THE NATIONAL QUALITY FORUM
+ + + + +
IMAGING EFFICIENCY STEERING COMMITTEE

MEETING

+ + + + +

TUESDAY
FEBRUARY 23, 2010

The Imaging Efficiency Steering Committee met in Suite 600 North of the Homer Building, 601 13th Street, N.W., Washington, D.C., at 9:45 a.m., Scott Gazelle and Eric Peterson, Co-Chairmen, presiding.

PRESENT:
G. SCOTT GAZELLE, MD, MPH, PhD, Co-Chairman
ERIC D. PETERSON, MD, MPH, Co-Chairman
MICHAEL BACKUS, Member
JACQUELINE A. BELLO, MD, FACR, Member
STEPHEN V. CANTRILL, MD, FACEP, Member
CARL D'ORSI, MD, Member

TROY FIESINGER, MD, FAAFP, Member
HOWARD FORMAN, MD, MBA, Member
MARY GEMIGNANI, MD, Member
RAYMOND GIBBONS, MD, Member
RICHARD GRIFFEY, MD, MPH, Member
LASZLO MECHTLER, MD, Member
PATTI RAKSIN, MD, Member

DONALD W. RUCKER, MBA, MD, Member
GAVIN SETZEN, MD, FACS, FAAOA, Member
REBECCA SMITH-BINDMAN, MD, Member
ROGER L. SNOW, MD, MPH, Member
KIRK T. SPENCER, MD, Member
ARTHUR STILLMAN, MD, PhD, Member
JUDY ZERZAN, MD, MPH, Member

HELEN BURSTIN, NQF
IAN CORBRIDGE, NQF
SARAH FANTA, NQF

T-A-B-L-E O-F C-O-N-T-E-N-T-S

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

9:39 a.m.

CO-CHAIR GAZELLE: Good morning, everyone. It is five minutes early, but everyone is here. So we are going to go ahead and get started, and maybe that means we can finish on time at least.

My name is Scott Gazelle.

CO-CHAIR PETERSON: And Eric Peterson.

CO-CHAIR GAZELLE: And we are the two Co-Chairs of the meeting. So on behalf of the NQF and us, thank you for agreeing to participate and for all the work you have done before coming to the meeting.

Helen, do you or Ian want to say some comments about format?

DR. BURSTIN: Sure. Happy to. We will talk a little bit further about the actual contents in a little bit. I just want to at least add my welcome.

Helen Burstin. I am the Senior

Vice President of Performance Measures at NQF.

In case you can't tell, we literally just opened up this conference room on Friday. They unpacked the table. There is still duct tape on the floor. We really wanted to try to have in-house meetings rather than always having to rely on hotels, and again get you some wireless to be able to get your materials in real time.

I apologize for our measure developer friends for being a little cramped. We will work on that next time. It has literally just been since Friday. So let us know if you need anything.

Again, I just want to add my welcome to the Chairs. This is, obviously, a very interesting project, very diverse, lots of expertise required, which is why, actually, the Steering Committee is a bit larger than some of our prior ones. We aim for 15 to 18, but just really felt, given the diversity of measures, we wanted to be sure we had the

right expertise at the table.

So thank you all for coming, and we will get into more details to follow, but in terms of just logistics, there is food, coffee right there at the side over here. Let Ian or myself know, or Sarah, if there is anything you need, and bathrooms are right out to the --

MR. CORBRIDGE: Women's are right out to the right, gentleman's to the left. You need a key. If the key is not there, you might have to do a handout as you go in there.

Just kind of some other housekeeping stuff: There is a coat closet in the back, if you want, and just wanted -- Before we move forward, I wanted to make sure that everyone was aware that all of NQF's workings are open to the public and recorded. So everything that is said within this room and discussed is actually being recorded. Donald over there who takes care of all our AV technical stuff is recording all the

information.

So individuals on the phone can hear as well as later on, if individuals from the public or the Steering Committee want to listen to the actual recording, and there is also a transcript available as well. So that is just one housekeeping thing to keep in mind, that what you do say today is recorded and will be available to the public.

Another housekeeping that I want to just bring to individuals' attention -- I just was aware of it. Across on the south side there is Toyota, and I think the hearings are happening. So if you see reporters and cameras in here, it is not because of this meeting right now. So we are okay at this time. I just want to bring that to people's attention now, that there may be film crews here today. Hopefully, I think they are going to be on that side.

One other thing, I guess, for individuals who want to access the Internet,

if you haven't already, it is the Homer Building. There shouldn't be any lock to it. So it should be free to get on line.

We would like to start off with introductions. I know not everyone was able to attend. There an introductory phone conference.

CO-CHAIR GAZELLE: So we should go around the room and introduce ourselves. I will start. My name is Scott Gazelle. I am an abdominal radiologist by training. My PhD is in health policy, and most of my research is new technology evaluation.

I was on the prior committee. This is my second time on the metrics effort.

CO-CHAIR PETERSON: Eric Peterson. I am a cardiologist by training, but have no imaging background whatsoever. I am the random assortment here. I also do outcomes research and I'm associate director at Duke Clinical Research Institute.

DR. SPENCER: I am Kirk Spencer.

I am a clinical cardiologist with expertise in echocardiography, and I do work on advocacy for the American Society of Echo.

DR. ZERZAN: Judy Zerzan. I am Colorado Medicaid Medical Director. I also do a little research on Medicaid prescription policy at the University.

DR. MECHTLER: Hi. I am Laszlo Mechtler. I am a trained neurologist with subspecialties in neuroimaging and headache and neuro-oncology, and I have been running a fellowship program in imaging for 20 years at the Headache Center.

DR. RAKSIN: Hi. Patti Raksin. I am a neurosurgeon with Critical Care at Cook County Hospital in Chicago. I am here as a representative of the American Association of Neurologic Surgeons Joint Guidelines Committee.

DR. BELLO: I am Jacqueline Bellow. I direct the Division of Neuroradiology at Albert Einstein and

Montefiore Medical Center, and I run a fellowship training program there, and I am on the ACR guidelines Committee.

DR. FORMAN: I am Howie Forman. I am a diagnostic radiologist practicing primarily in emergency room, trauma imaging, and I teach health policy and health economics at Yale.

DR. RUCKER: Don Rucker, Chief Medical Officer for Siemens. We, as I mentioned in our disclosure sheet, manufacture, I believe, all the devices under consideration here, and so I am, in some perverse sense, neutral, and I am also on the clinical faculty at the University of Pennsylvania, Emergency Medicine.

DR. FIESINGER: I am Troy Fiesinger, a family physician in Houston. I am on residency faculty at the program there, and I am here on behalf of the American Academy of Family Physicians. I have been on their Commission on Quality for the last four

years.

DR. SMITH-BINDMAN: My name is Rebecca Smith-Bindman. I am a radiologist at UCSF. My research focuses on outcomes and the benefits and benefits of a range of tests.

DR. D'ORSI: Carl D'Orsi. I am a diagnostic radiologist. I have been doing breast imaging for 20 years, and my research interests are basically in technology assessment, comparing various technologies for detection of early breast cancer.

DR. GIBBONS: Ray Gibbons, staff cardiologist at the Mayo Clinic, standard experience in national cardiovascular disease guidelines and cardiac imager, primarily in nuclear cardiology.

DR. SNOW: I am Roger Snow. I am internist and the Deputy Medical director for Mass. Health, which is Massachusetts' Medicaid program.

DR. STILLMAN: I am Arthur Stillman. I direct the cardio-thoracic

imaging at Emory, here representing -- at the request of American College of Radiology.

DR. CANTRILL: Steve Cantrill, emergency physician from Denver. I have been involved in clinical guideline development and also quality performance measure development, representative from American Academy of Emergency Physicians.

DR. SETZEN: My name is Gavin Setzen. I am a practicing otolaryngologist in Albany, New York, and am here as Chair of the Board of Governors of the American Academy of Otolaryngology -- Head and Neck Surgery. I am also involved in guideline development and on the Board of the Intersocietal Commission for the Accreditation of CT Laboratories, ICACTL.

DR. GRIFFEY: I am Richard Griffey. I am an emergency physician at Washington University in St. Louis. I did my MPH in clinical effectiveness, and do work in quality and safety.

MR. BACKUS: My name is Mike

Backus. I am with American Imaging Management, which is a subsidiary of Wellpoint. We manage radiology and cardiology preop for about 35 million Americans. I am in charge of analytics and medical economics.

DR. GEMIGNANI: I am Mary Gemignani. I am a breast surgeon at Memorial Sloan Kettering Cancer Center. My primary research interest is in screening for high risk women. I was on the previous NQF meeting.

CO-CHAIR PETERSON: Great. I think what we have heard as you go around the table, there is a lot of varying interests, and to the credit of NQF, they've got a diverse group of people who might, outside of here, be on opposite sides of various arguments, or most any argument. We could find some diversity of opinions around the table.

What I would like you all to consider, though, is why you might have got on

this committee, because you represented a certain group or a certain field or even have your own self-interest, unfortunately, in these fields.

Today you are here as a physician or a policy person who is trying to do the right thing for medical care, and I would like you guys to really keep that in mind as you think about the deliberations over the next two days.

We all have -- these have major implications in theory or in reality for American medicine. They can be remarkably positive effects in terms of creating a system of care that will improve major outcomes and make it affordable to do in a right manner.

We all realize there are certain things wrong and broken in the current system. It is our responsibility, and those for the next generation who will have to deal with these, to make wise decisions.

Sometimes you may have to make

compromises in things that would be near and important to your field or your profession or even sometimes your belief system, but today the main thing is come up with the answer that you believe is ultimately the right one when you leave the meeting.

DR. BURSTIN: We have some folks in the back.

CO-CHAIR PETERSON: Sure. Go ahead.

MS. STEPHENS: I am Sharman Stephens, and I am with the Lewin Group, and we are serving as a contractor for the Centers for Medicare and Medicaid Services.

MS. PETERSON: I am Laura Peterson. I am also with the Lewin Group.

MS. DaVANZO: I am Joan DaVanzo with Dobson, DaVanzo Associates.

MS. ARDAY: I am Susan Arday. I am with the Centers for Medicare and Medicaid Services.

DR. DEHN: Hi. I am Tom Dehn, a

radiologist, Chief Medical Officer of National Imaging and a consultant with CMS.

DR. BRUETMAN: I am Dr. Bruetman. I also work for the Lewin Group.

MR. PENTACOST: I am Michael Pentacost. I am one of the medical officers of National Imaging, subcontractor for CMS.

MR. BASSETT: I am Larry Bassett, director of Imaging at UCLA. I am here to represent for the American College of Radiology.

MS. WOUTERS: I am Ann Marie Wouters.

MS. COOMBS: I am Laura Coombs, I am the director of data registries of mammography at the American College of Radiology.

MS. BURLESON: I am Judy Burleson, Director of Metrics at American College of Radiology.

MS. GROMAN: Rachel Groman, the Senior Manager of Quality Improvement and

Research at the American Association of Neurological Surgeons.

MS. DUNLEY-GALLIGHER: Rita Dunley-Galligher, Senior Policy Fellow at the National Center for Nursing Quality at the American Nurses Association.

MS. FANTA: Hi. Sarah Fanta, Research Analyst at the national Quality Forum.

MR. CORBRIDGE: All right, thank you. I guess I would just like to just bring your attention, two individuals who were initially on the Steering Committee were unable to attend today. So that is Dr. Patricia Kunz Howard as well as Marilyn Kramer. So they were unable to attend today, just to let you know that.

In terms of just moving forward, I want to make sure that everyone has the actual paper copy of NQF's Measure Evaluation Criteria. I know I tried to pass that out as individuals came in the door, but if you are

missing it, we have copies here. I will just pass some. Do you know how many we need down there?

This is just a paper copy of the digital PDF that you were provided. It is just NQF's measure evaluation criteria. Hopefully, it will be helpful in terms of reviewing and reviewing the measures to be able to look at NQF's criteria.

It seems like we are way ahead of schedule. I know I was here at 8:00 o'clock, and people started showing up. So I was quite surprised. It is quite an eager group.

So we are ahead of schedule. I think at this point, we would really like to just touch on some of the points that we looked at in the introductory conference call, go over that just quickly, some of the key highlights of the project, and then we will move forward from there.

DR. BURSTIN: We are going to skip over a lot of the stuff we did on the call.

MR. CORBRIDGE: Okay. So as we mentioned, this is some of the information that we discussed as well as had on the webinar for our introductory call, just going over some background of the project.

It is part of a sub-task of the larger HHS Resource Use Project. This project is specifically with imaging efficiency, which makes it different from the other projects that are primarily within resource use across episodes of care.

Really, one of the main focuses of this project is to expand NQF's current portfolio of imaging efficiency measures. I indicated at the last project, which Dr. Gazelle participated with, I believe there was eight endorsed measures that came from that.

We are really looking to expand NQF's measurement domain in terms of imaging efficiency, as well as to identify gaps within the field which the Steering Committee identifies are key areas that we need in terms

of measurement moving forward, and helping to support health reform.

So just some goals of the project: As identified earlier, to identify and evaluate and endorse additional measures suitable for public reporting and quality improvement which specifically address imaging efficiency.

I just want to bring to your attention, as we discussed earlier key parts of NQF's process is the public reporting and quality improvement. So that is a lens that each member of the Steering Committee will need to look through in terms of evaluating the measures. Are they available for public reporting, and is the measure really intended to improve quality within a specific study or in cross-settings; and then as touched upon earlier, really to identify gaps within imaging efficiency domains.

So just the scope: These are kinds of specific domains. When we put out

the call for measures, these are some areas that we touched upon, trying to elicit some measures. We didn't get everything that -- All the responses didn't touch on these areas, but we got a very robust set of measures, I think, that came to us.

So some areas we focused on were overlap screening, patient safety. You can see here. So looking at past projects, as I talked about, we had an imaging efficiency project in 2008. At the end of that, we walked away with eight NQF endorsed imaging efficiency measures, and they went across different focus areas.

For the current projects, the measures that came to NQF for the call for measures, we kind of looked at them in different buckets. The review group kind of based on those buckets, and we tried to sit you with fellow reviewers within the specific group that you were looking at.

We had measures touching on

cardiac imaging, mammography, measures focused on the emergency departments, fine CT as well as the coordination of care.

So this next couple of slides will just go over the process of what the actual Steering Committees expect to do and participate with NQF, and then what NQF's role is within the projects.

At this point, you can look at the top kind of bar. In the center, the projects have really already been specified. We are moving forward. At this point, we are now really at the Steering Committee review of measures submitted to NQF.

Some Steering Committees -- there is a Technical Advisory Panel that supports them. Just due to the smaller set of measures that we received, we decided to just really have a Steering Committee.

Really, in some groups we have broken out into different review groups, and they have come back and reported, but for the

flow of proceedings with this Steering Committee, we are hoping just to be able to take everything at the table.

