability of the PRO performance measures. I'm going to introduce the panel and then I'll make a few introductory remarks about our NQF criteria just to put that in context.

So, we will be hearing from our Commissioned paper authors and RTI and Brookings are our authors. We selected them because they have experience in developing organizational performance measures and have brought performance measures to NQF for endorsement. And our speaker about reliability from the Commission authors is Laura Smith who's from RTI.
Our panel then will be Lewis Kazis from Boston University School of Public Health, Lori Frank from Patient-Centered Outcomes Research Institute and Jack Fowler. And his organization name has changed slightly so it's Informed Medical Decisions Foundation. Is that correct, Jack? Okay, great. All right.

So without further ado I'll dive into NQF criteria related to reliability. Next slide.

So, again, just to put us on the same page we refer to the PROM, that's the instrument or scale or single-item measure used to assess the outcome of -- or concept as perceived by the patient. And then that patient level scores or values on those PROM instruments will be aggregated in some way for a performance measure, aggregated for the healthcare entity providing services. And that's what NQF would be endorsing in terms of these performance measures.
So, although -- and actually in this session, reliability and validity, we talk about data electronic and performance score. We are -- our Commissioned paper authors will specifically be addressing the performance score. So let's go to the next slide.

So in terms of NQF criteria regarding reliability our subcriteria related to this is first of all that the measure is well-defined and precisely specified so it can be implemented consistently within and across organizations and allow for comparability.

Remember that NQF endorses performance measures that will be used not only for quality improvement which is of paramount importance but also in accountability applications. So we need to have some standardization. And this starts with having measures that are precisely specified.

And how this relates to
performance measures based on PROMs is that in order to -- you need to also specify what the PROM instrument is that's going to be used in the performance measure and if multiple ones are going to be used then they would need to have comparability or some equivalents as we talked about last workshop.

The other aspect of our reliability criteria -- so we think that precise specifications form a foundation to have reliability but it's not the only thing that we look for. So we do look for some reliability testing that demonstrates the measure data elements are repeatable and producing the same results, or that the measure score is precise.

So we do allow in our current NQF criteria for quality performance measures testing at either the level of the data element or the performance measure score. So next slide and I'll talk about that a little bit more.
So, in 2010 we had a task force that specifically looked at measure testing. NQF always had criteria for reliability and validity but we wanted to get more guidance on how our steering committee should evaluate performance measures on these areas.

And so I'm just going to mention, we have a whole report on this but I'll mention a few key things. And one is that reliability and validity require empirical analysis. And it should be based on the measure as specified.

Again, our task force and guidance at this point in time suggested that we allow for testing at either the data element level or the performance measure score. So an example of that is if we have a performance measure about percentage of patients that achieve a blood pressure below 140/90 the data that goes into that is the blood pressure value. And is that data reliable or your data for identifying patients who have a diagnosis
of hypertension, is that reliable.

And in this case, and then at the performance measure score we're looking at the signal-to-noise analysis that people have mentioned. If you have data on many providers do you have good signal to be able to distinguish among the providers who are being measured.

In this case the PROM data would be the data element. So we've already talked about basing performance measures on PROMs that would be reliable and valid, and I think we all agree in the context for which they were developed, but the question is should we require that there be also testing at the performance measure level whereas currently our criteria would say at either level. So these are some things that we'll get into as we get into discussions. Okay, next slide.

Is that the last one? Okay.

So that's kind of the background just on reliability in the NQF criteria. And
now Laura Smith is going to tee up some of the
issues and considerations in regards to --
that were pulled out in the paper.

DR. SMITH: Thank you, Karen. So,
in this section I am going to talk about some
methods for evaluating the reliability of the
provider-level performance measures and also
potential strategies for designing your
measure in order to address potential
reliability issues. Next slide.

Before I dive into that though I
just wanted to pause for a moment just to
acknowledge sort of the broader context that
we need to think about reliability within.
And so this first bullet is basically sort of
your basic psychometrics that the reliability
is necessary but not a sufficient precondition
for validity.

And when we think about
reliability of the performance measure there's
also a very direct importance relationship
with validity, especially if performance
measures are being used to rank facilities or providers for public reporting or other types of policies like pay-for-performance that if you have the performance measure scores largely determined by noise or measurement error you're going to end up with misclassification of the ranking of providers which would be a serious threat to validity. So just for that backdrop.

And then the last piece, I just want to acknowledge that any of the suggestions that I'm making about evaluation methodology, all of this needs to be taken in the context of a lot of the issues about meaningfulness and validity that have been discussed in the sessions earlier today. Next slide, please.

So here I've shown how you can basically depict reliability as an index. And starting with patient-level, the classic Streiner and Norman way of thinking about reliability as the ratio of the subject
variability to the total variability in the measure. And here in this first equation it's decomposed into two parts. So it's the subject variability, the true variability, and then measurement error. And so if you have a lot of measurement error then your reliability score index is going to go down.

And in the next equation you can see that actually this is an analogous way of thinking about reliability at the performance measure level. And so using the language from NQF that we have the ratio of signal-to-signal plus noise. And that noise can be decomposed into measurement error at the patient level and measurement error at the provider level.

So reliability can be quantified as this index which has a range of zero to 1, zero indicating that the entirety of the variation that you see among providers is attributable to noise, and a 1 indicating that the entirety of variation that you see among providers on the performance measure is
Attributable to true differences among providers.

And so the literature suggested that a good threshold for evaluating the reliability of your performance measure is having a reliability index of 0.7. Next slide, please.

So, to discuss further the determinants of performance measure reliability the components, the important components are the magnitude of the true differences among providers so that numerator value and then the magnitude of within-provider variation. So the measurement error that you see at the PROM level and also at the provider level. And then the size of the provider sample or that denominator for your performance measure.

So, one question that we started talking a bit about earlier today are the implications of how you might aggregate the PROM data into a provider-level performance
measure. And in this case we need to pause
and think about how choice of using an average
change or a threshold might have on the
reliability of the performance measure.

I don't necessarily have an
endorsement of a particular strategy, but
rather a couple of examples to weigh in this
decision. So, one issue with using an average
amount of change, and I'm not editorializing
on this particular measure but to give you an
example of where this approach is used there's
the -- excuse me, "Change in Basic Mobility as
Measured by the Activity Measure for Post-
Acute Care."

And so what this measure does is
look at the average change from a baseline
score for mobility to a follow-up. And so one
concern about using an average change is that
this type of measure is vulnerable to
measurement error at both the baseline data
point and at follow-up.

For the threshold measure, and
another example would be from the nursing home world the percent of patients who are able to self-report moderate to severe pain. And in this case one consideration would be what the reliability is around that threshold where the decision has been made that the patient would be counted in the numerator as having moderate to severe pain.

And then the last consideration for this particular issue that I wanted to touch on although I know there's lots more that could be discussed is that also that choice of threshold. So what I was just talking about with both of those examples is the impact on the within-provider variability, but this choice of threshold could also have an impact on the between-provider variability.

So if you chose a threshold that was either very low and very easy for most patients to clear, or very high and something that would be very rare you'll end up with very little variability for the most part at
the provider level and therefore reliability is reduced. Next slide.

So, to sum up some of that discussion, the performance measure reliability is therefore dependent on the characteristics of the set of providers but also the patients included in the measure. And also, another side effect is that reliability is not static. So you can have basically a set of reliability indexes for providers at the beginning of public reporting and if people's -- if providers begin to improve performance and everyone is converging on a similar score the reliability of the measure is going to be reduced.

Then lastly, estimates for smaller providers are more vulnerable to random error. Next slide.

So in the next few slides I'm going to discuss methods for reliability testing that would help provide evidence to support the endorsement of a PRO-PM score.
And for the most part what is included in the paper and in this discussion, we really haven't treated PRO-PM measures differently from any other in terms of methods.

However, there are potentially other conceptual considerations that I think have been brought up earlier about whether or not the decision to include the requirement of reliability at the measure level should be considered. I'm going to leave that panel to discuss though. All right, next slide, please.

So, here's a list of performance measure reliability testing methods. In the interest of time I won't go into a lot of detail, but point out two of the strategies here, so the two-level hierarchical model allows estimation of the signal and noise that I referred to in the initial slides. And it basically results in a reliability estimate for every provider.

The interclass correlation...
coefficient also allows the estimation of a reliability index for each provider which can be very useful in some of the subsequent discussion about trying to determine the adequate sample size for your measure necessary to have a reliable measure.

These two strategies have been used in some recent measures that have gone through NQF endorsement and were -- the two-level hierarchical model is explained in some detail in a recent paper by the Committee of Presidents of Statistical Societies. And so there's a lot of good information out there in terms of how to implement this testing. And also there's a report by RAND. All of this is referred to in the paper.

There's other strategies that have been used in the literature to examine reliability, to examining the overlap in confidence intervals calculated for every provider. You can give a visual depiction of the reliability of that particular measure.
Inter-unit reliability can be
derived from ANOVA and GLM that's derived from
the F test generalizability theory and Monte
Carlo simulation or other strategies. And
I'll refer you to the paper for references for
those.

So in the next section we talk
some about the issue of provider size. Just
remember that given the different components
that determine reliability provider size is
not the only determinant but it's an important
one. So actually we can go on to the next
slide.

So those two strategies that I
mentioned, the hierarchical modeling and the
intra-class correlation, because they give you
a reliability estimate for every provider
allows you an opportunity to look at what the
relationship between provider size and
reliability is and potentially identify a
threshold where your reliability estimates for
providers are 0.7 or higher. Next slide.
So if you find that a large proportion of providers in your target population have a reliability that are lower than 0.7 you may want to consider some additional strategies for improving performance measures.

It was mentioned earlier today one strategy would be to design a composite which is combining two or more performance measures. So at the provider level in order to increase the data points that are being used in the calculation of the performance measure.

Another strategy that increases data points would be to change the time window over which the measure is being calculated. So if the measure is looking at results within one quarter consider increasing the time period to two quarters or a year.

One concern about this strategy is that your performance measure is going to be less sensitive to changes in quality which can have multiple concerns, one having to do with
the use of -- the usability of the measure for quality improvement but also in terms of the acceptability to providers in terms of their being able to make changes that appear later and quickly in public reporting.

One other strategy is to work on improving the reliability of the underlying PROM which could have to do with changing the way that questions are phrased or instructions are made.

And then lastly which I'll spend a little bit of time on is to apply reliability adjustment. Next slide, please.

So reliability adjustment which again can be applied using the hierarchical modeling, in connection with the hierarchical modeling that I mentioned earlier and then also intra-class correlation coefficients.

Basically instead of dropping from public reporting or whatever format that you might be presenting results, instead of dropping facilities that have small sample
sizes from being publicly reported, you can actually adjust the provider scores by shrinking their estimates towards the mean value. And this mean value could be for all providers or if there's a wide distribution of sizes of providers you might consider stratifying into a larger -- excuse me, you might consider shrinking towards providers of a similar size.

However, there's some concern if you're using volume for provider size. This might be something that's particularly an issue if you're looking at physician practices where there isn't sort of a hard way of counting size other than patient volume, is that there's endogeneity of volume with quality, that current volume of patients that a provider might be providing services to can be affected by prior quality -- perception of quality by patients.

So in light of that shortcoming you may want to use in addition to volume
other provider characteristics. And nor
surprisingly given that smaller providers tend
to be more vulnerable to poor reliability the
smaller providers are more likely to be -- are
shrunk towards the mean values.

And then just in closing wanted to
bring back sort of my initial comments about
the importance of the relationship between
reliability and quality -- excuse me,
reliability and validity. And that you can
see in cases where there's poor reliability of
the provider performance measure that you can
have misclassification in public reporting and
other uses of performance measures. Thank
you.

DR. PACE: Okay. Jack? Or,
sorry, Lewis. Looking right at you and saying
the wrong name.

DR. KAZIS: Can you hear me? It's
indeed a pleasure to be on this panel today in
particular because of Jack Fowler. And in
fact Jack was my first instructor at the
Harvard School of Public Health more years ago than I'd like to think. And Jack and I haven't changed in our appearance since then. But I'm delighted to be on the same panel with him.

So, my charge today for this panel is to speak about the relationship between reliability and validity, or what is the connection between these two concepts. In some respect being in this field for more than 25 years has provided me with some historical perspective I think on this issue and on things going forward.

For a number of these years much emphasis has been on the patient-reported outcome measures with a view to assessing reliability and validity of the instruments. And this was clearly a mandate for a number of years with such organizations as ISOQOL, the Academy of Health and other organizations where you go to the meeting and you hear about reliability and validity of your particular
instrument or questionnaire.

More recently these meetings I think have taken on a different context and for that reason have become much more interesting given that the instruments, questionnaires and so forth are now for the first time being applied to the front lines of care in terms of how care is being rendered and also in terms of the organization and processes going forward to improve quality.

So Figure 1 is a pyramid. And what the NQF calls PROMs includes the focus on the consistency of the metrics using approaches that build on the signal and noise concept that Laura talked about earlier.

The precision of the measure -- and these assessments include such things as test-retest, internal consistency reliability which has not really been talked about here but that's the Cronbach alpha that we all know about. And other methods including hierarchical models, Markov, all kinds of
things that are fairly sophisticated and bring an appreciation for reliability.

  The precision of the measure in the world of legacy measures has been considered as important to a reliable assessment. The validity of the assessment or more generally does the measure measure what it purports to measure is really the bottom line in my opinion as one can have a reliable assessment but if it isn't valid then the exercise becomes pretty futile.

  More recently, the use of the HRQOL assessments that measure a range of functioning from the physical to the psychological have undergone a continued transformation with the advent of the PROMIS approaches using IRT and CAT. And David Cella did a great job at the last workshop detailing those approaches and applications.

  These assessments have important applications in the measurement world and now have begun to be applied in a range of
settings to evaluation healthcare. Both legacy and PROMIS measures are now being introduced to a new world of performance measurement applications. The use of these measures in this context is fairly recent in the United States and present important challenges as we begin to think about how such measures can be used in the context of evaluating the healthcare processes at the provider practice levels, among hospitals, between plan organizations, and so forth.

The relationship of the metric for performance measurement is well established from a statistical and scientific vantage point. The validity applied to a performance measurement, however, is really in my opinion very early on in its development. Clearly the validity issues will become established with more experience.

An analogy that I thought of this morning, and maybe it wasn't a good one, and it might be a bit of a stretch, but to use
this in the context of the space program, to launch an astronaut to the Moon. This took nearly a decade when John Kennedy announced this challenge and goal in the early sixties. Maybe being from Massachusetts I thought it would be appropriate.

The physics for this had historically been in place since Sir Isaac Newton, many years, actually centuries before. However, the proof in the pudding or of the physics was not established until we actually put a man on the Moon.

While we have many instruments and metrics out there, the proof of concept of the application and validity of these remains to be applied and tested in the use of the metrics in the real world, and then ultimately to improve the quality of care in the healthcare system. Perhaps we should develop a time line with goals as this moves forward.

For purposes of performance measurement and just to be a little technical,
reliability can be defined simply in terms of what I call the expected and the residual reliability. Expected reliability tests for the consistency of the case-mix differences in predicting outcomes among entities being compared such as physician practices, healthcare plans, et cetera. Case-mix is based upon the sociodemographics and clinical characteristics that are covariates in the model. So, expected reliability would be the consistency of those case-mix -- of that case-mix across different entities that you're comparing.

And Steve Fihn in fact you asked a question this morning about case-mix and this type of reliability in fact can get at that issue as to whether case-mix in fact might be changing amongst what's being compared. A very important point I think.

The residual reliability on the other hand which is actual values minus expected values is an indicator of whether the
signal can discriminate amongst the entities being compared which would be practices, plans, or whatever, and is based upon the magnitude of the signal among the entities and the sample size of each of the entities where there's baseline and follow-up data for the outcomes that are being compared.

And this can be calculated using intra-class correlation coefficients. Markov modeling techniques can also be used which was in the very nicely done report by Laura.

The residuals measure the extent to which a member's experience exceeded or failed to meet the expected health change. So you are comparing these plans in terms of whether in fact expectations are exceeded or whether they failed. The reliability of these residuals then describe the extent to which the entities being compared are truly different compared with measurement error or noise.

So the science for measuring
reliability for purposes of performance measurement is in place and the challenges as I see them are in demonstrating the validity or whether the measurements are resulting in a meaningful difference in the comparisons being made. So a couple of questions related to this.

Are we measuring correctly the appropriate domains for the purposes of performance? Are the populations being targeted correctly for this purpose? Are the populations sufficiently homogenous so that one is able to find meaningful differences that one can act upon in terms of the HRQOL outcomes being assessed? Is what is being measured truly measures of outcomes that are impacted on by quality or processes of care?

Could you go the next figure?

So, my charge was to look at the differences between reliability and validity, or to talk about the interface. And you'll notice that the arrow there is bidirectional
with a question mark because I think there are some that will think it should go only in one
direction, from reliability to validity.

Others that have a broader perspective I think see it as bidirectional. And the reason for this is that reliability will give us a metric that is precise and provides an understanding of the measurement characteristics. Validity on the other hand will determine if we're on target. Both can inform each other. In the event that the metric is not sensitive to the comparison being made among, for example, doctor practices, this could be a function of the content and construct validity.

The content validity may be that the domain of content being measured needs to be modified or added to, or that the construct being measured needs further refinement. Also, are we measuring what is relevant to the provider and patient? This then points to both issues of reliability and validity.
The organizations being measured which is at the top there are also an important consideration. Is the measurement being taken seriously within the organization? And how much -- or in fact are they just giving lip service to it? How much emphasis is being placed on interventions that will impact on the outcomes being assessed?

The measurement also may be so rigid and specific that the processes of care and interventions are not being reflected in the outcome metric and the dynamics of what is occurring in the entities being measured and compared are not adequately represented.

Does this measurement provide a depth and breadth of content so that the dynamics of the processes of care impact adequately on the outcomes being measured? Do we need to consider generic measures and disease-specific metrics as well? Is the overall net of what is being captured reflected in the signal in the entity being
compared with other entities so that interventions and processes of care are reflected in the variability and the outcomes?

And just a couple of last points. Gaming is something which is at the bottom there of that figure that can be defined as meeting the target or threshold but clearly missing the point. There are proprietary efforts in companies designed to consult with health plans or insurers designed to improve the Stars ratings that was discussed this morning. In fact, one of my former students works for such a company.

And basically they'll go into the insurers, to the health plans and they'll look at how in fact can we improve the Star rating. So some gaming clearly is going on, influencing the metrics and in the future it may be important to develop methods to better understand gaming approaches and how this can be revealed through the patterns of the results.
Is it possible that the metrics, for example, may influence the nature of the case-mix to provide the physician practices or plans with more favorable patient mix for purposes of outcomes? This goes back years ago to what we used to call adverse versus preferential selection. But I think we all know that some of that is out there.

To conclude, I think that the science of performance measures for reliability is far advanced in evaluating the reliability of the metrics. On the other hand, the validity of the metric as a performance measure is pretty much in its infancy.

Validity I view as when the rubber hits the road. It involves the success of the measure in adequately measuring entities among health plans or physician practices or whatever with sufficient precision and variability that the results are found to accurately reflect the outcomes of processes.
of care.

Are the measures in populations studied covering what needs to be measured for purposes of improving performance in quality of care? There are a number of methodological issues that need to be addressed I think related to the validity of performance measures, and this includes face validity, construct validity and criterion validity. And I think you all are aware of what that is.

The sensibility and consequently usefulness of what is being asked of the patients needs to be present. And similarly for the clinicians and administrators that they also view the metric as important and useful. So I think the clinical sensibility issue becomes really paramount in this context.

A last point is that much of this work goes back to the buy-in of the patients and providers at the front lines of care. That's where it's at. And I begin to see that
as I age and I see clinicians and you know, clearly that's where everything is happening. That view of the metric of what is being evaluated is very important. I think that more experience with validity of performance measures is needed as we move into what Aldous Huxley used to say, a brave new world, where performance measures become a mainstay of healthcare. Thanks.

DR. PACE: Okay, thank you, Lewis. And just also looking ahead we have two panels that will specifically be targeted on validity issues. But thank you for that relationship, that's what we needed.

Okay, Lori? Or Jack? Which one?

DR. FRANK: All right, I'll go next. All right, thanks very much. The outline for my comments since reliability relates so much to variation is variation on a theme. I'll discuss how the concept of patient-centeredness relates to reliability and as with the last time my goal is to
explore how re-framing patient involvement in PROs can improve measures and enhance the value of PROs for use in clinical settings and for performance measurement.

Communication, trust and perspective are my main themes as I think about performance measure reliability. And I had a visual slide up as well.

DR. PACE: Yes, Jessica will put it up.

DR. FRANK: Okay, great. First, the authors made a distinction, an important one between generic and condition-specific PROs. And the distinction has implications for psychometric assessment. How you approach psychometric evaluation differs by the type of PRO. I think that it's useful for us to think about performance measures and whether they are setting- or process-specific or setting- or process-generic which is a point Laurie Burke made in terms of context of use.

So my question to address is are
there any differences or unique considerations for demonstrating and evaluating the reliability of a PRO performance measure. Really the question is what's special about patient report. Psychometrically, nothing.

Is reliability of the PRO performance measure score needed in addition to reliability of the PROM? I would say yes but it connotes something different for the performance measure than it does for the PRO which is a point I think made very well in the paper.

Reliability is foundational which is why I chose an image that had a table with a strong base there. So we need to know for an individual PRO that it demonstrates adequate internal consistency reliability as well. Are the items all pears or are there some apples mixed in? Are there things there that don't fit in? So I picked in image that had different fruit in it as well.

And then we also need to know
about test-retest reliability. At a different
time and with no changes expected from time 1
to time 2 will we get the same results?
That's why I chose an image that also had a
clock in it.

And relatedly, at what time point
does it make the most sense to assess
reliability and for that matter, quality?
Have patients provided input into the timing
of the assessment?

For the theme of perspective which
is part of why I chose something that came
from a cubist tradition although this is
technically futurist which I think fits with
this meeting well, it makes us think about
perspective. Do we have trust in our PRO for
use with individual patients and do we have
trust that aggregating the responses will lead
to sound conclusions? So what's required for
us to have trust once we aggregate up? What's
required for the patients to have trust?

They're important reporters but also important
consumers of the information. What is it that they need in order to trust this information?

So what's different about patient report when we consider PRO performance measures? As I said earlier, psychometrically nothing but because we're dealing with information obtained from people and heterogenous people, everything's different. People are variable. There's inter-individual differences, there's inter-individual change and a reliable measure needs to work across different perspectives and faithfully transmit differences and faithfully report on those true changes.

The authors raised an important issue, that of the provider-specific care quality which requires that the provider is considered in terms of the match with individual patients. For this we're talking about sample means or ratings, and what could be termed goodness of fit between a provider or a system profile and a patient's
preferences. There was some talk this morning about the role of patient preferences in performance assessment.

A separate type of reliability to contemplate then is the extent to which the provider consistently scores well for patients with similar preferences on key variables. This is a perspective issue which I think those strange wavy lines on there help to communicate beyond what we usually consider. It's something we don't usually see.

How best can we communicate the reliability of the quality measure, quality or performance measure, to patients and to other non-clinicians? How can we do so in a way that builds and maintains trust? What else should be communicated? Which goes to what we were talking about this morning about squeezing every last bit of value out of a patient-reported outcome measure which would include considering the direct-to-consumer communication possibilities.
Creating trustworthy measures, measures truly worthy of trust, is a very important way to do this. Among the issues that are hard to address let alone communicate to a wide audience is when quality of the evidence and the ability to draw inferences from the data varies systematically by some variables of interest as with the example the authors gave of the shrinkage towards the mean strategy which works better for top hospital quality than bottom-quality hospitals.

How do we establish trust in our measures? In the section on aggregating data the authors rightly point out that means can be misleading when the population of interest is diverse in some important ways. Obtaining reliability estimates of a measure can be similarly misleading if a measure truly performs differently in different populations. So, getting back to perspective the shrinkage technique referenced earlier, the performance may differ by a specific variable.
In the case of measure reliability that differs by populations clinicians and also patients should have input into drawing the meaningful distinctions between groups and then the researchers can go onto hypothesis-test the differences in terms of impact on reliability. Cut points, for example, might differ by patient groups in ways that we haven't explored with patients helping to define what those groupings should be.

Part of the point of involving patients in performance measure development and selection is opening up performance measurement exercise to things that we as clinicians and researchers have not considered as important but in fact may actually be. So, things that patients see that we might not be able to until we view it from the patient's perspective.

The authors reference a notion of similar patients with similarity defined by researchers or clinicians generally. But
there's room there to consider other types of definitions of similarity, and even having patients comment on those definitions.

An example of patient preference given in the paper is an example of a caregiver proxy measure. So I'd be remiss if I didn't make the point that family and other caregiver reporters are important but we need to be mindful about loss of fidelity which might be more of a validity issue but missing the mark a little bit despite reliably hitting that same spot on the target. Proxy report, obviously not the same as a caregiver-reported outcome, a point I think that we're all clear on here.

The authors correctly point to inter-interviewer variability as a threat to validity. I would say it's a concern for reliability as well. The example they gave was a script used in the minimum data set to support inter-interviewer reliability.

It's interesting to consider when
the standard step that we've all used to
minimize variation might actually introduce
unwanted variation in some cases. So are
there patients who need a different script?
Are there those who need the interviewer to be
able to bury the script in order to collect
the most reliable and most accurate data?

I want to end with the concepts of
"in" and "of." We talked last time about
patients involved in the research as subjects
but also helping to define the outcomes,
helping with recruitment, et cetera. And then
patients of the context of research, helping
to define the topics and prioritize, and
helping with some of the funding decisions.
I think that same in -- way to conceptualize
it can be applied to performance measurement
too and it's very exciting as we think about
futurism for us to contemplate that here.

As Ethan said, how are patients
involved in defining meaningfulness? How do
we measure meaningfulness of scores reliably?
1 Are patients involved in this?
2
3     And I really appreciated what
4 Patti Brennan said in terms of listening to
5 the patients about how they think about their
6 own care. And it's somewhat analogous to
7 hearing the patient voice in development of a
8 PRO. All of this reinforces the
9 trustworthiness of the measurement enterprise.
10
11     Should performance measurement be
12 patient-centered? Not always and not
13 necessarily, but I would challenge us to think
14 about situations in which inclusion of the
15 patient perspective does not improve the work.
16
17     Last time I discussed the idea of
18 involving patients in discussions of sources
19 of measurement error in performance
20 measurement. For this topic today I stand by
21 that recommendation. We have to recognize the
22 patients who are not being asked about their
23 healthcare in the best possible way for them
24 have the power to bring measurement error
25 including reliability in.
So last time I referenced Donald Berwick's 2009 Health Affairs piece on patient-centeredness. These are radical and disruptive shifts as he would call them in control and power. And if they improve care quality then I think they're worth it.

Thanks.

DR. PACE: Thank you. And now Jack.

DR. FOWLER: It's neat to follow two people. You've covered a lot of things. And the paper was very good. And I'm not going to talk much about the calculation strategies which I think there are other people here better qualified to talk about that.

But as I thought about the reliability issue I keep thinking that there are these two steps here, and without the two steps working then it doesn't work. And so the question that was raised I think that was framed initially and -- is it enough to assess
the reliability of the patient-reported
measure or do you have to assess the measure
of the performance measure? And you've got to
do both.

And there are two steps, and there
are two sources of reliability that are here.
You've got whatever amount of error there is
in the measurement of how the patients turned
out and how well or badly we are measuring
what kind of shape they're in. And then we've
also got this relationship between whatever it
is the providers do and this outcome. And
that -- if that's got unreliability too, I
mean I think you reduce the reliability -- bad
reliability times bad reliability gets you
wherever you get.

And I think you've got to do both
steps. And without doing them both I don't
think you're there.

And so we do have quality measures
that are not two steps and I just want to
point that out, that these process measures
that folks talk about don't involve those same kinds of two steps. So the CAHPS measures involve saying did they talk to you in terms you can understand. How long did you wait in the waiting room to see a doctor. There are not two steps there. You had the clinical experience and you reported on it, and that's got its own reliability but it's one step, it's not two. We don't have to have a hypothesis between waiting time in the waiting room or in anything else, we just wanted to hear about that.

We've been working a lot on decision quality. And actually, the three things we think you need for a good decision are a process that involves interaction with a physician in a certain way, being informed and having a decision that is -- matches your goals and values.

So, the first one, the one about the process, again is a one-step thing. We can ask people did the doctor -- how much did
the doctor talk about the cons as well as the pros. Did he talk about alternatives and did she ask you what you wanted to do. Those are things with no steps in between, they're just reporting on what you experienced. The questions can be perfect or medium but the measurement problem is just the one thing.

But then if you go to the next step to say, you know, did you get a concordant decision then you've got to have some kind of link between the way the physician interacts with a patient or whatever you think the action is and the outcome that you're measuring. So you've got to have both of those steps.

And one of the things I sat around and thought about, building on what Liz Mort was talking about this morning is what are the kinds of situations or patients that we can think of where you could really say with some confidence we think that the way the doctor treats them could make them better, or keep
them from getting worse. Or at least you could predict what the outcome would look like or the change would look like if they got good care, whatever that means.

And I had trouble making a list of those things. You know, there are just a lot of things -- so you know, and I know others in the room have wrestled with this sort of thing. For example, you take low back pain and that's pretty cool but the problem is that about 80 percent of low back pain resolves itself in a couple of weeks. So probably no matter what you do to them a bunch of people are going to get better. So there you go.

So you say, well you've got to get the right back pain patients in order to study whether the intervention or the support they get. And that, I keep going back to the problem of how well can we do at identifying the patients whose outcomes we're going to measure.

I notice the Europeans, both of
their approaches were to take surgical patients. So they actually ducked the patient identification problem of who's eligible for this by just picking the ones that got intervened with.

Now, we thought a bunch about this and that is the easy one to do. If you want to do something reliable to sort of measure outcomes you can measure people who get cut or intervened with in some other serious way and find out if they got better or enough better. One of the nifty things for providers is that with a few exceptions there are a whole lot of interventions like that that make people better at least if you measure symptoms before and after. You kind of end up looking better though we could say whether the Hopkins people turned out as well as the MGH people turned out as well as the UNC people or something like that. We can work on that.

(Laughter)

DR. FOWLER: But the real problem
with that is the point that Liz Mort brought up also, is that the -- where I'd really like to start the evaluation of how do we treat people are with people who have got the back pain or who have got the knee pain or have got BPH or whatever it is they've got. And I don't want to -- and I don't want to give people bad scores just because they're less interventional. And if you just take the people who get the big interventions you're likely to get pretty big improvements but you don't get to give credit to the providers who let the guys live with their BPH symptoms or who let the guys say that I'm willing to kind of work with my herniated disk and treat it conservatively, and if you measure me 6 months later my back's going to hurt worse probably on average than if I got surgery for my disk but I didn't want to go through the surgery. And that's -- I got better care in some way even though my symptom score may be higher. And I say I would hate to build in
interventions, you know, incentives for interventions that will fix problems at some cost and you're not measuring the cost so much as just measuring how the symptoms come out. And the other -- so I worry about that and I keep coming back to the fact that maybe the most complicated problem we have is how to identify the patients whose outcomes we really want to track. And how do we do that. We've been working with a bunch of medical practices to work again in this case on quality of decision-making. And so one of the things that we need to do is identify people who were candidates for making decisions who are not -- and that's a pretty good group of the people that you might want to see how things turn out for. And boy, do practices have a hard time identifying people who have a certain symptom situation or medical condition. They have trouble identifying their diabetics. And if they do identify diabetics, maybe everybody
who's got a diabetes score at a primary care visit, they can be in all kinds of different stages and places in where they are.

Arthritis was another one that you picked out as maybe could monitor pain. But boy, the arthritis people, just because you've got a code of arthritis, you know, you could be somebody who's been managing it for a long time, had it under control and you're just sort of maintaining it, or you could be a new onset person for whom actually how things get managed could make a big difference.

And I think the patient identification thing it seems to me is one of the greatest challenges. And it is a reliability problem, not a validity problem because I think every practice is going to do it differently.

And I was again interested in I think it was the Swedes who seem to be managing to get patients to fill out sort of baseline questionnaires about what their
status is about things. And that would be a big help in terms of patient identification for these symptoms. But we don't do that in this country and there's nobody who's got -- or hardly -- it's unpredictable who might happen to have a back pain score or a BPH score or something on their general practice thing.

So the notion of who could you reliably identify so I could have the same patients with the same kind of clinical characterization across a bunch of practices is a problem I haven't been able to think of. And that is a genuine reliability problem because as we were talking about earlier today, you know, the impact and the validity of measuring performance, you've got to even the playing field about treating the same patients and how they're coming out.

And if there's differential information about who they are and different ability within practices to identify the same
people that really is an important source of unreliability I think in the measurement that's going to be really hard to wrestle with and get right.

So there are a bunch of challenges here and you know, it's such important and good work, and there are -- there probably are some cases where we can identify reliably some people the care of whom we could take as an outcome of care. But it's not a really, really long list I don't think and I think the problem of patient identification has got to be paramount, and whether you can make a reliable and meaningful assessment of whether or not medical care is actually better, worse or whatever for that set of people. Thanks.

DR. PACE: Okay, well thank you to Laura and our panel. And now we'd like to open it up for discussion with the rest of our expert panel and audience. So again those of you on the phone you can get ready. We'll take your questions in a little bit but I'll
start with any questions or comments for our panel. Patti. At the table they're green. Sorry.

DR. BRENNAN: Sorry. Hi, everybody. I woke you all up. This can be addressed by anybody on the panel. I see an oversimplification of the healthcare system that we're going to have patients within practitioners and practitioners within practices and everybody's going to stay where they're supposed to be.

And so can you help us think about how we're going to deal with me who has a primary care doc who I never see, a nurse practitioner that I always see, but sometimes I see the person who's covering that day because one of those people is sick or I've come on the wrong day. And I know that you can't solve all the problems but are there ways we can think about reliability and validity at the level of the practice that we could actually sum across clinicians?
DR. FRANK: Yes. So I just wanted
to raise the idea of the quantified self and
big data, and data are getting bigger and
bigger. So if we start from the patient up
then we can systematically review data for
patterns and come to some reasonable
conclusions about quality from that direction.
You're absolutely right, we're not there yet,
but it's to me an interesting notion. It has
to be top-down also. But I'm excited about
the possibilities for patient up.

DR. FOWLER: I mean I think the
only other thing I'd add is that the CAHPS
people have been wrestling with that for a
long time too. And partly it's sort of what
your unit of analysis is. And I think, you
know, a care team or a care provider or an
ACO. And if you get a little bigger then we
give the whole system credit for it and that
may be helpful. You still could have two
systems treating you too but that at least --
that solves the problem for a lot of people.
DR. PACE: Okay, Greg?

DR. PAWLSON: I think I've been abandoned at this table. They're all up there.

Two observations and a question.

One, I can't agree more with Jack about the need to really characterize populations very carefully. And even where we think we have accurate diagnostic information we don't for the most part. And there is so much range in the labeling that we use at the current time that it's almost meaningless. And so I think if we're going to minimize the huge variance that is introduced by patient-level variation rather than provider effect, if that's what we're actually trying to do I think that's sort of a numero uno problem.

The second I think poses a real problem for NQF. If we are to go beyond the characteristics that -- of reliability and validity that are sort of baked in at the measure level and get into the performance
measurement level in a way it would be
interesting to get the panel to reflect a
little bit on how far do we think NQF has to
go in requiring testing at that second level.

In other words, I think most
people would agree what we want to see in
terms of reliability and validity
characteristics of the measure piece itself.
But when you're testing as both the paper and
a couple of panelists pointed out the
particular reliability that you find is
dependent somewhat on the population that you
tested in. So what's going to be sort of the
practical advice for the NQF panels in what
should we require, how much testing, how broad
does the population have to be, what can we
accept as enough to give us a reasonable idea
that this will work in most populations let's
say, or at least if you don't go way far
afield?

My favorite example is actually
one from the VA. There was a paper published
slamming the colon cancer screening measure
and it was done in a VA clinic in San Francisco that was end-stage COPD. Well, we
never anticipated when we developed the colon cancer measure that that's where it would be
used and obviously that population the measure had no real validity or reliability for that
matter. So, the question is what do you think NQF should do.

DR. KAZIS: Yes, it's a very good question. I think that my recommendation would be to begin with what I call low-lying fruit and start with some success stories that are going to get the clinical community interested and involved in the process rather than going after something that might be initially viewed as being out in left field and more difficult to reach.

So, perhaps to target some populations very specifically and apply some assessment tools that have some track record already. And then provide the ability for
those metrics to be understood and recognized by those that are using them through recommendations and through other means. But I think what's also important is not just to provide a metric but to provide those that are using it with some understanding of what they need to do with it.

DR. PAWLSON: You were also suggesting sort of post-marketing surveillance after -- in other words, if we accept, you know, testing. And I guess to use an example that Jack was pointing out, if we sort of focus on surgical procedures I would guess that there is a fair amount of variance in whom -- to whom surgical procedures are even applied. So what level of testing would we start with and then what about the sort of post-marketing surveillance of, you know, further evidence of its reliability and validity as it's used more widely?

DR. KAZIS: Could I just mention one thing related to that which is
variability. And the whole idea of health services over the years was to reduce variability. And is that -- I think we could raise sort of a skeptical question, you know, looking at it as a skeptic is variability necessarily a bad thing. And are there situations in medicine and clinicians out here might clearly have some examples of this where in fact a procedure is done in a very different way by two different groups of docs but in fact both are success stories?

DR. PACE: Right. And then in that case you would have similar outcomes so I don't think people would quarrel with that as long as you had two effective processes. Jack?

DR. FOWLER: That is my question a little bit. And variability may not always be bad but I think we have tons of data that show that it’s physician-driven most of the time and is not related to either the patient preferences or their conditions. And so while
I would not want to say stamping out variability should be our goal by any means, I think making sure that it's related to outcomes and patient preferences and their needs should be a really high priority here.

DR. FRANK: And I'm glad that you raised the notion of post-marketing surveillance because I think the FDA regulatory pathway is instructive here. You know, a measure is never valid or reliable for that matter. We need to iterate and then it's just drawing the line, when is it good enough.

DR. PACE: Al?

DR. WU: It seems to me that there could be sort of a best case scenario for what the reliability of the measure might be. At least at the first phase I had two thoughts. One was that the procedures, the prescribed procedures for how to collect the measure include -- in order to attempt to achieve good reliability, acceptable reliability, should be specified very clearly.
The second thing I wonder about is it seems like this lack of reliability, variability results from lots of things. And I wonder if -- and a lot of those things could be termed metadata. And I wonder if at least when something is being rolled out if we should make a concerted effort to collect some of that metadata so we can parse out where the variability is coming from.

And you know, for example if you were to say have a checklist for an administration procedure and checking boxes 1 through 5 essentially went along with the score that you got and you could then calculate your reliability you could see what the relationship was of those boxes being checked to whether or not you got an acceptable reliability. And that sort of thing might be worth thinking about as we go from just PRO to PRO measure.

DR. PACE: Right.

DR. WU: I'm sorry, PRO quality
measure.

DR. PACE: All right, Laurie?

DR. WU: Whatever.

(Laughter)

MS. BURKE: This is Laurie Burke from FDA. And I appreciate that last diagram that showed the relationship between reliability and validity because I think that really is key in understanding what's going on here.

But yes, and I agree, Al, that there are many -- I have to think about four sources of variability and the first one is true heterogeneity amongst the people you're trying to measure. So, and I also agree, Lori Frank that there's never complete validity or optimal reliability. It's always something that can be improved upon.

However, if you're not measuring what you think you're measuring at all that's a big problem. So, I mean we can identify validity that's not adequate. And we can also
but the amount of reliability that's not adequate depends on what you want to use your measurement for.

So in clinical trials where we are always measuring very small differences between treatment groups we have to have very good reliability in order to be confident of the differences that we're measuring so that we can make a conclusion that's valid that a treatment works.

So therefore when we're looking at a measure to use as an endpoint in a clinical trial we look very carefully at the entry criteria for that clinical trial because companies that are developing drugs are very careful about who they let into their clinical trial. They're going to exclude all kinds of people and it's not, you know, and this is the whole real world, not real world controversy.

But the clinical trials are very clean. And they're going to exclude those with certain severity of illness, they're
going to exclude those with concomitant illness, they're going to exclude those with - - in certain age groups.

And so when we look at a measure to evaluation whether it is well-defined or reliable for the purpose of use in this clinical trial we compare the results of all that validity and reliability testing before the trial with this measure. We compare that population that it was tested in to the clinical trial entry criteria. So, because you want to minimize that variability so much that we can be confident of the effects that are demonstrated in terms of an effect size.

So, for performance measures that's what you're going to have to figure out, how much reliability is necessary to be able to use the population size that you're going to use to make some sort of conclusion and come up with a result.

I mean, we have, you know, the argument that Dr. Fowler, the discussion that
Dr. Fowler presented in terms of identifying
the right population and the right
comparability between -- this is -- you're
describing the whole reason for the evolution
of the clinical trial methodology is because
we have to be able to compare and know what
we're measuring.

DR. PACE: Right. And just to --

MS. BURKE: So that's not possible
in performance measurement, we understand, so
we have to come up with another standard.

DR. PACE: Right. And we'll get
into this more in one of the validity panels
but obviously in clinical practice people
aren't being randomly assigned or selected as
Laurie talked about in clinical trials. And
so you know, we rely on methods such as risk
adjustment, stratification, et cetera. And
that definitely is important in terms of
looking at the validity of the conclusions you
can make from a performance measure score. So
definitely important in terms of getting to
that. Ted?

DR. GANIATS: I'm thinking about cholesterol and it varies by season. We know that. We heard earlier today that mood can affect a patient-reported outcome and we know that and that's just all variation that's going to affect reliability. And we just have to hope that it's going to wash out across the groups.

So I'd like you to address something different and that is a controllable, a game-able reliability. I mean, when I go to my car, get it fixed, they say hey, we're going to mail you this satisfaction questionnaire. Make sure you mark them all fives. And we haven't talked about external threats to the reliability that would be important to the NQF. And I'm just wondering if the panel can think about it.

I mean predominantly I think we have satisfaction, information perhaps on satisfaction though I don't know it, but it
would seem that patient-reported outcomes might be game-able in a way that we're not used to seeing in most other performance measures.

DR. PACE: Go ahead, Jack.

DR. FOWLER: Actually we have -- again I'll use the CAHPS experience. We have pretty good data that they've experimented with having physician's offices collect the data by handing out questionnaires, et cetera. You get much better ratings if you do it that way it turns out, much.

So the standard is really that you have an external contractor that has to collect data and you do it in a way that's anonymous so that patients don't have to worry about you seeing my report, guessing who I am and knowing that I'm not pleased or something or that I'm not doing well. So we haven't talked about the protocols and there are all kinds of ways to do it.

I did notice again in all three of
the ones I think this morning that presented
have outside people collecting I think at
least the follow-up data. I'm not sure about
the baseline data in Sweden. But I think
outside data collectors are pretty important
and some way to protect the patients from
feeling like they're exposed to their
providers.

There's some other reasons you
might collect data that you would feed into
providers but performance evaluations are
probably not the right one.

DR. PACE: Could we see if anyone
in the audience would like to ask a question
and Evan can get you the microphone. And
Operator, would you also ask the people on the
phone to signal if they want to ask a
question?

OPERATOR: At this time in order
to ask a question press * then the number 1 on
your telephone keypad.

DR. PACE: Go ahead.
MS. MASTANDUNO: Hi, this is Melanie Mastanduno and I'm from the Dartmouth Institute and working with clinicians at Dartmouth in an observational setting as opposed to a research study on collecting patient-reported outcomes.

And so my question to the panel is in the interest of promoting patient- and family-centered care would not the panel of patients or the clinical team, the practice level be the right unit of analysis for evaluating the aggregate results in order to say we are not pitting one provider against another but rather the approach in this clinic is consistent because they share nurses, they share medical assistants, the secretarial and appointment staff are organized at the front desk. And it's not one clinician practicing at wide variation within a practice. And so could you please respond to that given the discussion of the very granular methodologic concerns that you've raised?
DR. KAZIS: I think what you're raising is an important point. The unit of analysis clearly becomes really important at the practice level in order to who you're evaluating. And whether in fact, you know, one can assume but maybe not always that if you're dealing with a practice with a number of clinicians and nurse practitioners and others, there's a culture within that group. And there may be more homogeneity within that group then if you compare that physician with somebody outside of that particular practice. So there may be some -- and there's been data to suggest that over the years.

There may be some evidence then to suggest that dealing with a unit of practice that includes a number of providers and others that make up that practice is an approach and one that might work.

DR. PACE: Okay, one more. Go ahead.

DR. JAMES: Hi, Tom James from...
Humana. A question that I have has to do with that of attribution. That's a word that was not on the listing of all the definitional phrases.

Attribution is something that at least in traditional medical measurement has wide variation and no standardization. And it's something which drives physician practices crazy. It's clear that we can expand the universe so that we don't have to deal with attribution but then we lose some of that accountability.

How do we get to the point of having some common definitions on attribution?

DR. PACE: I'll just -- it's something that comes up often at NQF and it is one of those areas of measure harmonization that we'd like to see some movement in terms of having more standardized rules regarding attribution. But it is something that plagues us. So thank you for bringing that up again.

It's definitely going to be appropriate in
these types of measures as well. John?

DR. WASSON: Just to follow up on Jack's point. This is just dirty laundry part 2 because a lot of issues being laid out are certainly reasons for caution.

Jack mentioned, well it's very important to have vendors. This morning we were asked the question about cost and everybody fudged it. The bottom line is that when we play a vendor game it's billions of dollars, everything we're talking about here when you multiply it across the physicians.

When you think of newer methodologies that are patient-driven from the bottom up like internet you're talking a fraction of that. And we haven't -- we'll maybe get into this in terms of administration but we're being rather glib about the cost side of things and I think we do have to be careful.

DR. PACE: Okay. Operator, do we have anyone online that wants to ask a
question or make a comment?

    OPERATOR: At this time there are
no questions.

    DR. PACE: Okay. All right. Al?
And I think we've got to wrap it up.

    DR. WU: Just in response to that.
This morning we had a terrific presentation
from Sweden and he mentioned the cost that
they had invested which was something like 20
to 30 million Euros. So, which doesn't sound,
you know, multiply it by 1.3. It's not that
much.

Then you remember that Sweden has
a population of about 9 million. So if you
scale it up it's $1, $1.5 billion U.S. for
what -- which would be an equivalent amount to
what they spent. Now, certainly there's some
economies of scale, maybe it would be a little
less than that, but it's still a big number,
not inconsequential.

    DR. PACE: Okay. All right. Any
other questions or comments? Yes, Chas. You
1 want to tell us who you are?
2
3 DR. MOSELEY: This is Chas
4 Moseley. I'm with NASDDDS. And I'm on the
5 long-term support side of the discussion which
6 is a little bit different than -- a lot
7 different than what you're talking about with
8 the very narrow acute care measures.
9
10 But I think it's important to note
11 that even when you're doing the patient-
12 reported measures for folks who are receiving
13 clinical care it's important to recognize and
14 to characterize the populations very
15 carefully. It's important for the tool to
16 characterize the various populations closely.
17
18 People with intellectual
19 disabilities and cognitive disabilities
20 receive acute care along with everybody else
21 and would be expected to respond to various
22 types of survey instruments. We found in our
23 research with national core indicators over
24 about 20,000 people a year that there are very
25 strong factors that influence the response.
Level of intellectual disability is one. Home residential situation, age, the type of care a person receives, whether it's in an institution or a community, a wide variety of variables that could be expected to influence how they're responding to a patient-reported outcome in a narrow clinical setting.

And I think it's important that whatever instrument is used be able to be adapted for people who have different types of learning styles so that they won't be excluded from the numbers.

DR. PACE: Okay. All right, one last comment.

MR. ROONEY: Hi, Ted Rooney from Maine again. I actually work for a lot of groups who pay the bills, employers and organizations, consumers and unions. And in Maine we have a $7 billion spend. If you believe most of the experts 25–30 percent is waste. We have a totally un-patient centered system.
So if you could say that we're going to institute PROMs in the context of the overall measurement system and the overall measurement system would implement measures that would help drive efficiencies and return on investment in the system and make it patient-centered I think you'd have people rushing to do it.

So I think it's -- the cost is a lot if it's done like we've traditionally done things, one-offs that are marginally effective. But if we can make this the centerpiece of how we do performance measurement going forward and use it to drive real changes I think it's a bargain and that's from the people paying the bills.

DR. PACE: Okay. Well, thank you to our panel.

(Applause)

DR. PACE: And we'll take a break. You can get a little bit of refreshment and we'll reconvene in 15 minutes.
(Whereupon, the foregoing matter went off the record at 2:50 p.m. and went back on the record at 3:08 p.m.)

DR. PACE: Okay, if everyone would take their seats we're going to get started here. All right. So, our next panel is about demonstrating validity of PRO-PMs. And this is part one. We'll have part two tomorrow morning. There's a lot that goes into validity and we've already talked about the relationship between reliability and validity.

I'm going to introduce the panel and then I will make a few comments about the NQF criteria, again, just to put that in perspective in terms of our mission to endorse performance measures. So, next slide. Okay.

And again, you know, most of this panel will focus on the performance measure versus the PROM because we dedicated last workshop to the PROM instrument. But obviously once again the validity of that PROM instrument for the context in which it will be
used for performance measurement is an essential building block to having a valid performance measure.

So, let me introduce our panel. And for our Commissioned paper authors this time we'll have Anne Deutsch from RTI and Barbara Gage from Brookings Institution. And then our expert panel members are Steven Fihn from Veterans Health Administration and Albert Wu from Johns Hopkins. So all of our expert panel bios are in your handouts as well. So, next slide.

So, I'll make a comment about a couple of things and get into our specific criteria. I know Laurie Burke mentioned this morning that in terms of testing they found it to be more efficient to start with validity testing and then go to reliability testing so that you don't have to repeat. From NQF perspective we're evaluating measures after they've been tested and we tend to look at reliability first and then validity. But I
think that's certainly an interesting strategy
for those who are going to actually be
developing and testing measures and probably
does have some efficiencies attached to it.

So, in terms of what NQF is looking for with validity is that our first thing is that we want measure specification that are consistent with the evidence that's been provided to support the measure focus. And again, we see this as foundational so that if the evidence -- if the measure is specified to be consistent with the evidence that's a foundation for validity, but then we do require validity testing.

And we want validity testing that demonstrates that either the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Again, in the context of NQF we're endorsing performance measures that will lead
to improvement but also will be used in accountability applications. And so it's key that you can make valid inferences about quality. If you see a list of providers and their scores on a quality performance measure can you say -- can you know that higher scores mean better quality versus lower scores, poor quality or in some instances vice versa. So, next slide.

In terms of NQF guidance on validity testing again we'd like to see empirical analysis. And again, currently we allow for demonstration of validity at either the data element or the performance measure score. And in this case the data that would go into a performance measure is actually the PROM value or score on that particular PROM.

The other thing is that we do currently allow for measure developers to submit a demonstration of face validity of the performance measure. We ask that this be systematically assessed but this is kind of an
area of weakness in terms of face validity
done by a group of experts and then another
group of experts may have a different view of
the face validity of that particular
performance measure. So, it's something that
comes up periodically in just the quality
performance measures.

And again, you know, with any of
the testing we want an appropriate scope and
method and acceptable results. I think as was
already talked about validity is really
something that's built on over time. And so
we do expect that that will increase over time
and that our criteria for testing, our really
initial entrance to get NQF endorsement and
for you to be thinking about, you know, what
is the minimum amount of testing or
demonstration of validity and as we talked
about in the last panel reliability that means
it could be endorsed as a performance measure
that could be used in accountability
applications.
So with that I think I'll turn it over to our authors and we'll go from there. Thank you.

DR. GAGE: All right. I'm going to start us off and then turn it over to Anne to talk about some of these issues on the validity. So, thank you.

So I'm going to kind of refocus us. As Karen just mentioned we've been talking a lot about performance measures, about organizations or providers and holding them accountable. We've also had a lot of discussion in here about quality improvement and about thinking about the measures that are necessary for ensuring quality in the organization or in the provider. And there's a lot of -- I like to think in terms of a 2 by 2 cell.

So you have the instrument as we talked about with the first paper gathering and all of the issues associated with that. And then we have the performance measure. And
what happens when you take that measure, that instrument up to the organizational level and start applying it? What are the differences in the validity and the reliability as we heard earlier?

But there's also the second row are the differences between the performance measures that are clinician-based and those that are patient-reported. And really what we're here to talk about today is that fourth cell, the patient-reported voice in the performance measurement.

And so it's a little bit different than traditional thinking in developing quality measures because you have to think about if you're holding an organization or provider accountable then you want to make sure that that measure is appropriate to that population as we talked about this morning. And tomorrow I think we're talking about some of the issues of risk adjustment so that you know who's in, who's out, how to apply it.
But keeping that in mind, that what we're really talking about is how do you incorporate that patient's voice in the performance measure and keeping in mind some of the issues that came up this morning in terms of if you're holding -- if you're trying to measure quality at the organizational level who -- what is the difference between the clinician's assessment of that outcome because we're talking about outcomes or the patient's assessment of the outcome. And how is that patient viewpoint affected by their preferences, their knowledge of the care.

Some of the examples that Anne will give come out of the rehab field and there you have a real disconnect between what the patient thinks will happen now that they've had that nice little hip surgery and what the physicians or therapists know is possible. So it takes measuring performance to a whole different level when you start talking about accountability and the patient's
voice.

On that I'll turn it over to Anne to go into some of the details.

DR. DEUTSCH: Great, thanks. Next slide. Terrific. So can everybody hear me okay? Great.

So one of the first questions that we were asked to address is what are the implications of various approaches to aggregating the PROM data, for example, an average or a medium amount of change percent to improve or reach a benchmark. And what is the validity of the conclusions about quality and the ability to discriminate performance among accountable entities. So next slide, please.

So last workshop we spent a lot of time talking about reliability and validity of the instruments and so now as Barb and Karen mentioned we're moving to aggregating that data up to the provider level.

So there's really two options when
we have these measures. So one option for looking at the information at the provider level is to calculate a change score. And so the example here is a decrease in pain or perhaps an improvement in functional status between the start of care and end of care.

Also, you could look at a threshold which is the level that the patient achieved at the time that care either was ended or at a certain time point after care was initiated. So here the example might be percent of patients with moderate to severe pain. Next slide.

So in many cases change is often thought to be a good way to look at things. And as Barb said I come from the rehab world and so a lot of times we've looked at improvement in function over time between start of care and end of care.

But there's some limitations with change. And I think this was mentioned a little bit last time by Dr. Ottenbacher but I
just wanted to reinforce some of these issues.  

So, an individual's change score can vary in terms of the magnitude and the direction. So you can have improvement of positive 10 units or minus 10 units or you can have an improvement of plus 2 or anywhere in between any of those or even more. And so individual differences can really be masked when you start doing an average. So 10 plus, 10 minus, the average is zero. And so you know, you had two very different outcomes and yet your average really doesn't tell you what happened to either one of those patients.

Change scores also tend to have a lower reliability than the baseline at follow-up scores. And Laura mentioned this this morning, that basically if you have error in your baseline and you have error at your follow-up you might actually be adding those errors up together.

Also, there's floor and ceiling effects. I know this was also discussed in
the previous workshop. And so you might actually have real changes that occur but your instrument is insensitive at the low end perhaps and so you don't see people gain when there's a floor effect, or you may have patients who are at the high end of the scale. They have real change but your scale isn't sensitive enough at the high end and so you might have some patients who the scale really doesn't fit them at either end.

And then also it's unclear sometimes what the clinical meaningful change -- what the change score really means. So again, from my world in rehabilitation we measure functional status. And so you can have an improvement of 10 units. What does that really mean?

And actually one of the research projects that I've done recently is we actually presented information to people in the community and said, you know, if you were trying to pick a good rehab facility which one
looks like it would be better. And we gave them fictitious data. And we actually did give them a change score and they looked at it and they were like, well you know, I can look at this but what does it really mean. So nobody really knew what it meant. And I will tell you a lot of rehab hospitals do actually put that information on their web pages. But it really is hard to know what it actually means. Next slide.

So just some examples. There are some patient-reported outcome measures that are endorsed by NQF and so I wanted to highlight some of those as examples of some of the issues that I'm talking about.

So, and Laura mentioned this one, the "Change in Basic Mobility as Measured by the AM-PAC" is an instrument that has been used in both outpatient and inpatient rehabilitation programs. And that basically is a change measure. And I'll get into more detail about that when I talk about some of
the other methodologic issues.

    There's also the measure that was mentioned this morning, the percent of patients with moderate to severe pain. And that is a threshold value.

    And then the last example I wanted to highlight was depression remission within 6 months. And that basically is a threshold value but the way that's designed, actually the patients who are included in that measure are people who at baseline have a PHQ-9 score that indicates depression, possible depression, and then the follow-up score looks at how many patients moved into the area of no depression based on the PHQ-9 score. So it actually is threshold but it's kind of an indirect measure of change. Next slide.

    So the next area I wanted to talk about is again aggregating this up to a provider level. You can calculate a number of different statistics. So you can calculate a mean, a median, you can calculate a percent or
a ratio.

So I know in school you know we were taught use as much data as you have. So if you have age, use it as a continuous variable when you're doing risk adjustment or when you're doing analyses.

And so I think a lot of the work that people initially think about is to calculate a mean. But a mean or even a median might not necessarily represent the diversity of patients when you have a pretty heterogenous population. And in many cases when we're looking at these outcomes data they really are heterogenous in terms of the population.

And then also again when we're looking at the provider level if the data aren't normally distributed a mean or a median may not really represent what's going on in that particular provider. Next slide.

So a lot of the performance measures that actually are endorsed by NQF are
percent measures. And so there's different ways of calculating percents. So basically it's how many patients reach a certain -- reach or exceed a benchmark.

And so one way to do that is to say, you know, what's the national expected value and what percent of patients meet that expected value, whether it's a threshold or a change. And again it should be similar patients. I think that was mentioned earlier.

The second option would be what percent of patients meets some kind of fixed amount of change. So if there's some kind of clinically meaningful difference that's been identified that's, I don't know, 10 units, what percent of people or patients actually meet or exceed that 10-unit defined difference. Or a minimal detectable change.

A third option could be a threshold value that's associated with a long-term outcome. So for example, if you wanted to look at balance confidence with somebody's
ability to -- with balance you might say well, this threshold is important because that threshold is associated with a reduced risk of falls.

So for PROMs that have established clinically meaningful thresholds or cut points I think it's easier to create quality measures out of them. For areas like functional status where there really aren't good clinically meaningful thresholds or stages it's a little bit harder I think or more challenging to develop some of these quality performance measures. Next slide.

So the last option I talked about was a ratio. So basically that's a score that may have a value of zero or greater and it's derived from dividing the count of one type of data by another count of data. And so the example would be the number of patients reporting a pain score of seven or higher divided by the number of inpatient days. So the number of days there is the bottom of that
metric.

So a ratio may be preferred when the amount of time, in this case the number of days that the patient is at risk for the outcome is important to consider. Next slide.

So some of the examples I wanted to highlight in this particular area. Again, the depression remission within 6 months that I mentioned before. Again, this classifies patients into clinically meaningful groups.

So the people who are depressed at baseline and then they change 6 months later into this category of not being depressed.

Second example again is the change in basic mobility a measured by the AM-PAC.

And this is a percent of patients who change and the change here is defined as a difference of more than one minimal detectable unit basically.

So a minimal detectable change for those of you not familiar with that term it refers to the minimal amount of change that is
not likely due to measurement error and that represents a true change. So one of the questions that I would have about using that threshold is whether there's a lot of variability, whether all patients should be expected to gain and you might have some providers who have a lot of gain and that wouldn't be reflected if you have this threshold that's the minimal detectable change. Next slide, please.

So the next set of questions were about validity testing. So the first question, what methods of validity testing would support the demonstration of validity of performance measure scores that are making a conclusion about the quality of care. Second question, are there any differences or unique considerations for demonstrating and evaluating the validity of PRO-PMs as compared to other quality performance measures. So, next slide, please.

I want to start off talking about
face validity. And I know that face validity
is not necessarily the strongest but I do
think it's a really important step personally.

As somebody who's been involved in measure
development and been involved in a lot of
TEPs, both participating and also running TEPs
I feel like you always learn when you hear
different people's points of view. And so I
do think it is an important step. So I
personally feel it would be important to have
face validity testing at a performance measure
level.

There is various methods that NQF
has put in their materials in terms of
recommended ways of systematically looking at
face validity including modified Delphi
survey, some kind of formal consensus process,
the UCLA/RAND appropriateness method, and then
there's also the American College of
Cardiology and American Heart Association has
a paper that outlines steps for considering
face validity, next steps.
So, of course we always talk about expert panels being experts. And so I just want to highlight here our experts here are probably our patients. So sometimes patients are included in expert panels but it seems to me that we could do a whole series of validity testing really using qualitative methods. So for example, focus groups, semi-structured interviews and cognitive testing with expert patients.

I also want to highlight that the patients would need to be people who are well informed what they were being asked to do. And there is actually a fabulous article that Dr. Judy Hibbard wrote, and I know she's here in the audience that I think could really help frame this. It's "What is Quality, Anyway?" And so she actually conducted some focus groups and asked patients what terms that they thought about in terms of quality.

And again, I've done research in this area where I went out to senior centers
and gave people information, percent of patients with moderate to severe pain and asked them to interpret the data. And I certainly gained a lot of insight into what people really, you know, when they look at the data what they're really looking for. And a lot of the people that I spoke to without any -- without much orientation were really quite sophisticated in terms of what they were looking at and were not sure that they wanted to make decisions just based on one piece of data.

So, I'll bring up some more examples I think probably tomorrow when we talk about threats to validity. But I do think that there is an important role for the public, patients, whatever word we're using to be more involved in validity testing for the performance measures based on the PROMs.

But I guess I also would say, just to follow up with what Barb said, there's no reason why non-PROM measures shouldn't have
more testing with actual public patients also.
So, next slide.

So the next area to talk about is criterion validity. So again this is the extent to which the measure agrees with the gold standard.

So, one potential example of testing here could be a PRO-PM being used and the data compared to a performance score based on clinician observation. So again the patient-reported information being compared to a clinician observation if it taps into the same construct, so for example, functional status. And let's say the clinician observation measure was really found to be valid and then you had this patient-reported outcome that agreed with that. That would be one potential way. Next slide.

The next area is construct validity. And so this speaks to how the measure performs based on theory. And so I kind of made up this idea that -- and this is
all very theoretical, but a way to test this
would be basically if you had national data.
So everybody collected PROM data.

So I know I'm dreaming here but I
have national PROM data and I had some
facilities who did quality improvement
projects and it was focused on whatever topic
we're talking about, maybe it's functional
status or pain. And I was able to compare
their data before and after. And then also I
had all these control facilities who I know
they didn't do quality improvement projects.
And so you'd be able to really look at whether
there was improvement in the places because
the quality intervention -- quality
improvement project really did work, by the
way. That was the other assumption. And so
there would be improvement that you could see
there.

So I think this is probably an
important point to highlight that I think in
order to really do validity testing we
actually need people to collect this data. We cannot test validity beyond the face validity unless we have multiple providers collecting this information and we're able to compare. If we just have two places collecting data and oh, aren't we good Karen and Barb, our three facilities, you know we really are not in a good position to make a judgment about quality. And so we really need a lot more implementation in order to really be able to test validity at the level that we would love to.

So I think it's ideal to obviously have a lot of validity testing but I think realistically at this point face validity is very important and we can certainly learn a lot with that. But I think the construct validity, we need more widespread data collection before we could expect to be able to have measures meet those standards. Next slide.

So the last question is is
validity of the performance score indicator of
quality needed in addition to the validity of
the PROM. So as I said I personally think
face validity is an important issue at the
performance measure level. I think the other
levels, it's ideal but we're not probably
going to be ready for that for awhile.

And I think that's my last slide.

Yes. Okay, great. So turn it back to Karen.

DR. PACE: All right. Steve, you
want to?

DR. FIHN: So, I had actually
prepared some remarks and slides but as we get
later and later in the presentations I think
the speakers are going to find that there's a
great deal of overlap in the comments. And
instead I decided to sort of abandon that
tack.

I was actually going to walk
through -- we thought with all the sort of
theoretical discussions that sort of a
practical story might be useful. And I was
actually going to walk through the story, a 20-year-old story actually of developing a PROM related to ischemic heart disease and how we went through the reliability test retest and so forth, the validity assessment, responsiveness, defining the minimal clinical detectable or important difference. And the use of this measure which is called the Seattle Angina Questionnaire now and something like two to three hundred clinical trials. And make the point that we went through a lot of this and it obviously overlaps with much of the material from the first workshop too. And then sort of pose the question given all that work would I then trust it because that's what we're talking about. Do we trust these measures to be a performance measure. And what would make me want to trust that aside from the fact that you know we did all that work.

(Laughter)

DR. FIHN: And you know, I think
all the previous panels have walked through a
lot of the concerns about rolling these up in
terms of statistical and methodologic and many
of the issues that I would have -- was going
to delineate.

And I think the answer to the
question really was even after 20 years not
yet. And this discussion for me has been
during both the earlier workshop and this one
has been I think very interesting and
exhilarating but in some ways also sort of
frustrating because we've been hearing similar
discussions now as I think back for two or
three decades.

And the question then that keeps
getting posed is sort of what do we need
actually to move these measures into clinical
or organizational use. And what would move us
forward.

And I think what I'm hearing is
again to sort of develop beyond the
methodologic and scientific basis a framework,
a use framework. And I was saying to Liz Mort earlier I particularly liked the framework that she presented which was I think, you know, in terms of the tight linkage between measurement and effectiveness of intervention I think that's a key piece and she walked through examples of that.

And why is that important? Well, in our own measure we just completed and reported a clinical trial in which we selected out people with extreme scores and subjected them to what we thought was a very effective intervention. And by provider, almost 200 of them, and found really no effect. And made me then go back and wonder about sort of the mutability of this measure in terms of linkage to our therapeutic interventions. And I think Liz pointed out there are some areas where we do have, you know, good linkages there.

I also particularly like the NHS examples. And you know, I think they are as was mentioned picking on surgical or
procedural interventions which might actually be a good place to start for a few reasons.

One is these are episodic episodes of care. And one thing that hasn't been talked about a lot today is sort of when you apply these to longstanding chronic illnesses you get into a whole `nother set of issues of repeated measures, change over time or lack thereof, what do you do in those circumstances.

An example would be we've had mandated screening for depression and PTSD in the VA for years now. And annual at a minimum and more often if it's positive. And we have a prevalence of chronic depression of about 20-plus percent. So we, every time I'm in clinic I have two or three patients who have got to be re-screened. I know they're depressed, they've been on therapy, they've been through most of our treatments and they represent sort of our residual chronic depressed population. It's mostly prevalent,
very little incident depression.  