We will have the primary review group really lead and elicit the discussion for a specific measure to which they are assigned, and then have the rest of the Steering Committee really add to that process.

The next step would be we are looking at drafting recommendations throughout this whole process at NQF. We are taking notes. Everything will be recorded. We will have transcripts. We will go back and record the conversations. We will have a meeting summary that will be provided online, and the Steering Committee's input will really be key in coming up with that meeting summary.

From that, we will move forward into actually drafting recommendations. They are put online for review and comment from the public, and then moving forward we will come up with actual recommendations for then voting

and CSAC and Board approval, and then we will come up with an NQF endorsed set of imaging efficiency standards. At the end of that, there is an appeal process.

So NQF has moved toward really trying to have complete transparency through our -- really, at each step everything is open to the public, as well as there are opportunities when information is put online for the public to respond.

So any type of public comment that we get, that will be forwarded on to the Steering Committee. So we hope that you guys will be able to help us respond to those comments.

So just going over a little bit further, I know we talked on some of these. Obviously, you are representing a diverse set of stakeholders, and really, I guess the main goal today is really to evaluate the measures that came forward to NQF, based on NQF's criteria, and make recommendations to move

forward.

Then the Co-Chairs are actually -- time permitting, will be there to represent the measures that are potentially endorsed for CSAC.

Then the role of NQF staff here: Really, the staff are here to support the Steering Committee and providing documentation, providing kind of a conduit to the measure developers, and providing access to information the Steering Committee needs to really make the rational and best decision that they need.

Then really, another function is to help along the process of drafting reports and posting that onto the web so individuals from the public can respond to it, and another key part is to just maintain the documentation in the documentation as it moves through this process, making sure that we have sufficient notes and documentation to capture what the Steering Committee recommended to move

forward.

Here is a brief project timeline that we are looking at. Obviously, December and January dates already took place. We've had the measures. We have formed the Steering Committee. We had introductory call, and then coming up in April and May, we are looking to move toward a comment period, then moving toward member voting, and then those measures which we may determine to move forward then would go to CSAC in July, and then NQF Board endorsements on July 28th, after which there is a 30-day appeals process.

So that is just a brief rundown of the project's timeline, as well as the project as a whole. Any questions from the Steering Committee about the process, timeline? Yes?

DR. D'ORSI: I don't know if it is particularly -- excuse me, Carl D'Orsi.

These metrics are meant to evaluate efficiency and quality for individuals, facilities, or both?

DR. BURSTIN: It actually depends on the measure itself. I think the majority of these measures are facility level measures. There is a specific part of all the mission forms that specifically ask the developer to note the appropriate level of analysis. That is a really important question, Carl.

So as you review those measures, please keep an eye on whether that is a measure that would be very appropriate for public reporting with QI at the facility level, and then consider whether rolling that up or down makes sense. It is a really important point.

DR. SMITH-BINDMAN: If there some back and forth period with the developers of the measures where we could provide some impact on how to improve them?

DR. BURSTIN: I'm sorry. I was just going to go through a couple of additional things, just to emphasize your role today.

So part of what -- again, really emphasizing the point Eric made at the outset, although you bring a very diverse stakeholder perspective, you are here because you bring expertise to the table. We want you to really help us evaluate the measures, see if they are the right set of measures to move forward.

The criteria that you were given in this handout -- we have tried very hard over the last few years to increasingly make them more objective, make them things that you could truly be able to rate overall and, again, because we are so transparent, give more information to the end users who are going to be able to look at this, evaluate it, see if they agree or not.

You should know that on all these projects, we are probably averaging, oh, over 300 comments that we will receive from the public and members. So there is a very alive -- which is a wonderful part of the process, but it means there will be a lot of back and

forth, even post this meeting, once we get through your initial process.

As much as possible, your evaluations are completely brought into these evaluation criteria, and I am happy to answer any questions as we move forward through those.

Your options after each discussion -- I want to spend a moment or two on that, because I think it is an important piece of this, and thank you for bringing that up. You have the option of, at the end of the discussion, if the reviewers who reviewed the measure, after the discussion of the Steering Committee, you can say we recommend this measure move forward. That is the role of the Steering Committee.

What that means is it will move forward through the rest of the process. Now all measures go out for public comment, not just those that are recommended. We made that change about a year ago. So we will get

public comments on even the measures you say shouldn't go forward, and you will be able to reflect on those.

Every once in a while, the Steering Committee sees the comments and says, oh, that is an aspect of this that we hadn't really thought about, and may make some changes, but in general, you will overall recommend the measure.

You have the option of recommending the measure with conditions, and this is really the point, I think, that you are trying to make. There may very well be clear opportunities to improve the measure, based on your expertise.

You can't rewrite the measure. That is not appropriate, obviously. You can't create a new measure. That is not appropriate either. But you can very much make recommendations to the measure developers.

They oftentimes can't on a dime say, yes, we can do that, but we give them an

opportunity. After the meeting we will write up all the details of what your recommendation with conditions are. They will then have a chance to respond to you. We will share that with you, and then you can make a decision as to whether you would continue to recommend the measure, if the conditions have been met.

If the conditions weren't met, you then have the opportunity to say, okay, we will accept it as is, or you could, in fact, make the decision to not recommend the measure.

The other opportunity I want to mention is that there are a fair number of measures, I think, within this dataset as well, within the set of measures that have not yet been tested. So NQF does have a time limited endorsement policy, which specifically allows measures that have otherwise passed all of the other evaluation criteria. This isn't endorsement lite.

This is really, you have done

every other aspect of this with the exception of the fact that you don't have adequate reliability and validity testing yet. Since the measure is brand new, hasn't been in the field perhaps, there hasn't an opportunity to do that yet, you also have the opportunity to recommend the measure go forward as time limited.

We, up front as staff, have actually gone through it and at least indicated is there testing here or not. It is not as if you can recommend a measure that could go forward fully if it, in fact, has no testing.

So those are your options, and we will work with you to be spelling out those conditions, but again we can't just say recommend with conditions and be vague. If it is really recommend with conditions, there has got to be two or three things: This definition isn't quite right; the denominator needs tweaking, you know, things that are very

discrete that we can hand back to the measure developer based on the guidance of the Committee.

CO-CHAIR GAZELLE: Helen, I just wanted to comment. One issue that came up, I know, in the past is where is the line between, sort of, recommending changes and rewriting? So a number of the measures that we reviewed had internal instances -- for this one, had internal instances where, for example, the title, the definition in that one sentence title was inconsistent with the numerator and denominator, where to clear up that, that doesn't count as rewriting the measure. That just counts as with conditions.

DR. BURSTIN: Absolutely.

DR. SETZEN: One question. Gavin Setzen. With respect to the handling of the comment period when we have the comments, what are the mechanics and logistics in terms of how those are dealt with, with respect to staff and the Steering Committee itself?

DR. BURSTIN: So what we will do, what Ian and Sarah will do is take all those. We will put them into a big spreadsheet for you. We will go through the recommendations initially. We will make some recommendations. Most of them are "thank you for your comments" or we will specifically highlight ones that say Steering Committee needs to review it and make a decision.

So we will highlight that. We will have a conference call with you where we will go over the entire comment table, highlighting the ones where there is clearly an issue where there is an expectation the Steering Committee would need to reflect on it, as opposed to more mechanical things that we can do back and forth for you with the developers.

So as much as possible, we will try to reserve your time for the areas where we think we need your expertise, and we will make more of the mechanics the work of NQF

staff.

DR. SPENCER: So the steward of the measure, if we think it needs some minor changes, can change it and still save it for this site.

DR. BURSTIN: Exactly.

DR. SPENCER: It is not like we say no, and then --

DR. BURSTIN: Right. No. So that actually part of the logic. You may have wondered why we are meeting in February, but it is not going out for comment until mid-April. That is to allow the back and forth with the developers. That will also be for us to draft the draft report that goes out with the measures.

So what goes out in our draft report will, in fact, be after the back and forth with the developers. You have seen it. You have agreed it met conditions, and that is what goes out. So that is why there is a little bit of a cushion in there for us to get

that work done.

MR. CORBRIDGE: And today there are opportunities, because many of the measure developers are here today and tomorrow, to actually discuss with them, kind of work out some of these issues up front, and then move forward, and we can have that back and forth comment period later on, if needed.

DR. BURSTIN: But again, we can't rewrite measures. We can't completely say this doesn't work, but if we did it this way. Now the one thing you will have the opportunity to do as well, which is actually becoming, I think, increasingly important, is that at the end of the discussion -- all through the discussion we will be kind of culling from your comments what are the measurement gaps? What are the measures that, boy, we really wish they had come to the table.

Then part of this draft report and final report that we will put out will

actually be a set of what we call research recommendations or measure recommendations. They may not have been in this set, or maybe if you had completely rewritten measure A, you would have really gotten this measure, and that would be in those research recommendations.

So keep in mind as you are going through it, as you can see, for those -- several of you who were on the first part that we did on this, you know, this is a fairly new area. Oftentimes, it takes a few cycles to really put out to the measure development field. There is really -- we are part of a supply chain.

So as much as we can help support the supply chain and say the experts say what we really need is a measure on why, we are happy to put that out there, give them time to let that work come through a process, which can take up to a year, especially for measures that are tested, and to then have another

opportunity in the future to bring back those measures.

So the other thing you should know is we didn't really talk about it very much, but we also are always trying to refresh the overall portfolio. So even if you endorse a measure at the end of this process, it is only endorsed for three years, and it is endorsed only for three years because the expectation is that evidence base changes.

Things happen such that, if you look at most guidelines, the recommendation is about three years is the general right amount of time when there is a -- you know, you are going to look at guidelines, and generally you would probably want to revisit them.

So even if that measure goes through, it is still going to get another look. Secondly, we also have an ad hoc review process. Again, just keep in mind the evidence, particularly for some of these areas and some of these guidelines change so quickly

that we also have the capacity that, if any member or any public or anybody out there says, you know, this measure no longer works, this guideline has changed -- the study indicates the evidence would suggest this actually leads to unintended consequences of measurement -- we have the chance to go back and re-review the measure off-cycle.

So one notable example was that a measure that had patients getting antibiotics within four hours of hitting the ED for pneumonia -- lots of unintended consequences with that measure, lots of little old ladies with PHF getting a good slug of antibiotics --

DR. FIESINGER: Antibiotic resistance.

DR. BURSTIN: Yes, antibiotic resistance, and we -- you know, as soon as a lot of those articles began, that evidence began coming out that there was a problem there, we quickly worked with the measure developer. We did an ad hoc review. A

revised measure was put forward that had a provisional diagnosis of pneumonia required as well as a six-hour window.

Again, so we can make those changes. We try to make it such that the portfolio really has currency and that we are trying to get it best in class.

Also, if a better measure comes forward within that period of time as well at the time of maintenance, we have the opportunity to refresh the portfolio as well, and say, okay, that measure may have worked for now, but it is all we got; there is a better measure down the road, and we will try to refresh the portfolio going forward. Long answer, sorry.

DR. SMITH-BINDMAN: I know I am going to ask this later. So I might ask it in a general sense.

If we feel the need for risk adjustment -- you used to have them. Is that a minor -- Is that a rewrite or is that as

long as they can accommodate the writer to change?

DR. BURSTIN: No, it really depends on what we are talking about. If you are asking, I think, somebody to add a risk model that doesn't exist, that seems like a pretty significant rewrite.

If, on the other hand, the data is already stratified and you are saying, you know, you should really add age and gender or something like that, that might be something they would be able to accomplish and put that. But you couldn't add a risk adjustment.

DR. SMITH-BINDMAN: I am having a hard time understanding what rewriting the measure means versus adjusting -- not to put work in our hands, but why can't we rewrite the measure a little bit? Is that not in our --

DR. BURSTIN: Well, first of all, you know, you need to respect the fact that the measure developers have often spent up to

a year coming up with this measure. They have had advisory committees. They have had lots of logic for the reason they put the measure together. So you want to give them an opportunity to go back to their advisory committees and say, okay, this is what the committee said.

And secondly, you know, if it is really a different measure, that is one of the sort of clear lines in the sand for NQF is, because we are part of the supply chain, we don't do measure development. I think we try really hard to stay on the side of saying, okay, the measure is before us. You know, it either works or it doesn't. Maybe there are some fairly minor changes, and again it all depends on the measure developer as well.

We have seen some measure developers being somewhat saying, okay, fine, we will take the changes; we just want to make it done. And if they can do it in the time frame, and even if they are sort of bordering

onto being more significant changes, that is fine. But again, it is a back and forth. We can't force the developers to make changes. They still have the opportunity to come back and say, no, and you have to make a decision at the end of the day.

Any thoughts from anybody who has been through this process want to comment?

DR. RUCKER: This is helpful. I think it is not well known that NQF doesn't actually primarily generate the measures, just as an out there in the world kind of comment.

CO-CHAIR GAZELLE: I think my experience on the last one was that there were a portion where we came to very clear consensus of what needed to happen to make the measure better, and on some of those the measure developers agreed and were able to respond, and those measures went forward.

In others, either the measure developers didn't agree or the changes were so large that they couldn't be accomplished, and

I think in that latter group we have seen some of them come back this time.

DR. BURSTIN: Yes.

CO-CHAIR GAZELLE: I think we will see that with the mammo measures where we have made specific suggestions that couldn't be accommodated in the review cycle, and so we are now seeing them in the next cycle.

So I would say that is indication that the process is working in all of the different ways that it is intended for.

DR. CANTRILL: Steve Cantrill. As was talked about before, I think you potentially get better measures if there is a larger lag time between the call for measures and when you start looking at them. Some folks may have been working these for a year, as you say. Many of us only found out about it in December, which is a very, very tough window to produce a quality product.

DR. BURSTIN: Right, and one of the things we are doing, which is a broader

sort of NQF approach, is we are actually trying to move toward more of an expectation of a slight goal of when measures will come up for both new measures as well as maintenance, and have come up with -- it scares me a bit, but there's about 28 committees that would need to meet over a three-year period of time.

The idea would be -- I mean, in some ways it may replace some of these sort of quick ad hoc, get these things in quickly, but if you knew, for example, that cardiovascular was happening in 2010 and is happening again in 2013, it gives a better window to say when you can prepare for the next cycle.

So that is definitely our emphasis as well. It also then allows us to have the same cycle to look at what is currently endorsed and what is submitted.

One of the difficulties we get at times is a measure may already be part of the portfolio. It is not up yet for maintenance. It has only been in the portfolio a year and

a half or two years, and yet a better measure came in.

So to really say at the end of the day we have best in class measures, we have to have that capacity to do those head to head comparisons with all measures being at equal footing, both new and currently endorsed. That is what that -- so the change in mindset is moving toward us. We are getting there.

CO-CHAIR GAZELLE: I know we are ahead of schedule. Is there any reason not to move on to the mammo measures?