So then you know how do you deal 

with that in the context where you want to see 

change and most of the literature really 

demonstrates change with incident depression, 

not chronic or sort of intractable depression. 

So, episodic. 

We largely have pretty good 

measures as we heard for hips and GU and 

cardiac disease. We know there's high 

variability as Jack pointed out in many of 

these areas. And the numbers I think actually 

will be tractable. 

One of the scary things in primary 

care we face for instance in the VA is that 

we've got 6 million people in primary care. 

And if we're going to start surveying all 

these people, you know, it gets back to what 

John Wasson pointed out, just the logistical 

and expense of it is actually I think 

daunting. Whereas if we were just to do 

certain limited procedures at least to start
like hips or cataracts or something that
that's probably tractable and a way to get
started.

And so I personally sort of like
the idea even though yes, it's biased, and no,
it doesn't get us to sort of the real
population base we want. It is a place to
sort of get going.

One of the interesting things also
I thought and Jack Fowler also brought this up
in terms of the issues of patient selection.
And I think there's a paradox here. In fact,
one of the uses, I think one of the goals of
using PROMs would be actually to influence
patient selection.

You know, in fact one of the
concerns I think we have, people have talked
about the waste in the system is the use of
procedures for individuals who don't
necessarily stand to derive a great deal of
benefit. And one of the things I think that
would be a positive if you started to do this
was to see a sort of up-shift of the patients
who started going, you know, who actually have
some -- who aren't floored out, you know,
already, or ceilinged depending upon your
perspective and could stand to benefit. You
know, you might actually see a bit of a Will
Rogers phenomenon where everybody sort of you
know gets a stage shift, gets better.

So again that was really the issue
I asked about case-mix was to sort of sense if
you're asking providers to demonstrate
improvement then they're really going to focus
on people who have an opportunity for
improvement as opposed to people who might be
mildly symptomatic.

And I think also just to comment,
I think Lew Kazis brought up the issues, he
brought up several important issues, but I
think the issue of gaming which we've seen a
lot in the performance measurement world of
ways in which these things can be gamed.

I think the PROMs actually present
us with some new avenues that will be interesting to observe as we get in here in terms of not only gaming by systems and providers but by patients. We talked earlier about the notion that occasionally there's the temptation to drill down to the patient level on these measures.

I mean, we're sort of in a reverse situation where traditionally we've had measures that we know work on groups and the question is can they work on individual patients. And now we're being asked let's think about measures that we now think work okay at the individual level. Can we roll them up but roll them up in a different way, not to the original populations but to the providers they're seeing or to the systems in which they're enrolled.

But nonetheless at the patient level you could see where patients actually now that they're contributing data to systems might have motivations for eligibility for
certain procedures or drugs or whatever might
actually now have some motivations to frame
their responses in some ways.

You know, I think I have, you
know, a lot of concerns about the broad-scale
implementation of these. Nonetheless I do
think it's time. We're actually as we speak
implementing the heart disease measures for
the 30,000 or so elective cardiac caths that
we do to sort of look before and after as a
start in our system similar to what I think
the Brits have been doing for a couple of
years it sounds like and the Swedes for
longer.

But I think in systems that the
larger systems can get started doing these
things. So thanks very much for the
opportunity.

DR. PACE: Thank you, Steve. And
Al?

DR. WU: So, Steve this morning
had all his slides prepared. I had scribbles
on some, on you know, my clothing and my hands
and index cards. And so we've done a little
role reversal because to clarify my thinking
I put together some slides and so now I have
a couple of slides. But they still resemble
things that you might scribble on your palm
while you're thinking. Next, please.

So, first of all this is, you
know, sort of going back -- we keep going
forward and back a little bit and there's a
little bit of a theme to all of that. But
maybe we, you know, two steps forward, one
step back, two more steps forward. I think
we're getting there.

Sort of a must-pass criterion is
that our performance measures should have
scientific acceptability in measurement.
Next, please.

So I just -- I'm here to tidy
things up a little bit. And this is a -- it's
sort of maybe the second or third most famous
painting by Seurat, you know, que sera sera.
But it's late in the afternoon. But in any event so my purpose is to sort of tidy up all of these difficult pesky little questions that we have almost like individual dots of pigment on a page. So next slide, please. So there we go.

(Laughter)

DR. WU: That's where we hope to be. But -- though I'm worried that I don't think we're going to get there this afternoon.

Next, please.

So here are the four questions that were posed to our panel and I'm going to actually only just comment on each of them. First, please. Next slide.

So, various implications for aggregating data. And I actually -- aggregating could be taken in two ways and as we're thinking about this it's probably worth thinking about on the one hand scoring, generating scores, and then separately on aggregating rolling up more than one score --
a score from more than one individual into a
score that is used for a performance measure.
And sort of maybe as we look back at the
report and so forth maybe it would just be
worth splitting it out like that because I
found that I was confusing myself which is not
that difficult really. Next slide, please.

A couple of issues. There can be
some problems with aggregating at the
individual level and then there can be some
problems with -- I'm sorry, in terms of -- I
realize this is actually not so much
aggregation but scoring so the slide's a
little bit mislabeled.

But if we look at an item that's
used very often and in fact some version of
this is I guess an NQF measure. If you take
a visual-analog pain scale or a 1 to 10 pain
scale, rate how much pain are you having, on
the individual level first of all there's some
problems because measurement tends to be very
course. Even though you might theoretically
have 10 or 11 or an infinite number of points between you know sort of 1 and 10 or zero and 10 in fact a lot of people are 5, a lot of people are zero, really a lot of people are zero perhaps legitimately. And then there's quite a few people who are 9 or 10.

And I saw someone the other day who was sitting very comfortably in the exam room chair, not crying, not grimacing, not wringing her hands and tearing her hair and she had indicated a pain score of 10.

And I said, "So are you in any pain?" She said, "Oh, you know, a little bit." And I said, "So, you wrote 10 down here. Why did you do that?" And she said, "For emphasis."

(Laughter)

DR. WU: And she did have some arthritis pain, she'd been a little stiff, and she wanted to make sure it got taken care of. And so at the individual level people are interpreting what this is for for different
reasons. We think of it for measurement, maybe for screening and she -- for making sure something got taken care of which is completely legitimate and probably more worthwhile than what the Joint Commission asks us to do, but nonetheless.

At the group level those biases actually probably tend to iron out. But I have observed and there's some data that suggests that provider panels differ with regard to the patients who they attract.

Another example. I recently, we had a very good and a very nice physician leave our practice recently, Dr. Rochelle Brown. Is this a HIPAA violation? Maybe not. So she's a terrific physician and she left and her patients loved her. And so I've inherited a bunch of them. And the average pain score for all of these people, what do you think the median is? Ten, yes.

And basically she is so nice that all of these people who are, you know, very,
who are in a lot of distress gravitated to her and stayed with her. And if she were examined for proportion of patients who had pain scores cross-sectionally of 10 she would look terrible when in fact, you know, everyone really would like to have her as their doctor. So, at the group level there are some other things to consider.

When we're testing validity of aggregation strategies I think we need to do a few things. One is first we need to look at the distribution of scores period which I think has been done but we probably need to look at them at several levels including at the level that we're going to be aggregating.

If we're looking by provider if we see that some people have very skewed distributions of pain scores, some providers have patients with a very skewed distribution and others with a very normal distribution we should worry about our ability to compare them fairly.
Of course there is genuine heterogeneity as we heard. And some of it is heterogeneity by provider. Some patients really are in a lot of pain and some people less so. And we do want to be able to detect that. Oh, next please. I don't mean to touch these buttons.

So, I actually shouldn't get into this very much but we know that measure of -- that asking people about change is unreliable, that in fact measurement of individual-level change you ask -- even if you measure it twice and then calculate the change, since there's error measured at both time points those individual change scores may have a lot of noise in them, especially if there are different things happening to the people at different time points. When you aggregate change scores, if you look at average change scores some of that noise gets taken out.

A question which I realized I didn't know the answer to is is it more useful
for the purposes of quality -- for performance
measures to measure mean change or median
change or the percent achieving a benchmark or
the percent with a meaningful change of some
sort. It seems like that could be done in
some data sets, maybe even the HOS data. I'm
not sure but I think that I would like to know
the answer to that question before I start
deciding that my measure is going to be based
on for example a same/better/worse for example
scoring system.

An example from several of you
were involved in a Medical Outcomes Study and
one of the -- maybe the most impactful study
that included actual results from the Medical
Outcomes Study was John Ware's study in 1996
looking at the 4-year outcomes for the panel
of chronic disease patients. And overall
there was no difference between HMO and fee-
for-service care. However, if you looked at
some subgroups and you used a
same/better/worse scoring method then people
who were in fee-for-service and particularly
people who were more disadvantaged or older
did better than people in HMO. And the
results are probably true but the same
differences were not as prominent when you
looked at comparisons of mean scores. And
someone please correct me if I'm wrong. I
just had glanced over this paper again
recently.

And that makes me think. And
these data in some ways are equivalent to
something, to data that we might use in a PRO
performance measure. It made me wonder and
worry a little bit about if you score things
differently do you get different conclusions.
Next, please.

Some patients can't improve. We
heard a little bit about that from Floyd
Fowler. If you look at an ambulation measure
and you've got people who are paraplegic. If
you on the other hand look at people who don't
need a surgery and who get it it's possible
that they won't benefit.

(Laughter)

DR. WU: So first of all it made me think that maybe we, in some of our validation studies we need to look at appropriateness too as another piece of metadata so that we can weed out those people who really didn't need the surgery.

This is a shameless plug for a book that I'm not connected to but my colleague just wrote, Marty Makary just wrote a book called "Unaccountable." He's a surgeon and he's talking about how the system is not very accountable. It's coming out I think on Monday.

And he told a great story of a surgeon who -- at a terrific institution. I won't, Liz, I won't mention what institution. And he -- who operated on sort of a VIP and who should have been a hernia. It turns out it wasn't. But he said oh, we'll fix it anyway. And so they fixed his non-hernia.
The guy had a bunch of complications. And you know, he was not better and he probably could not have benefitted because he didn't actually have a hernia as it turned out.

So, in any even if some people cannot improve then we need to understand the heterogeneity of people who we are measuring quality on. This just gets at defining your measure carefully, defining the specifications very carefully, are they chronic patients, are they acute patients, are they people who could possibly benefit. Next, please.

What methods of validity testing would support and are there differences or unique considerations. Next, please.

So, we were having this conversation at our table and a little bit -- even in between this meeting and the last, and we're really confronting a little bit of there and back again which is basically PROs were originally developed in order to be able to measure the effects of health services
 interventions at the level of Group Health Cooperative of Puget Sound or the Medical Outcomes Study or the RAND health insurance experiments or trying to measure sort of the utility of populations.

And we actually have quite a lot of data on the validity of PRO measures developed for group comparisons for these and later for clinical studies. So we've got actually loads of data. Maybe not enough, we could always use a little bit more validation, thank you, but we're sort of -- we're not in terrible shape.

PROs are now beginning to be applied quite a lot for individual patient care. There really is a dearth of data at this moment on the validity of measures used for that purpose. We know about the greater unreliability for individuals. There haven't been so many validity tests for individuals. We don't know if those measures are responsive for individual people.
And so I'm actually sort of diverging a little bit from Steve in saying that they have been used in individuals but we actually don't have enough evidence yet. So now we're at the point where we're going back to groups and we want to use PRO performance measures -- typo here -- again for group assessment.

And we have maybe some advantage of the fact that we're looking at groups again. But the data that we have developed for group comparisons of services research is not -- may not all be applicable because we have the added complication that we're defining these measures for specific populations, specific contexts, to answer really specific kinds of questions. Next, please.

So, we're looking at validity of PROs for group comparisons, PROs for individual use which I'm not going to talk about so much and PROs for quality
improvement. Next, please. PRO performance measures.

It's worth -- this is a little simpler than the measures that we -- than some of the figures that we were provided with today but I like looking -- I like this figure because it's really easy to remember. And this is from Ira Wilson and Paul Cleary's JAMA paper, the relationship of pathophysiology to symptoms, a relatively direct link. Most of our treatments, the things we're trying to measure the quality of are mostly aimed at improving -- reducing symptoms, improving pathophysiology. All of those things affect physical and mental health and all of those things affect quality of life, maybe social functioning, maybe role functioning.

The problem is that as you get further and further away from treatment and pathophysiology there are other variables that come in. There's lots of things that affect your quality of life, for example. And it
becomes more and more difficult to demonstrate that, the effect your intervention had on those more distal variables. So it is worth keeping this in mind. Next, please.

So, when we're trying to validate PROs for group comparisons we're going to do the usual things, content validity, construct validity, responsiveness which is maybe, you know, which another way of thinking of that is longitudinal validity, perhaps predictive validity which in my experience is very useful to convince clinicians that something is -- that a measure is worthwhile.

I'm not going to talk about validating PROs for individual use now but in -- and this is really my last slide. In validating PRO performance measures for quality improvement within groups I think that we're interested in construct validity but we almost immediately need to think about risk adjustment I think. We'd like to be able to discriminate one group from another and a test
would be to take known groups and see if our performance measures can discriminate one from another at a point in time.

    I think that's almost -- unless you do great stratification or are very, very careful about how you specify those groups I think you're going to have to immediately get into risk adjustment. And we can talk about this more tomorrow.

Another test of validity would be to look at the responsiveness of the measure to an intervention of known effectiveness. And I'm going to actually sort of disagree with Anne just a little bit because I think it's too optimistic to think that quality improvement efforts at a national level are going to improve quality. You know, every improvement requires change but not every change is an improvement. And a lot of times things get worse before they get better while you're sorting things out.

    And so I think that if we have
interventions that we know are effective and they don't have to be at a national level. They could be more focused. But we then again probably need to either randomize -- we either need a randomized design or we need to risk-adjust again in order to see if our PRO performance measures are able to detect the change that's caused by that intervention that we know is effective.

But I think that those would be sort of doable tests. We've got loads of interventions happening. It seems like it's doable.

We certainly need to test in populations that we're interested in and the narrower the better initially. And in the context that we're talking about, again, the better specified the better.

I think that's it from me. So, we've got some time. Good. Thank you.

(Applause)

DR. PACE: Okay. So we have time
for some comments and questions from our
expert panel and audience. So I'll open it
up. Kathy?

OPERATOR: At this time if you
have a question or comment please press * then
the number 1 on your telephone keypad.

DR. LOHR: This is a comment, a
great panel as the ones were this morning and
lots more questions for us to juggle. But two
or three particular things. And this may be
for you and Barb more than for Al and Steve.

You mentioned clinically
meaningful differences. You mentioned minimal
detectable change or differences. And I was
under the impression that at least for some
PROs if not PROMs and whatever we know some of
that information already. And it's possible
that trying to find certain kinds of measures
maybe in depression where those things are
already kind of documented and known might be
a useful step. Not necessarily for your paper
but more generally as NQF sort of moves down
this path.

But the other question that I had and I'm -- nobody ever accused me of being a statistician -- is whether the minimal detectable changes are not in fact driven to some extent by sample sizes. And whether that has to be taken into account about whether you have very large numbers of patients that are more or less alike, say have the same condition or something, or you have only a few. Or as you aggregate up to -- across practitioners to healthcare systems or whatever that concept and the actual measures of it might change.

The third thing that I had wondered about is your definition of ratio which puzzled me because it isn't the way I think about ratios. And so I was just sort of calling that out.

The other thing that I wanted to maybe pick up on with Al Wu is one of your questions, Al, is that you said you weren't
sure about the answer is which is better, whether it's mean or median change, for example, or percentage meeting or exceeding a threshold.

And you didn't mean it this way I think but that was cast a bit as an either/or kind of question. And in fact it's not an either/or kind of question across the board. It's definitely going to matter depending on the purpose of the measurement and so forth. And it's not that it's not a -- it's an appropriate question but it's not a generalizable one. And I think the purpose and context for the measurement may to some extent drive the answer to that particular question you had.

DR. PACE: So Anne, do you want to start and then we'll go to Al.

DR. DEUTSCH: So I think your first question about the PROMs where there actually are known clinically meaningful differences, so I agree with you. I mean
those are going to be perhaps easier to develop performance measures based on those so I agree with that.

Second question was about sample size. So yes, I would agree and I think Laura touched on this, the larger your sample size the more comfortable you're going to be that you really are being able to distinguish quality. So.

DR. PACE: Can I add something there? Because the other thing is that just because it's detectable statistically doesn't mean that it would be meaningful change to a patient. So I think there's some tradeoffs there.

DR. DEUTSCH: Absolutely.

DR. FIHN: I just would recommend Gord Guyatt's series on sort of how to calculate that which I think separates out the issues of statistical versus clinical, the sample size issues and tries to sort of get at what the underlying sort of construct of an
MCID is which I think still are true today.  
I haven't read those papers for awhile but I  
I think they're still right.  

DR. DEUTSCH: The definition of  
ratio, is that from NQF? I'm trying to  
remember. I'll have to get back to you for  
sure.  

DR. LOHR: It's just I read it and  
I thought, well, you know, I'm thinking about  
odds or you know cost-effectiveness ratios and  
that sort of thing. It isn't the way I think  
of ratio.  

DR. PACE: Right. What we've seen  
in the performance measures is often different  
units. So for example, you know, for example  
with adverse events, you know, the number of  
events to the aggregated time. So it's on a  
different kind of scale but I'm sure there are  
different ways to look at that and we can  
certainly get your definition.  

DR. DEUTSCH: So I think I'll pass  
it off. Barb, did you want to add anything?
DR. GAGE: I would only point out I realize since we got to the end of the presentation that it sounded like boy, there's nothing out there when in fact we know darn well that -- I know we've submitted measures on patient reports of pain and many of those things. So the world is moving along in limited ways.

DR. PACE: Okay, Al?

DR. WU: I would just say, Kathy, I agree with you entirely. It's not one, the other, either/or. Yet another thing to think about is other things that are important about what's the distribution of whatever it is really in the population and how well does your measure measure that. And another thing is what's the functional form over time of health. After surgery if you measure, you know, at 1 day, 1 month, 6 months, a year you'll get different answers and so that's also important to know.

And which measure you use could --
one way of rolling it up could be better or worse depending.

DR. PACE: Okay. Yes, Judy.

DR. HIBBARD: I appreciated the comments on patient face validity and earlier we talked about meaningfulness for patients. And Anne pointed out the need to sort of give people a context for thinking about quality because we do know that patients and consumers often don't have a context for thinking about quality. They don't share our assumptions and understanding. And so when you ask them about quality to give them some context.

But I would go further than that in thinking about querying individuals about their -- the face validity or the value, the meaningfulness of this in the sense that a lot of times different words that you'll get a different response because people don't have this context and understanding. So if you describe something one way you may get a very different response than if you describe the
very same thing with different words. So, thinking about how to get at face validity I think we need to be aware of that.

DR. PACE: Okay. Would you say your name.

MR. ROONEY: Tim Rooney from Maine. I'm thinking about when do we hold communities accountable, you know. Because if you look at the work that RWJ in Wisconsin has done on social determinants they suggest that morbidity and mortality is 20 percent due to clinical care, 30 percent due to health behaviors. And PROMs start getting into health behaviors, but that's only 50. And then there's the built environment.

And in Maine we're doing a lot of work around patient-centered medical homes, community care teams, whatever. We're doing some interesting work with area agencies in aging where they have people on Meals on Wheels who go into homes. And lo and behold you talk to the people that deliver the meals
and they're saying well, we find people with
10 pill bottles living alone and they had 5
before the hospitalization, 5 after. They had
no idea which to take and they're not sure any
of them have worked well. Well, we're trying
to connect that back to the PCP or care
management or whatever it is.

And then you've got some work that
you look what United Way agencies at least in
Maine do. They do a lot of what I would call
healthcare type of stuff. They run a lot of
behavioral health stuff, they do a lot of
stuff.

I think of Steve's comment about
intractable depression. Well, it's
intractable to the traditional medical care
interventions but is it intractable if you
start to look at community interventions to
address that like AA programs and things like
that.

So, I'm thinking at some point is
our PROMs really for medical care, or are they
for healthcare, or are they for care. And I'd love to think about because the thing that -- if you go to the RWJ website they have this terrific video on their project match where this community in Iowa or someplace like that, I apologize if someone's from Iowa but it's somewhere in the middle. They were the worst ranked county in the state. So the whole community got together and they built a grocery store in an area where there wasn't any food. So, that's the power of a community coming to address issues that affect it. So I wonder if we could think of PROMs or PROs as not just a way to hold this doctor accountable but to hold a community accountable in a way that makes them want to do something.

Any perspectives on that or is this too far afield for this discussion?

DR. PACE: Well, NQF does endorse population-based measures but I think that's been an interesting discussion in terms of, you know, then who's being held accountable.
And you know, will they ignite change. But we'll see if the panel wants to add anything else to that.

DR. FIHN: You know, I'm reminded, Ted, of the, you know, there was an ACC/AHA performance measure regarding time to PCI. And you know, the issue was what do you do about transfers, you know, which was I think NQF struggled with that one, the door-to-balloon time one.

And it's sort of a micro of what you said. On one hand no hospital wanted to be responsible for the fact that they said well, if we could do our part but the other, either the sending or receiving hospital might not do their part. And then the community-based sort of approach would be to say that doesn't matter, what really matters is does the patient get the procedure in the prescribed time.

And you know, I don't know what the right answer is to that. I guess it
depends upon what the goal is here, who, as you say, who's being held responsible. And you know, what's the purpose of the performance measure. You know, I think those are, you know, good and very hard questions.

DR. PACE: Helen?

DR. BURSTIN: Just an interesting side note to that. So actually what wound up happening is we do in fact have the time-to-thrombolysis measures. There's also a set of measures for rural hospitals that are actually the time with which they're able to package somebody and transfer them rapidly for their thrombolysis -- for their PCI.

So, it may be sort of in some ways almost a balancing measure that of course your end goal is to get the right therapy at the right time for the right person but that not everybody can play in that space. And so having measures that actually fit what everyone's role is may make sense.

DR. PACE: Ethan?
DR. BASCH: Yes. This is a great panel and I think really brought essential issues to the forefront of the discussion.

I'd like to say in thinking about the day so far I'm actually not so worried about reliability or about construct validity. To me the areas, and we've talked about this in the last meeting particularly, that are of greatest concern are first, what's I think sort of being alluded to here as face validity which others of us call content validity.

Others call it qualitative research.

But you know, the piece which involves going out to the patient population, assuring that what you're measuring is important and meaningful, number one, and number two, that the -- so that's up front. And then that the measures themselves are understandable to the patients, right. They understand what you're saying. And that the terminology within the measures maps to the underlying concepts of interest which one has
a priori identified as being the important concepts, i.e., outcomes that one wants to look at.

And to me really that's central. And without that, whatever we want to call it, you know, I call it content validity but others have argued that they'd like to call it something else. That is a fundamental initial step.

The other piece that I think is key that has been touched upon both in the reliability and the validity conversations is sensitivity to change over time. And that's sort of been wrapped into the conversation about reliability and validity but it really is separate. And I think again is so essential to this idea of evaluating performance or quality because really what we're talking about is the ability of a tool to measure change within or between practices over time because I suspect that's really what most of these measures are going to turn out
looking at. There may be some threshold measures but I really do think we're going to be looking at change measures. And without being able to detect change over time then, you know, as has been pointed out accountability really is sort of you know irrelevant if you can't change it. But you know, we're not sure is it that you can't change it or is it that you can't measure the change. And so I think to establish up front whether your measure is capable of measuring change is really essential. And I bet Laurie Burke has a follow-up to that.

DR. PACE: And I was just going to say, and that is one of the things in the key characteristics, the table from the first paper is responsiveness to change. So I think that's maybe something that will be emphasized in that way. Okay, Laurie, were you going to add something?

MS. BURKE: Oh well, I completely
agree, Ethan. Content validity is critical.
If you don't know what you're measuring then
how are you going to -- what good are the
reliability and correlations or lack thereof
with other measures.

So, and I also think that content
validity really alleviates a lot of the need
to do an abundant amount of responsiveness
testing because if you know what you're
measuring you understand how your measure
responds across the full range in your
population on that concept that you're
measuring then you have a better idea of what
change is in that continuum.

And we have psychometric methods,
a new, modern theory that can help us
understand that we're, you know, that what
gradations along a continuum of a scale mean.
And that we have comparable, equally spaced
intervals between scores. And I think that
that is this iterative approach to content
validity that holds a lot of promise right
now.

We're really able to -- we have a whole lot of hope that these things can be developed much more quickly and much more -- be much more applicable to the situation that we want to put them into.

Now, the validity, the content validity of the performance measure I think has got to be fairly similar except you need to know how the patients feel about that performance measure in terms of how you're going to calculate it.

But content validity isn't just patient input. If it's something that requires expert input of one type or another also is a part of content validity.

DR. PACE: Okay. Are there any comments or questions from the audience members? Evan's got the -- and Operator, what about on the phone line? Are there any people in the queue for comments?

OPERATOR: There are no questions
at this time.

DR. PACE: Ethan?

DR. BASCH: Sorry, I don't want to monopolize the microphone here. But you know, there's something else I was thinking about listening to this which is, you know, once one looks at the measurement properties of the PROM how much do you really have to do on the -- what you're calling the PRO-PM?

And it really seems to me that, you know, and I think you're getting to this, that you know it kind of depends. It appears that there are probably a lot of, you know, as Laurie would say it's a review issue. I think there are probably many settings in which one really does not have to repeat that testing if the PROM has been tested in that sort of target population or context.

For example, I'm an oncologist. There are good measures of nausea. If one is interested in looking at the proportion of patients who experience alleviation of their
nausea following a highly emetogenic chemotherapy, right, and one has a measure, a PROM that has been shown to be responsive to change, be reliable and valid in patients receiving chemotherapy there probably isn't a whole lot one has to do to then consider that as a performance measure.

DR. PACE: Okay. Jack?

DR. FOWLER: I guess I want to argue with that a little bit. I think this is the difference between effectiveness and efficacy. I think the trials that we're likely to have are that, you know, you give somebody a drug and it'll fix whatever it's aimed to fix, or you'll do surgery and you'll improve something or other.

But if we're going to be patient and holistic about this that's implying that doing the intervention which will fix the problem is quality care. And we've talked about it's not always quality care. And I think that because it's not always I think to
just intervene more and that that makes you a better -- that's a better quality of care.

In fact, that's one of the reasons that people are pushing for quality care measurement from a patient perspective is to get beyond just doing stuff to people as a measure that you're doing good stuff and to find out if you're really doing good for people.

So I think the idea that you also need to know what does a body -- what happens to a bunch of patients who are subjected to a particular provider in terms of what their overall outcomes are including the things they care most about which may or may not be that symptom that that drug with all the side effects actually fixes.

So I think that's why you need the studies of the interventions in practice and reality both the condition and also how it works in practice. And overall measures of how patients are doing as well as the efficacy
trials when you subject people with a particular problem to a drug and see the outcome.

DR. PACE: John? Could you use the mike? Sorry.

DR. WASSON: All I wanted you to do was just clarify a bit more. His comment triggered you to make a general statement. Could you just give us a specific? So he's happy with a nausea scale of whatever and you then reacted to it and said no, I'm going to disagree. I'm not sure, maybe it's just me. I'm dense and it's late in the day, but what exactly were you saying is wrong with that? Is it necessary but not sufficient, was that your point?

DR. FOWLER: Yes. I think that the idea that -- I mean, if you're really focused on nausea and if it was really that narrow, well even then I think. He was saying that the trial that showed that the drug works to fix the problem is evidence enough that you
-- no, I think you were. Evidence enough that
the intervention, you know, that the
intervention is a good idea.

DR. BASCH: I actually think we're
agreeing with each other. I think there are
two different pieces here. You know, I think
maybe what it sounded like I was saying is
that if a measure has been evaluated for its
properties in a highly restricted context, for
example, in regulatory trials it could then be
brought out into a CER, into a registry or
into a quality assessment context which isn't
what I'm saying.

I'm saying that if the measure has
been evaluated in a satisfactory way itself as
a measure then when you drop that measure into
a similar context you probably don't have to
go back and, you know, retest it. So if
you've gone out and looked at hard-to-reach
patients and different languages and all like
the good stuff that we care about that's
probably okay. We good? All right.
DR. PACE: Steve.

DR. FIHN: I was just going to take -- you know, and I realize some of the comments were for reaction. But you know, I do think we have to take a balanced view of a technical evaluation of some of the measures. You know, a good example would be the CAHPS satisfaction, outpatient satisfaction scores.

Those of you who are in the VA right now know we've been for the last 2 years struggling to put out patient-centered medical homes in all 1,000 sites of care that we have. It turns out when we started looking at the CAHPS for trying to discern which places were doing better than other places we didn't see any differences. And we went to Commonwealth and we went to NCQA and they said well, we know. That instrument isn't responsive. We've known that for a long time and we're in the process of developing a new measure. And we've got the new one now rolled out which we hope will do it.
But you know, I think we can look at these measures, they may look good, they may have a lot of the testing, et cetera. And when you actually put them into practice they actually don't perform. And I think we should make sure that actually before we do that that we understand that they do perform.

In this case for us it's a pretty high-stakes issue. We're investing $1 billion in this and we can't tell what's working. So, you know, I think there's something to be said.

I think you're right, if you just focus on the technical issues of the measures then we might be measuring something that we don't understand or don't care about. But on the other hand I think if we don't do that we're, you know, liable to perhaps not be able to measure things we really want to.

DR. PACE: Helen?

DR. BURSTIN: And perhaps just one very concrete example that we just went
through in our surgery projects. There was a
measure put forward that looked at improvement
after cataract extraction within 90 days using
a very well-validated tool. I know Al knows
this work well, the VF-14.

No question about it, this is as
validated a tool as it could be. When you
apply it to a performance measure all the
issues we've been talking about today came to
the fore.

The first is what does improvement
mean. In those research studies any degree,
any sort of increase plus up was good enough.
What's actually a meaningful difference of
cataract improvement? And so, there are a
whole series of issues.

It was administered in a very
structured way as part of the original
studies. Now it's going to be submitted, you
know, sent to patients by mail from their
ophthalmologist's offices. So you can see how
of course the tool itself is pristine, but
when you put it in a performance measure in
the real world it gets kind of messy.

DR. PACE: Okay. One last
question? Okay.

DR. LOHR: It's one last question.

Part of your paper in the validity section
dealt with how to cope with patient
preferences. And I wanted to sort of throw
out an alternative universe that asks whether
that is an element of validity per se, or
whether it needs to be considered but as a
somewhat different or separable measurement
activity. Or even a set of items or some
other way of getting at things related to
patient preferences that aren't categorized
inside validity. And that was just, that's
just my --

DR. GAGE: Well, it's a very good
question because the issue comes up. If you
take a measure like pain, you know, one of the
earliest measures of the patient's perception,
but their threshold for when it's impairment
is very different from patient to patient.

So as you're thinking about the use of the patient's information you know part of what we keep struggling with, this whole notion of taking a patient-reported outcome and using it for the purpose of accountability. Not just QI, not just modifying the treatment but actually holding somebody accountable is how much do you allow for that more subjective nature of the patient's preference and the patient's voice and all the other factors that are unmeasurable that might be affecting the patient's response on that particular day.

And yes, we can risk-adjust, we can stratify within different subgroups but you're still, there's still that issue out there. And when you talk about QI the clinical communities are all for giving the best care and taking into account the patient's view. When you talk about withholding payment or some of these other
accountability actions there's less consensus about the importance of the patient's voice. And it's really something we'd like to hear from this group about because it's a difficult issue.

DR. PACE: Okay. So we have one more thing to do before you leave and that is we thought we would spend this last few minutes of you people -- of our expert panel at their tables. And to just kind of as a group, as a table group kind of identify those unique considerations about PRO-based performance measures in relation to NQF criteria. And also thinking about the pathway that we're going to look at tomorrow in terms of the unique considerations of getting from a PROM to a PRO-based performance measure.

So we're going to just stay at your tables. We'll have an NQF staff person at each table to take some notes. And we'll just ask you to do that. And Patti is going to make a few additional comments before we
start that. Go ahead.

DR. BRENNAN: Yes. And could I invite all of our guests -- ladies and gentlemen, I am the second eldest of 10 children and I know how to make a room get quiet. You all have to do the dishes.

Could I invite our guests to please come up and just join a table because we're going to be spending the next 25 minutes or so getting everyone's input on the notes they've been jotting down all day of what has to be done uniquely or with the consideration of patients as a contributor to patient-reported outcomes in contrast to all of the other NQF work.

So if you'd come up and join a table I'll get you started on your activity. There are eight tables I believe and there's plenty of seats up in the front. Don't be shy. And there's going to be an NQF staff member at each table to take notes. There will be people coming up to join you. And our
experts, please be nice to the company coming to join your table. Don't make faces at them.
Perfect. Okay, everybody has a table to sit at and a chair to sit on.

Now, this is going to require a little bit of thought which is going to mean that you have to have oxygenation in your brain. Most of you don't have any left up there, it's been draining down all day so I'm going to ask you to do one chair exercise to get some oxygen back up in your brain. All right? So I want you to put your right arm over your head, grasp your right elbow with your left hand and pull slightly and bring it back down. Now, put your left arm over your head, grasp your left elbow with your right arm, pull slightly and bring your arm back down again. That's a kind of hydraulic pump. There's oxygen back in your brain now.