MR. CORBRIDGE: No, there is not, actually.

CO-CHAIR PETERSON: Since I am on the mammo group -- one thing we learned last time was it takes us a lot longer to do the first ones than the others, because we are all orienting ourselves to the process, to each other, and what-not. So I will try to do that with benefit of how this worked last time.

The other thing I will say is that

we are all here because of our particular expertise and background, but we are all here also to participate in the whole process.

So even though you may be a cardiologist or a neurosurgeon or have expertise in an area other than mammo, now is the time to become a mammo expert and to be engaged in the discussion about the mammo measures, because that is the idea of the process.

All right. So we have five mammo measures to consider today. Four of them are proposed by the American College of Radiology. One of them is proposed by CMS.

At the prior meeting of the Steering Committee, one measure we considered was the recall rate, and the short story from that meeting was that we felt the recall rate was not a good measure in isolation.

The specific discussion was lengthy, but we felt that, for recall rate to be a useful measure, it needed to be paired

probably with cancer detection rate and a PPV2, which we will get to. So the measure developers have -- because they really couldn't do that in the time frame -- have come back with a suite of measures that we are here to discuss.

Because they all relate to each other, I think how we should proceed is we will have a brief discussion from the primary reviewer of each metric, what it is, what its strengths are, what issues might either relate to its definition or its applicability, some comments.

Then we will move on to the next measure, if we could, because my suspicion is what we will end up recommending is that we can't approve one without some combination of others, but that we probably don't want all of them.

DR. BURSTIN: Yes. Just one qualifier. It would be very helpful for us, as the primary reviewer goes forward, to actually

give their ratings of the criteria. Again, you want to keep it very grounded and make that very transparent.

CO-CHAIR GAZELLE: But I think for each primary reviewer, as you go through, even though I know all of us who reviewed the mammo measures have comments about the others, we should try and focus just on a run-through, knowing that we will come back and go through them all as a suite.

So the five we have are Number 1, 2, 3, 4 and 9. In brief, Number 1 is the cancer detection rate. Number 2 is called the PPV2 for Screening, which I think some of us would say might have been defined differently as a PPV1. Number 3 is the PPV2. Number 4 is the recall rate, and number 9 is the follow-up rate.

So with that introduction, Carl, do you want to go first, measure Number 1?

DR. D'ORSI: Do I want to or do I have to?

The way I looked at this metric was to be used in isolation, and that is very important to what I am going to say. I think it is a good measure, but not in isolation. So my comments will be based on the what I was told to evaluate it for, which was a metric.

This, basically, is a metric that is asking, for all the agony you produce by recalls and biopsies and evaluations, what do you get back? So it is saying, for every positive mammogram you do, which includes Category Zero from a screening and includes 4 and 5s after the evaluation of the zero from a screening, and that woman goes to some kind of tissue diagnosis, i.e., needle core biopsy or, much less frequently, surgical biopsy, how much cancer is produced?

So that is what it is saying, and the way it is written, it is written as a percentage. We usually consider it as a rate, X number per thousand. So the way it is

written, if you multiply that metric by 1,000, you will get what the standard measures are.

It is very important to realize that this metric varies -- can vary widely, depending on the population you are testing, i.e., age is very important, whether it is a prevalent screen or not is very important, and these numbers can vary.

There is a wide range, if you include all of them, that will kind of include all these variables. Anywhere from two to eight or 10 per thousand is the range, but again within that range there is a big variability, depending on --

CO-CHAIR GAZELLE: Could I interrupt for a second? I think, in terms of procedure, it would probably be helpful for everyone else if we start by defining the numerator and denominator --

DR. D'ORSI: Oh, I'm sorry.

CO-CHAIR GAZELLE: -- as proposed for the measure, because not everyone may --

DR. D'ORSI: All right. Let me read right from the statement: The number of screening mammograms -- this is the numerator now. The number of screening mammograms where the BIRAD assessment of 4 or 5 plus the number of screening mammograms with a zero that result in a tissue diagnosis of cancer.

So, basically, it is the positive mammograms, including screening and diagnostic, positive being defined on a screening as zero, 4 and 5, positive being defined on a diagnostic exam as 4 or 5. That combination is the numerator.

The amount of screening exams you have read is the denominator. That multiplied by 1,000 is the cancer detection rate. So that is the metric, and it is a very good metric when used with others. In isolation, it doesn't tell you too much, other than you are in a huge range.

It is sort of like accuracy. I can -- if I define accuracy for screening

mammograms, which sounds like a great metric -- right? Accuracy is true positive, true negative over everything you do. Well, if they read everything as negative, I will have an accuracy of 99.8 percent.

DR. SMITH-BINDMAN: Could we put this into context, just so people have a ballpark of what this means? If you read 1,000 screening mammograms, there should be in the ballpark of six or seven or eight cancers in that group of 1,000 women, and the cancer detection rate is usually around five.

So you are expected to find about five cancers per 1,000. As Carl said, it varies by age. So if you are looking at 20-year-old women, there aren't that many cancers to find. If you are looking at 80-year-old women, there are a lot of cancers to find. If you are looking at women with palpable breast lumps, there are a lot of cancers to find.

So those things matter, but basically you are looking at about five or six

cancers that you usually find out of 1,000 mammograms. If you are really doing a lousy job, you might not find that many. If you are doing a great job, you might find more of them. So that is what this is trying to get at.

CO-CHAIR PETERSON: Just another thing, just a little perspective thing. Radiologists' view of the world is, the patients I do, how did I do on them? From a more societal perspective or a hospital perspective, you might say, well, are you screening the right people, as you sort of indicted here.

If you, obviously, are screening a remarkably low risk group, 20-year-olds, you are going to have a low score on this, but it is not reflecting anything the, quote/unquote, radiologist did right or wrong. It is a reflection of who is going to the test.

DR. SMITH-BINDMAN: So just taking it one step further, a measure that

radiologists like to think doesn't matter so much about the prevalence of the group is a measure called sensitivity.

What that means, among the people who had cancer -- I said there would be about seven or eight cancers -- if you find five of those, the sensitivity gives you a sense of how you are doing proportionately that is not influenced by the prevalence of disease.

It is really hard to get at sensitivity. You have to learn about your misses. Cancer detection rate, you don't have to find out your misses. You know that you found five cancers. I don't know how many there is supposed to be. So cancer detection rate has a measurability tool that sensitivity does not.

CO-CHAIR GAZELLE: Yes. I guess my sense is -- I am just trying to ground and make sure I am correct on this. This is not a measure of anything to do with how good the reading was. It is a reflection of how we use

the technology itself. Did we screen a population who was at reasonable risk?

DR. SMITH-BINDMAN: It turns out that cancer detection rate is highly correlated with cancer prevalence. So even though it is imperfect, because it strongly depends on the prevalence, and even though my major problem is that it is not risk adjusted to the population -- so I don't know how useful it is without that, but in general it is highly correlated.

So if you are doing a terrible job in terms of finding cancer at a low sensitivity, you will also have a low cancer detection rate. They go hand in hand. So it is used as a measure of gross quality. So in facilities that provide care to underserved, turns out the cancer detection rates are lower.

DR. SNOW: One point of clarification. The word screening I take to mean an asymptomatic individual. So someone

who is there for a breast lump is not being screened. There is something there or believed to be there. So that is a different bucket.

CO-CHAIR GAZELLE: The denominator here is the number of screening mammograms.

DR. SNOW: Okay, so specifically asymptomatic subjects.

DR. SMITH-BINDMAN: But the age is hugely important.

CO-CHAIR GAZELLE: Yes. yes. They are asymptomatic, but there is still a difference in prevalence as a function of age. Yes. So I think, let's try and get back to Carl's review of the measure in terms of giving your evaluation of it, remembering that it is likely that we would recommend this be paired with other measures or combined with other measures.

DR. BURSTIN: Just one more point of clarification. The measure developer did put the measure forward to be looked at as a

group. So there was not an expectation on the part of the measure developer that this measure would get looked at in isolation. It was supposed to be paired with, on the first page there, the positive predictive value and the abnormal interpretation of the recall rate, just to put that in context.

DR. D'ORSI: Okay. Well, I called specifically about this, just to bring up a point, and I said should I evaluate this in isolation or with the others, and I was clearly told to measure it in isolation.

DR. BURSTIN: Clearly, evaluate the measure as it stands on its own, but keep in mind at the end of the day, the developer is recommending they get looked at together. So at the end we can put them together.

DR. D'ORSI: Okay, that is very difficult to do. It is a great measure not in isolation. That is all I can say. The way I evaluated it, I gave it an N only because I was told to consider it in isolation, and in

isolation it is relatively useless unless you have something else to define how the leader is obtaining these numbers.

CO-CHAIR GAZELLE: Could we go through the specific points, though, the specific areas in terms of its validity and reliability?

DR. D'ORSI: Sure.

CO-CHAIR PETERSON: But, again, I am just going to question right off the bat here. Are we talking about a measure -- you gave it an N because, as a radiologist, do I think this reflects my quality.

DR. D'ORSI: Alone.

CO-CHAIR PETERSON: If the goal isn't to reflect your quality as a radiologist, the goal is to reflect how is the ordering hospital screening patients. Then it may need a different criteria.

DR. SMITH-BINDMAN: I think it needs to be assessed within the strata of risk groups, just like we assess risk of other

ones. So now we state what you are saying: If the strata are 40 to 50-year-old women, or 50 to 60-year-old women, that will be our measure of the radiology quality.

CO-CHAIR PETERSON: Right. You are getting back to the radiologist again. I don't really care about the radiologist -- just for a second. Let's imagine we want to do this -- the analogy would be --

DR. SMITH-BINDMAN: At the hospital level.

CO-CHAIR PETERSON: -- in cardiac disease where you wanted to see, you know, did you order testing the right patients, is what it basically comes back to. I am just curious if the measure itself couldn't be seen under that light. You know that the radiologist has a quality measure, but --

CO-CHAIR GAZELLE: But it is not intended as an individual physician measure. It is intended as a facility level measure.

CO-CHAIR PETERSON: Right.

DR. SMITH-BINDMAN: That is a different -- we've got those in HEDIS already.

MR. BACKUS: To what degree does this facility really define who their screening, though? I mean, essentially, in a straight screening mammography -- right -- asymptomatic patients, and this is much more patient directed than the facility having a substantial amount of influence over the asymptomatic people that they get to show up in the door.

CO-CHAIR PETERSON: Now, see, this is where the world also -- the degree to which the center who gets the test -- people I will refer to you, you have the responsibility of being a screener of, are the tests coming into me the right ones. Are we getting the right patients in to do this test?

CO-CHAIR GAZELLE: But, I mean, screening mammography is at least something that is fairly -- the eligibility requirements are fairly clearly defined, notwithstanding

the November --

DR. SMITH-BINDMAN: But this is completely separate from that. This is once whomever comes in comes in, is the quality that those patients are receiving at some minimum level?

DR. D'ORSI: The problem, I think, that you are actually touching on there is a problem of, are we dealing with something like a blood test where it doesn't take any cognitive input, and then you can say, oh, the facility or, you know, the testing of this metric is good. Their method is very good, and it works.

There is a cognitive input to screening. So you can't separate it as opposed to, okay, the facility is doing it. Well, the facility is also the people who are leading it.

So, indirectly, it is a measure of the people working at that facility. So if you have people who are -- again, my apologies

to any surgeons who read mammograms -- who are all surgeons, they might have a cancer detection rate of 3 sitting in the group, but they should have had one a day, if we take into account the age and if we take into account all these other things.

The problem is it is very difficult to stratify by age, very difficult to stratify by prevalence. They can do this in service screening countries where they have that data right off the bat. You can't do it here. So you have to get a range.

CO-CHAIR GAZELLE: So when that -- I think it might be mentioned in the next measure, but what if they are rated 16, and --

DR. D'ORSI: Great.

CO-CHAIR GAZELLE: Well, but are they really cancers or are they not, and is there a lot of --

DR. SMITH-BINDMAN: Are there a lot of cascades of tests to then, say, those extra three maybe not being cascades?

CO-CHAIR GAZELLE: I still like to let Carl get through his ratings of this, and let's get through the discussion and ratings of the measures, and then have a discussion, if we could, because I think we need to at least get to that point.

DR. D'ORSI: So, basically, as I said, I ran through them in isolation, and I said a No for the reasons that a lot of everyone brought up.

CO-CHAIR PETERSON: Which did you give a No?

DR. D'ORSI: The first one, the first evaluation, that it shouldn't go further. We are not supposed to evaluate it as a pool.

CO-CHAIR GAZELLE: Let's go through all of them, and then we will have a discussion.

DR. D'ORSI: All right. As not a pool. I don't know how to say this anymore clearly. As not a pool, in isolation as one

metric, it is a No for me.

CO-CHAIR GAZELLE: For which one?

DR. D'ORSI: For each one, for importance, yes.

CO-CHAIR GAZELLE: All right.

DR. D'ORSI: The reasons are what I discussed already, that it varies so much on factors that it is difficult to assess. It doesn't tell you anything about what you are getting. So that is --

CO-CHAIR GAZELLE: So that is fine. So for discussion, how about the other metrics?

DR. D'ORSI: The other metrics --

CO-CHAIR GAZELLE: In terms of reliability, evidence to support, those scientific --

DR. D'ORSI: The reliability is excellent. There is a lot of evidence to support its use, and there is the article by Rosenberg that everybody is familiar with from the BCSC that has a huge number of mammogram

screenings, and it is a very solid individual metric. Its calculation is good. Its definition is good, and what it gives you is good alone.

CO-CHAIR GAZELLE: I think Helen is pushing us. We would like to get for each of those, if we could -- we need to record it.

DR. D'ORSI: All right. Let's go back to process.

CO-CHAIR GAZELLE: We are going to need to do that for every measure.

DR. D'ORSI: All right. So 2 is the definition of the detailed measure specifications, can they be attained? Yes, they can be attained. It is much easier to attain these electronically.

CO-CHAIR GAZELLE: Would you give it a C then?

DR. D'ORSI: I would give that a C. All right, the next is 3, which is --

CO-CHAIR GAZELLE: Helen, you want us to do 2(a), 2(b)? You want us to do each

one? Yes. We would like to have each one, if we could.

Just for process, let's see if we can get through the primary reviewer's comments, because I think from the NQF standpoint, we need to get the specific evaluation.

DR. BURSTIN: And, certainly, if there's any ratings that would differ from Carl's.

DR. SMITH-BINDER: I didn't know I was the secondary reviewer.

MR. CORBRIDGE: There was not a primary and secondary, really. It was review group, just in terms of dividing up, because we really didn't have enough to -- in terms of efficiency. So there is a review group. So, really, it should be in tandem, if individuals can really work together.

CO-CHAIR GAZELLE: 2(a) is a C.

DR. D'ORSI: 2(a) is a C, and for the reasons I gave. Let's go to 2(b), which

is reliability. I gave that a C as well, because it has been reliably tested in this large group.

Let's go to (c), validity testing. I gave this a P, only because the analytic method that's used to establish the validity requires a little more description. The current domain, I gave as a C. So it is a combination. I gave this a Partially Described.

Let's go to 2(d), exclusion is justified. That is not applicable. The next one, 2(e) wasn't applicable. The next one 2(f) wasn't applicable. The comparability of multiple data sources method: I gave that a C, because they clearly in this portion stated that they included PPV2, and the cancer detection rate, and the recall rate, which I think is a beautiful set of metrics. They are what you want to get at.

2(h), which is disparities in care, I gave an NA, Not Applicable. So, let's

see, Steering Committee -- again, I only gave it an M, because I was thinking of individual use.

Why don't we go to 3? Okay, 3 is in use. Couplet reporting of this initiative: Alone, I gave an N. No one would know what this means in isolation, especially for public reporting. Look at us here discussing this, and we fighting back and forth, and we are going to put this on public information. So I gave that an N. That is 3(a)(2).

3(a)(3), used in other programs and initiatives: That I gave an N because of the isolation.

3(b), which is -- what is 3(d)? Harmonization. I gave that an Not Applicable. I gave 3(c) an Not Applicable, and the Steering Committee overall, to what extent was a criteria of usability met? I gave that an M. As a sole indicator, it really isn't significant for the above reasons, but the M came from the fact that it was well

constructed as an individual metric. So instead of giving it an N, I popped it up to a M, because its definition was very clear and precise, and it is in use, not in isolation.

4 (a): Data generated as a by-product of the care process. I gave that a C.

4(d): Electronic sources. I gave that an A, because I don't have a -- in order to get this metric, the easiest way is if you have what is called a mammography module where you prospectively, as you read each exam, you put in the data, and it generates a clinical report and saves the data. If you don't have this, the usability is much, much, much more difficult to do this by hand. So that is why I gave it an A.

I don't know how many facilities have a mammo module. I don't know if the ACR knows this, but it is very difficult to get without a mammo module. So that is my reason for it there.

Exclusions were, for (c) were Not

Applicable, to me. Susceptibility to inaccuracies, errors or unintended consequences, I gave a C. I believe there could be unintended consequences with that.

Data collection strategy, 4(e), I gave as a C. I think the points that were brought up are very good.

To what extent was the criteria of feasibility met? I gave that a C.

I think that is it.

CO-CHAIR GAZELLE: Thank you. So you can see what a challenge we have in front of us. These measures are hard to evaluate. One of the things that -- and then I am going to ask Rebecca, since you also are with the group, to comment on the measure, even if not item by item.

One of the challenges: This has been proposed as a suite of measures, if you will, with two other measures, but we have been given no specific instructions on how they might be interpreted as a suite. So even

if all three were approved, the question is what happens if you are high on one and low on another. So there is no guidance yet there.

DR. BURSTIN: Just as one comment. Again, this notion of pairing it -- we don't actually know exactly what that means. We do have clear guidance on composite measures where multi-measures come together with the idea of getting a single score at the end of the day.

CO-CHAIR GAZELLE: Right.

DR. BURSTIN: And at least from that perspective, because I think that might aid Carl's thinking of, again, they didn't present it as a composite, is that we individually evaluate each of the measures and then make a determination of whether that measure could stand alone or should really only be used as part of a composite.

So I think, at the end of this discussion, that would probably be the right piece. I still think it will be helpful -- we

are not going to go through the whole measure again, each of them separately, and then make the decision overall, but we probably do need guidance from the developer as well as this group about what does it mean that they would be reported together exactly.

CO-CHAIR GAZELLE: Yes. And in fact, there is some ambiguity as well, because they say they should be paired with cancer detection rate, recall rate, and PPV2, but then this measure has proposed two measures that are both called PPV2. So we will need to, as a group, come to clarity on that.

Rebecca, do you want to give a --

DR. SMITH-BINDMAN: Thank you, because I think I have a very different take than Carl.

I would just start out by saying it is -- There are programs that use these measures together. So the best example would be the National Screening Program in the UK, which uses cancer detection rates, PPV, and

recall rate together.

Basically, you have to have a minimum cancer detection rate and, if you don't -- you are not doing well -- then they try to balance that cancer detection rate with a recall rate that is acceptable.

It is not that easy, the way they do it, but they combine them together. They don't use it as a composite. They basically plot each facility and each radiologist in this space that includes both PPV and cancer detection rates. I think it is a very nice model that you guys could adopt.

I actually like this measure a lot. I think the measure -- If you had to ask women what the single most important thing about a mammogram was, they would say to find cancer, and this tells you about finding cancer.

So I think that this measure, if I could pick one, it wouldn't be an inefficiency file. That is not efficient, but you would

want to find cancer. So I care about this measure more than any others, and I would be happy with this measure by itself. So I really like cancer detection. So I rate it as a C in terms of the importance of this measure. I think it is extremely important.

Going through the numbers -- Helen, do you want me to just give you my results or do you want me to say them out loud?

DR. BURSTIN: If you just want to probably just say them out loud, especially the discrepancies with what --

DR. SMITH-BINDMAN: Okay. I highlighted those columns. So for: Was it important for the measure to report? I would say yes, which is number 1.

Going down to number 2 in terms of the specification of the measure, I think it is very good. In terms of -- and so C. In terms of harmonization, I am not sure about other measures that you guys have. I don't

think there are any others.

DR. BURSTIN: No.

DR. SMITH-BINDMAN: So that was kind of easy. Going into: Was the extent usability met? I gave it a C.

Going to 4(b) Electronic Sources, I think all these data are available electronically. So I gave it a C.

I am actually looking for the width. I keep going past that. So --

CO-CHAIR GAZELLE: I think it was not listed.

DR. SMITH-BINDMAN: Right. I'm sorry. So I am going back up to 2. So 2(a) 12-13, the people who submitted this measure said no risk adjustment was needed, and then gave an explanation of breast cancer risk from Gil Barlow's paper, which is not relevant. Risk adjustment is for this measure, and I think risk adjustment is absolutely needed for this measure.

So I think it is a fabulous

measure. I think risk adjustment absolutely needed to make it a useful measure, and it doesn't need to be risk adjustment. It needs to be risk stratification, which is easier to do. So there isn't a model to do risk adjustment, but there are models to do the stratification.

CO-CHAIR GAZELLE: And you propose stratifying it by age?

DR. SMITH-BINDMAN: It needs to be stratified by two factors. It needs to be stratified by age, and whether exams are first or subsequent.

The relevance of that, I can't really emphasize enough. There is a two to threefold to fourfold difference in these variables based on age and first and subsequent, and you can imagine that facilities have a very different distribution, whether they see younger patients or older patients or they see patients who come in every year at Kaiser for a mammogram and they

are subsequent screenings versus a population that is an underserved population, and they are trying really hard to get everyone to come in once. Those variables are different.

So I think it is a great measure, but I think it needs stratification.

DR. D'ORSI: By risk stratification, you are not referring to breast cancer risk, are you?

DR. SMITH-BINDMAN: That is correct. Thank you.

DR. D'ORSI: They did. Okay, that's the problem.

DR. SMITH-BINDMAN: Well, they are talking about a breast cancer risk model, not a model of a measure. They both have risk in the name, but otherwise they have nothing to do with each other.

DR. D'ORSI: Correct.

CO-CHAIR GAZELLE: And I think what you are saying, if I could paraphrase, is that if you have a facility that is actually

doing a really good job of getting everybody in at their recommended intervals, they are going to have a lower cancer detection rate.

DR. SMITH-BINDMAN: They are going to have a lower cancer detection rate.

CO-CHAIR GAZELLE: And that facility that is doing the right thing would be --

DR. SMITH-BINDMAN: I would say that the range of allowable values to this cancer detection rate include tolerable care and off-the-chart good care. So that range needs a little more narrowing. The reason they gave this range is because they haven't done the stratification. It is in a useless category at the moment. The range is too wide.

DR. D'ORSI: The fine tuning on that range, which is more difficult to obtain but is really important, is minimal versus non-minimal cancer. You can be in that range and be finding Stage IV. You know, that is

useless for a mammography range, but -- and as you alluded to -- you may be at the lower end and be finding early cancer. But minimal cancer versus non-minimal is a very difficult metric to get.

DR. SNOW: There is another element to this. A feature of this is that the numerator requires a biopsy diagnosis of cancer. Now what happens -- one, that is a whole separate step, and there are other cracks to fall through, but probably not a large crack.

The one that is larger is what do you do if it is -- in a place like the Sloan-Kettering, everything gets done in the same shop, but what do you do if the initial four or five is done in a little community hospital, and immediately the patient is referred to the Sloan-Kettering for the biopsy? There is a big gap.

I know for sure that our record keeping isn't 100 percent in that area. That

is why we are spending billions of dollars to get there. That contaminates the result. I just don't know how much.

DR. SMITH-BINDMAN: It is also a very relevant point when you are talking -- I was going to get to it when I got to 2(h) -- disparity, in fact. So facilities that are underserved are much less likely to either find the cancer or to know about the cancers that they have found.

DR. SNOW: Should there be stratification for ethnicity, too, was the question. I don't know.

DR. SMITH-BINDMAN: Cancer detection rates vary a lot by underlying race and ethnicity, but not in the way that you would necessarily think that they varied. So to do what you are saying, there aren't data out there to create metrics, but in terms of this measure biasing against facilities that have less resources, which is what you were raising, is a -- to get at the racial and

ethnicity one.

DR. GEMIGNANI: But is it not the responsibility of the primary place that orders are issued to follow up on those results, even if that biopsy is not done at that -- I mean, that is part of reporting what your --

DR. D'ORSI: Right. The way that verbiage is stated is a reasonable effort. If you have -- if you are a small facility and you are sending a lot of your things out, that becomes a big problem to get -- order biopsies done somewhere else. This was a good example. That is not an issue in countries that have service -- because they are all attached. So, easy. We don't have that.

DR. GEMIGNANI: So that facility would get a lesser rate, having used a measure like this, because they are --

DR. D'ORSI: Correct, because they don't know, or they don't know, if they can't find it.

DR. GEMIGNANI: But isn't that something that you want to know about that facility, that they are not able to track?

DR. D'ORSI: Yes, but that may be an unintended consequence. They may be doing something very correct in defining a four or five, but they may not have the resources to search.

DR. GEMIGNANI: So they can't detect those cancer rates.

DR. D'ORSI: Well, that is a problem.

CO-CHAIR GAZELLE: So let me take a stab, then, at summarizing the discussion on this measure to this point, because I think it will be important to go through all of the mammo measures and then come back to a global discussion -- is that the general sense I am getting is that there is some value in measuring cancer detection rate, probably in combination with other measures.

There's issues about

stratification by first screening or subsequent screening and by age. There's issues about how the data would actually be collected, registry data, claims data, etcetera. But I think, as a group at least, we have -- is it fair to say we have a sense of what this measure is trying to accomplish and what some of the issues are, and it would be all right to move on to the next measure?

CO-CHAIR PETERSON: I just have a few clarifying questions. Question number one: since you like the measure, I will direct it your way, but anybody can click in.

I am getting a relative magnitude. It appears that this rate would vary much more depending on the strata that you are talking about, age of patients, ethnicity, first versus follow-up screening, than anything to do with the quality of the reader, meaning that, in fact, the degree to miss -- if your concern is that this is a reflection of missed cancers that were there that were missed, that

rate would be, we would imagine, relatively low relative to the magnitude of two, three or fivefold variation, depending on if you are first or second, or very young versus very old population.

So if this is to reflect quality in terms of the reader, I would argue that this probably is to work without this stratification by the underlying population. That is one clarifying question, and as it is written, it doesn't stratify.

CO-CHAIR GAZELLE: But we could propose that.

CO-CHAIR PETERSON: I am not so sure that that isn't a remarkable rewrite of this.

DR. D'ORSI: How is that not a remarkable rewrite when there is a fourfold difference?

CO-CHAIR PETERSON: We don't need the answer right now.

CO-CHAIR GAZELLE: Well, we don't

need to answer it. But, for example, we could say the measure would be acceptable if it was reported by decade-age strata, and first or repeat screening. We don't need to have a model.

DR. SMITH-BINDER: It turns out that those variables that would be needed in this case are available for everyone. We know the age of the woman, and you know if it is first or subsequent, pretty much. You know, that is pretty good. So it is not a fancy model.

CO-CHAIR PETERSON: We can maybe take up some discussion about whether it gets rewrite or not.

DR. D'ORSI: One other point on the stratification. You need number of hits for it to be valid. When you start teasing decades of age out, you are going to need a lot more in that age group to make a meaningful data analysis. That is why it is done as a group, and may not be as stratified

and useful for a single facility.

CO-CHAIR PETERSON: Great. Just one more clarifying question, and then I will stop.

CO-CHAIR GAZELLE: Before we leave stratification, the argument against stratifying, which is probably not valid, but if you assume that everyone has the same general mix, if you aggregate up against large enough -- some people have argued that, and we could reject that. I would reject it, but that has been proposed as, well, you know, if you look at facilities, everyone has got about the same mixture across a large enough group.

So just for perspective, that argument has been proposed by some people.

DR. BURSTIN: I just need to point out that Dr. D'Orsi and anybody else may still have a chance to respond.

CO-CHAIR PETERSON: And then the other is an unintended consequence question, because actually, you are ranking that, which

is going to include -- I thought, if I heard you right, you said it had potential unintended consequences, but you gave it a C. So that is just a positive-negative thing, I guess. I would have said it the opposite. If it does have unintended consequences, then it should be ranked as not scoring.

DR. D'ORSI: Let me look again. I may have been wrong.

CO-CHAIR GAZELLE: I am going to propose that we take a break. We are scheduled for a break. We will take about a 10-minute break. We can come back to conclude -- do you have one other?

CO-CHAIR PETERSON: So the unintended consequences portion of this that you were concerned about are that, in fact, if you do mark -- let's take it to the extreme. Every one of your tests are positive, and you send every woman on to a biopsy.

Your score here would be good, because you would, hopefully, find every

cancer, assuming the system worked, at the downside of every woman having now the negative effects that we have heard in the news so much.

So that, in fact, this measure has the very strong potential of encouraging over-reading as opposed to -- you know.

DR. SMITH-BINDMAN: When people use this measure -- just to sort of put it into context, there is a very nice breast cancer program going on in Chicago to figure out -- it is a unified effort across the city for everyone who provides breast cancer care.

They found that their cancer detection rates at their hospitals were really, really low. They were missing all the cancers. So it is more of something that we think about at the extreme of they are providing services, but they are not finding cancer. Is there a major quality problem at the low end, rather than at the high end, pushing so many recalls that you will find

more cancer?