Okay, if you look on the front boards in front of you -- we're not doing the hokey-pokey till tomorrow. If you look on the
projectors in front of you you'll see that there are the four major endorsement criteria for NQF. Your job in the next 25 minutes is just to brainstorm a little bit about do any of these four require special consideration for PROMs. If you want to follow along in your electronic handout it's on page 43, 44 and 45. If you have the full packet from today. And if you have a paper handout it's on the same pages, they just happen to be on paper. But mostly discuss these four. There will be no report out. At 5 o'clock we will call a quick stop and we'll see you all in the morning. Thank you very much.

(Whereupon, the foregoing matter went off the record at 4:37 p.m.)
<p>| unpredictable      | 175:20 345:19   | 255:5        | 339:12,20 |
| unwanted           | 244:3          | 308:18       | 359:3      |
| un-patient         | 279:21         | 330:15       | 359:3      |
| up-shift           | 313:1          | 353:4        | 359:3      |
| urgency            | 33:8           | 150:18       |           |
| urinary            | 150:17         | 166:4        |           |
| urologist          | 150:17         | 170:20       |           |
| urologists         | 188:5          | 216:3        |           |
| usability          | 15:16          | 236:17       |           |
| use                | 4:11           | 322:22       |           |
| useable            | 11:10          | 330:11       |           |
| usefulness         | 161:15         | 234:12       |           |
| users              | 130:12         | 240:10       |           |
| uses               | 37:6           | 337:5        |           |
| usual              | 69:6           | 286:8        |           |
| usually            | 192:20         | 299:17       |           |
| utility            | 79:14          | 327:5        |           |
| UTIs               | 32:11          |              |           |
| U.S.               | 100:22         | 227:15       |           |
| variable           | 196:8,12 286:8 | 299:17       | 304:3     |
| variability        | 99:14 188:9    | 121:1          | 123:15    |
| variance           | 220:17 221:15  | 222:30 223:8 | 305:10    |
| varicose           | 42:13          | 48:2         | 61:8      |
| varies             | 241:7          | 270:3        |           |</p>
<table>
<thead>
<tr>
<th></th>
<th>(X)</th>
<th>(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.256.294.127.9</td>
<td>1.721.6.217.222</td>
</tr>
<tr>
<td>2</td>
<td>1.154.15</td>
<td>1.154.15</td>
</tr>
<tr>
<td>3</td>
<td>1.503.15</td>
<td>1.503.15</td>
</tr>
<tr>
<td>4</td>
<td>1.364.36</td>
<td>1.364.36</td>
</tr>
<tr>
<td>5</td>
<td>1.216.120.16</td>
<td>1.221.160.16</td>
</tr>
<tr>
<td>6</td>
<td>1.432.43</td>
<td>1.432.43</td>
</tr>
<tr>
<td>7</td>
<td>1.667.66</td>
<td>1.667.66</td>
</tr>
<tr>
<td>8</td>
<td>1.898.89</td>
<td>1.898.89</td>
</tr>
<tr>
<td>9</td>
<td>1.145.14</td>
<td>1.145.14</td>
</tr>
<tr>
<td>10</td>
<td>1.290.29</td>
<td>1.290.29</td>
</tr>
<tr>
<td>11</td>
<td>1.435.43</td>
<td>1.435.43</td>
</tr>
<tr>
<td>12</td>
<td>1.671.67</td>
<td>1.671.67</td>
</tr>
<tr>
<td>13</td>
<td>1.906.90</td>
<td>1.906.90</td>
</tr>
<tr>
<td>14</td>
<td>1.111.11</td>
<td>1.111.11</td>
</tr>
<tr>
<td>15</td>
<td>1.226.22</td>
<td>1.226.22</td>
</tr>
<tr>
<td>16</td>
<td>1.341.34</td>
<td>1.341.34</td>
</tr>
<tr>
<td>17</td>
<td>1.456.45</td>
<td>1.456.45</td>
</tr>
<tr>
<td>18</td>
<td>1.571.57</td>
<td>1.571.57</td>
</tr>
<tr>
<td>19</td>
<td>1.686.68</td>
<td>1.686.68</td>
</tr>
<tr>
<td>20</td>
<td>1.796.79</td>
<td>1.796.79</td>
</tr>
<tr>
<td>21</td>
<td>1.901.90</td>
<td>1.901.90</td>
</tr>
<tr>
<td>22</td>
<td>1.101.10</td>
<td>1.101.10</td>
</tr>
<tr>
<td>23</td>
<td>1.210.21</td>
<td>1.210.21</td>
</tr>
<tr>
<td>24</td>
<td>1.319.31</td>
<td>1.319.31</td>
</tr>
<tr>
<td>25</td>
<td>1.428.42</td>
<td>1.428.42</td>
</tr>
<tr>
<td>26</td>
<td>1.537.53</td>
<td>1.537.53</td>
</tr>
<tr>
<td>27</td>
<td>1.646.46</td>
<td>1.646.46</td>
</tr>
<tr>
<td>28</td>
<td>1.755.55</td>
<td>1.755.55</td>
</tr>
<tr>
<td>29</td>
<td>1.864.64</td>
<td>1.864.64</td>
</tr>
<tr>
<td>30</td>
<td>1.973.73</td>
<td>1.973.73</td>
</tr>
</tbody>
</table>

**Notes:**
- \(X\) and \(Y\) are coordinates on a graph.
- The table represents data points for various years and corresponding values.
C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Patient Reported Outcomes Workshop 2

Before: NQF

Date: 09-11-12

Place: Washington, DC

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

[Signature]
Court Reporter
National Quality Forum
+ + + + +
Patient-Reported Outcomes
Workshop #2
+ + + + +
Wednesday
September 12, 2012

The Workshop met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Patricia Brennan and Joyce Dubow, Co-Chairs, presiding.

Present:

Patricia Brennan, PhD, University of Wisconsin-Madison, Co-Chair
Joyce Dubow, AARP, MUP, Co-Chair
Richard Bankowitz, MD, MBA, FACP, Premier Healthcare Alliance
Ethan Basch, MD, MSc, Memorial Sloan-Kettering Cancer Center
Jim Bellows, PhD, Kaiser Permanente

David Cella, PhD, Northwestern University Feinberg School of Medicine
Anne Deutsch, PhD, RN, CRRN, Brookings Institution
Stephan Fihn, MD, MPH, Veterans Health Administration
Jack Fowler, PhD, Informed Medical Decisions Foundation

Lori Frank, PhD, Patient-Centered Outcomes Research Institute
Barbara Gage, PhD, MPA, Brookings Institution
Ted Ganiats, MD, University of San Diego Health System
Kate Goodrich, MD, MHS, Centers for Medicare & Medicaid Services

Judith Hibbard, DrPH, University of Oregon
DENNIS KALDENBERG, PhD, Press Ganey Associates
IRENE KATZAN, MD, MS, Cleveland Clinic
LEWIS KAZIS, ScD, Boston University School of Health
UMA KOTAGAL, MD, Cincinnati Children's Hospital Medical Center
KEVIN LARSEN, MD, Office of the National Coordinator for HIT
KATHY LOHR, PhD, RTI
ELIZABETH MORT, MD, Massachusetts General Hospital
CHARLES MOSELEY, EdD, National Association of State Directors of Developmental Disability Services
EUGENE NELSON, DSc, MPH, The Dartmouth Institute
KENNETH OTTENBACHER, PhD, OTR, The University of Texas Medical Branch at Galveston
GREG PAWLSON, MD, MPH, FACP, BlueCross BlueShield Association
ELEANOR PERFETTO, PhD, Pfizer
COLLETTE PITZEN, RN, BSN, Minnesota Community Measurement
CHERYL POWELL, Centers for Medicare & Medicaid Services (via telephone)
DAVID RADLEY, PhD, MPH, Institute for Healthcare Improvement
TED ROONEY, RN, MPH, Maine Quality Counts
DEBRA SALIBA, MD, MPH, UCLA Borun Center/VA/RAND
MARCEL SALIVE, MD, MPH, National Institutes of Health
LAURA SMITH, PhD, Brookings Institution
BARBARA SUMMERS, PhD, RN, University of Texas-MD Anderson Cancer Center (via telephone)
KALAHN TAYLOR-CLARK, PhD, MPH, National Partnership for Women & Families
MARY TINETTI, MD, Yale New Haven Health System
PHYLLIS TORDA, MA, National Committee for Quality Assurance
JOHN WASSON, MD, Dartmouth Medical School
ROB WEECH-MALDONADO, PhD, MBA, University of Alabama-Birmingham
LINDA WILKINSON, MBA, Dartmouth Hitchcock Medical Center
ALBERT WU, MD, MPH, Johns Hopkins Health System

NQF STAFF:

KAREN ADAMS, PhD, MT
HELEN BURSTIN, MD, MPH
SHEILA CRAWFORD
EUGENE CUNNINGHAM, MS
KAREN PACE, PhD

JESSICA WEBER
EVAN WILLIAMSON
## T-A-B-L-E O-F C-O-N-T-E-N-T-S

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION TO DAY 2</td>
<td>6</td>
</tr>
<tr>
<td>Joyce Dubow</td>
<td></td>
</tr>
<tr>
<td>METHODS THAT CONTRIBUTE TO TRUST - ADDRESSING THREATS TO VALIDITY</td>
<td></td>
</tr>
<tr>
<td>Overview of NQF Endorsement</td>
<td>25</td>
</tr>
<tr>
<td>Criteria on Threats to Validity of Conclusions about Quality and</td>
<td></td>
</tr>
<tr>
<td>Differentiation between PRO &amp; PRO-PM</td>
<td></td>
</tr>
<tr>
<td>Karen Pace</td>
<td>25</td>
</tr>
<tr>
<td>Commissioned Paper Authors Tee-Up</td>
<td></td>
</tr>
<tr>
<td>Key Issues and Best Practices or Strengths/Weaknesses of Approaches</td>
<td></td>
</tr>
<tr>
<td>to Aggregating Individual-Level PRO data and specifying PRO-PMs</td>
<td></td>
</tr>
<tr>
<td>Reactor Panel:</td>
<td>33</td>
</tr>
<tr>
<td>Anne Deutsch</td>
<td></td>
</tr>
<tr>
<td>Kenneth Ottenbacher</td>
<td>48</td>
</tr>
<tr>
<td>Robert Weech-Maldonado</td>
<td>63</td>
</tr>
<tr>
<td>Expert Panel and Audience Engagement</td>
<td>73</td>
</tr>
<tr>
<td>IDENTIFICATION OF UNIQUE CONSIDERATION RELATED TO NQF ENDORSEMENT OF PRO-PMs</td>
<td>99</td>
</tr>
<tr>
<td>REVISIT PATHWAY FROM INDIVIDUAL-LEVEL PRO TO NQF-ENDORSED PRO-PM</td>
<td></td>
</tr>
<tr>
<td>Ethan Basch</td>
<td>104</td>
</tr>
<tr>
<td>Jim Bellows</td>
<td>110</td>
</tr>
<tr>
<td>Eleanor Perfetto</td>
<td>125</td>
</tr>
<tr>
<td>Expert Panel and Audience Engagement</td>
<td>133</td>
</tr>
<tr>
<td>Closing Remarks and Next Steps</td>
<td></td>
</tr>
<tr>
<td>Patti Brennan</td>
<td>183</td>
</tr>
<tr>
<td>FUTURE DIRECTIONS</td>
<td></td>
</tr>
<tr>
<td>Moderator: Patti Brennan</td>
<td>186</td>
</tr>
<tr>
<td>Expert Panel and audience Engagement</td>
<td>191</td>
</tr>
<tr>
<td>Wrap-up and adjourn</td>
<td>238</td>
</tr>
</tbody>
</table>
(8:30 a.m.)

DR. PACE:  Okay, good morning everyone. Thank you for getting with us bright and early. And Joyce Dubow is going to come to the front. I heard that yesterday some people were having trouble hearing us. So please waive your hand if we aren't speaking into the microphone so we can try and make an adjustment. And I am going to turn this over to Joyce for some introductory remarks and then we will go to our first panel.

MS. DUBOW:  Thank you, Karen and good morning everybody. You know I want to thank everybody for your participation yesterday -- is this loud enough -- because this is a big hunk of time from, I know, every busy schedules. And I thought the conversation yesterday was very, very helpful. I know it is going to be helpful to the staff and to the committees that get formed when
everybody has to get down into the business of finding measures on patient-reported outcomes. So I think the work that you all are doing here is really, really important and I want to say thank you and I want to thank the staff. They again did a remarkable job and I am about to go through some slides that Karen and Karen created. They are just amazing. So thank you, both, and to Helen, of course.

Well I know they did it but you know it is always good to have a spine that sort of keeps everything straight.

So if we could have the first slide, we are going to spend about ten minutes -- and Gene, thank you, too. I just eyeballed him.

We are going to spend about ten minutes or so just reviewing some of the highlights of yesterday's work and these are what occurred to us as being very significant but I am going to ask you to think about whether there is anything we ought to be
adding to this list.

So thinking about the main issues that we discussed, the overarching considerations, one of the first issues that jumps out is the need to assess meaningfulness and how to demonstrate evidence that stakeholders think the PROM is meaningful. We talked about the importance of getting consumers involved in this. And the brilliant Patti Brennan helped us think about this in terms of the three C's. Patti's presentation in my view was masterful yesterday. I think everybody else did, too. And if I mangle this, Patti will clarify.

She identified three C's, the conceptual which helps us identify the PRO by engaging people in a dialogue to hear from them what matters to them to define the concepts. So that helps us identify which PRO to think about.

The next phase would be the contextual, how the information is captured.
This, the contextual takes into account not only patients, which is -- who are people or individuals, remember I am using shorthand here, but for clinicians as well. For them to consider how they capture and use the information as well. That would be the PROM.

And finally, to think about this in the contextual sense -- in the consequential sense -- oh. Well, here, we have some illustrations of the contextual. How people will participate in the large social enterprise in using this information. For example, if an individual using the information selects a provider or to understand the information with respect to one's individual health situation and then finally to consider the consequential. What happens when the information is used? This is the PRO-PM, to assure that good quality is available and to understand its impact on the availability of services.

I think I mangled that Patti. Do
you want to do anything to clarify a little bit?

DR. BRENNAN: I think you translated me lovely and thank you.

The one thing I would say first of all is these are concepts that are evolving. So these are not written in stone and there is no citation for this. What was important for me to do is to stress that we begin first with what matters to the patient before we can event define a PRO, we have to figure out what really matters. Second, we have to think about both the capture and use of information relative to an individual, patient, or patient clinician engagement. And third, we have to think about the impact on practice and policy.

So that is really how I see them. And I think you did a nice job laying those out. Thank you.

MS. DUBOW: But each one of these things speaks to which level we are talking about, either the PRO, the PROM, or the PRO-
PM. Could we have the next slide, please?

So for additional considerations that we mentioned, there was a lot of emphasis on actionability. And again in the same panel, Liz helped us understand that we would think about the spectrum of actionability because actionability is a criterion, an attribute that we talked about last time as well as yesterday. This is an ongoing theme that we have acknowledged to be very important.

A hot, something, a PROM that will be highly actionable, will be subject to intervention and it is suitable and able to be demonstrated outside of a clinical trial so that it can be actually implemented in practice. Something that is highly actionable will have high credibility in the clinical community, of course, and it will have an impact on patients as well.

If it is moderate, someplace in-between, those that have low actionability
probably should be off the table because they
will not necessarily be useful to patients.
You know, it may not reflect symptoms that
matter to patients, for example. And if we
cannot demonstrate this type of actionability
to clinicians, it won't be particularly
credible either.

So this spectrum is very important
and when we start thinking about PROMs to
select, we ought to be going to the low
hanging fruit, which would be those that are
highly actionable.

Next slide, please. I'm sorry.
It's so hard for me to understand that I can't
be heard. My children would never agree with
that. Closer? I'm breathing into it. We are
going to need to -- okay.

So then we talked also about the
business case, the ROI. And you know, this is
a very pragmatic consideration and we had a
couple of people talking about this.

We heard from Larsson, Dr. Larsson
in Sweden, that they use their registries to
do CER and to demonstrate appropriateness.

There are opportunity costs. There are
benefits to using this stuff and we know that.

But there are also costs to it. The cost of
administration, the vendor-driven
administration expense. The cost of CAHPs for
example. Liz didn't talk -- Liz Goldstein
didn't talk about the cost of fielding the HOS
but obviously there is a cost to that.

And John Wasson talked about the
issue around getting consent from patients.
So we need to take these issues into account.

But it was interesting to hear Dr.
Larsson talking about the ability to make
assessments of appropriateness of care using
their registries. Of course, they have a
completely different system. The idea of
having 64 relatively compatible registries in
this country is mind-boggling but it certainly
is a system that would lend itself to some
economies.
We heard talk about phased implementation to link the systems and mechanisms for care improvement before we rush to market. I think Steve talked about this a lot and others did, too that we want to express a sense of urgency, at the same time recognizing that we may have to think about processes that relate to the outcomes that we want to achieve.

And we talked about looking at the impact of implementation of measures by doing some kind of post-market surveillance of these new measures. But you know I think that we also acknowledge that we need to do that generally with all measures, but in particular learning from implementation of these patient-reported outcome measures will be very important.

Next slide, please. So we heard a few of our colleagues talk about the iterative nature of reliability and validity and the suggestion from Laurie that the validity
testing actually take place first, you know, to get the concept down before going on to the reliability but that this was an iterative process. I can't remember whose slide it was who showed the -- Lewis' slide for me was really very helpful to see that iterative process. He showed it with arrows.

We talked, again, we kept emphasizing the importance of engaging patients and determining the face and content validity. I think that got mentioned a lot. Identifying the patient populations whose outcome you want to track. I think Jack gave us important insights here. And that if you only measure those who get the intervention, you could be penalizing those clinicians and those patients -- well probably the clinicians who are engaging not in the intervention but in watchful waiting.

And Jack I don't recall that you actually had good solutions for how we do this. Okay, well we will need to think about
Unfortunately, Jack, the mike wasn't on but Jack was talking about having records so that we could identify these patients just the way they do in Sweden. But this is going to be a challenge, I think, for us.

Is there a next slide? I can't remember. That's it.

So could we have a quick conversation about things that we neglected to put here on these slides just to remind us or any other observations?

There is somebody from -- could you use a mike, please?

MS. OKUN: The two characteristics of importance to the patient and actionability. What if something is important to the patient but not actionable? And for example, fatigue. For a long time no treatments for fatigue. Fatigue due to cancer treatment, very important to the patient. If
it is not measured, then it is not -- it doesn't become a focus and ultimately treated.

MS. DUBOW: Yes. You know, thank you for that. You know, we also had our last half hour where we had an opportunity to look at the NQF evaluation criteria to see whether we had any ideas about whether or not they had to be tweaked. I know I was in a conversation at our table where we didn't get past the importance criterion. And I haven't seen the notes from the other tables but we had a kind of lively conversation about the need to consider the audiences when thinking about importance, including figuring out how to engage consumers in the process of making a determination about importance.

Ethan?

DR. BASCH: You know, I think it is a great question. I think certainly there are contexts in which there are non-actionable pieces of information about the patient experience that may be valuable for patients
to understand but may not necessarily be appropriate for this kind of use.

But to directly address the question, there actually are some contexts in which fatigue or tiredness would be appropriate to measure. For example, for one you brought up the cancer example, one could look at inappropriate use of chemotherapy in patients who are too fatigued to baseline. And that would be one potential example. Another would be there are certain kinds of cancer where patient's fatigue does actually improve with active treatment. So I think that those are two potential examples.

MS. DUBOW: I thought the point here was things that were clearly not actionable --

DR. BASCH: Example.

MS. DUBOW: -- but still important to patients. And I think --

DR. BASCH: Absolutely.

MS. DUBOW: -- that is really
important.

DR. BASCH: Right. You are absolutely right.

MS. DUBOW: Because there may be some things that are not actionable. Phyllis?

MS. TORDA: Good morning. In going through the small group exercise, I actually was struck by how often I could think of an issue that applied actually both to the reported outcome measures and other measures as well.

So in some cases, things like these measures are more similar to other measures than different. In other cases, it seems like maybe there is an issue that is really magnified for these kinds of measures and then there may be some ways in which they are very specifically different.

But I think it behooves us to make those distinctions.

MS. DUBOW: Yes but you know -- I
think that is important. But I think Helen
and Karen and Karen, I think Karen Pace
actually said that it might be an opportunity
to tweak the criteria so that we could broaden
the applicability to embrace these measures as
well.

MS. TORDA: Yes, I think a lot of
the issues that we discussed could have broad
application to considering measures in
general.

MS. DUBOW: Right. Gene?

DR. NELSON: I think Patti Brennan
mentioned patient-defined outcomes and
patient-generated outcomes. And this idea
about me as an individual patient, I may have
certain health goals and certain health
outcomes in mind that are very important for
me. I want to go to sit in the bleachers at
the Red Sox game. That is what I am hoping to
do with my grandson. And then we have very
important general measures of health status
that most people would wish were good;
physical health, mental health, well function, et cetera. So there is this tension between individualized outcomes of importance and general outcomes of importance and how we try to understand that and make this operational in the real world, the general measures as well as individualized measures. I think that is one of the like motifs that we keep hearing and thinking about.

MS. DUBOW: Thank you. Ted, did you want to make a comment?

DR. GANIATS: Ted Ganiats from San Diego. And just two comments related to the actionability issue. I mean first of all in my mind --

I'm still Ted. Two comments on the actionability. One of them is the cost of doing this is so great that we have to be careful about it being actionable. I mean, just because it is important, the resources it requires to gather and try to act on the information is so great we want to be careful.
But more important, I think it is important for us not to be comprehensive. If I am thinking of heart failure or diabetes and all the guidelines and all the things that I could measure to make sure that good care is being provided, but we only have a couple of quality indicators and I think that the same thing has to hold here, that even though patient-reported outcomes are, in my mind, the most important outcomes, we don’t want to be comprehensive. So we don’t want to list all actionable ones. We don’t want to make sure that we do everything because we are going to be spending too much time measuring and not enough time providing the care. So these should be indicators, not comprehensive.

MS. DUBOW: I’m going to give David Cella the last word and then we are going to have to go to our next panel.

DR. CELLA: Good morning. I’m Dave Cella from Northwestern and I was at some of yesterday but not all of it. So I
apologize for not being double-booked.

    Most of you know, maybe all of you
know, that there is a teacher's strike going
on in Chicago. And the holdup on the strike
isn't actually money, at least pay raises.
That part is settled. The holdup is that
teachers don't want to be evaluated based on
standard scores. And there is a parallel here
that I think relates to providers who I think,
not all providers, but many are probably
afraid of being evaluated based on standard
scores or PRO scores. And so I really
strongly endorse the issue of the spectrum of
actionability and I was here when Liz pushed
that and I was persuaded by that that it is
important to be careful about setting up
expectations that something can be improved
when there isn't a whole lot of control in the
hands of the provider, similar to the teachers
that are complaining that they shouldn't be
judged by the quality of their work by the
standard scores and yet if the district wants
federal funds, it has to apply that directive.

    So I guess I would like to push
this group to answer the question where do you
want to jump in on assessing meaningfulness on
the conceptual issues? Because we could dance
around those issues and talk about those
issues and convince ourselves that PROs are
not perfect or they are not quite right for
this setting and, therefore, we had better
hold back or we could decide to jump in and do
it cautiously but to jump in.

    And you know, not hearing all the
discussion yesterday, I may be off target but
I hope to see a continued commitment to start
somewhere, start with high actionable areas
and jump in with measures that you know are
important to people, perhaps not proven to
everyone's satisfaction with that particular
group of patients in that particular setting.

Because we could go down that road and reject
everything every time if we go too far down
that road.
MS. DUBOW: Well before we adjourn, I will just inject a personal opinion. And that is that the horse is out of the barn and that we are moving down this road and we want to do the best possible work we can that is fair and that gives us valid and reliable information to inform decisions that reflect patient input. So with that, I think we should let the day begin.

DR. PACE: So if we could have our panel come forward, we will get started.

Okay, so we are moving on to validity part 2 and I am going to introduce the panel and then I will do a little overview of NQF evaluation criteria that relate to this area.

So for this panel we have Anne Deutsch from RTI, who is one of our commission paper authors. Next we have Ken Ottenbacher and Ken will also be addressing some of these thorny issues about validity. I'm sorry. Let' me find my place. Ken is with the
University of Texas Medical Branch at Galveston. And then we have Rob Weech-Maldonado from the University of Alabama at Birmingham who also will be addressing these issues.

So I am going to start with just a little bit about NQF criteria that relate to this and we are calling this validity part 2 because yesterday we talked about validity in general and about the actual performance score that will be used to make some inferences about quality. And today what we want to talk about additional aspects of validity but these are things that can kind of threaten validity or throw a ringer into what we are trying to do. So next slide, please.

So again, just to orient ourselves, NQF is not endorsing the individual PROM but I think we have all agreed that the PROM needs to be valid for the context and the target population it is being used in. That is definitely going to be a foundation to have
a valid performance measure. But we are
talking about using those patient or
individual PROM scores or values and trying to
use that for a particular healthcare provider,
whether it is a hospital, a physician
practice, accountable care entity so that we
would have a score on that accountable care
entity in terms of how they are performing.

Okay, next slide.

So we have talked about some of
these threats to validity and certainly we
have talked about conceptual, which can occur
at either level of the PROM or the performance
measure. We have talked about the
relationship of reliability to validity. But
some of the other specific things that we get
into this section are very much part of how
the performance measure will be defined. So
what patients end up being excluded from the
performance measure and is that appropriate?

So just again, outside of the PRO-
PM, NQF often sees performance measures that
come in with very broad general exclusions. And the question comes up are too many people being excluded that you are really not knowing what to make of the actual performance measure. Certainly we have talked about differences in patient mix for outcome measures that need to be adjusted for because patients are not randomly assigned to healthcare providers. And if we are going to use this to make inferences about quality, we need to account for those difference in patient mix that come up. Measure scores that are generated with multiple data sources or methods. So if we are going to say that you can use two different PROM instruments for the same performance measure, do we have evidence that they are really equivalent and comparable so that again we can use these in an accountability framework. And then certainly systematic
missing or incorrect data affect validity.

And we know that with these types of surveys,
we have talked about response bias, et cetera.

So all of these things, even though we have a
good idea about the performance measure, when
we actually go to implement this in the real
world, in real clinical situations, we have to
at least consider these and ask for some
assurances that these have been addressed.

Next slide. So an NQF has some
very specific criteria about each of these.

So we have very specific criteria about
exclusions. That first of all they should be
supported by the clinical evidence. So you
don't want to exclude patients unless -- if
the clinical evidence indicates that a certain
patient subgroup should be excluded, then that
obviously should be done. But and this is one
that I think we will have to grapple with
here. It is one that comes up a lot is
patient preference. And you know, some people
see that as kind of a catchall, a quick way to
check a box to exclude patients from a measure. Oh, the patient doesn't want it. Others mentioned that the provider intervention can actually affect patient decisions. And we have had lots of those discussions here about how much time you spend with the patient, how much they are informed. And so high exclusions because the patients are rejecting something may also indicate a quality problem.

So you know, this is a delicate balance here and I think where we are right now with NQF criteria is that if patient preference is specified as an exclusion, that we have to have some way of making that transparent, so that everyone is aware of differences across providers about patient preference but I am sure we can have some more discussion about that.

Next slide. So this next one is specifically about outcome measures that we need to have an evidence-based risk-adjustment.
strategy. This should be based on patient factors that influence the measured outcome but not factors related to disparities in care or the quality of care. That is what we are trying to make -- see differences in.

These should be present at the start of care, not things that develop in the middle of the care process. And we want risk models that demonstrate adequate discrimination and calibration. You know, sometimes we see risk adjustment handled through risk stratification versus a statistical risk model. Sometimes we see measures that are not risk -- outcome measures that are not risk-adjusted but again there would have to be adequate rationale and data to support that no risk adjustment is necessary.

In terms of risk factors, I think one that has come up frequently in our discussions about PROM or PRO-PM is patient baseline scores in terms of is that a risk
factor that should be considered.

Okay, next slide. Okay, another one is that computed measure scores demonstrate that methods for scoring and analysis allow for identification of statistically significant and practically or clinically meaningful differences in performance.

So generally again, NQF is endorsing performance measures for not only improvement but also for accountability applications. So if a performance measure really can't discriminate good and poor quality, and again, this relates to validity, then maybe it is not an accountability measure. However, the exception to that could be that we may have all decided yes, there is not very much discrimination and it is because in general we are doing a really poor job across multiple providers of a particular area of interest.

And then the last one in this area
is again about the multiple data sources or methods. That if a measure is going to be specified, that you can use multiple PROM instruments, then what is the demonstration that you would get comparable scores?

And I believe that is the last one or one more? Okay. All right, so from there, I am going to turn it over to Anne and then we will get our panel and your comments and questions. Thanks.

DR. DEUTSCH: Great. Can everyone hear me, including the back? Okay, great.

All right, so I will just wait for this slide to come up here. Great, thank you. So next slide.

So one of the first questions that we are going to address as part of this whole threats to validity section is are there any differences or unique considerations for risk adjustment for a PRO-PM as compared to other quality outcome performance measures? Next slide please.
So the short answer, I think, is no. I don't think there are differences. Certainly patient factors are important and those should be based on evidence that those could affect outcome. Evidence can certainly include peer reviewed research, clinical expert opinion. I would also say just in line with my presentation yesterday, that informed patients could certainly provide some very valuable insight into potential covariates also. And I am not sure to what extent that has been done with other performance measures but it certain applied to other non-PRO performance measures also.

The covariates would be very different, based on the different PRO concepts. And in the paper we give a couple of examples. And actually I would like to highlight the area that I work on is functional status. And functional status can be clinician observation as well as patient self-report and I would say the risk factors
or covariates you would consider for either
self-report functional status or the clinician
observation would probably be the same. So in
this case, I don't think there is really
differences between the PRO versus the other
kinds of performance measures. Next slide.

So in terms of examples, patient
demographic factors that are often adjusted
for are age. One of the areas of controversy
and we talk about this in the paper related to
race, ethnicity, and limited English language
proficiency. And that is a controversial area
and perhaps we can get into a conversation
about that as part of the panel discussion but
I would say the issues that are a concern for
other measures are equally a concern here.
The different SES, race/ethnicity variables
may be associated with outcomes but it may be
related to disparities and so in general those
are not adjusted for in performance measures.

Patient clinical factors that are
present at the start of care would also be
important, obviously, and typical factors included are things like diagnosis, severity of illness, comorbidities, and baseline scores.

When we put the outline together for this, it was a suggestion from somebody on the expert panel to include psychological factors like adherence, motivation, understanding, engagement, and readiness for change. Certainly those may be important for patient-reported outcomes performance measures but I want to highlight that if those are being included, then it probably would mean some additional data collection. So in addition to the PRO outcome, there may be some additional data that would need to be collected on these things and patients or persons may have questions about why this information is being collected. But typically, this is not information that is available in a medical record already. So it might be something that in addition needs to
be collected.

And I would also say that motivation is certainly something that we have talked about when we have been working on this measure for functional status because certainly patients who are more motivated might actually do better but therapists who are very good might actually motivate patients a little bit more. And so you don't want to remove that effect that a clinician may be really good at motivating their patients. And so their patients actually get better and we want to give them credit for that as part of their care. Next slide, please.

So as Karen mentioned, there is various ways to adjust for these covariates. So a very simple way is to stratify by risk groups. And out of the current performance measures that our patient-reported outcomes endorse by NQF, I don't think any of them actually do that at this point. Certainly others do but in general, you would be able to
stratify based on kind of one factor or a factor that can be split into two or three groups.

DR. PACE: Wait one second.

Helen?

DR. DEUTSCH: Oh, and the cataract. Thank you Helen. Sorry.

Regression modeling is another alternative and then a third option would be that you both stratify and then use regression modeling with your strata.

There is definitely some controversy in terms of regression modeling and so there is quite a debate about whether we are using these hierarchical generalized linear models is better than using fixed-effect regression models. And in the paper we do talk about the paper that recently came out that was commissioned by the Committee of Presidents of Statistical Societies called statistical issues in addressing hospital performance. So if anybody is really
interested in this topic, they should
definitely see that paper. Next slide.

So one of the issues that we
talked about is incomplete or missing data,
the next topic. So the question is what are
the implications of exclusions,
incomplete/missing data, and response rate
bias on validity of the performance measure
and the testing needed to assess impact on
validity?

And I mentioned yesterday that I
had a project where we actually presented some
fictitious quality data to some people in
senior centers and asked them which facility
is doing better. And I just want to bring up
one example that is pertinent related to
missing data.

So one of the measures that we
tested on -- asked people about was percent of
patients with moderate to severe pain. And
one of the seniors that I interviewed said
that she would pick the place that had the
higher percentage of patients with moderate to severe pain. And so I asked why. We asked why, regardless of their answer. And she said well I think probably they probably asked a lot more people. I think the places that had lower percentages, they probably didn't ask everybody. So I want to go to the place where they really care about it and that would be the place with the higher percentage. So I thought that was kind of an interesting answer and probably correct. All right, next slide.

So there is kind of two categories in my mind in terms of why there is missing data. So for measures that have self-administration, people may just decide they don't want to respond. For interviewer-administered measures, basically the clinician didn't ask the question. So that is -- you know, it was just not done but there is not necessarily a reason behind it other than it just wasn't done.

There are obviously more
challenging issues where the person is unable
to respond, due to cognitive limitations,
young age, language barriers, other things
like that. Next slide.