At some point, recalling more women, you don't tend to find that much more cancer. It becomes a random.

DR. ZERZAN: But do you think that, in trying to figure out what the inefficiency is, it's both under- and overuse that we are trying to get a better -- what is that middle measure, and then --

DR. SMITH-BINDMAN: This particular measure doesn't show much push-through overuse. The other ones, the other four measures --

DR. SNOW: I don't think this would cause over-reading, because you have to have a confirmed diagnosis. If you screen everybody and send them all to the pathologist, that doesn't mean that they are all going to come back positive. If you over-read, you are going to have a lower rate, because your numerator will go down, because you won't be able to get sufficient diagnoses.

CO-CHAIR GAZELLE: I think, really, this is a balancing measure against recall rate; whereas, if we want, say, to achieve recall rates below 10 percent, for example, one way to do that is to miss a lot of cancers. So if you --

DR. SMITH-BINDMAN: This is a fail safe on the low end.

DR. D'ORSI: If you look at an ROC curve, it is very clear. As your false positives go up, what happens to your false negatives? It goes down, and that is exactly what is being said here. As you get close on an ROC asymptotically to the top, the price you pay to get one or two more cancers is massive.

So most people operate in the middle of an ROC curve, because they realize that, if I operate here, I am going to miss; if I operate up here, it doesn't pay for what I am doing to get the cancers.

CO-CHAIR GAZELLE: And, in fact,

from the last meeting when we did consider recall rates, the feedback that came from the Steering Committee as well as the mammography community at large was you can't possibly have recall rate unless you also have cancer detection rate.

That is why it is hard to discuss these alone, because they really do need to be considered together.

DR. FIESINGER: I just wanted to throw out a vignette. I think the measure is important. The unintended consequences, I think, are really significant. On one hand, you could just throw the measure out there and see what develops, but I was Medical Records at MQHC, we had a breast cancer graft.

Texas Medicaid doesn't cover undocumented women for cancer treatment or biopsy. So if you get the mammo, detect it, we would have low cancer detection rates, a barrier to citizenship status, and then you add financial resources on top of that.

Grant funding depends on measures for compliance standards; whereas, like 95 percent want us to track every patient. Therefore, health care which funds that case sees this big push for tracking quality metrics, has no time for funding yet, maybe down the road.

So how it is interpreted can really impact the safety net system quite severely in the wrong way.

DR. SMITH-BINDMAN: Because your patients couldn't find out about cancers, because they were not documented?

DR. FIESINGER: Because we couldn't get funding to get a biopsy. You can get the mammograms through a charitable organization, but getting emergency -- you have to get a biopsy and, if they have cancer, get a emergency Medicaid to have cancer treatment. But if they are not documented, meaning not citizens, they can't get Medicaid. So how do you get the biopsy?

DR. SMITH-BINDMAN: So they really don't need a mammogram.

CO-CHAIR GAZELLE: Yes. If they are not going to get care anyhow.

I think we could go on, on this measure, forever, as a base. I know you said the most -- you know, the thing that a woman wants when she goes to get a mammo is that cancer is found -- cancer detection. My question would be is it that cancer -- you know, it is a place that has a high cancer incidence or is it a place that is better on PPV2, so that she has faith in the radiologist's judgment? Right? You are balancing the concern of a negative.

It seems to me that what I really want to know is that, when they say I have cancer or say I have an issue or say I don't have an issue, they are right; as compared to this wild population here.

MR. BACKUS: That gets into our next couple of measures.

DR. SMITH-BINDMAN: I agree with you. Women don't, for better or worse.

CO-CHAIR GAZELLE: Let's go ahead and take a 10-minute break, if we could, because I think otherwise we will just spend the rest of two days on this first measure.

(Whereupon, the foregoing matter went off the record at 10:58 a.m. and resumed at 11:14 a.m.)

CO-CHAIR GAZELLE: Okay, could we get started again, please. Because the other measure in review group 1, which was Number 9, Rebecca's, is proposed by CMS and not the ACR, we are going to go on to the other three that were proposed by the ACR.

We will discuss the four total from the ACR as a group after we go through each one individually. Then we will allow Larry Bassett from the ACR to comment after we have all commented, and then we can talk about our feeling of those four as a group.

DR. SMITH-BINDMAN: Do you think

the ACR might be able to say a word or two about this measure before we go on?

CO-CHAIR GAZELLE: No, they just want to go through all of the four first, and we talked during the break about that.

All right. So the next one, which is number 002-10, titled Screening Mammography Positive Predictive Value 2, and it is described as being the percentage of screening mammograms with abnormal interpretation that result in a diagnosis of cancer within 12 months.

It is actually defined in terms of the numerator and denominator slightly different from that. So the numerator is the number of screening mammograms with the BIRADS 4 or 5 or BIRAD zero associated with a 4 or 5 on a diagnostic mammogram, so basically a positive screening mammogram that results in cancer within 12 months.

The denominator is defined as the number of screening mammograms with a 4 or 5

or zero, and the zero has to be associated with a 4 or 5 on a diagnostic.

So it is basically the positive screening mammograms denominator. Numerator is the subset of those that have cancer.

So the first thing I will say is that in the literature this might be called the PPV1, and so there is going to be some confusion about that for those of you who are familiar with the literature on those measures.

So, in terms of my evaluation, I thought for 1(a), Importance to Measure and Report -- let me make an overall comment first. There are two very similar measures, this one and the next one. They are both called PPV2. I think this is really PPV1, and the next one is PPV2.

I am going to score this in isolation, but as a preface I am going to say that, if I had to choose between the two, my choice would be for the next one. But I am

going to score this in isolation.

So I thought for 1(a) I gave it a C in terms of importance to measure and report. For 1b I gave it a C, and for 1(c), the relationship to outcomes, I gave it a P for partial, because I think -- for all the reasons that we have discussed before. 3 is only partially collected outcomes.

In the text of the proposal, the measure developer suggests that it should be combined with other measures, and we have already talked about that, though there is no clear guidance on what that would mean. I don't think we are envisioning a composite measure so much as reporting of the three individually, but that hasn't been addressed.

Then for the global one, importance to measure and report, I said yes.

Then for measure specifications: 2(a), Precisely Specified, I said yes. 2(b), reliability testing, I said partially, because it was my impression that the text in the

measure was talking about, really, the reliability of BIRADS and not the reliability of the proportional measurement. So I would give that a P.

For validity, I gave it -- I'm sorry, for 2(c), the validity meaning the relationship of this measure to outcomes, I gave it an M for minimal, because I didn't see that there was a connection between this measure and outcomes of concern.

Then for exclusions, NA, and data sample, NA.

Identification of meaningful difference in performance, 2(f), I gave that as M. They do cite ranges from the literature, although I think there is a typo. They cite a range for PPV2, not withstanding the comments I made about the confusion between the two measures labeled PPV2 of five to 10 percent, and from the article that was cited, it is 25 to 40 percent. So I believe that is a typo in this one and some of the

other measures.

DR. SMITH-BINDMAN: This is a screening measure?

CO-CHAIR GAZELLE: Yes.

DR. SMITH-BINDMAN: Then it should be the lower number.

CO-CHAIR GAZELLE: Right, but the screening measure would be PPV1. So that is the confusion.

DR. SMITH-BINDMAN: But we are assuming that this measure is PPV1.

CO-CHAIR GAZELLE: Right. I think we have to.

For 2(g), multiple data sources, I am not sure how to evaluate that. So I gave that an N, but it could have been an NA, and for disparities I gave that an NA.

So for the overall: To what extent was the criterion scientific ability of measure properties met? I gave it a P for the reasons I just stated.

Then for 3: 3(a), the current use

one, I gave it a C, although there was some question I had as to whether or not this could actually be done everywhere as opposed to at the sites participating in the ACR net for mammography database and the BCSC.

For harmonization, hard to evaluate, because I think the proposed -- so the way I interpreted that question 3(a) was that it could be used in a public reporting initiative, and there is a lot of text there about BCSC and the National Mammo Database, but there is no text to indicate what percentage or what proportion of sites in the country participate in one of those two. So it wasn't clear to me that this is usable --

DR. SMITH-BINDMAN: But I think it could be. They don't cite the right literature.

CO-CHAIR GAZELLE: Right.

DR. SMITH-BINDMAN: But I think it could be.

CO-CHAIR GAZELLE: I gave it a C.

I did give it a C. It is just that I raise that question based on the text.

Now let's see. For 3(b), harmonization, I gave it a P, and it was hard for me, because it is really not harmonized with the existing measures so much as harmonized with others that are proposed, but I think it is harmonized with the intent of -- or there is the intent of harmonization.

For added value, I gave it a C. I thought that it was clear that it did.

Dataset, data generated -- so my overall for 3 -- what extent was the criterion usability met? -- was a P, again for the reasons I said. In my view, you got to get a C on everything to get a C for the overall.

Okay, and then for 4, Data Generated as a Byproduct, I thought it was: 4(a), clearer, that the data elements could be generated as a byproduct of the care process, but it may not entirely be now, based on the issue of the cancer rates. So I gave that a

P -- cancer detection.

Electronic sources, I gave that, again, a P, because I think the feasibility of using those existing electronic data sources is there, but I don't think everybody is using them yet.

Exclusions, NA. Strategy --

DR. D'ORSI: You mean C, right, not A?

CO-CHAIR GAZELLE: NA.

DR. D'ORSI: Oh, NA, I'm sorry.

CO-CHAIR GAZELLE: There weren't any. So then I think there were a lot of -- To what extent were the criteria on feasibility met? I gave that a C as well. I gave it a P leaning towards a C, to be honest with you, because I think that it may be close to feasible. I am just concerned about some sites that may not have access to the full panoply of electronic data registries and sources.

Then for my overall -- do you

recommend it for endorsement? -- I gave it a Yes with the proviso -- I know we are not allowed to give this proviso on an individual measure, but with the proviso that either this or the real PPV2 -- my preference would be that real PPV2, the next measure -- should be paired with recall rate and cancer detection rate. A quick run-through.

Now leaving all these boxes and scores, here is my gestalt on it. It is a valuable measure, not in isolation. If it is being paired with other measures, I think it does add value; but if it is being paired with other measures, I would rather see us use the next measure, the PPV2, and not this one.

So let's see. Mary, comments?

DR. GEMIGNANI: Yes. So I am going to be the primary reviewer for --

CO-CHAIR GAZELLE: First, any other comments on this measure before the next one?

DR. GEMIGNANI: I have no

comments.

MR. BACKUS: My only thing is how much are we looking at one being a measure of screening mammography and one being a measure of diagnostic mammography, and those are, to me, really two different target audiences amongst -- if we operate within the context of this is information for the public, then they may be thinking much more about going and getting a screening mammogram; whereas, as health care professionals are thinking much more about PPV2, which is how good are you at picking it, once you get it.

So to me, it is just two completely different populations that you are looking at. In one, you should be hitting one out of four, so to speak, and in the other you are hitting one out of 20.

DR. GEMIGNANI: I think that the previous -- the measure we just discussed with the PPV1 sort of leads into the PPV2, because it takes all comers of the pie; whereas, once

you move all the true diagnostic mammographies, it is a purer measure.

MR. BACKUS: Right.

DR. GEMIGNANI: So I am not so sure whether excluding the other one, if we were able to tweak it a little bit, is necessary, because they are actually targeting two different things.

CO-CHAIR GAZELLE: Are there any other comments from the group on this measure? I forgot to mention, please give your name when you are commenting, if you could, for the recording.

DR. SMITH-BINDMAN: Just a question. You skipped by -- instructions are hard. This is Rebecca Smith-Bindman.

Just a comment on whether or not, to the degree that cancer detection rate needs to be stratified by age, should I just comment on whether that needs to be the case for PPV1. I think it varies by age.

So the PPV1 of mammography in

women who are in their forties is about two to three percent. The PPV1 for women in their seventies is about eight to nine percent. So there is a pretty big range in that. It is not as important as for cancer detection rate or for recall rate, because they go a little bit in tandem. So they both go up together.

So when you are dividing them, there may be a little bit less error, but --

CO-CHAIR GAZELLE: So that wasn't addressed in the measure.

DR. SMITH-BINDMAN: No, it wasn't.

CO-CHAIR GAZELLE: And the only thing I would say -- I am not sure that I could comment on it from a sufficiently educated viewpoint, except to say that, if we are proposing these as a group, three or two or four or whatever, and if we are saying at least one of them needs to be reported by, for example, strata, that they all probably ought to be. It would seem reasonable to me.

DR. GIBBONS: Ray Gibbons. Just

to follow up on that point, I am having a hard time understanding when you are describing what seems to be a known narrow range, how this will spur quality improvement.

If you now start talking about risk stratification, how many patients do you have to have to have a reasonable precision to every use that it is required?

CO-CHAIR GAZELLE: So those data were not presented. So I am not sure we can answer that question based on data. However, an average site would do what number of mammograms?

DR. SMITH-BINDMAN: I can address that based on the data. Your point is very well taken. So the average facility size in the U.S. is between 1,000 and 2,000. It is a medium size.

So in the -- and there are a fair number, 25 percent of facilities who are very small, and the very small facilities won't possibly have enough cancers to get at cancer

detection rate, let alone cancer detection rates of five.

So I think this has to be limited to facilities of a certain size, and that will, by definition, throw out at least a quarter of the sites.

DR. GIBBONS: So out of the one or two thousand, how many are positive, because that is the denominator in this study?

CO-CHAIR GAZELLE: Right.

DR. SMITH-BINDMAN: No. The denominator is easy, because out of the 2,000 mammograms there will be 300-400 that are positive. It is the numerator, the number of cancers, that is the trick in this.

CO-CHAIR GAZELLE: The denominator is positive screening mammograms.

DR. SMITH-BINDMAN: That is easy.

DR. GIBBONS: So 300-400, you are saying, is --

DR. SMITH-BINDMAN: Will be positive. The denominator will get a 10-15

percent positive rate, the denominator. There will be about 150 per thousand to 300 in the 2,000 example. So that will be about 2,000. The numerator would be something like 10.

DR. GIBBONS: Well, I am just trying to work through the math. We are down into single digits.

CO-CHAIR GAZELLE: Yes.

DR. GEMIGNANI: I think this measure is -- this is Mary Gemignani. I think this measure is also getting at how often are you calling it an abnormal mammogram just on any facility that comes in, and how often are you really having a cancer out of you calling a BIRADS 4 or 5.

So if you use it in isolation probably to the point that has been discussed, probably not such an effective number. But if you are using it in conjunction with your cancer detection rate, then you are getting more at how many abnormal tests are you really -- false positives are you really doing?

DR. SMITH-BINDMAN: But it is a question about how applicable this is for small facilities and how many facilities are small. It is quite a lot.

CO-CHAIR PETERSON: But I guess I am just missing why is it not a reasonable measure in extent here, by itself, because this is a meaningful number to patients. I want to know how many times -- if you call me again and tell me I have a positive study, how many of those will really end up being cancers?

DR. SMITH-BINDMAN: If your target is four percent -- that is the target, or five percent -- you need to have a large enough sample size that my estimate of your four or five percent is valid.

CO-CHAIR PETERSON: And the target is four or five percent, because?

DR. SMITH-BINDMAN: That's because that is as good as it gets. That is the number. That is the average PPV across

mammography.

CO-CHAIR GAZELLE: So if you think of an ROC curve, one way to get a really high PPV is to operate toward the specificity side of your ROC curve, which is to have too high a positivity threshold. So, basically, if it takes an awful lot to get you to call it positive, everything you call positive is going to truly be positive.

So in isolation, you might have a high positive predictive value, but you have a really low cancer detection rate.

DR. D'ORSI: When you are looking for something that is potentially lethal with a very small client probability, almost by definition, when you are screening for that, you are next going to have to pull in a lot of things that are not related to that.

If you had -- if the prior probability of cancer was 50 percent, you can have a very wide net, and you would have a pretty good pickup rate. When you go down to

three or four prior probability of malignancy per thousand, your net has to be very, very large to catch a reasonable sample of those malignancies. So there is no way you are going to drop false positives and do that.

CO-CHAIR PETERSON: So I'm just trying to get this again. So a good score here is 96 percent wrong. A bad score is what?

DR. SMITH-BINDMAN: Say it is 92 percent wrong, if you are going to say really good. I mean, the best of the best. The best of the best.

DR. D'ORSI: But it is not wrong, Eric. It is not wrong. It is not wrong.

CO-CHAIR PETERSON: Yes, it is. It is a miss. It is a miss.

DR. D'ORSI: It is a miss by statistics, but it is not a miss for what you are doing.

CO-CHAIR PETERSON: I am just asking. So this is the range -- so we're

talking about 92 percent to 100 percent wrong. That is the range we are talking about measuring. Let me get this down.

CO-CHAIR GAZELLE: Basically, low prevalence.

DR. D'ORSI: So if you went to a facility and your wife went in and said, hey, Eric, this place is wrong 90 percent of the time, so the other place is wrong only 98 percent of the time, I would say go to the place that is wrong more often. That is what I would say to my wife.

DR. GIBBONS: The probability of detecting the cancer is higher.

DR. D'ORSI: Correct.

CO-CHAIR GAZELLE: Depending on whether on whether or not they are moving on the same ROC curve.

DR. D'ORSI: I am assuming that they also -- all the same line.

DR. GIBBONS: Ray Gibbons. I am sorry just to keep harping on this point, but

if the numbers are going to exclude 25 percent of centers, facilities, in the country, do we have any data as to where there are quality problems with respect to facility size; because much of what else we have in medicine suggests that volume helps drive quality, and low volumes helps lead to low quality.

So I am concerned about a measure that might exclude 25 percent of facilities in the country.

CO-CHAIR GAZELLE: What is the volume that is required for certification?

DR. SMITH-BINDMAN: The volume is only at the radiologist level, not the facility level. So the radiologist level is just about 500 mammograms per year, and it turns out the facility averages are about 27. So your question about whether or not there is an association of volume and facility, there hasn't been strong data to look at that.

I have two large papers on my desk that are looking at that, and the answer is it

is not clear. But your concern that those facilities, where there could be a problem, we don't have a tool to measure the quality, is inherently more in the statistical sample size.

DR. D'ORSI: But you bring up a very good point. There are several articles that are trying to relate experience with performance metrics, and what they found overall is that there is not that close a relationship. But it appears that, if you are reading about -- this is data from Linda Warren Burhenne in British Columbia who has a large screening population there.

If you reading about -- each individual is reading about 2,000-2,500, they are doing better in that group than the ones who are reading less.

The UK requires 5,000, and there is no real solid data of a linear orientation with number of performance other than that British Columbia reported about 2,000-2,500.

But that is another country. It is another -- whole set of circumstances. So it is not a linear relationship.

CO-CHAIR GAZELLE: Are there other comments on this measure, number 2, before we go on to measure number 3, which is a very similar measure? Hearing none, Mary?

DR. GEMIGNANI: This is Mary Gemignani. I am going to review measure number 3-10, and I think a lot of the points that we brought up for the previous measure are definitely applicable to this measure, and this measure is actually probably the easiest one of all, because we are working off of diagnostic mammography as opposed to the screening in general.

So it is the subset of patients that already have an abnormal mammogram, and you really want to determine biopsy proven cancers within this subset.

So the numerator is cancer, and the denominator is anyone who has a BIRADS

score of 4 or 5 mammography.

So having said that, I will move on through some of the reviews. Looking at number one: So as far as eliminate overuse or ensuring delivery of appropriate care -- So that is 1(a).1 through 3. So 1(a) is Completely Agree.

For the opportunities for improvement, I think that this one also gets a C.

Outcomes for evidence to support measure focus: The writers of this do mention that sometimes we use recall rates in comparison with this, and how using a recall rate individually can cause controversies for the evaluation of mammography in centers.

So they do bring this up, and I think that that was a good thing to sort of bring up in the measure. So I put it as a C.

So was a threshold criterion, importance to measure overall for measure, quality measure number 1 is Yes.

So scientific acceptability of measure and properties, which is number 2, I put C for 2(a), which is basically looking at, again, the target population in the denominator. Then 2(b) was a C for the testing and analysis that they used, and for validity testing I put C.

Exclusions justified: There were really no exclusions for this. So we put it as NA. Then there was really no true discussion of risk adjustment on this here. So I put it as an NA, and it sort of comes back to what our discussion was. It should be looking at some stratification in this.

So for 2(f), it is C, and then comparability of multiple data sources and methods -- that was NA, and there was no disparities in care statement with this. So that was an NA. So overall for the scientific acceptability, I put it as a C.

Usability: Most centers do have data on this, on how many that they actually

have biopsy tissue on. So I think that being able to obtain this data should not be unfeasible. So I gave it a C.

Then for harmonization, I didn't see any harmonizations that I could sort of find. So I gave it as an NA, and then again we have had a lot of discussion so far about whether we should be using these in relation to each other. So as far as its individual value, I think out of all of them, this is probably the one that could most likely stand on its own, but would be best in conjunction with the other measures we talked about. So overall for usability, I gave it a C for feasibility.

For 4(a), I gave it a C. Then I had some questions, and it came up in discussion for 4(b). I gave it a Partial, a P, because if we came up with this discussion a few minutes ago about whether we would be able to track patients who went elsewhere. If you gave them a BIRADS 4 and 5 and then they

went to another place and they had their biopsy and we had a reasonable attempt at getting their pathology but we couldn't, how is that going to really affect this measure? So I put it as a Partial.

Exclusions were NA, and that is 4(c). Then unintended consequences: I gave this a Partial, because I think that, without knowing the volume of the center, without being able to incorporate the detection rate and the other rates, it may be difficult to interpret this value by itself.

Also, if it is a small center and you don't have access to get the additional pathology results from the biopsies, you might not have complete data collection. So I gave the data collection aspect support a P, too.

So overall, even though I kind of dinged it a little bit for the data collection and being able to get that pathology, I think this is a good measure, and so for feasibility and endorsement: feasibility, Complete, and

then recommendation would be Yes.

That is the primary.

CO-CHAIR GAZELLE: Okay. Thank you, Mary. Carl?

DR. D'ORSI: Carl D'Orsi. Can I make one comment? This is PPV2, which is a recommendation for biopsy, not the actual performance of biopsy. So if we do PPV2, that is an added difficulty for a facility to go find their 4s and 5s who actually haven't gotten anything in their own facility, and it is over and above those who have a biopsy somewhere else.

So it is a little more difficult. They are probably pretty close in this country, but it is a difference.

CO-CHAIR GAZELLE: Can I ask for a clarification on that, because that is not how it is defined here, I think. The denominator is a BIRADS score of 5.

DR. D'ORSI: It should be recommendation -- the BIRADS is a

recommendation, by and large. It does not mean that they are going to have the biopsy. That is PPV3.

CO-CHAIR GAZELLE: That is right.

DR. D'ORSI: And that is a difference, though.

CO-CHAIR GAZELLE: But the denominator is defined here as the number of diagnostic mammos that are 4 or 5, and the numerator is the cancer. So --

DR. D'ORSI: Right, but 4 or 5 is a recommendation. It doesn't mean that they have the biopsy. The denominator of PPV3 is biopsy obtained.

CO-CHAIR GAZELLE: Right. No, this is PPV2, though.

DR. D'ORSI: Right. I am just making that slight difference, that it is going to be a little bit harder. People have to follow up their 4s and 5s in their own facility who decided not to have it.

DR. GEMIGNANI: Yes.

DR. ZERZAN: This is Judy. I would say that the outcome, whether the labs actually have been done is more important than whether it's recommended, because that's what's really going to change patient health. You can recommend things, but that doesn't get you to better health.

DR. D'ORSI: Carl D'Orsi. That could be important to see how follow-up is, but you are right. As far as this is concerned -- that is mandated for the FDA that we present, not two but three.

DR. GEMIGNANI: But this is also getting at the BIRADS. So all BIRADS are recommendations for physicians. So I think the way it is written, it is still getting at the recommendation, not the --

DR. D'ORSI: I just wanted to make sure that everybody understood the three levels of definitions, that's all. They are very close, if not identical.

DR. BURSTIN: We are not talking

about -- but one of the other measures is trying to get at what we have actually done versus what was recommended.

DR. D'ORSI: Right.

DR. SNOW: This is Snow. It is worth making the point that, for that small facility, being able to document electronically the recommendation as opposed to the completion is much, much easier. So from the standpoint of feasibility, taking a PPV2 and saying, well, they are going to get it, right, I would have a little hope for that last bit. This makes it easier to do. I am not saying that you should stop there, but --

DR. D'ORSI: Well -- Carl D'Orsi -- you have two layers now. You still have to find out who's got cancer in the 4s and 5s that you recommend. So not only do you have to find out who goes somewhere else; you also have to find, out of your own group, who didn't do it. So it is a little more work.

MR. BACKUS: This is Mike Backus.

Do we have any sense for what proportion of people that are a 0, 4 or 5 don't come back for follow-up? What group of people drop off, five percent, eight percent, one percent?

DR. D'ORSI: It varies by area. It varies by the population you are looking at. Most people, when you recommend a biopsy, will get it done. I don't know what "most" means.

DR. BASSETT: In our practice, every one you recommend basically gets done. There are some other practices where you might recommend it, but the surgeon won't do it.

DR. SMITH-BINDMAN: It is an extremely hard question to answer. What you have to do is ascertain it. So the CDC National Breast and Cervical Cancer Early Detection Program first published Mays' paper, and they have in their underserved population 25 percent lack of follow-up to recommend it.

So that number was huge, and most of that has to do with assessment and

ascertainment problems that they got down to about 10 percent. So it is a really hard question to look at, and the way they deal with this issue on two papers that are going through the Breast Cancer Surveillance Center, a big dataset, is they cut off the time period at six months and say, if we can't find you by six months, you kind of didn't have it done; and they are getting about a 90 percent, 92 percent, but that mostly is a data issue.

So you are looking at the underlying rates, and there is no way to do it. It hasn't been done.

MR. BACKUS: Well, we know it -- I mean, it is not half.

DR. SMITH-BINDMAN: Less than 10 percent.

CO-CHAIR GAZELLE: All right. Are there any other comments on this particular measure? See, we are getting better at this.

Okay. So the next one -- I think we have time to do this one. Let's do IPE-

004-10, which is the recall rate.

MR. BACKUS: I am Mike Backus. I was assigned primary review for this. I don't have the benefit of what appears to have been substantial discussion about this measure the last time the NQF met, but I will go through, once again, a little bit in isolation, and my comments are obviously tinged with it coming in a set.

So the measure is recall rate, which is, you know, how often you are calling it for a unknown. And rate is strictly the percentage interpretive is 4s or 5s, and it does look at screening mammograms here, not diagnostic.

CO-CHAIR GAZELLE: Zero, 4 or 5.

MR. BACKUS: Zero, 4 or 5, right -- and not diagnostic mammograms.

If you come down, you know, from an importance, I gave that a C. Obviously, the impact is pretty well understood. It has been discussed before for 1(a).

1b, the opportunity for improvement: The same thing. It is a pretty straightforward measure and a way that compares centers.

1(c), outcome or evidence to support the measure focus: Once again, I think it is fairly important, although on its own, I would say it might be a Partial. In conjunction with everything else, I would give it a C.

So overall, I think it does meet the importance criteria. The scientific acceptability of the measure, 2, that I give a C. It has obviously been around the block.

Reliability, I think, is C; and the same for validity. The exclusions: I gave that a P, only because there might be some issue about stratification of the population, if you are working in a different demographic. So if you could stratify it, that would be a little bit better.

The analytic method is 2(e). I

gave that a C.

Meaningful difference in performance: I went back and forth here between a C and a P, and I ended up on a C, once again just because of the stratification issue. You will get differences in the centers, I thought.

2(g), the comparability of multiple data sources: I put this as an NA. One thing I did think about using the multiple data sources is -- and the reason I asked the question about dropoff before is you say a BIRADS 0, 4 or 5.

Assuming that it almost always goes to follow-up, taking the perspective of a health plan instead of the perspective of the imaging center, if you have continuously enrolled members, it is pretty straightforward to look at who had a screening mammo. You paid a claim on it. Then who came back and had either a diagnostic mammo to follow it up or a biopsy, and actually out of the

pathology, you would see a cancer diagnosis coded on the pathology.

So I do think that from the plan perspective there is a pretty good way to get at alternate data as compared to from the imaging center where you are kind of going to chase down that path. That might have happened in a different place.

Disparities of care: I put that as an NA.

So overall, I like the measure, and even within the realm of the patient population, once again, from a health plan perspective you've got a much narrower band of membership or a demographic. You might have like a full Medicaid plan, a full Medicare plan or a commercial plan. So I thought that that might help take out some of the stratification problem.

It is meaningful. I gave that a C, and then harmonization gets between a C and a P. Obviously, I think it should go with the

other measures, and I think it has some additional value, and the feasibility for 4: I thought it was -- given that there is a low dropoff rate, I think the data is generated.

I think the electronic sources are there from the plan perspective. I don't think electronic sources are there from the center perspective, because as soon as it is outside your center, you have to go get it. But if we have -- you know, the EMR eventually comes to be, there are electronic sources available.

Then for exclusions, I put NA.

4(d), susceptibility to unintended consequences: I gave that a Partial, just because of the things that we have talked about where you could bias your sample set.

Then data collection and strategies: I gave that a P. From the health plan, it is pretty good. From the center, it is not as good. There is possibly a manual component there.