So I guess one of my thoughts is
that as part of the testing of performance
measures when they are being put forward to
NQF, during the testing that is done, pilot
testing or whatever it is being called, the
response rate for the proposed PRO-PM should
be reported as part of the testing results.

I think that is important because oftentimes
the testing is kind of an almost ideal
circumstance. And so if you start
implementing things in real life, you are
probably going to have a lower response rate.

So it would be obviously very helpful to know
if there is a low response rate in the first
place, in practice you might actually even
expect a lower percentage. So I think that is
available information.

The PRO-PM description should
describe the mode of administration. And this just, one example, one of the measures that is currently endorsed, a lot of the testing was done as a research project and research assistants, research project managers were out collecting the data, interviewing patients. But in terms of the implementation, it had been implemented basically with patient self-report. And so the staff working in the clinic have been the ones who had to decide were there cognitive limitations for the patients who couldn't be interviewed or trying to get patients to fill out the form. So the response rate can really vary. And so I think knowing what the expected mode of administration is is important both for the testing and the implementation. And obviously, they should be consistent as much as possible.

For the PRO-PM description, it should address the use of proxy responses and methods of data collection. So I think in
general that is not something that explicitly
is asked for at this point but I think going
forward that would be really helpful to have
that information.

Some of the measures I think
explicitly address missing data issues and
others don't. So I think that would also be
very helpful as part of the review.

So one of the examples I just
wanted to highlight as part of this
administration issue is the percent of
residents with moderate to severe pain, which
is the performance measure that I mentioned a
few times yesterday. So those data actually
are collected based on an interview from the
Minimum Data Set. So for those of you who are
not familiar with the Minimum Data Set, it is
a mandated instrument for skilled nursing --
well nursing homes. So skilled nursing
facilities and nursing facilities. And
it has resulted in relatively low missing
rates because it is actually part of this
mandated assessment tool.

I think the other thing I like about that instrument is that the actual script is written on the instrument. And so clinicians really know what they are supposed to ask. So Deb Saliba, who is in the back, developed the MDS-3 and so she might be able to give us some comments about that performance measure later when we have the discussion. Next.

So another issue is the use of proxies. So the question we were posed, what are the implications of using proxies on the validity of the performance measure and the testing needed to assess impact on validity. Next slide.

So in order for the use of proxy responses within a performance measure to be pooled with the other data, it would be important for obviously the proxy responses to be reasonably accurate.

Proxies have demonstrated
acceptable reliability for some PROs, like functional status where there is actually an observation of the person but proxy responses may be a little bit more challenging to include for the more subjective patient-reported outcome concepts like pain, nausea, depression symptoms. It is just really hard to be able to know what somebody is feeling in those areas accurately. But certainly functional status could be included. Next slide.

Proxy responses are reasonable to consider for child health measures where parents are proxies and the research has shown small differences in patient-child -- parent-child reports. Use of proxies may minimize missing data but it may introduce errors, obviously if they are not compatible or not easily crosswalked. So it could definitely be a threat to validity. Next slide.

Another question is what are the implications for specifying more than on PROM,
so more than one instrument or scale in a performance measure and the testing needed to assess impact on validity. Next slide.

So the use of different PROMs to measure the same construct could certainly be done. Research demonstrating the agreement of the assignment to clinically important groups. So for example, if the depression measure was being used, it would be important to know that the sensitivity specificity of the two measures were very similar so they could be crosswalked. So it is not, I think, -- let's see. Some of the work on -- well that is actually another topic.

But anyway, there would need to be agreement in the way that it is being classified. If assignment into clinically meaningful groups is not well aligned, this may introduce systematic errors for the instruments that are selected. Next slide.

So the example I want to use here is the percent of residents with moderate to
severe pain, which I mentioned before. And within this measure on the MDS, there is actually two different options for the data collection of pain. One is the numeric rating scale, which goes from zero to a hundred and then the verbal descriptor pain scale which allows the patient to describe pain as mild, moderate, severe, very severe or horrible. So within that performance measure, the clinicians and patients have the option of completing one or the other. And within the performance measures, those are basically, they were crosswalked and, again, Deb Saliba can probably address this during the discussion.

But just as an example, for people with severe pain, that is basically linked up to the ten and the very severe and horrible. So there has been research to basically link up and crosswalk those two categories. And so they are included in the currently endorsed performance measure.
I know Rob is going to talk a little bit more about that. So I will let him go on to that a little bit more.

I think that is the end. Right?

Yes, okay. Thank you.

DR. PACE: Ken?

DR. OTTENBACHER: Okay, well good morning everyone. Can you hear me okay? I want to make sure everyone can hear before I get started.

I would like to thank the NQF and the conference organizers for the opportunity to participate in the workshop and also acknowledge Anne and her colleagues at RTI for the intellectual work in doing a very comprehensive job in their paper on a difficult, complex topic.

My task today is to comment on issues associated with a litany of PRO performance measures. Specifically, I have been asked to address the following questions.

Are there differences or unique
considerations for risk adjustment of PRO performance measures and what are the complications of exclusions, incomplete and missing data and response rate bias on validity of PRO performance measures?

I will make a few general comments about validity and then address the two questions. Defining the context is an important first step in examining both reliability and validity, as we have heard from the previous speakers. Context is particularly important in considering PRO performance measures. The approach, the methods, even the conceptual frameworks may differ from one context to another.

One important challenge in determining the context for validity is variation in language. The terminology regarding validity can be confusing, even contradictory. Similar concepts can be defined using different words and, at times,
different ways by individuals from diverse disciplines.

The terms used to classify types of measurement validity include content, face, criterion, concurrent, predictive, discriminant, convergent and construct validity. The NQF Measurement Testing Task Force Report uses another term to describe validity, the correctness of measurement.

Correctness is a term used by Dr. Deutsch and her colleagues. In the NQF context, validity of a performance measure refers to the correctness of conclusions about the quality of the facility provider that can be made based on the performance score. That is, a better score reflects higher quality.

This definition of correctness is linked to other more commonly used terms such as criterion and construct validity. And these are described by Dr. Deutsch and her colleagues. The use of the term correctness illustrates the importance of clearly-defined
and operationalized language in the context of performance measures, which Dr. Deutsch and her colleagues do very nicely.

In dealing with conceptual issues such as validity, it is essential that the context and relevant definitions be made clear. If they are not, we may find ourselves in a situation similar to Alice in her famous conversation with Humpty Dumpty. "'When I use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means just what I choose it to mean -- nothing more, nothing less.' 'The question is,' said Alice, 'whether you can make words mean so many different things.' 'The question is,' said Humpty Dumpty, 'which is to be master þ that is all.'"

NQF has been a very good master in defining our terms for us related to PRO performance measures. If we are not careful in our definitions of validity and related terms, we will find ourselves in a Wonderland and, like Alice, we will be hopelessly lost in
a rabbit hole of our own construction. This is particularly true in dealing with unique challenges of risk adjustment in PRO performance measures.

One of the key lessons learned during the development of quality indicators and performance metrics through the 1990s is that appropriate risk adjustment must be content-specific.

Lisa Iezzoni who has written extensively on this topic argues that creating appropriate risk-adjustment strategies requires answering four questions. Risk for what outcome, over what time frame, for what population, and for what purpose.

A fundamental distinction regarding the purpose of risk adjustment is between risk adjustment at the individual patient level versus the facility provider level. Risk adjustment at the patient level is designed to better target interventions and resources to individual patients. In
contrast, risk adjustment at the facility provider level is used to develop quality metrics for public reporting, understanding financial incentives, and to provide benchmarks for performance comparisons.

Dr. Deutsch and her colleagues provide an excellent overview of the important issues relevant to risk adjustment for PRO performance measures. These include selecting factors for risk adjustment, data collection sources and modes, and the technical methods of generating risk adjustment models.

Selecting factors for risk adjustment presents some interesting challenges. In creating models for a PRO performance measure, typically factors are selected using previous literature, theoretical models, clinical expertise, and pilot research or other analyses showing statistically significant relationships between potential covariates and the outcome measure.
Dr. Deutsch and her colleagues note that patient factors used in risk adjustment modeling can be categorized into patient demographic factors and patient clinical factors present at the start of care. They state that informed patients could provide very valuable insights into potential covariates.

Along this line, Iezzoni suggests asking clinical experts or panels of practicing clinicians to participate in the risk-adjustment model building process. She states involving clinicians in developing risk adjusters helps achieve essential clinical credibility. The same argument could be made for soliciting input from knowledgeable patients and consumers in selecting factors to include in risk adjustment models.

Soliciting patient input to help identify factors for risk adjustment is consistent with the patient-centered approach to quality assessment. A challenge facing the
NQF and other healthcare organizations and providers is how to facilitate the evolution of patient-reported outcomes to include patient-centered outcomes. The Affordable Care Act and the creation of PCORI have highlighted the role of stakeholders, not just in the assessment of outcomes, but as partners in the decision-making process regarding the content of what should be assessed.

Examples of strategies to actively include patient input are emerging in several areas of medical care. For example, the work on activity limitation staging by Steinman and colleagues that assigns consumer values to functional daily living skills across different impairment groups and settings illustrates a systematic approach to incorporating stakeholder input into complex healthcare processes.

In the widely referenced text Risk Adjustment for Measuring Healthcare Outcomes, Iezzoni and colleagues list eight dimensions
of validity that should be considered when evaluating risk adjusted measures. They include face validity, content validity, construct validity, convergent validity, discriminate validity, criterion validity, predictive validity, and attributional validity.

Attributional validity refers to the degree to which a change in outcomes can be attributed to the care being evaluated. Iezzoni notes that in the context of using risk-adjusted measures to motivate practice changes or to monitor provider performance, attributional validity is the key dimension.

There are many issues that must be addressed in achieving attributional validity in selecting factors for risk adjustment. For example, whether to include patient characteristics such as race, ethnicity or socioeconomic variables associated with disparities. Do we really want to adjust or control factors that may potentially mask
disparities in care?

Another complex issue in establishing attributional validity is how to deal with missing data. Missing data is a common problem in clinical research. The impact and approach to dealing with missing data once again is dependent on the context. The question of how to address missing data is one that a priori has no correct response.

There are multiple approaches to addressing missing data from relatively simple, such as substitution of the most common missing value, to complex imputation procedures. Each approach has advantages and disadvantages. For example, substitution of common or expected values referred to a single imputation might appear to be a weak strategy; however research on risk adjustment using the APACHE and ICU studies suggest that single imputation can be a useful method. The assumption made with the APACHE is that unmeasured parameters are likely to be normal
common values.

Using substitution of common values to manage missing data would probably not be satisfactory in other clinical context.

The approach to dealing with missing data has important implications for creating risk adjustment models. For example, the amount of information available will depend on how missing data are managed. Some statistical software programs drop an entire case or patient record if any values are missing, referred to as list-wise deletion. This means that many cases could be eliminated during statistical modeling and data sets with a large number of variables.

Risk adjustment models using a list-wise approach to managing missing data may produce different results than a risk-adjustment model using pair-wise deletion of missing data. Pair-wise deletion only removes a specific missing value from the analysis, not the entire case.
Over the past several years, a number of sophisticated statistical methods for imputing missing values have been developed. The robustness and limitations of the newer imputation strategies are not completely understood. It is important that the methods used to manage missing data be clearly described and justified, since how missing data are handled will influence the final model.

For PRO performance measures, an important potential missing data issue is non-response rates to surveys or health questions or questionnaires. And extensive research literatures exists regarding non-response rates and a wide range of potential methods to improve rates are available.

The approach to addressing non-response bias will depend on the outcome of the performance area or the performance area being examined, the setting, the population, and a number of other patient factors.
Deutsch and colleagues acknowledge the problem of missing data as a threat to validity. They discussed the issue of non-response bias in assessing PRO performance measurement and suggest that an important first step is to adopt consistent definitions and methods for calculating response rates, cooperation rates, refusal rates and contact rates based on recommendations from the American Association for Public Opinion Research.

I would like to include with a final comment regarding risk adjustment and missing data and that is a comment for transparency. Valid PRO performance measures based on risk adjusted models must be replicable. Replication requires transparency.

In this widely cited, public knowledge the British philosopher of science John Ziman states "the ability to reproduce observations and replicate experimental
findings is at the very heart of the scientific method."

When performance measures are either mandated or de facto required, policy makers, professional organizations in the scientific community should work to ensure that details of the methods are available to the public or subject to external evaluation. One way to examine the validity of risk adjusted methods would be to compare different models by applying the same data set. Proprietary organizations, health information vendors, and others have developed and promoted risk adjustment methodologies for a range of purposes. They would argue that putting their models in the public domain would harm the ability to market their product. That concern has merit. Carefully designed policies are needed to balance private sector interests with public needs. Iezzoni and others have suggested the establishment of an external, independent, and
objective body that would operate an accreditation process and develop standards of evaluation to ensure that risk adjustment methods meet established, explicit criteria of clinical validity and scientific soundness. This is not a solution but it is a potential step to a solution. Such a task is outside the purview of the NQF. But the NQF and other agencies involved in quality measure could certainly contribute to ensuring the future transparency of risk-adjusted methods associated with PRO performance measures.

In his book The Man with a Thousand Faces, one of my favorite authors, Joseph Campbell, explores the role of myth and legend in the development of culture. He makes the observation that as an individual or a society, we can only have those adventures in life that we are ready for.

Based on the discussion at the past two NQF meetings, it is obvious that we are ready for the adventure of figuring out
how to use patient-reported outcomes to
improve the quality of the healthcare that we
all receive. Thank you.

DR. PACE: Okay, Rob?

DR. WEECH-MALDONADO: Yes, hi,
everyone. Rob Weech-Maldonado at the
University of Alabama at Birmingham. I also
would like to thank the NQF for the invitation
and also congratulate Anne Deutsch and their
colleagues for an excellent paper.

I have been asked to address two
particular issues in the paper; one of them
dealing with proxy use and the other in terms
of having multiple PROMs in developing
performance measures.

Most of my comments will probably
be centered more on CAHPS, since that is where
a lot of my experience, since that is where a
lot of my experience has been. Next slide,
please.

Just to remind you, you know in
terms of the use of proxies, very important
especially in addressing the reports of vulnerable populations, at least on the research that has been done with Medicare and patient surveys of Medicare beneficiaries, they tend, those that use proxy, they tend to have lower education, more likely to minority, and have poor physical health and slightly worse mental health.

So definitely, they have a very important role in patient surveys or patient reports. However, we also know on the other hand that they do have an effect on survey outcomes. And this may be because a proxy has different cognitive perceptual strategies in addressing the questions. There may be issues, I think this was brought yesterday, that the person serving as a proxy may be of a different age category than the intended respondent. So there is definitely differences.

Now one good thing about the studies that have been done is that the proxy
effect tend to be smaller when you have more
objective reporting items versus global
ratings. For example, in CAHPS we have the
ratings that ask the patient or the person to
rate their healthcare, rate their physician in
a zero to ten scale. So those tend to be
"more subjective."

Then you have the more objective
that would ask for specific experience. For
example, how often was it difficult for them
to get an appointment in a reasonable manner.
How often do they have to wait beyond 50
minutes beyond the appointment time? How
often did the physician explain things in a
way that was easy to understand? So those we
tend to call them reports of care. They tend
to be more objective versus, again, kind of
the more global ratings. Next slide, please.

Research also finds, especially in
the CAHPS literature that it also depends is
the proxy actually responding for the person
or is the person or the proxy assisting?
Perhaps the intended person may only need assistance in reading, completing the questions, but they still have an active role in completing the survey. And actually the CAHPS have tried to capture that data and distinguish between whether they are proxy respondent versus assisting and actually has found that those that are proxy respondents have even less positive evaluations than those that provide assistance. So that may be important in distinguishing that.

The other thing is that it also depends on who the proxy is, the relationship of the proxy to the intended person. Spouses and those that live with the person tend to provide responses that are closer to those of the intended respondent and that may make sense because this person may have more of a day-to-day interaction with the person and know exactly how they interact also with the healthcare system versus non-spouse proxy that tend to be less positive than the intended
person. Next slide.

So some of the ideas in terms of addressing proxy effects. You know, in CAHPS we do use case-mix adjustment to adjust when it is, say respondent, a proxy respondent. Beyond there is only very few variables that are used in case-mix adjustment, age, gender, education, and this is also one, especially the Medicare surveys. The other alternative is kind of called propensity score matching. It is a little bit more complex but the idea is that there is selection bias in terms of the people that actually use a proxy. So that would be a better way of actually differentiating or getting a better sense about how different the assessment of proxies are versus the intended person.

The other key thing is that we may want to emphasize more objective reports. I think Anne alluded to that in terms of some of the measures that she was talking about, especially when you are serving the population.
that you may expect to have a high proportion
of proxies, perhaps administering more by
survey by phone, even in person.

Also paying particular attention
to the health literacy. You know, sometimes
people just have problems understanding the
measures and so that is always something that
we tried to emphasize in CAHPS. And you may
even want to consider alternative measures to
capture the proxy perspective. And an example
is CAHPS with their family member survey for
the nursing homes, where they have the one for
the resident and then they have a parallel one
for the family member. And they tend to ask
very similar questions but then you get those
two different perspectives. Next slide.

Now we get into the whole issue of
multiple PROMs and how to deal with them.
Anne did an excellent summary already. I just
want to reiterate a couple of things here. We
are talking about when you have basically
substantive prompts. And a great example is
where you have two screening tools for depression; PHQ-9 and BSI, the brief symptom inventory or index, one of the two. And so you have those two alternatives and they have been found equally valid, reliable. So which one do you use or you may have some people using one versus the other. So that is what this is trying to address.

And so Anne alluded to how the MDS 3.0 team, Deb Saliba, will be able to provide more information on this, how they dealt with one particular area in terms of pain, the intensity of pain. So in the current MDS survey, they tried to provide alternatives. So some people may be better able to answer on a zero to ten scale while others it may be easier to provide more of a verbal description of pain. And I was asking her this morning, I was thinking that maybe those with more cognitive impairment may lean more towards the verbal versus the zero to ten that requires perhaps greater cognitive skill but that was
not the case. Apparently they function fully
well regardless of cognitive function.

Can I go to the next slide and
then come back to this one? I just wanted to
provide you -- so I was looking at -- this is
really not my area. But in terms of pain
management or pain intensity of pain, so this
is yet another scale, the Wong-Baker faces.

And as you can see, there is the English
version and the Spanish version. Two
interesting things about this one, that the
Spanish is not just a translation into
Spanish, they also use different faces. And
apparently this was trying to capture not only
the linguistic adaptation but also cultural
adaptation that a level of pain in one
language may have a different connotation in
terms of the face that you see and how you
relate to that type of pain.

So that is something I guess I was
trying to bring also again that cultural
linguistic differences that may sometimes may
have to be capturing these measures, as well
as -- so we have these alternative measures
that we could use and Deb was telling me that
this one was problematic because we have the
Spanish and English but it didn't necessarily
translate as well into other languages or
other cultures. So they stick more with the
numeric and the verbal.

But assuming that we have again
these different measures, going back to the
previous slide, please, basically you would
use some type of IRT methodology to create a
crosswalk between the two or more scales that
you have and developing what is the right
threshold in one scale versus the other. So
perhaps ten being extreme pain and what would
that represent in the other scale, you know,
severe/horrible. So that would require that
crosswalk so that we can actually then have
comparable measures. Next slide.

And this is the last slide.

Another way that I guess I was thinking about
the multiple PROMs is that you could actually create a composite score with some type of performance measures. And the one that came to mind as overall health status, you have the physical and mental health scores but you may want to combine those two into one overall score. So then you have to think about the weights that you provide to each of the different scales.

You know, the first thing that you would probably think about is using equal weights but that may not always be desirable. And this is where you may want to capture the values preferences of those using the measures, depending of patient versus providers. You may want to use regression-based weighting if you have like a gold standard that you can -- that some of the scales predict better, that gold standard. And one interesting thing about combining measures is that if there is some of them that have a greater standard deviation, they will
have a greater influence in the performance measure.

So you have to think about a way of standardizing those scores or having a weight that would be more for reciprocal of the standard deviation of that domain.

I just wanted to kind of bring the two options that may be possible when combining the multiple PROMs. Thank you very much.

DR. PACE: All right. Thank you again to another excellent panel. So we will stop here and open it up for questions and comments from our expert panel and audience. And Operator, you can queue up anyone on the phone line also at this point.

OPERATOR: At this time, I would like to remind everyone in order to ask a question, press * then the number one on your telephone keypad.

R. PACE: Okay, so why don't we -- we have a lot of food for thought here. Deb
Saliba, do you want to?

DR. SALIBA: Thank you. So a lot of mention was made today of the minimum data set, work that we did. Let me start by saying this is one of 450 items on the instrument. So minimum is a bit of a misnomer.

But to start with this item, it is sort of a case study in patient-reported outcomes. And I think we started with the fact that in focus groups, patients and families told us that pain was a very important construct to them. When we talk to ombudsmen that hear complaints in nursing homes, this is a big source of contention with families and residents in nursing homes. So it really started from the fact that patients and families feel that this is a very important area.

But pain is multidimensional. I mean you have heard today just about the severity items and we tested other items as well to go into the minimum data set. It had
already been identified as a fifth vital sign in a lot of healthcare systems, not just in nursing homes, but also in hospitals. In the veteran's administration, our entire healthcare facility is a fifth vital sign and everyone has asked about it or supposedly asked about it.

The challenge is that it was being asked in different ways in different organizations across different providers. There was a lot of variability in how it was being asked, as well as whether or not it was being asked systematically. Some nurses will just look at the patient and say I can tell and fill out the item. So we had to face that challenge as we were thinking about putting that into an instrument.

And we had the problem, as Anne mentioned earlier, of detection bias because those facilities that were systematically screening better for pain tended to have higher pain reports than those facilities that
are being less systematic in how they were looking for pain.

And then we also found that when they were having to report it in the old minimum data set, they were using various scales and it didn't crosswalk into the scale that they were being asked to report. So it was very problematic for them as well because they were having to translate it and didn't really have the instruments for translation. So even those facilities that were doing it well.

So we looked across the instruments that were out there and saw that the two that were the most common and seemed to have the least operational problems with the zero to ten scale, and when we say zero to ten scale, we are really talking about the visual analogue scale, where you actually show the scale anchored verbally at zero, anchored at ten, and you ask the items as -- you show the scale at the same time that you are asking
the items. And the verbal descriptor scale.

And there was a lot of debate in the pain community. There are multiple scales that various providers advocate, often the ones that they developed. So we sort of tried to work through that with stakeholders and worked through that with the pain community as well.

Cognitive came up today. And we started, to be honest, with the assumption that there was an absolute cognitive cut point below which people could not answer these items. And we were told again by stakeholders that that was wrong, that we couldn't do that for multiple reasons, that people could self-report their symptoms, even with some cognitive impairment and also that it would send a signal of disenfranchisement of an entire segment of the population that was not appropriate.

So we said okay, well this is empirical. We can test that, as opposed to
going in with assumptions. So we actually
tested everyone who was capable of responding
and looked at response rates, consistency. We
also went in and looked at reliability of
responses, asking daily for five days every
morning with one interviewer and then going
back at the end of the third day and back at
the end of the fifth day and looking at recall
ability and saw that residents were, even
people with moderate cognitive impairment,
were able to recall moderate to severe pain
that had occurred in the prior period.

So we were surprised by these
findings. They were right. We were wrong.
And it wasn't the first time in the study that
that happened.

So then we tested it. A lot of
the people in the pain community felt the
verbal descriptor scale was better for persons
with cognitive impairment. Again when we
tested in a sample of 3,000 nursing home
residents with nursing home staff asking the
items, there was no systematic difference in
whether cognitive impairment were able to
answer zero to ten versus the verbal
descriptor scale.

So we ended up sort of at a
quandary because a lot of people were using
verbal descriptor, a lot of people were using
zero to ten, and we didn't want, as much as
possible, we didn't want to change people that
were already doing something that was
appropriate just for ease of use.

And as one of the stakeholders
said to us when we sort of when to them and
said please make a choice. They said, you
know, you are at RAND. Can't you figure this
out?

So we used item response methods
to -- item response theory methods to
crosswalk the two instruments and found that
we were able to crosswalk the zero to ten
visual analogue scale to the verbal descriptor
scale for this population when staff were
So it is sort of a case study in how we ended up with a severity measure in the minimum data set that includes both types. 

Was it, you know, has it been embraced necessarily? Nursing home staff aren't particularly -- were not initially enthusiastic about asking patient reported items. So I ultimately had to take off my researcher hat as a UCLA person and put on my trainer hat and work with staff to help them understand that you can talk to your residents and you can get this information from them.

So it is really a multi-step process to look at how just this one item out of 450 ends up being part of a standardized instrument. And you have to look at it -- we have to look at it as measurement people all the way from identifying the importance, testing its performance in a very specific population, and then how it is actually doing to be used by the providers when they have it.
So it has been referred to a great deal today. I mean when you just see it as first and you think, God they are doing it with cognitively impaired people and what are they thinking. But we really went through an empirical systematic process to decide on that item.

DR. PACE: Okay, thank you. Gene?

DR. NELSON: The discussion about risk adjustment factors and demographic or clinical and psychosocial, a tricky area of course, but one thing that created red flags for me was what was labeled as psychosocial but then the examples were more things having to do with activation or engagement or motivation, which are often mutable, as you said, by the care before or after and so sometimes if viewed as at least a proximate outcome for good outcomes, health engagement being generally valued by consumers or patients.

So my sense is this is not a good
thing to consider for risk adjustment. Others might wish to comment on this; John Wasson or Judith Hibbard, or others that are actually experts in risk adjustment methods.

DR. PACE: Lewis?

DR. KAZIS: Can you hear me?

DR. PACE: Yes.

DR. KAZIS: I had a couple of points. The first was, in terms of the --

DR. PACE: Now we are kind of losing you.

DR. KAZIS: Can you hear me now?

DR. PACE: Yes.

DR. KAZIS: On the issue of imputation of missing values, we have gone through some extensive work for CMS related to the HOS study, which you heard about yesterday, and this includes the modified regression estimator which in effect would allow us, based upon our outcomes, using in the past SF-36 and now the BR-12, we are able to capture 90 percent of the missing values on
the basis of this particular approach, which was developed by Bill Rogers and is available in the public domain and we have extensive documentation and reports that have been done that are available on the HOS website with publications that we have subsequently shared -- basically that have been published. So that is available and this approach tends to be quite useful, does not require high computer power, and might in fact be an approach that folks want to consider, whatever the outcomes, that are used that are patient-centered.

The separate point is in terms of the risk adjusters. We have used what is called a cascading approach where we have developed a number of different models, with a minimal set of variables that might be required. So one starts with all of the risk adjusters that are in your model and then you cascade across different models until you get to a minimal set of variables that would be
required in your risk adjustment. So you could literally start off with, for example, 12 variables that become your risk adjusters, and then in the minimal data set, it would be as few as say three. That would still allow us to risk adjust adequately for a particular case.

So those algorithms have been worked out and are also available through this HOS website.

The last point has to do with in the report there was a mention made in the paper that we have all reviewed of an initial covariate adjustment using the baseline value of your outcome. And this is quite controversial. And in fact there is a correlated error problem that unless the design of the study is randomized, a randomized clinical trial design, if you are dealing with an observational naturalistic data set, it can become quite problematic with that initial covariate adjustment. And I
think that that needs to be considered.

DR. PACE: Okay, thank you.

Collette.

MS. PITZEN: This is Collette from Minnesota Community Measurement. I had a couple comments. One was on the whole idea of missing data. And I wanted to make the point that, and I will use the PHQ-9, our favorite tool, as an example. We don't have much of a problem with patients not completing that simple instrument. So less of an issue of the actual missing items within that tool. But a bigger threat to the validity of this measure is the longitudinal measure over time and the ability to connect with those patients at the measurement points in care.

For example, our current follow-up rate with these patients at six months is about 25 percent and we keep working on that to make it better but that is the reality of implementing these tools in clinical practice.

The second point that I just
wanted to add on to Lewis's risk adjustment.

When we are developing new measures, we are working with our expert measure development workgroups. And again, we are picking variables based on evidence and literature and expert opinion that the group thinks would be important variables for risk adjustment and then we start collecting that data and running those variables through the models to see which, indeed are good risk adjustment variables.

Thanks.

DR. PACE: Okay. Ethan.

DR. BASCH: Thanks, just a couple of comments. This was a great panel. I have three comments.

The first, regarding bias. You know, I do agree with the comments about response bias or what you called detection bias. I think it is really an essential adjustment that needs to be made in this kind of work and one that the NHS PROMs initiative
has a fair amount of experience with and I
mentioned that before.

And I think this really touches on
the issue of optimizing response rates for
minimizing missing data. And in work that we
have done in cancer populations, we have found
that the principle reason for missingness of
data in nonclinical trial populations is that
people were too busy or forgot. And then when
patients become very ill, they are unable to
complete, unable or unwilling to complete.
Although, patients have to really be quite ill
not to complete. In general, we found within
a month of death in cancer populations.

What we found in general is that
there are strategies that can optimize missing
data, which obviously is preferable to using
imputation in post-production. One way to
optimize response rates is by making
completion of questionnaires a standard part
of operations, rather than sort of a voluntary
or a carve out in a sub-population. We found
that response rates are boosted about 30 percent when it becomes mandatory. And I think that is probably reflected in both the Swedish and the English experiences.

The other strategy is using backup data collection methods. In particular if there is a mailed questionnaire, an electronic questionnaire to have actually have a human person call people who don't complete their questionnaires and that boosts response rates by about ten percent.

And in general what we found is that it really only takes one call. And after the first time people feel, oh you know, somebody is watching, so I had better complete my questionnaires in the future or I am going to be bugged about it, which is sort of interesting.

Regarding proxy, I actually had a question for Rob, which is whether one would advocate for proxy reporting as a substitution at a point where there is lost data or if you
advocate for longitudinal collection of both,
you know, patient and proxy reports, that you
can impute based on the trajectory when you
have a population at risk, because that would
substantially increase the burden, although my
understanding -- it is not my area -- is that
imputation is improved when you actually have
the trajectory of the proxy reports.

My other question for Rob about
multiple PROMs is about multiplicity problems,
which we think about in the registration
context all the time but that we really didn't
touch on here.

DR. PACE: Rob, do you want to
respond to that?

DR. WEECH-MALDONADO: Yes. Well
basically for the first question that Ethan
asked about if I am understanding Ethan, you
are saying whether it is recommended more for
at the first point or as well as over time.
Because if you use it over time, then you may
have also kind of perceptions right in the
follow-up surveys and all that.

I think to the extent that you are able to somehow take that into the adjustment, you know, the case mix, it should be okay, especially if you are thinking about, you know, when we are thinking more about performance measures. You know, as long as you are able to case-mix that over time, I think it would be definitely preferable to have that proxy in the follow-up, rather than having missing data.

So if it definitely improves your response rate over time, it definitely would be preferable.

But in terms of the CAHPS surveys, they do discourage proxy use. They are really more limited to the high-risk populations like Medicare in populations like that. So it is not universally recommended but when you have again an at-risk population. I don't know if that answers your question but yes, it definitely would depend on the population.
DR. BASCH: And the other was multiplicity issue.

DR. WEECH-MALDONADO: The multiplicity of --

DR. BASCH: If there are multiple PROMs.

DR. WEECH-MALDONADO: If there are multiple PROMs and --

DR. BASCH: That are being used to individually score.

DR. WEECH-MALDONADO: Oh, that are individually scored, yes. Yes, so that is where figuring out the right combination of those PROMs, in terms of the weights that you provide, it is where it gets to be critical and that would have to be subject to appropriate testing to determine what those weights would be in order to aggregate them, ultimately. Because you want to aggregate them, depending on what the performance measure, the idea would be to be able to aggregate them.
The first case was more about using one or the other, which is it is a little bit easier but if you are actually going to create a composite of them, then the main issue is how to address the waiting that you are going to be doing.

DR. PACE: Before we take any more in the room, Operator, are there any questions from those on the phone?

OPERATOR: At this time, there are no questions.

DR. PACE: Okay and what about in our audience? Evan will you -- and we are just about out -- we are a little bit over time but we will take a few more questions.

MS. MASTANDUNO: Melanie Mastanduno, Dartmouth Institute.

I just want to add to what Ethan was saying. While I do not have the statistical basis for this statement, when providers -- a boost in response rate will also come when providers and patients are
talking about the results of their survey and
if the provider goes as far as to say this is
important because I now understand you better.
Thank you for filling out this survey. And we
do see that across 14 different specialties
where Dartmouth is collecting these data now.

DR. PACE: Okay. Evan, behind
you.

DR. ROSS: Hi, I'm Clarke Ross.
I'm a new member of the MAP Workgroup on
Persons Duly Eligible for Medicare and
Medicaid and I represent the Consortium for
Citizens with Disabilities, which is a
national coalition a cross-disability
organization which has most of these
organizations as one of those members.

And I raised this issue at the
July meeting. I wanted to just get on the
radar a supplemental approach to documenting
the consumer person, sometimes called patient,
experience with the system. And this is
approach that is used in four states and it
has been used in two of those states for over a decade, paid for by health plans mandated by state Medicaid programs for persons with mental illness. And these are third-party independent consumer- and family-operated monitoring teams and their whole approach is to interview consumers and their families where they are, not on the site of delivery but where they are; where they live, where they participate in the community, where they work sometimes.