Overall, I do think that it is feasible, and overall I like it as a measure even on its own basis, and I think it is a little bit better if you put the other stuff with it.

CO-CHAIR GAZELLE: Thank you. Are there other comments, first from the group that reviewed the mammo measures, and then from the group as a whole?

DR. GEMIGNANI: This is Mary Gemignani. The only other additional comment is I wouldn't endorse it on its own, this one, because I think that it has the unintended consequence of being able to provide a rate that is really meaningless.

So the question becomes, if you have a high recall rate, is that a good thing or a bad thing; but if you don't really know what your cancer is within that population risk, if you are just having -- you know, an individual woman wouldn't know whether to go to Center A or Center B, if you gave her two

recall rates. They are going to say, well, maybe I don't want the extra radiation from mammography. So I am going to go to Center A that has a 12 percent recall rate. But she should really be going to Center B that has a higher cancer detection rate, and they may have an 18 percent recall rate.

So that is the only caution I have when I reviewed this one about this measure.

CO-CHAIR PETERSON: I am still at a loss. I can't quite get how -- it seems like there is such a uniformity of views, if this measure has meaning. There is a good high number or a low number here?

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. Just to put it into context, if you look at how individual physicians perform, the variation in the recall is two percent to 27 percent.

So the example that you gave of going to a facility that has an 18 percent recall rate, I would strongly disagree that

that is a place to go. There is no overall benefit above a certain level, but you don't find those cancers if you have a low recall rate. So at the extremes of recall rate, I think it is clear that you are spending a lot of money. You are doing a lot of tests, and you are not getting much bang for your buck.

So at 26 percent, it is easy to say that out of 1,000 mammograms, we are looking for five cancers, but you are calling back 250 women to find them. That is a lot of recalls.

So at the extremes of recall, it is very expensive, and you are not getting much.

CO-CHAIR PETERSON: Let me just try it this way. Two centers; both have rates of 10 percent recall. One of them is sending the right 10 percent on recall. The other one is sending the wrong 10 percent. Do you know which 10 percent is good or bad?

DR. D'ORSI: That is why everybody

is saying this is no good as a standard.

DR. SMITH-BINDMAN: That is the other measure. That gives you the bang for your buck. I think the example you gave of 18 percent -- that is a pretty high number. That wouldn't be acceptable to me.

CO-CHAIR GAZELLE: Rebecca, two comments that I think you probably know a lot about, but my reading of the literature suggests that, one, there is variation between initial mammogram and subsequent mammograms at the recall rate.

Two is -- and we got hung up on this at the last cycle of this committee, setting the threshold at 10, which is the least stated here, when the average is 9.8 or 11, depending on which study you are believing, and sort of the range from the -- whoever published this study -- the range from the big Rosenberg study was something like 6-14 percent for the middle 50 percent. So --

DR. SMITH-BINDMAN: But, I mean,

Rob focuses on the interquartile range. So the standard that are set for the ACR don't really make sense. The purpose of this guideline is not to identify half a facility is just not doing a good job.

So I think, separate from is the measurement good, what threshold are we going to define quality. I would sort of question this because it's the only thing I keep raising, whether or not you need stratification of the recall rate. The recall rate goes up two or threefold with age, and even within a HMO well defined screening population, that range will go from 40 to 80, and that is where the recall rate goes up substantially. Well, I actually take it back. It is higher, and then it goes down, some factors, but what's the big difference?

MR. BACKUS: If I look at the population of 40-65, how much does that recall rate move?

DR. SMITH-BINDMAN: A factor of

two.

CO-CHAIR GAZELLE: Everybody has a blend. This is Scott Gazelle. The real question is not that. The real question is what is the extreme of variation due to different age make-ups in different practices?

DR. SMITH-BINDMAN: And in this one, to argue -- this is Rebecca Smith-Bindman -- about what is said, the recall rate, I think that will be driven by the quality of the mammography rather than the patient mix, because now we are twofold to threefold difference.

CO-CHAIR GAZELLE: Carl?

DR. D'ORSI: Carl D'Orsi. Let me bring something else up that clinical mammographers know. About 25 percent of recalls are due to what is called fake densities. You look at a 2(d) image, and you don't know whether it is real or not -- 25 to 30 percent.

Those are drastically diminished

when you have a prior exam to study. So if you have a facility that doesn't have a closed population, that tends to get people from various sources, they are not going to have as many prior exams, and their recall rate is going to be up much, much more than the age stratification.

So that is just something you don't realize until you do this.

CO-CHAIR GAZELLE: This is Scott Gazelle. That is the value of stratifying -- at least considering stratifying both by age and by first versus --

DR. D'ORSI: It is very high if you don't have prior exams.

CO-CHAIR GAZELLE: So other comments on this measure, in particular? Ray?

DR. GIBBONS: I am like Eric. I am baffled by the mathematics. So if my recall rate is slightly higher but within the acceptable range, but my earlier measure of PPV2 is slightly lower, is that good or bad?

CO-CHAIR GAZELLE: I would say that is what you expected.

DR. SMITH-BINDMAN: By definition.

DR. GIBBONS: Okay. But what are the magnitudes that you would expect, or do we know that? In other words, from a quality improvement standpoint, if those are my measures year one, and then year two, is that good or bad? Am I getting better or am I getting worse?

CO-CHAIR GAZELLE: So could I ask for clarification? Our role is not to define the threshold or the standard so much as to define the measure that would be used for reporting. Is that correct?

DR. BURSTIN: It actually varies very much by the measure. I am still struck by -- the question is how useful is a continuous measure if it is uninterpretable? So I guess the question would be acceptable -- I am being hyperbolic intentionally, just not about this measure specifically, but just at

times that is when measures get -- you have been trying to identify -- you guys keep repeatedly talking about that tale where there is potential for quality.

The question would be, if you put all these -- and I agree, my head is spinning from the math as well in terms of the small numbers here. But is there a tale of poor quality here that you are really trying to identify, in which case a threshold might be something to consider. Again, it might be something we would like to hear from the developer.

CO-CHAIR GAZELLE: So I could imagine that we would say -- we could come to the point, perhaps not today, where there would be three measures, and they would be taken as a suite of mammo measures, for example, and to obtain a passing grade, you had to be within range from all three, for example. Conceptually, I could imagine that.

I think the data exists for us to

get to that point, but it is the discussion of individual measures versus combining them that may be a challenge. I'm sorry, Judy. Go ahead.

DR. ZERZAN: This is Judy. But what happens when two of those measures, as the example that you just gave -- when you get better at one, you also get better at the other one -- what is the utility of having two measures that you expect will change in the right direction together? What you really want is something that is going to get at a different piece of that to try and get at both accuracy and reliability.

CO-CHAIR GAZELLE: And that is why you need all three.

DR. ZERZAN: If one going up always means the other one is going to go down, assuming that those are the good directions, then why do you need both?

DR. SMITH-BINDMAN: They don't necessarily go in that direction.

DR. D'ORSI: Yes, they do. The false positives and false negatives vary indirectly. So what you have to do is get a balance. Obviously, if you call everyone back, you are going to have a little higher cancer detection rate, but if you are working where normal people work, in the middle, in order to get that little extra cancer detection, you are going to have to call a hell of a lot back.

So you cut it off there. Okay, you are now, yes, doing better for cancer detection but, boy, you are calling back 800 women to see two cancers. So it is a balance, and so the edges are important.

CO-CHAIR GAZELLE: Scott Gazelle. Carl, that is only correct if you assume everybody is operating on the same ROC curve.

DR. D'ORSI: Correct. That is true.

CO-CHAIR GAZELLE: And they are not. We know that they are not. That is why

we have multiple measures to get at the people who are not on the same ROC curve.

DR. D'ORSI: But that is an indication of education, not metrics, to get people on the same --

CO-CHAIR GAZELLE: Not necessarily.

DR. D'ORSI: Sure it is.

CO-CHAIR GAZELLE: It's an indication of people's ability to perform.

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. We studied several hundred doctors who read several million mammograms, and we plotted them all in this ROC space, and there were a few doctors who recalled everybody and found most cancers, a few doctors who recalled nobody and found no cancers. The vast majority of doctors were in the middle. There was no threshold association. Some were good, and some were bad.

So we want to identify the doctors

who were bad, and I would argue the main way I want to find them is they are not finding any cancer.

MR. BACKUS: And that is why you -- Mike Backus. That is why you want cancer detection rate on the bottom?

DR. SMITH-BINDMAN: Right. Then I get past cancer detection rate, and I say, okay, you've met the threshold, but you are doing two times or three times as many tests for the cause; let's see if we can move you. But I think Helen's idea about the extremes are very clear. There are people who are just not operating at a safe level, and that is what it would be great if these metrics could identify. Either they are finding no cancer or they are doing too many tests.

DR. D'ORSI: That does relate exactly to what I said. If the false positives and false negatives vary internally. If you are not finding a lot of cancers, you got a lot of false negatives; and if you have

a high false positive rate and a low false negative rate --

DR. SMITH-BINDMAN: Those are 10 doctors out of the 270.

DR. D'ORSI: Well, that is who you want to cut out.

DR. SMITH-BINDMAN: No. You want to get rid of them, too, but you also want to do a better job of figuring out who is not coming up with a minimum standard.

DR. D'ORSI: That is an education thing. That is moving along a curve. That is not moving the curve up or back. That is moving along a curve.

CO-CHAIR GAZELLE: So, Helen and Ian, we are at noon. Should we take a lunch break now and then come back to the developer comments? Is that a logical break point?

DR. BURSTIN: Do people feel like they are ready for that yet? Or do you want to just -- food's here. It's right there. Be easy enough to grab a plate and come back.

CO-CHAIR PETERSON: Work through lunch?

CO-CHAIR GAZELLE: We are scheduled for an hour for lunch. Why don't we take 20 minutes or half an hour to get lunch, do whatever anybody needs to do in terms of catching up, and then try and continue the discussion as we are eating lunch.

(Whereupon, the foregoing matter went off the record at 12:03 p.m. and resumed at 12:45 p.m.)

- - -

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

(12:45 p.m.)

CO-CHAIR GAZELLE: All right. We got an extra 15 minutes for lunch. We are ready to go again. To bring us all back to focus on the mammo measures, we have reviewed four of the mammo measures, number 1, 2, 3 and 4. We are going to leave off number 9 for a moment to consider the four that were proposed by the ACR.

I think what I would like to do is take about a minute to summarize what I think I heard, which was that we had positive things to say about each of those four measures. We felt that there is probably greater value in some combination of them, not necessarily all four but possibly three, than any of them alone.

We had some concerns about exactly how to interpret the four measures, either alone or in combination. So I think what we should do now is take comments from the

measure developer.

Larry Bassett is here from the ACR to respond, I think, to the discussion we had this morning, make any other comments about the measures or how you would like to see them taken together. Then we can have some more discussion about those four measures, and then we can go on to discuss the CMS measure, which was number 9.

DR. BASSETT: Okay. This won't be a long time, but I wanted to just review some of the things we put forward and what you all said, and then maybe add something else to that.

So I just am not sure you are aware, but in 2005 the Institute of Medicine published a recommendation for a more comprehensive medical standard than required by the mammography quality standard.

Currently, the MQSA now only requires a report on the positive predictive value for biopsies, and so this is really very

minimal. So they recommended to revise and standardize the requirement by the QSA.

Now the question has been why not just the recall rate. The developmental studies have shown the recall rate alone is not a reliable standard. While very high recall rates can reach more cancers, as we talked about, there are negative effects such as the quality of unnecessary biopsies, and this has been in the public attention, particularly when it was published in the Preventive Health Service report.

It is also important to know if a facility's very low recall rate is associated with too many missed cancers. So this again is a balance. We will talk about what that balance should be in just a second.

So what else do we need to know except just the recall rate? We probably want the cancer detection rate, as was discussed here, percent of cancers detected for the number of biopsies recommended in PPV2. That

can be based on screening exams or diagnostic exams, and I don't want to get into this, because this is something that was brought up by Dr. Rosenberg, and it is really complicated.

I could just say briefly that it turns out most of your high end facilities, at least the ones that are recognized nationally and so on, do not get a 4 or 5 on the screening exams.

DR. SMITH-BINDMAN: The way it is used by most people is not the way it is used by what you are calling your high end something.

DR. BASSETT: Yes. But we don't know for sure how many -- I think that the BCSC had problems with this, too. A lot of places recommend biopsies on the screening exam.

We don't do it, have never done it, for a lot of reasons. One is we want to work it out carefully. We may want to do an

ultrasound, and we don't like to inform the patients by telephone. We want to talk to them one on one and show them what we are looking at. But it is not standardized.

Other information that we are not recommending but in the long run is probably reasonable is what is the size of the cancers detected. If you are detecting a lot of cancers in your population but they are all large, then that is not really a good sign.

Also, for example, most of them today should be a centimeter or less, if it is a screening exam, and that is why the whole staging system was changed only a few years ago, because most cancers now have moved from the larger sizes to those that are in the centimeter or less range, which is Stage 1. So they had to restage Stage 1 into A, B, C and D, including carcinoma in situ.

That is a good sign this is working, but it also means that we have to look at that as well when we are evaluating.

Are they detecting little cancers, like we are hoping, ones that are curable, or are they just finding big ones?

The stage will also determine how the treatment is. Since we now have mostly at Stage 1, we can subdivide that and then determine do all these really need the full treatments we have been giving for the advanced things? That allows us to do some research in that area, too.

Also, these particular metrics that we talked about, the cancer detection rates, positive predictive value of 2, and the recall rates -- they are in the literature. They are recommended, I think, in the literature, including the Agency for Health Care Policy Research Guidelines for Mammography, which was published almost 15 years ago now, had some ideas for what those numbers should be as a consensus of the people on the panel.

Subsequent studies by the Breast

Cancer Screening Surveillance Consortium, and a new publication that is going to come out in Radiology from the Breast Cancer Surveillance Consortium are going to give some guidelines, again, on what those metrics should be.

So we do have stuff in the literature to look at that say what it should be. We don't have to develop those. They are there.

Then in addition, I should mention, because I think I have been hearing at this table something over and over again; that is that not everyone is collecting their data, that how do these certain facilities collect the data if they don't have the data systems or the mammography modules that are currently made by private companies.

In addition, I told you that we don't have patients who don't get their biopsies done. Why? Because we have a special person, a quality assurance person, who tracks them down, finds out where they

are, why they haven't done it, did you forget it? We talk to their referring physicians. We have very few that don't get done. However, you said, I think, earlier that there is a large number that don't get done in large practices. So there's lots of issues that affect these patients' metrics.

The other thing that the IOM said just based on what I just mentioned and what we have all been talking about is that they suggested a proposal for a voluntary advanced medical audit on a national level.

What they want to do is make it accessible to people to find out, okay, well, what about a community like mine? What are the rates in that community, in those communities, and to be able to find out how they are doing compared to other people.