And so what I am asking is that the National Quality Forum recognize and acknowledge that there are supplemental methods used around the country to supplement the core of what you have been talking about. And again, these are paid for by state Medicaid organizations. In Massachusetts, the managed mental health company is mandated out of its contract with the state to locate and finance an independent consumer and family monitoring team. These are not ombudsman
programs. These are monitors. That is their sole purpose.

And so I just wanted to, as you focused on where these things are delivered and who does it and how to measure it, all fundamental questions, there are alternative -- not alternative -- supplemental efforts going on and I think it is important for the quality forum merely to recognize that there are some of these supplemental strategies going on, used today to ensure accountability of health plans to their enrollees. Thank you.

DR. PACE: Okay, thank you. Can we do one more question or one more comment? Kathy.

DR. LOHR: Two quick questions, don't necessarily need an answer or discussion. On the concerns about proxies, we have heard lots of really interesting options but the point that I didn't hear is whether we know much about whether there are systematic
directionality issues that can be dealt with
or is that just in its infancy in terms of
understanding which kinds of proxies will
overestimate or underestimate and whether that
can be taken into account.

On the multiple PROMs, I think
that is a very complicated issue for NQF and
for Anne and everybody to deal with in the
paper. But on one particular thing, I
wondered about whether we need to know more
about the -- and this was from Roberto's
slides -- the pluses and minuses of doing
composite scores. Because that just seems to
me to make life even more difficult and
whether that is a direction for NQF to go in
moving on towards performance measures and so
forth. Is that sort of a bridge too far to
try to do that, rather than just keep it more
simple with individual measures and cope with
all the measurement problems there?

DR. PACE: You're right.

Composite measures have a whole other set of
methodological issues and we do have composite performance measures, some, but it does increase complexity and it may not be the place to start. But again, I think it is something to consider in terms of addressing these issues.

Okay, we are going to take the next 20 or 25 minutes until break time to again, at our tables -- yes, and Patti you want to -- okay. So we will ask the people in the back to come up front. Again, as we did yesterday evening. And you can go to the table -- I guess if you want to rejoin the group you were with, that might be useful.

We are going to kind of continue on -- and Jessica you may want to put up that slide again -- to look at the NQF criteria and are there unique considerations when NQF starts looking at these PRO-PMs in terms of our criteria.

We have talked about and you have seen that our criteria map to a lot of the
issues we have talked about but we want to really have you pull out any unique considerations where we may need to tweak or think about unique ways of applying these criteria. And this will also apply to our next panel after the break where we will be talking about the pathway from PRO to an endorsed measure where we also kind of identify the NQF criteria along the way.

So we will continue with that. And Patti, do you want to add anything?

Yes, the audience in the back, please feel free to come up and join a table, even if you weren't at a table last night. We welcome your participation.

And we will go until 10:30 and then we will take a break and then resume.

(Whereupon, the foregoing proceeding went off the record at 10:09 a.m. and went back on the record at 10:52 a.m.)

DR. ADAMS: Okay, may I ask
everyone to rejoin their table? And if we
could queue up the slides.

So for this panel, we are going to
try to tie it all together. And so not only
will it incorporate our thinking from
yesterday and today, but really it takes us
back even to the first workshop, where we
started to look at characteristics for PROMs
and what would make them most ready for
primetime for performance measurement. And we
had all agreed that we thought that a visual
or a flow would be helpful, not only to inform
the field but I think to kind of collect our
thinking.

So this panel today is going to,
based on the input that we received at the
first workshop as well as today and with our
expert panel and our panel prep, so lots of
thinking went into this flow diagram. And of
course we are going to be further refining it
based on your feedback. But we are going to
try to tie it all together right now and we
would like your input in that regard.

In regards to our reactors, we have Ethan Basch with us, thank you. And he is going to cover what we are calling steps one through four on the diagram, which where we are going to kind of think about more how we would frame this up-front with the conceptual basis.

We have Jim Bellows and he is going to walk us through the process performance measure steps. And Eleanor Perfetto and particularly with experience with Pfizer around the outcomes element of this.

But I am just going to kind of give us a high-level overview before we take a dive into that.

I am really excited about but anxious to get your input on this flow because I think it will be very useful, not only for a performance measure committee that is going to be looking at endorsement. But I think as we start to play this out with the field and
we talked about yesterday perhaps, you know, would we have a few use cases? Would we put a few things through this to kind of test what we have been saying? And I thought that is really brilliant advice. So not to put a lot of pressure on our pathway panel but I just think that I really appreciate the effort that has gone into putting this now into a diagram, which in some ways you could get overwhelmed with all the things we discussed but in other ways, we are starting to see a path along this journey. So I am really taking a glass half full versus half empty because sometimes the methodological considerations can seem a bit daunting.

Okay. So if we look at this pathway, I mentioned the first part, which is going to be the conceptual considerations and I think that this, when we talked about what Joyce had recapped this morning, it relates back to meaningfulness. And certainly we want to identify PRO-based performance measurements
that are meaningful and important up-front to
patients and their families and caregivers,
and to all end users.

And I wanted to thank Ethan
because when we did our prep call, you really
-- you know, we dived right into the process
and you really offered us the opportunity to
step back and say hey, let's think about what
some of the prequel or preamble to this. So
I did want to acknowledge that.

And if we could have the next
slide. Now on this slide we have lots of
discussion around implementation or state of
readiness and should we first think about how
this applied into practice and should we go a
process measure route. We weren't advocating
for maybe just checking a box, but would we
look at this from the process and what we have
learned, I think particularly when we talked
about actionability and how this would
influence.

And Jim, I am going to call that
the Jim Bellows box. I know I did that on our call. But number five, particularly from the first workshop, particularly struck -- and we got that from our different levels of high medium and low from Liz Mort yesterday around actionability and the process there.

And then we get to the next slide, which is what we are always aiming for, ultimately is around the outcomes. And I think that Steve Fihn, I know he is not with us today, but he spoke that maybe there are some lower hanging fruit. Maybe there are some are ready for outcomes if we think about things like HIP where got these urology example yesterday. Maybe we can pass go and go to outcomes. But ultimately that would be where we want to be.

So with that bit of an overview, so we are trying to put together that picture of the pathway and we are going to break down each individual part and our reactors are going to solicit your feedback there.
So I'm going to turn it over to you, Ethan. Thank you -- yes.

Oh yes, thank you, Karen.

So in your package you received this. We also put these on the table. Here is the color diagrammatic what we are going to be talking about. We also, as a supplement, put the NQF criteria and how this maps to it, and also some of the characteristics we talked about yesterday. So this is kind of your cheat sheet, so to speak, as we go through this. Thank you, Karen.

Ethan?

DR. BASCH: Great, thanks. So before we start, I think we have a special slide to put up. Do you guys have that?

(Pause.)

DR. BASCH: Do the presentation first and then the special slide. Okay, that's fine.

So if you all have the -- I'm going to borrow this -- the diagram from your
packet, each of us is going to go to --

thanks, Vanna.

(Laughter.)

DR. BASCH: So each of us is going
to go through several of the steps. So I was
asked to go through the first four steps.

So as you can see, these initial
four steps really are what Karen called the
prequel. So first identifying outcomes that
are important and meaningful to the target
population; determining of patient reporting
of the outcome of interest is appropriate;
identifying existing measures to evaluate the
outcome of interest; and then number four,
applying characteristics identified at the
last workshop and in this workshop. A lot of
them in the technical paper that was
developed. And this includes looking at
measurement properties. And I am just going
to march through these. Could I have the next
slide.

So if one thinks about this, I
have tried to emphasize the importance of patient engagement or consumer engagement, person engagement and I would argue that patient engagement begins with box number one. This is really the first time point at which one touches the population of interest. Because without doing probably a qualitative and quantitative research -- you can go to the next slide.

So without doing qualitative and quantitative work, which may involve surveys, focus groups, key informant interviews or longitudinal observational research or cross sectional research, it is very difficult to know if those outcomes of interest are actually important in the target population. Go to the next slide.

Now I would argue that box number four is a potential second time point for patient engagement. And you know in evaluating the characteristics of a PROM, this would usually begin with a literature review,
a landscape overview. One could take a look at the measurement properties, whether the measure has been evaluated in the comparable population. But as was pointed out yesterday, very likely a second step of patient engagement would occur here. So you can go to the next slide.

And this would involve qualitative work to establish what I call content validity. Other people call it face validity or other kinds of -- you know, patient understanding in the target group of what is being measured. And this would include a diversity of patients within the target population, to make sure that people who may not have been a part of patient interaction during development of the tool are now a part of it. This may include patients whose literacy is different from the initial population, their performance status is different. Their linguistic orientation is different, cultural orientation and so on.
You can go to the next slide.

So I thought I would try to make this real with a hypothetical example. So looking at men with prostate cancer one can think about two different populations; say a post-prostatectomy population. Let's say hypothetically we went out to a group of men who had just had their prostates removed and asked them about what was most important to them. This is all hypothetical because I haven't done this. Right? So I'm just saying if we went out and talked to patients based on actually existing literature, these are the two most important domains to patients, although admittedly the target population may differ, urinary symptoms and sexual function. If you look at a metastatic population, actually pain and tiredness or what we classically have called fatigue are the most important domains. You can go to the next slide.

So now we have identified the
outcomes of interest. You know, we conclude that probably patient self-report is the most appropriate approach. That is number two.

So number three, we want to identify PROMs. We go out. We review the literature for post-prostatectomy. We think about the IPSS. And for number two, pain and tiredness, we think maybe we will use some PROMIS items.

So now we get to number four, which is we want to evaluate the measurement properties as applicable to our target population. You can go to the next slide.

And so we do our literature review. We find both were developed with patient input. Both have been evaluated in similar populations, looking at construct validity, reliability, sensitivity to change and so on. And we ask ourselves a question, do we now need to conduct additional qualitative research. I'll stop there.

And now we have a special slide.
No, we don't? Okay.

(Pause.)

DR. BASCH: What do you guys think, are we going to get it? You don't think so. Okay, well we were going to have a birthday cake on the slide because Laurie Burke whispered in my ear that somebody has a special day today. It's Eleanor's birthday.

(Applause.)

DR. BASCH: So we are all going to sing.

(Singing of "Happy Birthday.")

(Applause.)

DR. ADAMS: Not all methodologists are serious. Right?

So Jim, we will pass that on to you now.

DR. BELLOWS: Okay. So first of all, thank you so much for naming a box after me, box 5. I am sure there is people in this distinguished room that have items and scales and instruments named after them, which I will
never have, but at least for one day I will be
happy to have a box.

So and I want to come to a couple
of things. One of the questions is, does it
make sense to start with process measures
first before we go on to outcome measures? Or
I think the way the question is really written
is should process measures proceed. And my
answer to that is going to be with respect to
an experience we are having inside Kaiser
Permanente with the PHQ-9. Both the process
measure, namely did you do a baseline testing
and did you a follow-up at approximately six
months, and the outcome measure of whether you
treated to remission.

And so my answer to should the
process measure proceed the outcome measure is
no. It is could proceed. And it is based on
it is informed by our experience.

In treatment of depression, it
really could involve the entire system. We
are not talking one specialty. We are talking
about everybody in primary care, presumably
many specialties. So for us in Kaiser Permanente, that means getting 15,000
physicians to change their behavior and to change the data stream that they are tying into, and changing the behavior of 15,000 physicians in any direction is a gigantic change management challenge. And I guess one thing that I have learned about change management is that there isn't a simple formula.

And within our Kaiser Permanente regions, different ones have made different choices. There are some who have elected to implement -- we have made it clear across the system that our intent by the end of 2013 is to be driving performance on both measures, both the process of did you do initial and follow-up testing and on the outcome are we treating people to remission. But our individual regions, roughly states, have taken different pathways and some have elected, as
a matter of their change management, to focus on the process measure first, and some are focusing on both. And it really represents different hypotheses about how individuals change. So asking people to do the process measure first really reflects a belief in the sustainability of provider-driven change.

Namely, if you give a bunch of physicians the right information about whether their patients are actually getting better by requiring them to do the initial and follow-up testing, then smart and well-meaning people will begin to ask a bunch of questions like hey, are we doing the right thing? What is it we are actually doing? What did I do for that patient? And to set in motion a change of improvement activities driven from the ground up by providers.

There is, of course, another theory and hypothesis that accountability is required and that putting in not only the performance -- not only the process measure
but the outcome measure is actually going to
drive people to final results. I suspect many
people in this room are familiar with the
paper by Brent James and Molly Coye and others
about two pathways to improvement and
accountability and local process improvement
and their mutual reinforcement. And
ultimately I think we have to get to both but
to me it would make sense for NQF, as these
things come along, to offer both a process and
an outcome measure for people to be able to
make that choice about how their change
management will proceed.

There is a question about whether
the boxes are in the right order. And
certainly the 6, 7, 8, 9 boxes make lots of
sense to me in their placement with respect to
other boxes. I would like to suggest that
there might be two other boxes that might
either or semi -- well, one might be part of
a box and another is another box somewhere
else that might be missing.
One is that box five basically asks us to use the measure in clinical practice or, as Ethan put it so well, pilot it. And this is pilot it in representative practice. Obviously, many of these measures have been piloted. That is where the information about reliability and validity comes from. This is about piloting it in a representative setting to find out stuff about the actionability, about the meaningfulness of patients and clinicians and really, ultimately, about the value of the measures. Is asking this work the perturbation of the system it causes?

And a box that I think is missing, which is specify how you think this measure is going to be used. And we have talked about this in a number of settings. I think sometimes it has come under the label of fitness for purpose. And five is really about piloting fitness for purpose but a person can't pilot fitness for purpose without a
specification of what the purpose is. Is this a measure that we think is going to be introduced into clinical practice, namely, asked in the exam room or in a clinical workflow versus a measure that is being used like in a Swedish example where people are getting queried at home as a part of a national system? They have really different implications. So it seems to me that a person couldn't do box 5 of piloting and testing for not only implementability but actionability and all those other good things in a representative without really specifying what that representative setting is and how we envision that the measure will be used.

And I guess I think what I need to do is call attention to my own belief and bias that these measures, as they come through NQF will have to be quite a bit more specific about what is the setting and make distinctions between measures that are going to be used in a clinical workflow versus those
that are going to be used in other kinds of contexts.

Is that little bell my time bell?

No? Okay.

And the other box that I think might be absent from this diagram is one that has come up a couple of times, which is the post-market surveillance box. And how is it that once new measures are out of the gate we know when they are existing real life whether they are working or not.

And I don't know that any of us really know what the post-market surveillance could look like. I imagine from the NQF perspective it has something to do with what comes up with measurement, measure maintenance because there is a process where how measures are actually functioning in the real world could exist.

What I do know is that when I go to my car dealership and they give me their thing about how I have to check box five or
their kid left to drop out of college and all that stuff, I dearly wish that whoever created that measure was doing their own post-market surveillance in that somebody who is not part of that process was asking me occasionally is that thing working. And I would be able to tell them, no. Your measure has been corrupted. It is full of garbage people are distorting it. And the fact that it has been turned from what was probably a really great, wonderful, well-intentioned improvement measure has been bastardized in the process of turning it into an accountability measure. I really wish somebody was checking up on that.

So, I guess my recommendation to NQF is that the measure maintenance process includes not only stuff about what have we learned additionally about validity or reliability and about harmonization, but really goes out and collects data about how this is being used in practice and is it creating unintended consequences and so forth,
but to put some rigor into that part of the process.

There is also a question about whether the criteria are right. So the criteria, that is the stuff on this accompanying second page that Karen just called our attention to. And I think with respect to the process measures, many of the criteria seem exactly right and are really terrific.

There is perhaps a little bit more of an emphasis on how the data are collected and collecting the outcome data as well as the process data that I am not sure quite belongs there. But also I would like to just suggest a couple of other possible criteria that might go in here.

One has to do with systemness. and there is a way in which PHQ-9 process measure, for example, asks for what proportion of patients we have collected, the baseline and follow-up PHQ-9. There is a whole bunch
of other stuff that is implied and really important about how we are actually embedding that into a system, how we are embedding that into a clinical workflow in returning the information back to physicians at the right time and place and how we are embedding that into a performance improvement system to learn from and improve on that measure over time, all of which is outside of the answer to the measure itself.

So I would love to see some component, if we are going to have process measures, not only about the performance on the actual measures, as specified, which is, of course, important, but the sort of reporting out on the context and what is the system in which that information is being used and reported.

So there is almost a level about disclosing as another kind of metadata, how is this being used. It is a really important accompaniment about what the actual rate is
and I think we haven't often paid much
attention to that in our external reporting.
And then a second aspect that I
would love to see and I know it is part of the
NQF process about harmonization. So first of
all, bless NQF for the additional work it has
done on harmonization. I know this is coming
up in many contexts. But I believe in this
PRO context, it is especially important for us
to get harmonization right so that we don't
get a bunch of measures that have come in
especially from the research environment and
from a bunch of disparate and wonderful
measure developers who have each taken their
own approaches but where collectively we could
have a lot of noise and a lot of different
kinds of scales and different kinds of look
and feel. I don't think that is the process
we want. I think we want something that, over
time, looks more like a design process and
looks more like repeated use of the same
response scales and looks more like the same
kind of lead-in in introduction or the same
number of items, so patients know what to
expect and they don't say yes, thinking that
they are going to get four items and instead
they get 60 or whatever. So some kind of
uniform look and feel. And I think the way
that might come in is in the harmonization
process where if someone wants to introduce a
new measure that is measuring the same concept
but has a different kind of a scale or a
different number of items, that there be a
really, really high bar about what is going to
be allowed, just to avoid that possibility we
have talked about of having a bunch of noise
in the system.

And I think I will leave it at
that. There is many other aspects that I
think if we pay attention, not only to the
individual measures but to the system this is
going into. Actually I am tempted to go one
step further.

There is many ways in which we are
at risk of talking to ourselves here in this room. We have got a bunch of measurement experts and measurement nerds and we know a bunch about how to use item response theory to do crosswalks between measures and all that stuff, which is all stupendous but we also know from the world of improvement and improvement science that having multiple ways of doing things makes improvement really, really, really hard. So just to take a screwy little example, we know there is many ways of measuring weight and they are all perfectly cross-walkable. We know exactly how to relate pounds to kilograms to stones and however many other things are used in other parts of the world. But in many parts of improvement and maybe nowhere better than anesthesiology, people have taught us the importance of using the same measures all the time so a person doesn't have to translate. So we don't have to give people the additional cognitive burden of understanding how this kind of pressure
measurement relates to that kind of pressure measurement.

So I would suggest that in addition to, if we want to consider multiple measures, in addition to saying whether we, as measurement nerds, can crosswalk them and relate them that we also subject this to empirical testing with real providers. And if you put real providers in an environmental with multiple measures that overlap, whether you give them a crosswalking opportunity or not, whether they really understand when there is multiple different measures, whether they really are actionable or whether they create confusion.

And you mentioned the word multiplicity and I think you were using it in the measurement sense, Ethan, but I think there is also the multiplicity problem in the cognitive burden sense for providers. If there is multiple different instruments going on, can people really understand it and act on
it in the same way as if we had a more coherent and simpler, cleaner system?

DR. PERFETTO: Thanks. After the last meeting, I walked up to Karen and I gave her this diagram that I had been drawing in the back of the room. And she looked at me like I can't read this. And I mean, this has come such a long way from a scribble. And I am pointing that out because I am going to make some suggestions and I don't really have answers for some of the questions that I am going to raise. But I figure if she could do this with the doodle, then she can resolve any of these things that I am going to raise. So thank you, Karen, because I am very visual and I have to have pictures to look at.

On the orange part of the outcome performance measure part, I want to get back to the point, some of the points that Patti made. And they are going to overlap a little bit with what Jim was talking about.

The first part is on contextual.
And we have had a lot of discussion over the last couple of days about context of use. And so I think what happens is when we get to the precise specification, when we get to that specifications document, I think one of the things that we might need to see is some more detail in what the specifications document will have to have in it, in order to be able to capture context of use in the way that we have been talking about it because I think we are going to need a little bit more on the setting, a little bit more on the patient population, all the kinds of things that you were talking about in that testing phase. When we start to operationalize, we have got to make sure that the population is the right one and the specifications can get at that.

So the one example that I can think of where in a process sense where this isn't quite as detailed is a measure that exists now where if you are a patient who is over the age of 65 and you are hospitalized
for any reason, you are supposed to be asked
whether or not you had a pneumonia vaccination
in your lifetime and if you your answer is no,
you are supposed to get a pneumonia
vaccination. So the specifications there are
pretty straightforward and pretty simple.

But if you are getting to the
point where you are going to be asking
hospitalized patients a PRO that has to do
with their pain or their function, then the
specifications document is probably going to
have to be more detailed than you might think
about for something that is a process measure.

And the next step is in the
conceptual aspect that Patti was talking
about. And when I was thinking about this, I
was thinking about in the area of what is
important to measure and report. And I think
this actually comes out in a couple of ways.
And it is getting to that. What is really the
rationale for why we wanted to measure this
particular thing and thinking about that early
on, and trying to figure out if when we are
describing in this endorsement process what
our rationale was for why we think we need to
do this and why it was selected. That we
start to think through just kind of what
Laurie and I might talk about in terms of the
benefit risk sense, what are we going to get
out of this that makes it worth doing that
helps us answer the question that we started
off to answer or try to resolve the problem
that we started off trying to resolve.

And that made me think, along with
some discussions that we had at our table,
that maybe we had something that was missing
on the diagram. Right up front where we have
kind of gone to this outcomes that are
important and meaningful to the target
population. But we didn't really start off by
specifying what the problem was that we were
trying to solve. What is the quality or
accountability problem that we are targeting?

And then get to that point of what is
important to patients and can those things help us solve this problem in terms of the quality issue or the accountability issue that we are trying to gather the data to resolve.

And that if we can do that and articulate that clearly as part of this importance of the measure and the rationale why we decided to go this route, then we actually solve that problem of having that blank up-front because then we can make sure that these two things are connected when this information is being put together.

I think that also has a spillover on the consequential that Patti was talking about. And it also overlaps with some of the things that Jim was talking about in terms of what is our post-marketing surveillance. We kind of have another missing box at the bottom for that post-marketing surveillance consequential piece. And that ties into this issue of what is the problem we are trying to resolve. If we want to have a feedback
mechanism where what we have done at the end
starts to feed back into the process, then we
could feed back into well are we running into
some methodological issues? Maybe we are
running into a logistical problem in terms of
implementing or a validity problem. That
would feed back into different parts of that
methodological piece. But I think what we
really want to know in terms of that missing
box is if we have implemented this, has it
really helped us solve what this problem was
up in the front that we defined, the quality
problem or the accountability problem.

So I think that helps us by
connecting those two pieces, the up-front
piece and that downstream piece and having
that feedback mechanism be able to go through
and in that feedback look that can be
strengthened.

One of the things that Laurie is
famous for saying is we are not going to just
measure for measurement sake and use these
PROs for measurement sake. And so I think what I am saying is this probably applies to not just these PRO outcome measures but it probably jumps forward more because this can be burdensome. It can be costly. It can cause confusion when there are lots of different measures. And so it is probably one that needs to be thought through more clearly in this instance because of the complexity of going about gathering this data and making the choices that we would want to make when we are gathering these data.

So those are the three pieces that I thought tied in well with where Patti was going but kind of help round out this picture of being able to say we have got a diagram that not just walks us through endorsement but gets us through in trying to resolve the issue of when we have got this endorsed measure, what is the problem it is helping us solve and is it really doing that if we start to use it.

DR. ADAMS: First I would like to
thank our reactor panel for lots of thought-provoking questions. I think Eleanor, you went back to the beginning, as Ethan reminded u, and it is like defining the problem. What is the magnitude of the quality issue. And then Jim, lots of practical guidance to NQF but I think to the field, too, around measure and measure use and context and how look at that.

And I think twice we heard about the importance of this surveillance or the feedback loop and how I know we have a regulatory incentive, particularly with the FDA. And Laurie, you shared some of those insights, too, around we do this in that realm but we don't routinely do what I would call surveillance of measures and use. I mean certainly there are mechanisms but I think we are hearing for a stronger call for that feedback loop and that rapid kind of cycle learning.

So lots of great take-aways. I
I want to open it up for questions to the group. I am going to ask the operator to please queue up those joining us virtually. And certainly our audience participants, please join us.

So I might start -- oh, I see. Lori, would you like to? Okay, great. Thank you. And then I will work my way around.

DR. FRANK: Hi, thanks. And thanks to the whole panel. That was excellent.

My question is for Ethan. I appreciated how you brought patient engagement in and showed the specific points of contact. Who gets to determine if the patient report of the outcome is appropriate? And should patients be involved at step number two on that part of the flow chart?

DR. BASCH: You mean who determines if the patient is in the best position to report on a particular outcome?

DR. FRANK: So step number two is --
DR. BASCH: Yes, determine if the patient report of the outcome is appropriate. I don't know. What do you think?

(Laughter.)

DR. FRANK: I think that could present an opportunity for patient engagement as well. But it is a really important question about the pathway. Who gets to decide before we go on to the next step in the flow chart? Right now the way it is framed, it seems like it is limited and that there is not consumer and patient involvement.

DR. ADAMS: Lori, do you recommend -- I know from the first workshop you identified certain touch points. You looked to the PCORI model of where there would be touch points for research and you identified some touch points. And I think certainly number one was in that touch point. So as we think of this pathway as Ethan, I think you did on your slides, you started to identify touch points for the patient or person. So
maybe we should be more specific about that.

DR. BASCH: I would actually make
a comment, too.

So one initiative I have been
involved with is the PRO CTC, which is the
NCI's initiative to create a patient,
basically an item library for patient
reporting of harms of treatment in the context
of cancer trials. And in selecting what
adverse events are appropriate for patient
reporting versus clinician reporting versus a
lab system, a set of criteria were created
based on the level of subjectivity of the
outcome. And then a modified Delphi process
was used with multiple stakeholders to figure
this out. And there are actually five
categories. I can't recall exactly offhand
but assuming like totally subjective, you
know, experiential with no observable
component, and then a little of both, and then
totally observable. So maybe starting all the
way over here with nausea which really only
the person experiencing it can truly know over
to something like rash or alopecia which a
patient or an observer could potentially
report on. In fact, in the case of rash, the
conclusion was perhaps a patient can report
the incidence of a rash but to characterize a
rash actually requires some sort of expert
training.

So I think that there are actually
methods that have been used to try to
determine this.

DR. FRANK: Yes, to determine
that. So I would just want to add that
clinician interpretation can be a really
important step. So depression is a good
example. It is an internal experience and yet
a lot of people make the compelling argument
that you must have clinician report about some
aspects of the experience of depression
because they are able to gather the evidence
in a way that the patient cannot. But another
example would be cognitive impairment and it
was already raised.

So it would be a big no, for many people, step number two. Oh no, we can't ask cognitively impaired patients to report on this. But I think we need to question a lot of those assumptions and so I would just encourage us to all to think about that as a point of contact as well.

DR. PERFETTO: Also had another way of thinking about this in terms of re-framing number two a little bit. Because I think if we think about this issue of framing a problem and figuring out what is important to the patient, and then we have to have that step of saying okay well maybe these things that are important to the patient might be patient-reported outcomes that we might want to get to that are important to the patient may not be the ones that would resolve that problem. And so when we say appropriate, it may be is there a match between the problem and the measure because it could be a process
measure that resolves the problem and then you
would make the determination.

But to your point, I think it
would be good to have patients involved in
that discussion and that decision-making
because even if the end conclusion is oh no,
a PRO is not needed, a process measure will
do, it is good to have patients understand
that decision and be a part of that decision.

DR. ADAMS: Okay, I am going to
start in the back. And so Ted, you and then
Phyllis, and then I will come back up to
Kathy.

DR. GANIATS: I'm Ted from San
Diego. This is great. I really appreciated
everything that was said. And it stimulates
two ideas; one an editorial suggestion and
then second I will be selling tickets for the
fight between Jim and Eleanor later today.
I'm joking.

A dotted line from the green to
the orange. I think the way the diagram is
listed now, all process measures will become outcome measures. And it might be something -- we have said they don't all but the diagram isn't consistent with what was said.

Now Eleanor brought up a wonderful idea that there is a box zero stating what the problem is. And Jim brought up that not all measures, this is box five, not all measures may be incorporated into the practice. It might be that there is a survey that was done. And I think that box zero and box five have to interface somehow and that box zero may tell us that we are interested in a process measure. The problem is that doctors aren't asking about pain. The process measure is are they asking about pain, in which case we might want to incorporate it into practice. But there may be some times when we want to just survey that having something incorporated into the practice is too burdensome and all we want to know is as an outcome how do the groups compare. And we
could do a survey of a small number. So I think extrapolating or expanding box five to allow for a survey instead of implying, it is only going to be incorporated into practice allows Jim's comment to be incorporated. And it feeds back into box zero.

And obviously I am joking about you two fighting but the love fest will be --

DR. BASCH: I was just going to comment that in box five, we talked about box five a little bit and what does it mean exactly. Does box five actually mean pilot testing in practices where the target population is represented in order to assure that the measure is performing the way that you believe it will perform, based on all of your assumptions? Or does it actually mean integration into practice as a process measure?

DR. GANIATS: I go back to the diabetes foot exam measure where an option was to ask patients did the doctor examine your
feet or did you take your shoes off? And that would not be in clinical practice as box five words it. And so it is just allowing an expansion of how we think we might test them.

DR. ADAMS: Any other response to that?

DR. BELLOWS: So that was an excellent comment. I totally agree that the box zero and the specification have to be tightly related. And I also agree that some of the language of box five, as specified, is incomplete in terms of how I would hope -- to me it is not just about whether the measure is performing as anticipated, it is about how is the entire system responding. What is the impact on the key stakeholders? If it is a clinical measure in particular, what is the overall in the interaction on the key stakeholders? And that is bigger than the performance of the measure itself. It is what are the expectations created? What are the expectations realized? All that kind of
stuff.

DR. ADAMS: Phyllis?

MS. TORDA: So I have several reactions, mostly to Jim's presentation.

I think it is fine to do process and outcome measures simultaneously. I will say that our experience at NCQA is that whether you should start with one or the other, in large measure, depends on the sophistication of the organization. In organizations that have more experience with quality improvement, and that is how I am using the term sophistication in this case, can go to the outcome measure and do the analysis necessary to figure out the paucities that will lead to better outcomes.

But organizations that don't have that expertise really find it useful to have the process measures that provide the roadmap to good outcomes. So that is just an elaboration.

With regard to the remarks about
harmonization, a couple of remarks. One is it is much easier to achieve harmonization in virgin territory than it is once people are really attached to one tool or another. And I think telling the whole nation that it has to use one tool is a difficult thing.

I think some of your remarks, Jim, went more to the disadvantages of using different tools in the same organization as opposed to the disadvantages or advantages -- as opposed to using tools in different organizations that can be mapped. And some of the waste and inefficiencies go away if one organization is at least using the same tool. And then those results could be crosswalked or mapped across organizations.

I think we have given a little bit short shrift to the unit of analysis issue. And there is always feasibility issues and statistical reasons why larger units of analysis are easier than smaller ones but I don't think we focused very much and it is
worth thinking about conceptually what is the right unit of analysis. And some of these discussions about box five and whether it is clinical practice or it is really a bigger group of people get to the unit of analysis issue. You know, which individuals are making contributions to the outcomes that we are measuring?

And finally, a lot of talk involving patients, which is great. A little, I certainly would appreciate guidance about how to do stuff because in real situations, it is hard to recruit patients to participate in many of these activities. And so any wisdom that can be offered from the experts or the audience and communicated by NQF around that would be welcome.

DR. ADAMS: Any response?

DR. BELLOWS: Just a couple of things really quickly. I think sophistication is one really important aspect in how to go but I don't think it is the only thing, by
far. It also has to do with priorities and portfolio management and all that sort of thing.

And we tend to think of these one measure at a time but actually people in operations in our delivery system are working on 20 different things at the same time and it is partly a matter of, on this particular, with the use of the PHQ9, how far up it is in their priority scheme and how hard they want to push it. So I think there is a bunch of factors of which sophistication is one.

On the harmonization, you are totally right. A part of it is about within institutions but also there is the thing that our patients are crossing across many different settings. And as they go from hospital to primary care to skilled nursing or whatever, they are going to be touching many different institutions. So I actually think that even for the sake of consistent expectations for our patients, it makes sense
for us to have harmonization across institutions as well. You know, I know if I am going to be asked about pain, I know what that question looks like. And I know if I am going to be asked about symptoms, I know how many questions to expect, that sort of thing. So it is just a small thing.

And with respect to how to include patients, to me there is a really important distinction in the sort of box one through three stuff. We are asking patients rather hypothetical questions, in a way. What do you think would matter to you and what do you think we should measure. It is kind of like asking people to reveal their preferences, as if they had preferences that they could just reveal. And to me, that is why I partly put more stock in the box five, where you can create an environment where something real happens and then do the kind of qualitative stuff that everybody in this room in some ways knows how to do about what was this experience.
like. What happened? How did it work out?

Did we miss something? That kind of stuff.

So I think the methods are clearer in box five. Do it and learn, as opposed to the methods in boxes one through three that are somewhat more abstract and that I think maybe none of us understand quite as well because it is more conceptual.

DR. PERFETTO: I think to add to that, I would turn to someone like Laurie Frank and say you know, this is kind of where PCORI is headed, to try to flesh out and further develop those patient engagement methodologies. I think to date my own experience has been that we recruit patients depending upon the circumstance and the question from a variety of places, anyplace from online having them submit information online to various kinds of things to focus groups. And they could be people who are patients who are being seen by particular kinds of physicians to any kind of general
focus group. So it is any and all at this point and I think we are still going to be tweaking those methodologies as we go.

DR. BASCH: I would just add very briefly I have participated in a number of panels that have recommended that a variety of different measures could be used in the same context. Many of those panels have assumed in the future that there will be good approaches for crosswalk and there are a few of those initiatives going on but the truth is that either there are a lot of problems crossing between measures.

I actually personally would strongly advocate for recommending a single measure in a single context here and that the bar be, as Jim I think aptly put it, very high to unseat that. And if an investigator wants to come along and demonstrate that a new approach performs better, then they could maybe unseat the first comer.

But you know, I think that it will
avoid some confusion. Some may disagree with that.

The other thing that I would say about patient engagement is again, you know, I think boxes one through three rely on qualitative research that would include, as Eleanor pointed out, focus groups, key informant interviews, cross-sectional surveys, longitudinal surveys and there are fairly well established approaches for aggregating and analyzing that information.

DR. ADAMS: So I'm going to talk about the queue because we have lots of hands up. I know Kathy you have been waiting patiently. And we do have one of our panel members who also queued up earlier, Barb. So I am giving Barb the signal now that after Kathy she will come because she queued up after you. And then I will get the four that raised their hands in these two tables. Thank you.

Kathy.
DR. LOHR: I thought this was a wonderful discussion and thank you all.

I think the box zero is really important. And boxes one through three I would just reiterate what Ethan and others have said. We have been doing some of this kind of thing for 20, 30 years.

What might be missing from one through three is some understanding that you may have to, if you will, pay your patients back. There may have to be something that is given back, whether it is a pure incentive or some other pieces of information, to thank them but perhaps to make, if you will, worth their time. So we need that.

On box two I wanted to say this is another one of these perhaps it is not an either/or question. And it can be phrased as you want to determine whether and to what extent the patient reporting on the outcome is the appropriate thing because there is a spectrum and gradations there. And I would
I recommend that you change that.

I also have to say coming out of long sort of Donabedian quality of care triad, I could not work with this diagram because I immediately got off onto process measures, as one of structure process outcome, and I did not understand why you would go from a process measure of any sort to an outcome when this is about patient-reported outcomes.

And so I am just wondering whether at a minimum box six and the other ones, some other term that is a synonym if you will, for process, might be used. I mean there is operational, there is event, there is use. There is a bunch of words that maybe would serve you better so you don't send people like me off thinking why would we be talking about process measures and then only after we have done all those process measures am I getting back to outcome measures.

Ted or somebody said maybe a dotted line down from your green back to your
orange might help but following the way this is here, and then talking about process measures before you ever get to outcomes, I think it runs a risk of misleading some of the people, certainly in the quality of care, maybe not so much accountability, but in the quality of care world.

DR. ADAMS: I think, I know Barb's been waiting on the virtual line for us. So I am going to ask the operator to please queue her up so she can ask a question.

DR. SUMMERS: Hi, I have my mute off. Can you hear me?

DR. ADAMS: Yes, we can.

DR. SUMMERS: Oh, great. Thanks so much for the opportunity to comment. It has actually been lovely to be able just to sit back and listen to the rich dialogue that has been going on yesterday and today.

I had just three quick comments/questions. The first really goes -- and they relate to measure use and the context
of measure use.

    First going back to the notion of risk adjustment. And one question that occurs to me is how do we, how could we, and should we include the patient perspective as it relates to the relative significance of a PROM in their overall care experience? For example, can we somehow develop some patient-reported risk adjustments?

    And the example that I think about could relate to an individual with prostate cancer who, in treatment, sequence number one may experience relatively little fatigue as a consequence of the therapy being used. But if their disease progresses and they advance to a different therapy, that therapy could cause significant fatigue. So although patients would generally prefer to have no fatigue, I think most patients would prefer some fatigue to the alternative of death.

    And then following on some of Ethan's really excellent points, I think we
really should continue to focus on the
necessity of incorporating the patient-
reported outcome performance measures into
clinical workflow. Because it is not only
going to improve our response rates, it is
also going to hard wire incorporation of PROs
into clinical practice.

And then following on some of the
most recent discussion, the use of the PRO
performance measures in clinical practice is
also meaningful in achieving improvement. So
I believe that outcome measures should also
have utility in clinical practice as
clinicians look at their data and aggregate in
an attempt to determine how is it that they
can, as a practice group, achieve improvements
in their PRO performance measures. Thank you.

DR. ADAMS: Thank you, Barb. Any
comments for Barb?

So I am going to go back to our
table and work our way around. And Ted, did
I see your hand go up? Yes, and I will go to
this -- and David, I did see you, yes.

MS. WILKINSON: Hi, Linda Wilkinson, Dartmouth-Hitchcock, Coordinator of Patient and Family-Centered Care. And I mention that because I think there are lots of people addressing the desire to include patients in either the formation of or the evaluation of these PROMs. And I am heartened to hear that from people whose focus is often measurement itself and the protocols and so on.

What I am hearing that I would like to address very briefly is I am hearing the concern that it might be hard to, or might be confusing to, or dangerous to include patients in the process sooner. I would like to debunk that.

But what is also true is that people have said where are we going to get them. I mean, they have said it in many languages. How are we going to do this? How are we going to administer it? I guess we
could say you would us, meaning our
institution, as a test case. I can't tell you
how eager are patients are to share their
impressions and how amazingly revealing of
things that skilled clinicians and skilled
designers of programs have been to find the
things that they have not turned their
attention to that they realized was important.

And I would invite anyone who
wants to explore how we mechanically have gone
about this, what standards we have used for
who we are asking to join us and the like. We
are very happy to share. But I would, at
least for this meeting, urge people not to
discount it because we are not yet used to
doing it. It is in fact a lot easier. It
takes work. It takes attention. It is
exceedingly awarding and I would encourage you
to contact anyone like us who has had some
experience doing this and have found ways to
make it workable and very profitable.

So ollie ollie umphrey. Thank
DR. ADAMS: I would like to thank you for that offering to us because I think within many of our processes, how we can engage patients authentically at those various levels. Any comments in regards? We are grateful to your expertise that you could offer for us there.

Dave?

DR. CELLA: Thank you. Dave Cella from Northwestern. I agree it was a great panel discussion.

I have a couple questions that sort of I will follow I think, if there is enough time, with a suggestion or a thought. One question is box zero. Could you help me differentiate that from context? Because when people have been talking about context of use, I think about box zero, what you described as box zero but maybe I am missing something. I just want to make sure.

DR. PERFETTO: I am just simply
saying that we have to articulate what box
zero is.

DR. CELLA: Which is the context
of use.

DR. PERFETTO: It could be the
context of use but it could be the reason why
we want to do this in the first place. And
that could be part of the context of use.

DR. CELLA: Yes, okay. All right.

DR. PERFETTO: But I think what we
have been talking about in terms of context is
describe the patient population. Describe the
setting. We haven't really said describe --

DR. CELLA: So as long as we have
a broader --

DR. PERFETTO: -- the question you
are trying to answer.

DR. CELLA: All right, so broader
definition of context.

DR. PERFETTO: Yes.

DR. CELLA: We are basically in
that -- okay.
DR. PERFETTO: Yes.

DR. CELLA: Okay. I just thought it was missing something from the earlier discussions.

The other question is about what I think are Jim and Ethan's suggestion that we pull back from standardizing and equating and focus on one measure. Partly I am confused and then I have a comment to make. But the confusion is that I thought I had it nicely and very positively drilled into my head that NQF endorses performance measures, not PRO measures. And as such, why would we be even talking about the idea that only one measure, PRO measure, unless I misunderstood that, should be endorsed because NQF doesn't even do that. Maybe I missed something there.

And then I have a comment about PRO measures and equating and stones and kilograms and pounds. Did I miss something there?

DR. ADAMS: So Karen, do you want
to respond to the we don't endorse the tools
but why we would --

DR. PACE: So you are absolutely
right. We don't endorse the individual PROMs.
But by virtue of specifying a performance
measure that is going to be standardized, we
need to identify which PROM will be used to
collect the data.

So you know, although we don't
just endorse the PROM by virtue of the
performance measure, obviously we have some
relationship there. And I think the
discussion was should we endorse performance
measures that say you can use only one of the
instruments or two or three. But I think that
is where your equating comes in and maybe you
want to add something.

DR. PERFETTO: Can I give an
example here? Can I give an example on this
one?

There is a measure that exists now
and I am not sure I am remembering whether it
is NQF-endorsed, but there is a measure right now that is used in rheumatoid arthritis. And I don't remember off the top of my head if it is endorsed or not but it is a good example.

There is a measure that says on an annual basis, if a patient has rheumatoid arthritis, they should have a functional assessment done. It doesn't say how. It doesn't say what tool to use. So it is really a process measure. It is a check the box yes or no, on an annual basis did this functional assessment measure. I think the discussion is in the process stream on here, does this go the next step to say is a functional assessment done with XYZ tool, which still would be a process measure, versus it turning into an outcomes measure that says functional assessment -- function for this particular group improved by Y based on the tool that gets specified. And I think your question is a very good one. Would NQF go that next step to actually list which tools would be included
or excluded from that measure, whether it is
process or outcome. And that is the crux of
the question because right now, they are not
specified in that kind of detail.

DR. CELLA: So if I could then
with the comment, because it does kind of
follow -- I know those people want to talk.

So we used the -- Jim you used the
analogy of kilograms and pounds and then you
threw in stones. So one thing that got me
thinking was if today's measure is in stones
and yet, the one that is used in practice and
it is out there but yet we are aware that
there are better measures, you know, scales
that can measure in kilograms and pounds that
can truly be equated to one another with a
simple look-up table, why wouldn't we push for
the kilograms and pounds over stones?

And I think that is what you were
saying, Ethan, is that there ought to be an
ability to prove something is better and
switch to it.
But if the current could be a metric that is indifferent to how you get that metric as long as you have a way of getting to that metric, like a kilogram or a pound, we don't care what scale is used to come up with that. I mean, usually I don't think NQF or CMS asks which scale did you use or who weighed the patient; or even was it self-report weight or -- you do ask that. Okay. Well, thank you. I'm glad to hear that.

But maybe I am making the point and maybe not. It seems that there are areas, not all of them but there are areas of health-related quality of life like depression, which is in many of the current guidelines where you can link across different instruments. And I thought I was hearing sort of an argument against doing and I guess I disagree with that argument if that I heard it correctly because it seems like we care more about the pound than about the scale and we care more about the depression than we do about the
questionnaire.

And I am not advocating this in areas where it is not ready but there are areas where it is ready. So I guess that is my comment.

DR. ADAMS: Now I am going to have the panel to reply and I know Helen, you had some response to the direct NQF question.

DR. BURSTIN: Could I just respond very briefly?

So I think those are really great questions. And I think the key thing is we are trying to get to a set of standardized measures that really allow people to compare apples to apples.

So if there is information, for example, and you talked about this, it was very excited at the first workshop, David, that there might be opportunities to provide sort of almost equivalency tables, and I forgot the exact term that you used, that would say in fact that PHQ-9 versus the PROMIS
depression scale or equivalent. Use either and there is a way to make them both work. That's wonderful. But I think what we don't want to have is to have a thousand flowers bloom and not at the end of the day, allow the measure itself to provide the comparability that drives what we hope is standardization and improvement.

Karen, does that work for you?

DR. ADAMS: Ethan.

DR. BASCH: I would just add we also don't want to have -- since patients have multiple conditions, maybe seeing multiple providers have them answering multiple depression scales in different context.

DR. ADAMS: So Ted, I know you have been waiting patiently. And then I will -- Ted Rooney, sorry. And then I will go back and then Gene. Okay, go ahead..

MR. ROONEY: Okay, thanks. First a quick question for Jim. On your extra two boxes, it seemed like you were pleased enough
to have one box named after you. Now you want
three?

(Laughter.)

MR. ROONEY: But the real question
is, I live in the world of implementing
measures, both for accountability and
improvement. And we have been using the
Berwick NQF diagrams about ten years ago,
talked about the box, and QI on one side and
COM on the other. I don't believe one works
without the other. And we have been working
with this group for ten years now that meets
six times a year or five times a year or half
days with 14 PCPs, employers, health
planners, consumers, whatever. And we are in
the throes of this stuff. And I can't
emphasize enough the importance of
harmonization and benchmarking because if we
can't get one stand -- like if we could have
a PROMIS database like we have the CAHPS, it
would be phenomenally helpful because what our
docs want to know, they want to know how do
they compare to other areas. And then what 
can they do about it to improve it? 
And then in addition like if you 
can get the PROMIS or something like that, you 
know, standardized data base, you have no 
problems with copyrights, it is easy to do, 
and then have some technical assistance, I 
don't know what it would be for us, but with 
AHRQ they sponsor this CAHPS database. And we 
are doing a state-wide project, in physician 
practice and CAHPS is pretty straightforward 
but I can't tell you how important it was to 
have someone like Dale Shaller who we get to 
do some technical assistance who meets with us 
and talks with us about how we implement that. 
And we involve our providers in that, too. 
And it is just so important. And we don't -- 
it is so important to have some guidance on 
how we can implement something so that at the 
end of the day it makes it work in Maine but 
we can compare Maine to Minnesota and other 
places.
And then right now in Maine we do really well on quality. Like we are some better quality in the country. We are the worst in cost. It works both ways. And then sometime providers will say well we are already better than the national average in quality. Why are you pushing so hard? Well the national average stinks but that is besides the point.

We wanted to get to the point to say well it is not just a look at your quality compared by a state but is there a place in Minnesota or Colorado or Texas that really has superb quality? And then we can benchmark to that and then show what the difference is. Our providers would respond because they are working as hard as they can. And when they think they are doing really well, they sort of, you have got to understand why are they working so hard? Or when they find out that over here they have figured out a better mousetrap, we can get that imported.
But you can't do that unless you have harmonization and standardization. And you can't build a database with technical support unless you have the harmonization. So I know NQF just does the harmonization piece but I implore us, if we can get like a PROMIS database like we have with the CAHPS and get some of the support technical assistance, it would drive tremendous improvement.

DR. ADAMS: Jim, did you have a comment to make?

DR. BELLOWS: Well first of all, Eleanor and I have conferred on this and the post-market surveillance box is going to have her name on it and not mine.

(Laughter.)

DR. BELLOWS: So we do have equanimity here.

DR. ADAMS: He's one through three.

DR. BASCH: Always the bridesmaid, never the bride.
(Laughter.)

DR. BELLOWS: David, with respect to your questions, I totally get it but NQF doesn't have the charter or the mandate to design an entire system. But I guess it is just my feeling that they have their hands on an incredibly important control that could move us either towards the thousand flowers direction or more towards the greater consistency direction. And the more those that control is shifted in a way that brings us towards harmonization and consistent use of the same scales, I think the better off we are in some ways, I think. Not on everything but on some basic things.

And I know that there is one thing for us is one really great aspect of use of PROs within our system, for example, is as people transition across settings and return from specialty care back to primary care, from primary care to skilled nursing, that it can give a common language. And as people move
through our spine pain care for example, they have consistent metric of what their pain was. And the more we go into different scales and different places, we just erode that.

So I think there is just things about little preferences. If you use one of the three established response scales you automatically get a pass on some aspect. But if you want to come bring in some other different response scale, you have a higher bar, it is just things like that, little preferences to bring more coherences is what I am hoping that they can do.

DR. ADAMS: David do you want to respond to this, and then we will have time for two more, which will be Ted and Gene.

DR. CELLA: Yes, although you might also need to respond back to his points, Ted's points.

And I actually I am breaking in and I apologize for that because I think we actually agree. And we are using language,
interestingly, somewhat differently but maybe saying the same thing. So let me try again.

I am against a thousand flowers, unless those flowers relate to a common underlying unified metric. And when those flowers relate to a common underlying metric, I am in favor of those flowers because they give people choice. But the metric is what gets reported. And we can do that in some areas and not in others. And so I have been working on that for many years to get common metrics and common language, as opposed to all the flowers and people selling their wares. So I think we actually agree on that and NQF is in a good position, I think. And it works well with that performance measure certification as opposed to the PRO measure certification because they can say we certify on the metric and you can use the PHQ-9 or the CESD or the PROMIS depression. It doesn't matter because you are reporting the metric.

DR. ADAMS: Thank you, Ted.
DR. GANIATS: This is Ted and I have a corollary or a slightly different take. I think it depends on if it is a process measure or an outcome measure. And so I co-chaired a heart failure performance measure panel. We recommended two or three process measures. You should check for function.

But to be able to compare those, there is just absolutely no way because the theoretical constructs behind the three we chose were so different you couldn't crosswalk them.

And so if it was a process measure, I like a thousand flowers. If it is an outcome measure, I want to have one flower. And I have an exceedingly high bar for crosswalk. If it has a R square of 0.8, I don't care. That is not good enough. Because if it is an accountability measure and it is only a 0.8, I think that is too much variability.

I want one flower and I want the
one flower to have one bloom. And if you have
two instruments that aren't -- I mean the VA
version of the SF-36 and the RAND version of
the SF-36 and the quality metric version of
the SF-36, they are probably close enough.
But almost anything else, you are not going to
be able to do a close enough crosswalk for
accountability.

DR. ADAMS: Any response from our
panel? Gene do you want to -- or yes. I was
going to say, Lewis, with that thread, did you
have a comment? I saw your hand go up.

DR. KAZIS: Thank you. I just
want to indicate that David Cella is really on
the right track here. A common metric I think
is what we are after. It may take some time
but there is a body of literature out there I
think that began a number of years with Danny
Fryback who had an NIH funded grant to begin
to develop bridges across a number of
different assessments. Many of them were
generic in those days.
But I think that the objective here is the metric. You know, that is what we are looking at, a common metric. And the methodology, I think, is there. It is just a question of application of that methodology across these many instruments.

DR. ADAMS: Please respond and then we will go to Gene. Thank you.

DR. GANIATS: Denny and I, along with a few others with the HUI, EQ-5D, SF-36, and the QWB did simultaneous administration of all of those in random digit dialing across the country and disease-specific over time.

It is crap. Okay? It is really sad. You can crosswalk all day long between those instruments and their responsiveness to change in two different conditions are completely different; where sometimes the SF-36 will show a change, sometimes it won't depending on the condition. Sometimes the QWB, sometimes not. So the fact that you can crosswalk doesn't make them equivalent. And
so you have to be incredibly careful.

So there is one thing to be able to do the arithmetic, the arithmetic of a crosswalk. It is another thing for it to be good enough for an accountability measure.

And I mean it is published out there. I just think we have to be incredibly careful.

DR. ADAMS: So Gene, I am going to ask you. I know you have been waiting. And then we have an audience comment and then I will wrap us up. Gene? And anyone from our panel? Sure, of course.

DR. NELSON: So this is just very brief. The flow chart, once it gets revised a bit will be really helpful. And that would be really helpful to actually test it with a measure coming through and there is some opportunities with the ONC meaningful use PRO measures, et cetera that we should really take the flowchart as revised and try it out.

DR. ADAMS: Thank you, Gene. I
think that as you are raising here do we have
a few use cases that we want to put through
here would be very helpful.

And you mentioned you had a
comment from the panel in regards -- no?
Okay.

So there was someone from the
audience that wanted to comment on this, too.
And then Al, I'm going to give you the
pleasure of wrapping us up.

MS. POTTER: I'm D.E.B. Potter
from AHRQ. There are several of the PRO
measures that are included in our national
surveys. And so you have a benchmark of the
non-institutionalized civilian population.
And a lot of times you can cut them by various
sub-populations.

So I guess I urge people to think
about the use of that data. And should we
start to think about that as a way to build a
national benchmark? Because we are not going
to have the resources to build a benchmark
database for every single one of the PROs in use.

DR. ADAMS: Thank you.

DR. KELLER: San Keller, American Institutes for Research. And I wanted to merge the two positions of the common metric and the implications of using something that you have shown statistically to be similar. You know that there should be sensitivity analysis to the effect of the differences and whether or not they make a difference in the application of the measure.

And I am reminded of Dana Safran's work on the six response versus the four response and how it orders doctors and so on. So you can do that and those hypotheses should be stated up-front when you are making those translations to address that potential criticism.

DR. ADAMS: Albert, I think it is you. Yes, great. Thank you.

DR. WU: So I have a question
about I think that coming up with some use cases is absolutely the smartest thing to do. And to really specify the heck out of the first use cases, so that we are really, really focused on whatever it is we are focused on.

I sort of then became a little unsure about how this is sort of going to get rolled out because the first use case will then be generalized somewhat if we are looking at hypertension in African American men between the ages of 50 and 55, what is the next case -- what is the next use case going to be? Or how are we then going to move from that to all hypertension in all men, hypertension in genders, hypertension in all ethnicities, in primary care, in long-term care, in acute care, in rural, in urban, and so on and so forth.

So we wind up in a way -- do we wind up with a family of related flow charts for hypertension measures? Because in some cases, the evidence is going to be there for
African Americans but not for some other ethnicity. How is it going to -- how are the family of flowcharts going to relate to each other?

DR. PERFETTO: Our point was just rather than just start with the outcome, the hypertension outcome, it is thinking through the context, as we were talking about earlier, what is the problem. So maybe there is only one population of hypertensives that you want to improve care for and zero in on them and the others you may not need to because you don't have any indicator that there is an issue there. Or you start with the one and then you go to others. But you start with the one where the problem is most significant. It is really articulating the problem.

DR. ADAMS: Okay, so one other comment and then we will wrap it up. But Al, thank you, because I think the flowchart is already serving its purpose. Because you are already seeing how it could be used in various
ways with implementation. So thank you.

And Ted you want to wrap it up for us, please?

DR. GANIATTS: Yes, I just have a question. I would think of a PRO as an outcome. I would think of blood pressure as an intermediate outcome. I don't care about - - I mean from a PRO point of view, I don't care about blood pressure. I care about stroke, heart attack, et cetera.

So the PRO is conceptually much closer to what we are interested in -- it is what we are interested -- than essentially all the other performance measures.

DR. ADAMS: So on that closing remark, I would like to once again thank our reactor panel. And shall we give you a round of applause? Yes.

(Appplause.)

DR. ADAMS: Okay, so we are going to move on to lunch. And then there have been several topics that have percolated in
addition to our future direction. So when we come back from lunch, we will be discussing that and Patti will be leading that off.

And we will be coming back from lunch at 1:00, so in 45 minutes.

(Whereupon, at 12:13 p.m., a lunch recess was taken.)
A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

(1:03 p.m.)

DR. BRENNAN: Well, good afternoon and welcome back to the last session.

So I have learned that as you go through your career, in the beginning of your career they ask you talk about the future of a discipline. In the middle of your career, they suddenly start asking you to give retrospectives on your work. And when you are at the end of your career, they put you back into giving future perspectives again. So I guess this is a sign that either I am at the beginning of my career in quality out of informatics or I am at the end of my career in informatics, I'm not sure which.

I want to begin our last session here today by actually doing a lot of the things that we do often in the very last minute and don't have a moment to reflect on them, and that is to thank the staff and the participants and the audience for the work
that they have done to make this Patient-Reported Outcomes Workshop so successful.

Please join me in a round of applause for everyone here.

(Applause.)

DR. BRENNAN: Thank you everyone.

The double Karen or whatever we have been led by have just been inspirational.

But I also want to thank each of you who have given up time from work, each of you who has children at home starting school this week, giving up time from big family events, to be here to work through this.

In our last hour, we have one last piece of work to get done today and then I am going to give you a brief overview of what to expect. Over the next couple of days, if you begin reflecting on a conversation you had here and you have an inspiration or a clarification, please send it to Karen Pace or Karen Adams. They will be pulling together a draft report based on our conversations, the
background reports that were provided to the meeting and the two workshops. That report will be written within the next month. But if you have thoughts in the next couple of days, please send them on.

There is also you received for this particular meeting a background document that was prepared by RTI. If you have any comments on this document, it would be critical to get those here within the next couple of days to early -- at the latest Monday or Tuesday because the authors of that document will be taking that feedback and making a final copy of that report.

There will be a call with the expert panel, those of you who are in this room, sometime before October 26th. That will be before the report goes out for public comment. So you can expect to be hearing from Gene and the scheduling folks.

The comment period begins October 26th. Please, as you know, encourage your
constituencies to review the report and make
comments on it.

   There will be a meeting with the
entire expert panel via a webinar after the
comment period closes. That webinar will be
open to people, to the general public. So
those of you who are in the audience and would
like to listen in to that presentation, you
will be able to hear that probably before
Thanksgiving it is likely that will happen.

   And then the final recommendations
are going to go to the CSAC on December the
10th. So there will be lots of work going on
in the background, lots of opportunity for
interaction. Your primary contact points are
Karen and Karen. We need to leave them with
some guidance and some instruction.

   So I am going to ask everyone to
please pull out our colorful guide and box
here and we are going to review the
recommendations that were made in the previous
hour. If you have a pencil nearby, you might
actually want to update and mark up your
document. There will be a new document
circulated.

And then we will have a time for
discussion until about ten until two and then
we will start wrapping up the session. I
guarantee we have a hard stop at two o'clock
because I have a phone call, as well as some
of you.

So, thank you very much for the
panel right before lunch really gave us a
great deal of content to work with, as well as
commentary to work on. And we have several
adjustments. And so I want you take -- oh, I
can move these, can't I? I can't.

Please go to the next page. At
the very top of this diagram I want you to put
a black -- just a box that says box zero. Box
zero is what we are using as a placeholder
proposed to be the space where we would invite
the proposers of a new PRO-PM to explain the
motivation for it. Why is this happening now?
Is it opportunity? Is it clinical care problem? Is it a sense of an absence of a metric in the system?

So box zero is the first change that we heard. You are not committed, even if you are writing in ink. This is just I want to bring you up to date to where we are with our thinking about this. Next slide, please.

On the next slide, you will see in box five that we have redlined the phrase clinical practice and, instead, changed that to be real world. What this means is to clarify that we want -- before a PROM can actually go through the evaluation process and be applied to the criteria, someone has to use it in the real world, where it is supposed to be used. And if it is supposed to be used in the clinical care setting, fine, then clinical practice is fine but it might be a plan that is using it. It might be a community that is using it. It might be a large integrated system. So it is not meant to imply that you
must use every prompt in a clinical care service experience with a patient to have it reviewed. It is meant to say you must have real world experience with this, which then leads us to what I am calling box five prime.

Next slide, please.

Just below box five, you should make a small box called five prime and call that fit for purpose. And I think that was Rob's suggestion that we made sure we had a point there where we figured out was this PROM actually doing what we thought it would do? Does it fit the purpose it was intended to do?

The last major change we heard in the morning, if you go to the next slide, please, is it goes at the very bottom of the page, box 14. We are placing in there a feedback, a box for feedback that probably is going to have tentacles and arrows coming out of it, going back into going back into different parts of the process.

So now you see on your screen and
you see in front of you. And if the audience at home is following, on your papers you should see a slightly revised pathway that will help the NQF create a systematic way to start examining PROS and taking into the process of PROMs and get up to the PRO-PM.

Our responsibilities in the next half hour, 40 minutes are to look hard and think hard and talk actively about this document. We would like to get, if possible, to a point of consensus for the staff, not specific to individual words, maybe a box would be bigger or smaller, but a conceptual consensus that this is a pathway that we believe as a group will be useful for going from a PRO to a PRO-PM.

And so it will take a few minutes of conversation. Now you had two sets of conversations already to talk about whether the current criteria for approval need to be modified for the PRO-PM process. And the sense that I got there was a lot of interest
in making clear the importance criteria number one wasn't as much consensus around other of the three criteria what changes need to be made for the PRO.

But I would like you to reflect for a moment about the discussions you sat in and about the idea of what must be or should be considered for modification to ensure that the criteria the NQF uses to approve the PMs is not going to present an unusual or insurmountable barrier for the patient-reported outcome.

And I will take comments and I see that Dave is ready for a comment. Okay, Dave.

DR. CELLA: Well, this comes back to the first thing I said this morning. So it is kind of like repeating myself but in the context of this flowchart.

I think it is great. It is great to put all these things down on paper. But then to me the question is how will it be used. How will it be applied? If it gets
applied strictly, nothing will ever pass. You can always find a way to say something is not ready. So there are several points along this continuum where a reviewer could say this doesn't make it. This doesn't make it.

So it is possible to use this diagram to reject everything.

DR. BRENNAN: So we would like this diagram to be an enabler. So based on Dave's comments, would you take a look through here and just right now circle or place a checkmark about where you think this pathway might be particularly vulnerable to a capricious or perhaps non-supportive response by reviewers so that no PRO ever gets through. Because I am going to take some time in a few minutes to ask you what is missing. And it might be what is missing is where the judgment calls. Where is the point that we need to have risks. So I think your point is very well taken.

So take a look through here and
just mark off on the box or jot some notes for yourself about do you see a point of risk where in too rigid or inappropriate application of one of these boxes or of the commentary, remember that is on the second page that might help you understand that might make a PRO particularly vulnerable.

Other opening comments for now?
Yes.

MS. TORDA: It just strikes me that for broader use really being more explicit about the criteria for a PROM and the criteria for a PRO-PM and how they relate would be useful. Because I think we were continually have to explain it to ourselves. And if we had to explain it to ourselves, being more explicit about it would be a good idea.

DR. BRENNAN: So that would be occurring in boxes 10, 11, and 12 or 10 and 11, I guess.

MS. TORDA: Well or even when you
are selecting --

DR. BRENNAN: Oh, no. I see.

MS. TORDA: -- the PROM that you

are looking for and then how you are using the

PROM and then how you are turning the use of

the PROM into a performance measure.

DR. BRENNAN: Yes, Karen?

DR. PACE: I was just going to say

the second note page unfortunately the number

boxes that those notes go with got left off. But we can do more work there to make that

distinction to clarify what characteristics go

with selecting the PROM versus the PRO-PM if

that is useful.

DR. BRENNAN: Would it be possible

for someone to state what the PRO-PM might

look like so we are all on the same page with

that? So we have heard about PROMs. We know

that here is various phases to depression or

hypertension. But a performance measure, does

somebody feel energized by lunch?

MS. PITZEN: Hi, it's Collette
A performance measure would be depression remission at six months. So the outcome is having that remission event occurring at a certain point in time and the PROM is the PHQ-9.

DR. BRENNAN: Okay. Would you put like a threshold like 100 percent of the patients demonstrate depression remission at six months?

MS. PITZEN: No.

DR. BRENNAN: Isn't the measure the percentage of patients?

MS. PITZEN: Correct, it is the percentage of patients that achieve remission at six months.

DR. BRENNAN: So no threshold, just the statement of the percentage. Okay.

MS. PITZEN: Correct.

DR. BRENNAN: Okay. All right so --

MS. PITZEN: Could I back up and
add a comment of just something that would be prohibitive, especially in this measure but maybe for measures going forward?

The requirement to have a process measure endorsed before you are going to that outcome measure. I think it is very important to have those processes in place but it is almost like it is supportive of having the outcome that you desire.

DR. BRENNAN: So I think that what Karen is saying is that under box five prime that we added, you really should have, those are two pathways that could occur simultaneously. But what you are recommending is that they be made explicitly simultaneously so there is not a dependency from one to the other.

MS. PITZEN: I missed part of a discussion this morning. So my apologies.

DR. BRENNAN: That's fine. And someone had also suggested that we actually might want to -- you might want to even invert
the green and orange box so that you do one
first and then the other. But the issue
clearly this diagram was not meant to imply
that you must go through the process before
the outcome.

Yes?

MS. DUBOW: Are we saying or
Collette did you say we have to have a process
measure before we can have an outcome measure?

DR. BRENNAN: No, she wanted to be
sure we didn't have to.

MS. DUBOW: Oh. Oh, that's fine.
I'm sorry. I misunderstood. Thank you.

DR. BRENNAN: Yes.

DR. PACE: I just want to clarify.
That was not the intention and we can fix the
diagram but the idea was that you could go
directly to an outcome measure but in some
cases you may want to consider the process
measure.

So we obviously need to make that
very clear on the diagram and we will work on
that. But I don't think the intention was
that it has to go through that pathway.

DR. BRENNAN: Right. And once
things get immortalized on a website, though,
they can be complicated.

Jack, yes please.

DR. FOWLER: This, I guess, has to
do with box 11. And we talked a little bit
about, I mean we talked off and on what
validation of this performance measure looks
like. And when Anne Deutsch was talking
yesterday afternoon, I think she was the one
that put up an example of saying if you had
some practices that you thought were exemplary
and then some practices that were usual care
and you could demonstrate a difference but
that would be an example of validity
information.

But I think it should be clear.

And again I can't see how you could get
evidence of validity without at least having
practices that varied in what they did in some
ways that you thought were credible that in fact had different outcomes on your measure.

You know, we have talked around about whether that was really essential or not. And I just find it hard to believe that if you have got an outcome that you want practice to change, that you don't need an example of several practices that behave differently, that get different outcomes. Because that is all I could believe would constitute valid evidence that the way you practice medicine could affect the outcome that you are after.

And maybe you don't need to elaborate on what that means but --

DR. BREN NAN: So let me just restate it. We are taking all comments right now. And so I think that if I am understanding you correctly, what you are saying is for box 11 where there is the performance measure should be tested for reliability and validity, you are suggesting
that that be made explicit on the level of
practices that we show variability across
practices and variability in what they do as
well as the outcomes they achieve.

     DR. FOWLER: That's right.

     DR. BRENNAN: Okay. And so that
is it. That is an example of how one might
demonstrate that kind of a test. That is
helpful.

     Yes, John?

     DR. WASSON: Two comments; first
is on Jack's comment.

     What we also heard though earlier
today is that in some cases that would be the
highest level, I guess. And then there might
be a lower level, which is this is something
that is so important to measure for patient
reasons, et cetera, that although we need a
lot of validity and reliability criteria, so
to speak, we may not have an intervention that
is going to show an effect yet.

     So I don't know if we want to
throw those out but at least they might be at a different level.

DR. BRENNAN: If I can say that back to you, what you are suggesting is that there may be PROs, there may be concepts that we want to assess the performance of a practice on, that we don't have any way of intervening with right now. And someone brought up the idea of fatigue and chemo. And those of you who are really good a fatigue and chemo, I am sorry if I underspecified that one. And so it may be appropriate to include as another kind of validity evidence the fact that practices or groups of patients or plans vary in terms of this PRO, in terms of this phenomenon, not that the care leads to changes in the phenomenon.

DR. WASSON: At this time.

DR. BRENNAN: Okay. So you are arguing in a broader sense to a higher level — I'm sorry -- various approaches to validity evidence.
DR. WASSON: Yes, and intermediate.

And then because Helen didn't laugh at me for this last comment, I feel free to bring it out.

DR. BRENNAN: Okay.

DR. WASSON: And you can blame Helen for doing it.

DR. BRENNAN: All right.

DR. WASSON: Number 14, your feedback.

DR. BRENNAN: Number 14, yes.

DR. WASSON: Yes, I was suggesting to Helen, she didn't laugh, that I really think it shouldn't be just feedback. It actually should be sunset. That every measure at the end of a three-year cycle automatically will sunset unless there is positive evidence that it is the best of the current measures or that it has good evidence that it should be continued. Because otherwise, we are going to proliferate ourselves to death or kill our
patients.

DR. BRENNAN: We are going to have a million measures.

DR. WASSON: Yes.

DR. BRENNAN: Yes. Helen is not really laughing just yet but she is into a work mode up here.

So I heard a friendly amendment to box 14 that we expand the concept of feedback to include such things as earlier today there was a mention of post-market surveillance and now John is suggesting that we actually have a hard stop on all measures over a set period of time, or at least a reevaluation of the continued value of the measures. I think that is reasonable.

I would like to ask you, is that unique to PROs or is that for all of the NQF indicators and PMs?

DR. WASSON: I don't think anyone should be automatically exempt.

DR. BRENNAN: Okay. You can tell
you are in this business.

DR. WASSON: Yes, because the

market is going to -- I mean we really are

going to change in our knowledge over the next

X number of years.

DR. BRENNAN: Okay. I mean I

think that will come for consideration and we

will see when it comes out of the staff mixer

what it looks like at the end but I think we

have got a pretty clear documentation of it.

And I personally actually like it very much

because we are, at the university, over-

assessed right now to the point if one more

person puts a yardstick up next to me, I will

spit on them.

Yes, Kathy?

MS. PITZEN: This is Collette. I

am probably oversimplifying but part of the

maintenance endorsement for the NQF measure is

you have to demonstrate that your measure is

still valuable and that there is variability

among the people that you are measuring and
that there is still opportunity to improve.

So I am a little bit opposed to the automatic sunset setting. And I know Phyllis had a comment as well.

DR. BRENNAN: And it sounds like you have some knowledge about the process that I don't that says there might be a mechanism built into that already.

We are all not accepting John's idea of no tenure. We are just including it as part of the comment.

Kathy. Oh dear, you need longer arms. There is a quality performance measure going on up here; stretch the guest.

DR. LOHR: On the sort of box 14, I don't know if you could take a page from what the National Guidelines Clearinghouse does but guidelines are essentially dropped if they are not updated and otherwise some step done to ensure that they are still, if you will, valid and all up to date every five years.
So I think three years and
sunsetting might be too soon.

DR. BRENNAN: Okay.

DR. LOHR: But a five- or six-year
window particularly because you are looking
for -- you are giving PROs at least or PRO-PMs
up to six years to be used in accountability
applications and so forth.

So conceptually, I think the
sunsetting idea might be a good one. Three
years I think is too short a term. And there
are models out there for what one might do to
sort of help with, if not sunsetting, at least
updating.

My other question, though, is your
bottom 10 through 13 or 14 are all orange or
will be orange. But I thought that everything
up through 12 is essentially the
submitter/developer's responsibility and that
13 and perhaps 14 will be NQF's responsibility
and the developer sort of disappears.

So should 13 and 14 be a different
color or a different part of the whole thing?

DR. BRENNAN: I think that would actually just make the display useful. It may be as the staff is redoing the display. Because remember, this started with Eleanor's doodle. So we have come a long way from Eleanor's doodle. But I think what Kathy is saying is that actually box 9 and 13 --

DR. LOHR: Yes, and nine.

DR. BRENNAN: -- are part of the NQF internal process.

DR. LOHR: Yes, and everything else belongs external to NQF because it is the developer people and current users and so forth who are going through this whole process and then finally submitting something and then NQF does its thing.

DR. BRENNAN: Right. That is really very helpful.

Other comments at this point?

Okay, now we are going to come to the interesting part of the conversation. Look
carefully at either your modifications or the boxes here having heard the kind of changes likely to come, so lines will move and there may be 12 boxes instead of 14, but essentially we are looking at a process.

And let us know if there is anything that you absolutely cannot live with. And this is the point in time to go back to thinking about where are the trickster spots that could actually capriciously kill basically good measures because they have got some weird performance.

Kathy and then the back table with --

DR. LOHR: This is just really a question. But is there a possibility of having list A and list B of a lot of these criteria such that some would be utterly mandatory and others more desirable --

DR. BRENNAN: Okay.

DR. LOHR: -- and might move a measure along more quickly or something like
that. But to me, this is a lot of criteria that have to be met simultaneously with a lot of data and maybe a lot of landmines, given ten or 15 or 18 or so criteria. And there might be some way of trying to say some are just absolutely utterly required and others might let us discriminate among or across similar PROs or PROMs or so forth, instead of having absolutely everything be completely required.

DR. BRENNAN: Okay.

DR. LOHR: And it is a question.

DR. BRENNAN: That is really a helpful question. Let me ask Karen or Karen. It seems to me that that is consistent with the evaluation criteria already. Yes, you want to just comment on that for a minute?

DR. ADAMS: I mean, Karen, you can speak to it. But what came to mind to me was the NQF criteria around importance. It is a must pass.

So if the measure is not deemed
important and a lot of this tied into our
meaningfulness and naturally scientific, et
cetera evidence, that you don't get pass go
for that. So I think it is looking through.

Some of the comments that came
back from the survey when we were looking at
the characteristics mentioned this and I think
it builds on what David said that nothing
could beat every criteria. So we want to be
careful.

So which one of these, when I was
thinking about this, some of these are very
helpful guidepost, and it is what we hold true
and where we want to go and other things are
hard and stern as I think you are saying and
additional guidance there would be helpful.

DR. BRENNAN: Additional guidance
like you would like some recommendations for
that now or over the next couple of months you
would like people to be thinking about it?

DR. PACE: Right, I think there is
more flexibility in those characteristics of
choosing the PROM and this pathway. But in terms of NQF criteria, as Karen Adams was saying, you know, measures have to meet our criteria for importance to measure and report for scientific acceptability and measure properties. Usability or feasibility and usability in use are more judgment calls to a certain extent.

So I don't think we want to say that. I mean we have heard a lot of indication that maybe we want to be stricter, I don't know that we would want to eliminate NQF criteria. But I think there is definitely flexibility in this pathway and also in choosing those PROMs.

DR. BRENNAN: There is a comment in the back. Is your comment directly related to the Karen conversation? yes, go ahead please and then Ted.

MS. DUBOW: I just wanted to make the observation that this pathway provides useful technical assistance, as opposed to
being a set of an NQF requirements. The way
the evaluation criteria are, for example,
where those are criteria that need to be
addressed.

This pathway to me suggests to
developers, measure developers and to
everybody else in the process what we think
needs to be wrapped into this process and it
provides -- it's useful. And I'm concerned
about eliminating anything in order to come up
with something that looks more palatable if we
don't bless it with the notion that this is a
mandatory set of must do every item kind of
thing but rather to suggest that indeed it is
a pathway. I just don't want to lose the
useful guidance that we have spent so much
time thinking about.

DR. BRENNAN: So if I can
summarize what Joyce is saying. Your hope is
that the pathway serve as a model for
proposers to know how to go through the
process and that we not mandate or become
rigid about the many pieces of it but simply say this is a pathway. You need to follow this. If you have got to make a change, you need to explain why you are skipping something.

Okay, actually both of you can speak, that's fine. And then we will come up to Ted here.

DR. PAWLSON: On this point.

DR. BRENNAN: Greg is speaking.

Sorry.

DR. PAWLSON: Greg Pawlson. On this point, I think the issue that we are kind of grappling with here is not whether I don't think whether these criteria are the right criteria but providing, especially in a new area like this, more guidance than usual on precisely what each of these mean and how far we expect the review panel to go.

Because in the scars that I have accumulated from the NQF review process, which I have to say have gotten much more refined as
time has gone on. So I really want to
acknowledge that but this is a new area. And
what we don't want is somebody coming in and
saying oh, well that means you have got to
have absolutely everything in the reliability
thing absolutely nailed, you have got to have
I already used in 500 different practices,
because that will never happen because it
often takes NQF endorsement to get some people
to use the measure widely enough.

Now that doesn't mean you can't
test it in a couple of sites but I think
really making sure that the review panels
understand these are going to be new, they
haven't been broadly -- most of them have not
been broadly used as yet, at least in this
country. If we are going to take evidence
from Britain and Sweden we might do better.

But I think it is the degree to
which these -- and I would see especially
number five and number 11 as being potential
huge stumbling blocks if they are
misinterpreted or over-interpreted. And it
doesn't mean that those criteria shouldn't be
in there but it does mean that I think the
review panel is going to need some very clear
instruction on how to balance. And it always
is. I mean everybody who is in the room who
has been on a review panel knows that good a
review panel is always balancing off how
important is this, how critical is this to get
forward, and can we come back. And I don't
know whether you are planning to have
provisional approval of this or at least the
option of saying we passed this but we are
going to actually review it in a year and you
need to come back with additional evidence.

DR. BRENNAN: So Greg is calling
for some judgment throughout but actually
raised a brand new point which I am going to
ask Helen to address. Which is, in the
process after box 14 -- or rather maybe I
guess it is in box 13 when NQF makes the
endorsement -- is it feasible, is it possible
to consider a provisional endorsement. Like we will endorse this for a year and then you have got to come back?

DR. BURSTIN: This is a complex issue.

DR. BRENNAN: Sorry about that Helen.

DR. BURSTIN: So essentially we have a very, very limited applicability for untested measures coming through our process which might be what people think of as provisional. We don't allow it for outcomes actually because I think outcomes are something, frankly, that need to be tested before they come in.

However, one of the things we have been experimenting with, we have a pilot right now, is actually trying to create a two-stage endorsement process. So for some newer areas, we probably may need to consider bringing in the measure concept first, understand a measure concept, the importance around the
concept and then allowing the measure to go 
back out and get fully tested and blessed.

So I think there are different
approaches here. And again, just as an
interesting point, on the Minnesota measure
around PHQ-9, there were a lot of concerns
about the lack of risk adjustment, for
example, that there was no risk adjustment as
part of that measure. And there was an
understanding this was important enough. Put
it out there. We will learn. We will add
that in as it goes forward. So I think it is
important to also note that measures are
iterative and we recognize that and we are
happy to take updates to measures as they go
out in the field and we learn. It is really
through implementation that we gain a lot more
understanding of these measures.

DR. BRENNAN: So one of the things
that Helen has introduced is a possible
intermediate step, which I might call three
prime that would happen before a lot of work
went on the field, which is getting to understand the concept early on. So there might actually be yet another box added in. Did you have another comment to make, I'm sorry, Barb?

DR. GAGE: Yes, I was just going to raise the concern, and I think it has been hit upon in solutions in different forms, is that it is an iterative process in getting anything through. You don't want to make it too tight or you won't be able to get to the final points.

DR. BRENNAN: Excellent. Ted, you have been patient.

MR. ROONEY: I pretty much want to amplify what Greg said because the number five, I mean if you want to do a new measure on clinical data or hemoglobin A1Cs or BPs or cost of care, there is a wealth of existing data that you can get access to the model but there doesn't exist a lot of the data that we want to test with outcomes. So as long as
five is seeing that you are not going to have all the data, testing validity and reliability, and all these different population segments because you don't do it, it doesn't exist. So that is where I think you get really tripped up if someone starts comparing this.

The other quick comment was that again I live in the world of both improvement and accountability and Gene came over, Gene Nelson came over to us a year ago to really talk of PROMIS. We had huge excitement for it. We cannot get one PCP practice to even test it because they are so overwhelmed right now.

But whereas if you told me if I can go back and say, look, this is going to be an accountability measure in a year, I have more practices that I could handle to do it.

So that is the thing, if it makes it too hard to get to an accountability measure, it will never happen or I would be afraid it would never happen.
DR. BRENNAN: So I heard two things in what you are saying and I want to make sure I got them correctly. The first one was there is a great absence of test data. And so maybe one corollary activity might be to think about how a test bed of data could be developed that would allow people to begin to test measures. And the second is that in the absence of a mandate for action and in the test bed, the clinical engagement is really tough. Did I get that?

MR. ROONEY: Yes, and that is only a problem then if you need the clinical engagement to then approve it.

DR. BRENNAN: Right.

MR. ROONEY: So it becomes circular.

DR. BRENNAN: Yes, I got it. This gentleman and then over here. Yes, sir? I'm going to call you Mike. I know it is not Mike you just look like a Mike to me. Jim.
DR. BELLOWS: Hi, this is probably such a simplistic observation that it is not worth saying. But to me the boxes in green and the boxes in orange are incredibly similar. And actually the more we talk about it, it seems like they are almost the same. There is stuff in green about it has to be well-specified and there is stuff in orange about case-mix adjustment and so forth. But actually both of those things go in both of those boxes. So I wonder if we might want to just simplify it all by not repeating the green boxes and the oranges boxes.

And then for people who are just scared by the number of boxes on the page, there wouldn't be as many boxes.

DR. BRENNAN: One of the things that I have learned that is really important when you read an NQF document is to read the notes. So this might be arguing for fewer boxes, more notes, maybe. That is good. That is very helpful. Yes?
MS. TORDA: To go back to the kinds of other tests or the data that you threw out, I think we need to be clear. It needs to be data that was collected for non-research purposes.

DR. BRENNAN: Yes.

MS. TORDA: And what might be even more useful is somehow to identify sites that are actually using the PROM so that we would know where to go for testing, as opposed to a test bed of data.

DR. BRENNAN: I see. I see. So one of the things that you would find helpful is during the process of building up through box one, two, three, four, five, it would be useful to have some public list that you could refer to. So if a site in Maine is working on something you might collaborate with them.

The second thing I heard is that data collected from real world people in a real world activity is really scarce and important.
MS. TORDA: Yes.

DR. BRENNAN: Okay, I am going to go to the phone for just a moment, please. I understand we have a comment or a question.

OPERATOR: At this time, in order to ask a question press * then the number one on your telephone keypad.

Again, to ask a question, press * then the number one on your telephone keypad.

DR. BRENNAN: Well, all right.

OPERATOR: At this time, there are no questions.

DR. BRENNAN: Thank you so much.

Okay, we are going to come back to the room here. We have been talking about what you can't live with. And what I am hearing in the tenor of the conversation is people can't live with rigidity. There needs to be judgment, maybe there needs to be some use cases that demonstrate various types of PRO-PMs that come through and where their evidence will be presented and how they would
be treated in this model.

Anything else -- and somebody
doesn't like the green and orange being
separated. I got that.

Anything else? It's a show
stopper. It is just going to be not a good
thing. Yes, Ted?

DR. GANIATS: I think this is sort
of obvious, at least to me, but I haven't
heard it explicitly stated. Most, if not all,
performance measures come from a guideline.
There is a guideline that states X should be
done. We are de novo creating criteria for a
performance measure and have nowhere stated
that it should be from a guideline or that
there should be evidence that it should be --
there is evidence that it makes life better.

And I just think it is a little
late to work through all of that but it is
something for NQF to think about. You know,
just because it is a patient-reported outcome
doesn't, by itself, mean it is good fodder for
performance measurement.

DR. BRENAN: Very, very important point, Ted. And Karen is going to respond to that. Karen Pace.

DR. PACE: Just a clarification.

NQF does not require that for endorsement that something be identified in a guideline. Our criteria is about evidence. Many guidelines are evidence-based but what we are really interested in is the evidence behind a guideline a recommendation and that evidence doesn't have to be connected to a guideline recommendation.

I think the whole discussion we had here about actionability really relates to that evidence. What you are saying and many, that we have right now in our criteria that health outcomes don't have to present a whole body of evidence because they are kind of evident on our face that we should measure those. I think what we have heard here is we want some evidence of actionability, which
really gets that again, is there is evidence
that there are interventions that affect the
PRO.

And again I think in this
discussion someone brought up again some
flexibility because there may be PROs just as
health outcomes that by measuring it we will
hopefully stimulate improvements. But I think
that is a judgment call.

But I think we have heard pretty
loudly about looking this actionability,
especially at the start of this endeavor.

DR. BRENNAN: Yes, Kathy and then
Ted.

DR. LOHR: Another question back
to the pathway. If I were a developer or
wanting to do something to develop a PRO that
would be a PROM that would eventually get
submitted, is the implication here that
somehow or other I have to start de novo with
boxes one and two? Or is there a possibility
that I can take something that I have been
working on for the past ten years, has a lot of information behind it, I can pull together, if you will, proof that I did the things that I was supposed to do in box zeros, one and two, and short circuit both in time and effort some of what this implies would otherwise would need to be spent? And can people start with three?

DR. BRENNAN: I understand what you are saying. I think what you are asking for is to have boxes one, two, and three, be able to be operationalized in any order. And so I think that is a very interesting, very reasonable suggestion because there may be really, really good PROMs that haven't had enough patient input that get some patient input. There might be things that have had a lot of patient input but are fairly long down the track.

DR. LOHR: Right. And if I can prove to you that I have done a decent job with one and two, maybe then I can just go
ahead with three.

   DR. BRENNAN: This speaks well, though, to what one would consider proof or evidence, whether it is done de novo or it is historical. And I think that is very important guidance. Ted?

   DR. GANIATS: Yes, just in response to Karen. Thank you for making me relax and feel better and for causing me to get quite nervous.

   (Laughter.)

   DR. BRENNAN: Boy, is she a powerful woman.

   DR. GANIATS: The first part regarding evidence I am just ecstatic and you can see me dancing with excitement.

   On the other hand, I really don't like the other statement. I think that the performance measure or a measure that is supposed to go through the NQF should be a measure that is supposed to assure quality or assess performance or something like that.
For an NQF measure's purpose to be to be an intervention that is going to improve quality is something that I think we would have to talk long and hard about because I mean you said, gee whiz, maybe this measure would get people to do things better. I think that is an intervention in the practice and, personally, I didn't know that was part of NQF's charge.

DR. BRENnan:  Helen?

DR. BURSTIN:  It's not part of our charge, per se, Ted. I think it is just as we have gone through, particularly the issues around evidence, with our Board. And this is a multi-stakeholder group. We get a lot of different perspectives. It is what makes NQF very special.

You know we have heard clearly from consumers and purchasers in particular that there are clear instances where outcomes have been put out there. We don't always know what the right intervention is. And that by
putting it out there, publicly reporting on it, interventions begin to emerge.

So I think it is circular. It is not something linear. I think there is something about the process of public reporting, the process of learning about an income that then leads to interventions, that leads to -- I mean the classic example people give is the public reporting on central line-associated blood stream infections, which actually in fact preceded a lot of the actual interventions of what to do.

There are other examples like that. It is not part of our charge but I think we have to recognize that quality measurement and improvement is iterative. And we hope it is. And in fact, if putting some of those outcomes out there drives some of that improvement, I think that is a reasonable hypothesis.

DR. GANIATS: Quality improvement might be iterative but accountability, it just
makes me nervous to have an accountability
measure whose role it is to --

DR. BRENNAN: Albert is going to
fix this for us. Right?

DR. WU: Okay, Ted, I'm going to
fix you.

DR. BRENNAN: Uh-oh.

DR. WU: No. It does seem to me
that there are a number of sort of
aspirational outcomes.

DR. BRENNAN: Okay, I'm worried
about those already at other tables in this
place. So, yes.

DR. WU: No but I think that
insisting on accountability in three years or
five years or whatever for those measures,
particularly if all the processes aren't
specified is overreaching. But if there were
sort of a different -- since outcomes are
different, if the goal were simply to reduce
pain in cancer or whatever it winds up being,
to relieve dyspnea in chronic obstructive lung
disease, I wouldn't insist that that be an accountability measure in the immediate future. Maybe you get a longer time period. This is the sort of thing that would drive innovation, as opposed to pull people up to speed.

DR. BRENNAN: That is helpful.

Well we have gone through the things you can't live with and we have moved a little bit into the things that you would like to see added to at least the interpretation and application of this pathway. I want you to take one last look through it and see if there is anything that is unspecified or under-specified that you believe would be important to include in this pathway and our guidance about it to the NQF staff.

Yes, is that Lewis?

DR. KAZIS: Hi.

DR. BRENNAN: He looks like a Harry to me but I didn't want to say that.
Go ahead, Lewis.

DR. KAZIS:  Sorry?

DR. BRENNAN:  I re-baptized you.

DR. KAZIS:  Oh. So this goes along the lines of a box 14 again. And in addition to the concept of sort of a continued endorsement or recertification after a couple of years, perhaps to have something along the lines of on an annual basis to provide an annual update.

DR. BRENNAN:  Yes, I think that is part of the process now.

DR. KAZIS:  Okay.

DR. BRENNAN:  So users of the PMs have to state what they are doing.

DR. KAZIS:  Right and then maybe a recertification beyond that.

DR. BRENNAN:  Yes, okay.

Excellent. But making it explicit that there is a public trail. That's good.

Anything else under specified/unspecified? Yes, Mike -- no it's Jim, I
DR. BELLOWS: Well I made this point earlier but I would love to see the word harmonization somewhere on the second page. And I know it exists elsewhere in the NQF stuff but with respect to it being particularly important for stuff that we introduced into the system, I would love to see that word stuck in in some appropriate place and there is many appropriate places but we could figure it out.

DR. BRENNAN: We spend too much time at ONC harmonizing standards for me to really want to see that word ever again in my life. But what I would like you to be a little bit more specific and say if what you are meaning is that apropos of the discussion earlier of one measure/multiple measures, you are suggesting that we make explicit the need for if there are multiple measures that there be harmonization of some type.

DR. BELLOWS: That is correct or
maybe it also relates back to box zero, which
is what is the reason we are doing this. And
if we are just doing it because there is
another measure out there somebody wants, then
you could write that in box zero but whoever
is reviewing then measure might say is that an
adequate. But then if there is multiple
measures, then the harmonization kicks in.

DR. BRENNAN: So harmonization
explicitly. That's great. That's good.
Anything else underspecified/unspecified,
translated into three languages? Kathy.

DR. LOHR: I am not sure whether
folks would agree but to me it is still a
little unclear whether you are expecting
developers, when they have submitted a PROM to
have not only indicated whether risk
adjustment is needed for X, Y, or Z
applications but also specify anything about
the method for risk adjustment. I am not sure
that that is clear in here that all those
sorts of pieces of information would be needed
as well.

And I have some question in my mind as to whether PRO developers would necessarily have all the right information to say do your risk adjustment this way rather than that way. Or whether just saying the use of this across -- for accountability maybe not so much for internal quality improvement but across sites or across plans, you are going to have to risk adjust and leave it unspecified as to how. That is a question.

DR. BRENNAN: Okay, so the question is do we need to require that the proposer make explicit whether risk adjustment is needed and how to do it.

DR. LOHR: And how, yes.

DR. BRENNAN: Or is it acceptable for them to give a nod to it must be there but no specific plan. And Karen is going to make a comment about that.

DR. LOHR: And then I have one
other quick question.

DR. PACE: Okay, I just want to answer that question. In order to be a specified performance measure, that includes the risk model that goes with it. Because remember measures that are endorsed for NQF are endorsed to be suitable for accountability purposes. And so people couldn't implement an outcome measure that needs risk adjustment without having that risk model already specified with the measure.

DR. LOHR: Do they get more than one way of doing it? I mean Anne's paper has several risk adjustment models. Are you saying the developer has got to pick one or they can say you can do this, or this, or this?

DR. PACE: Pick one.

DR. LOHR: Oh, pick one, okay. I think I would think that might be tough but if that is an understood requirement, fine.
another box exactly but maybe it falls into
the same kind of bucket as harmonization and
that is whether in all of this process NQF is
in a position to say there is a research
agenda here and turn some of these kinds of
methods or application kinds of question,
whether technical or more political, back to
sort of point to some research funders -- it
might be PCORI, it might be AHRQ, it might be
others -- and say there is a substantial
research agenda here that somebody else has
got to follow through with. Because I think
this is another generation's worth of research
here.

DR. BRENNAN: But I certainly
think that what we have learned at least this
particular part of the NQF process is that
there is a lot of intersections with a lot of
different communities and the unique
perspective that NQF has to offer in offering
up new agendas is really quite important.

I am going to actually -- we are
at five of and I want to make sure that the
staff feel like we have got enough of a sense
of where to go and then I am going to let you
all have the last six minutes for comments;
two minutes apiece.

So Helen you can go and Karen and
Karen you can come and say your goodbyes.
You're fine? You're fine?

All right, then let me be the one
to say to the group speak now or it is going
into stone on the internet and it will be
there forever. Anything else you want to add?
If you think on the plane on the way home oh
gosh, they should have, make sure you let the
Karens know.

I hope to see you in a colorful
future and I thank you all very much for all
that you have done to get us to this point in
time. And thank you again to the staff. Safe
travels everyone.

(Whereupon, at 1:55 p.m., the
foregoing proceeding was adjourned.)
Page 244
29:22 51:12 53:12
RN 1:17 2:13,16,19
road 23:20,22 24:4
roadmap 141:19
Rob 2:25 25:2 47:1
62:4,6 87:20 88:9
88:14
Robert 4:13
Roberto's 95:11
robustness 58:4
Rob's 188:10
Rogers 82:2
ROI 11:19
role 54:6 61:15
63:10 65:3 230:2
rolled 178:8
room 1:8 91:8
109:21 113:3
115:4 122:2 124:6
145:21 184:17
214:6 222:15
Rooney 2:16
164:18,20 165:4
217:15 219:12,16
Ross 92:9,9
roughly 111:21
round 130:15
180:17 183:3
route 101:16 128:8
routinely 131:16
RTI 2:5 24:18
47:14 184:8
running 85:8 129:3
129:5
runs 151:4
rural 178:17
rush 13:3
S
sad 174:15
Safe 238:19
Safran's 177:13
sake 129:22 130:1
144:21
Saliba 2:17 43:6
46:13 68:10 73:1
73:2
SALIVE 2:18
sample 77:21
San 1:22 20:12
137:14 177:4
sat 190:6
satisfaction 23:18
satisfactory 57:4
saw 75:14 77:9
173:12
saying 73:4 88:19
91:19 100:4
107:11 123:5
129:21 130:2
136:15 157:1
161:20 171:2
195:11 196:7
197:13 198:20
206:8 209:15
210:3 211:19
213:4 214:13
219:2 220:3
224:16 226:10
235:6 236:14
says 160:5,17
186:18 204:7
scale 45:1 46:5,6
64:6 68:16 69:8
70:15,17 75:6,17
75:18,19,20,22
76:1 77:19 78:4
78:21,22 121:10
162:5,7 21 164:1
170:10
scales 70:13 71:9
71:19 75:6 76:3
109:21 120:17,22
161:14 164:15
169:13 170:3,7
scarce 221:21
scared 220:15
scars 212:20
Scb 2:2
schedules 5:19
scheduling 184:20
scheme 144:10
school 1:17 2:2,24
183:11
science 59:20 122:8
scientific 60:2,6
61:5 209:2 210:5
score 25:10 26:7
49:15,16 66:10
71:2,7 90:10
scored 90:12
scores 22:8,12,12
22:22 26:3 27:14
30:22 31:3 32:5
35:4 71:5 72:4
95:13
scoring 31:4
scornful 50:11
screen 188:22
screening 68:1
74:21
scribble 122:10
scribbled 124:8
script 43:4
see 228:12
second 9:12 37:4
84:22 105:19
106:5 118:6 120:3
137:18 192:5
193:9 219:8
221:19 233:4
section 26:17 32:18
sectional 105:14
sector 60:20
see 9:17 14:6 16:6
23:14 28:22 30:5
30:11,13 38:2
45:13 69:9,18
80:2 85:9 92:5
100:11 104:7
119:11 120:4
125:6 132:5
153:22 154:1
187:9 188:22
189:1,3 190:13
192:2 193:2
197:20 203:8
213:20 221:12,12
227:16 231:11,14
233:3,9,14 238:16
seeing 164:13
179:22 218:1
seen 16:10 96:22
146:21
sees 26:22
segment 76:19
segments 218:4
select 11:10
selected 45:20
52:17 127:4
selecting 52:9,13
53:17 55:17 134:9
193:1,13
selection 66:12
selects 8:14
self 39:14 41:8
76:15 162:8
self-report 33:22
34:2 108:2
selling 137:18
171:13
semi 113:20
send 76:18 150:16
183:20 184:5
senior 38:14
seniors 38:21
sense 8:8,9 13:6
65:18 66:15 80:22
110:5 113:9,17
123:18,20 125:19
127:7 144:22
187:2 189:22
200:20 238:2
sensitivity 45:10
108:18 177:9
separate 82:14
separated 223:4
SEPTEMBER 1:6
sequence 152:12
serious 109:15
serve 150:16
211:20
service 188:2
services 1:24 2:8,14
8:21
serving 63:17
66:22 179:21
SES 34:17
session 182:4,17
186:6
set 42:16,17 60:11
73:4,22 75:5 79:4
82:18,22 83:4,21
95:22 112:16
134:12 163:13
202:13 211:1,13
sets 57:14 189:18
setting 22:16 23:9
23:19 58:21 114:9
115:14,20 125:12
157:13 187:18
204:3
settings 54:16
114:18 144:17
169:19
settled 22:6
severe 38:20 39:2
42:12 46:1,8,8,17
46:18 77:11
severe/horrible 70:18
severely 35:2 73:21
79:3
sexual 107:16
SF 174:18
SF-36 81:21 173:3
173:4,5 174:10
Shaller 166:13
share 155:3,13
shared 82:6 131:14
sheet 103:11
SHEILA 3:17
shifted 169:11
shoes 140:1
short 33:1 142:18
205:11 226:5
shorthand 8:3
show 75:19,21
167:15 174:19
199:2,21 223:5
showed 14:5,7
132:13
showing 52:19
shown 44:14 177:8
shrift 142:18
underlying 171:5,6
underspecified 200:11
underspecified/under-specified 234:11
understand 8:15
understanding 35:9 52:3 67:6
understood 58:6
use 8:5
useful 11:2 56:20 82:9 96:14 99:19
Vanna 104:2
variability 74:11
variables 34:17
varied 197:22
variety 146:17 147:6
vary 41:14 200:15
vendors 60:13
vendor-driven 12:6
valid 24:6 25:20
validity 4:4,6 13:21
value 80:20

Neal R. Gross & Co., Inc.
202-234-4433
written 9:7 43:4
51:10 110:7 184:3
wrong 76:14 77:14
WU 3:9 177:22
230:5,8,14

X
X 203:5 223:12
234:18
XYZ 160:15

Y
Y 160:19 234:18
Yale 2:22
yardstick 203:14
year 165:13,13
214:14 215:2
218:11,17
years 58:1 149:7
165:8,12 171:11
173:18 203:5
204:22 205:1,7,11
226:1 230:15,16
232:8
yesterday 5:6,17,20
7:12 10:9 21:22
23:13 25:9 33:8
38:11 42:14 63:16
81:18 96:12 98:6
100:1 102:5,15
103:10 106:4
151:19 197:12
yesterday's 6:19
young 40:3

Z
Z 234:18
zero 46:5 64:6
68:16,21 75:17,17
75:20 78:3,8,20
138:6,11,13 139:6
140:9 149:3
156:16,19,20
157:2 179:11
186:18,19 187:4
234:1,5
zeros 226:4
Ziman 59:21

---

2 4:2 24:13 25:8
20 96:8 144:7 149:7
2012 1:6
2013 111:16
238:4:25
25:4 84:19 96:8
26th 184:17,22

---

3 3,000 77:21
30 68:10
30 87:1 149:7
33 4:12
36 174:19

---

4 40 189:8
45 181:5
450 73:5 79:16
48 4:13

---

5 5 109:20 115:10
50 64:12 178:11
500 213:7
55 178:11

---

6 6 4:2 113:16
60 121:5
63 4:13
64 12:19
65 125:22

---

7 7 113:16
73 4:14

---

8 8 113:16
8:30 1:9 5:2

---

9 9 113:16 206:8
9th 1:8
90 81:22
99 4:16
CERTIFICATE

This is to certify that the foregoing transcript

In the matter of: Patient Reported Outcomes Workshop 2

Before: NQF

Date: 09-12-12

Place: Washington, DC

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

__________________________
Court Reporter

NEAL R. GROSS
COURT REPORTERS AND TRANSCRIBERS
1323 RHODE ISLAND AVE., N.W.
WASHINGTON, D.C. 20005-3701
(202) 234-4433
www.nealrgross.com