That is not acceptable to all of them, as you all mentioned, because they can't always find out if the biopsy was done somewhere else. If we did have a national

mammography database, we would be able to find out if that patient on follow-up did have a breast cancer or not.

So this is something we are lacking in this country that we have in other countries that we think would be a better solution in terms of giving an incentive to a facility in terms of their payment, if they belonged to a National Mammography Database. I think that would be an incentive that would really help create an improvement in the overall managing of these patients.

So that is basically just my summary, but how we look at this, and just to tell you, the ACR National Mammography Database metrics are the same ones that we recommended here and the same used by the BCSC databases.

They could provide access to national and regional aggregate data for the participants. They are a quality improvement tool for physicians and practices, and some

facilities may not understand when you ask them for metrics, and they need to be provided guidance from some kind of a group that they are working with, whether it be the National Mammography Database or another organization, so they could get the right information in, because sometimes they are sending the wrong stuff.

We all have problems even understanding the recommendations in the centers, but think about these people who are not physicians or the quality assurance person in that practice. Many radiologists do not collect data, cannot evaluate the outcomes relative to the BCSC or other benchmarks. So it is essential in order for them understand how well they are doing.

Again, I think I would recommend the work group joining the National Mammography Database with the goal of improving overall quality of mammography, as much as any other incentive.

Lastly, but not least, this is not mine. Carl has been mentioning this over and over again, and that is that there is a relationship between sensitivity and specificity and recall rates and low recall rates. It is very complicated, but it has been mentioned. Carl, did you want to comment on that?

DR. D'ORSI: Just that you don't get something for nothing. That is the no free lunch curve.

DR. BASSETT: And that is it. Thank you very much.

CO-CHAIR GAZELLE: Thanks very much. I think this would be a good time for anyone to ask questions of Larry, representing the measure developer, if there are specific questions about these measures that are still unanswered that we like. Don, then Rebecca.

DR. RUCKER: I think you mentioned the IOM report at the very beginning, but I am trying to understand the overall magnitude of

the problem here.

I am a little puzzled, because as far as I can tell, mammography is the most heavily audited activity, just about, in all of medicine, and maybe cardiac surgery and some of the CAD stuff being runner-ups. So in that environment where there is already a ton of oversight as opposed to almost everything else, I am just puzzled, or not clear, that this would add on top of all of that.

DR. BASSETT: It is very highly regulated, but the regulations in terms of a medical audit are pretty simple. You just put your positive predictive values for the biopsies you did and, as we all know, one of those metrics alone doesn't work usually. It can depend on -- I mean, I could get what sounds like one of the numbers, but my community may provide that because the patient population is so high and the fact that they are very good about coming for their exams and at a higher level socioeconomically.

Lacking that, somebody who is in the countryside doesn't have a place to look and see what the metrics are for their kind of population.

CO-CHAIR GAZELLE: Rebecca and then Howard.

DR. SMITH-BINDMAN: I participated a little bit in the IOM report, and I think what the brunt of it was, is there is this test that is being used a lot. There is pretty high quality for the technical aspect of this test, but there is much less consistency in the quality of the interpretation. There are still gaps in terms of under represented groups not having access to it. So it really focused on how to improve the quality of that.

So if you looked at some of the other points, it was on how to we improve the quality.

DR. BASSETT: Yes, and the technical part, as you just mentioned, the

referred ledgers have to be reviewed by an on-site entity, and they have to be pretty perfect in order to be accepted for presentation. They've got to have the medical tests done on a regular basis. There are all kinds of other reasons. But the medical audit request is very minimal, basically one metric.

DR. FORMAN: I was on the committee that did the MQSA reauthorization report, whatever you want to call it at the time -- I think it was the Committee on Improving Mammography Quality Standards.

Our charge at the time -- We were doing this because MQSA was coming up for reauthorization. It actually got reauthorized, and then this report came out. Subsequently, some of it has been put into place in a regulatory way.

The concern that was raised in the committee, and a big part of the committee report that is not necessarily reflected in these standards, was the access issues as

well, and the fact that the higher the regulatory hurdle in probably on the most low reimbursed parts of imaging was actually -- could adversely impact access to care, while not necessarily connecting to improvement in imaging outcomes, because one of the things that we observed and we really were able to slice whatever available data there was at the time, and find that, despite what we might anecdotally or even in small empirical fashion identify as being quality improvements with certain high quality mammographers and mammography sensors, it wasn't linear at all.

I mean it wasn't linear at any point in the curve, that if you had higher volume, you are necessarily going to be better. These were great concerns to be able to try to regulate or mandate the use of measures or mandate a mandatory audit at a higher level as opposed to a voluntary audit, that it would actually drive out access to mammography at that time.

That is why, I think, the ultimate report was a lot softer than a lot of us thought it should be going into it. I think sitting here and listening to us talk about these measures, I feel like I am at the exact same meeting just seven years later or six and a half years later, because it is -- you know, I think what we felt back then and what a lot of you are implying right now is it would be great to get this data.

We are not sure we know what to do with it, once we get it. We are not really certain that any of these metrics on its own or even if you could come up with a scoring system would allow you to know who really is a better performer or not, because you can't plot out their entire ROC curve. All you know about is a couple of points.

I just wanted to give a little back-story for that. Having sat through this for, I think, 18 months in 2003 and 2004, I feel like it is deja vu.

DR. BASSETT: IOM actually has been involved here.

DR. FORMAN: That is right.

CO-CHAIR GAZELLE: Arthur and then Rebecca.

DR. STILLMAN: A sort of similar sort of comment. I am sort of struck in the conversation this morning that we have had several reasonable metrics for quality, but none of them are useful in isolation, and that there needs to be some sort of combination.

Yet I have not heard any articulated concept of how they could be combined to develop a true quality metric. I am concerned about making a recommendation without that piece.

CO-CHAIR GAZELLE: Okay. Thank you. Rebecca.

DR. SMITH-BINDMAN: Rebecca Smith-Bindman. My question is not dissimilar to yours. It's two-part. I am wondering, and I think I know the answer, if the ACR would be

interested and willing to come up with some simple stratification schemes that might make some of these measures a little more reliable in terms of being age or possible first and subsequent mammograms. That would be the first part.

The second part: Helen sort of raised the possibility of thresholds. I think, in some ways, it would be much easier to apply a crude threshold where, not so much getting people in the range but identifying people who are far outside what would be acceptable, if that might be allowed and if that might get at what Dr. Forman is suggesting, the need to improve this, but maybe -- we can't do it in subtle ways, but maybe we can put a sledgehammer to this and say above this, you can't assess it.

DR. BASSETT: And that is why it is important to get as much data as possible and, like you say, stratify it.

CO-CHAIR GAZELLE: A number of us

talked over the lunch break. One possible way to think about combining -- so let's say we have three, and we were able to establish threshold or ranges that you had to be within for all three of those, and we actually got, say, a passing score if you were in range on all three.

So if it would be possible to, say, have an upper threshold for recall rates, a threshold for PPV2, and a threshold for cancer detection rate, and you had to be within the range on all three, at least conceptually that could be a way to combine the measures.

DR. BASSETT: Measures and guidelines are out there. One of the problems we talked about was, if you are in an unusual population, that probably would be an issue. But those guideline numbers are there. They are in the original AHC policy and research guidelines for mammography.

DR. SMITH-BINDMAN: None of these

guidelines reflect any of the Breast Cancer Surveillance Consortium 30 publications. So I think those standards need to reflect the literature --

DR. BASSETT: Yes. We have just finished -- I served on a committee, and we came out with a method to try to come up with some recommendations. It is kind of a consensus type of method. It's considered scientific but it's mainly a bunch of experts. That is going to be published in the journal Radiology in the next couple of months. But the metrics are out there. The guidelines are there.

CO-CHAIR GAZELLE: Well, they are. The question is they aren't proposed within -- They are not proposed within these metrics. They are cited, but they are not proposed. So the procedural question is would we -- could we ask the measure developers to come back with thresholds, and then would that count as something that could still be approved within

this cycle or would the approach be to say let's approve these as reporting metrics and then anticipate down the line setting thresholds? I don't know the answer to that.

DR. BURSTIN: Some of it depends on how complex that task is. I am still left at the end of the day wondering -- I mean just to remind us what we said early on. The intent of NQF endorsed measures is that they are only for public reporting.

I guess the question would be: In this current form, are these measures in isolation or in some combined way appropriate for reporting. If the answer is, well, maybe if they are combined, then, obviously, that is a pretty big if. I don't know how a big a reach that is without knowing how easy it is.

There is a fair methodology in coming up with composites, all or none, however the case may be. So I don't know how -- not being an expert in this field, I guess my feeling would be I can't answer that

question without knowing how big a list that is in terms of coming up with something.

CO-CHAIR GAZELLE: But, for example, the easiest way to consider it is to say -- I don't want to throw out numbers, because we will get caught up in the numbers -- but we have a threshold for cancer detection rate, recall rate, and PPV2. So you have got to check all three -- You have to report all three, and to get a passing grade you have to be within range for all three.

That is not, for me at least, too big of a stretch, if we had the data to set those thresholds, and I would think the strategy would be to set them fairly broad, to start with, and then consider through this process of public reporting, collecting more data and relooking at it in three years. But at least it is conceptually something I can grasp without needing to have a composite score that somehow weighs each of the measures, and we would calculate the lineal

number. Eric.

CO-CHAIR PETERSON: I think the concept thing is going to be a little -- just a little challenge. It may be doable, but I would have to think through it, because these are measures that are partially quality, partially efficiency, and how you -- I mean where you sit is complex.

Think about how that might play out and the degree to which there would be validation of how many -- do they have enough data and enough time to do this in a short window to both develop the measures and provide me back data to say that this would identify X number is good centers and these many bad.

CO-CHAIR GAZELLE: Don, Carl, then Ray.

DR. RUCKER: Maybe the question is for Carl and Rebecca. If we did a composite, are all of these sort of essentially gatherable from the same stream of information

or is it really sort of, you know, you need to go to one bucket for one set of the composite and another bucket for another? Because I think there is just an economic issue here. It is a very poorly paid, litiginous prone activity. As far as I can tell, most radiologists run away from mammography faster than summer lightning. I mean, as a non-radiologist -- if we are going to do that, we ought to have something that meets some sort of simplicity test as well.

DR. D'ORSI: I can incorporate my comments with that question. I think the ACR and Larry are absolutely correct. We have to start collecting data. When we collect any kind of data to compare, we need a gold standard.

The gold standard is going to be what you are finding pathologically, not only cancer but what kind of cancer you are finding. Once you get that, then you can start setting gross metrics against that gold

standard. A recall rate 2 does not relate to 50 percent or minimal cancer, but this does.

Once you get that, then you are able to make some sense out of a composite metric, but until you do that, you are only estimating, which is okay. What I hope does not come out of this is some rushed measure to come across, just to get something across and it has no validity even on a composite level.

I think the big thing is to start collecting data and working on this, getting what a composite metric means with an X recall, and it doesn't necessarily -- it is not necessarily as simple as you think, Scott, because if you are here, your cutoff may be good here or here or in the middle somewhere on another metric. It may not be in a range. It may be good in the middle, and you may be at an outlier here, but you may be in the middle here. So what do you do with that?

You have to compare all these metrics to some gold standard, which is what

you are finding, stage-wise and curability-wise. That is the bottom line.

To do that, you need tons of data, and I hope these metrics are not going to be yearly evaluated. They should be evaluated over a longer period of time so you have enough hits in each facility to do a valid comparison.

I don't know if that answers.

CO-CHAIR GAZELLE: Ray?

DR. GIBBONS: Ray Gibbons. I think I can understand the concepts of setting acceptable ranges, but I would just offer the caution that, as part of the process of deciding on what those are, you need to look at the precision of the estimates for smaller volume facilities, because working in an area of the country where there is a lot of rural health care, the unintended consequence here would be very severe if you penalize centers out in western North Dakota, who are the only option for women in that area, because of the

statistical noise in their small numbers.

This would be a very bad consequence. So that has got to be part and parcel of this effort.

The second thing is I would amplify the point that Eric made, which is I think this process should be developing measures that facilitate quality improvement for everyone.

Having listened to this discussion, once you have met the acceptable threshold, it sure isn't clear to me what you are going to aspire for the next year with respect to those numbers, from the discussion. It would seem to me that has got to be part of the context as well.

DR. SMITH-BINDMAN: You want a continuous quality improvement?

DR. GIBBONS: Well, something to aim for. In other words, once I am acceptable in those three numbers, does that mean I am good, I'm done, or is there something I should

be aiming for the following year?

CO-CHAIR GAZELLE: You need to do it again next year.

DR. GIBBONS: Well, but aside from just being it again, am I going to be better? Can I be better, and can I facilitate quality improvement in the country in some way, which seems to me ought to be a goal for any measure.

CO-CHAIR GAZELLE: Okay. Others?

DR. BASSETT: Just relating to that, I think it is also important to remember also the facility. So it also helps the facility evaluate their own persons as well as that person evaluate himself.

CO-CHAIR GAZELLE: In response to your comment, Ray, the existing NQF measures -- I don't think any of them have that sort of continuous quality improvement component, which is to say that they have -- As far as I can think of, they have -- They don't have a sort of, if you made it this year, it gets

harder next year component to them.

DR. GIBBONS: Rate of aspirin use post-myocardial infarction is an NQF --

CO-CHAIR GAZELLE: I am talking only about the imaging ones.

DR. GIBBONS: I know, but --

CO-CHAIR GAZELLE: I am just speaking of so far the eight approved imaging ones.

DR. GIBBONS: Right.

CO-CHAIR GAZELLE: There are reporting percentages, but there is not a -- What you are suggesting needs to be there is not there in any of the eight that already are approved. So I don't know that that is the bar we need to pass here today, or else, if we did, we would have to throw out all the others, too. Right? I mean, none of them have that kind of context.

DR. BURSTIN: There certainly are with continual variables oftentimes or your readmission rate may be X or your time to

license may be Y.

CO-CHAIR GAZELLE: But I am talking about the imaging ones.

DR. BURSTIN: Not within the imaging. This is a fairly new area. That is part of what we are seeing here, is it is not tons of measures and years of experience. I think this is a newer area, and the question is still are these measures really at this point appropriate for QI, but are they not yet ready for public reporting, I think, is my major question.

I think even the fact that NQF endorsed measures is the ultimate intent, that they are okay for the use of public reporting, I think that is the question I want the committee to think about, either alone or in combination; and if in combination, I don't think we still have a -- I don't feel like I have a comfort level on what that means, if they are paired and how they would be interpolated.

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. For other measures, not imaging, what proportion of the U.S. population should they be applicable to, for your other measures? So aspirin use -- you know, everyone who is admitted with an MI should be in the denominator. How big a chunk do you need to consider it?

DR. BURSTIN: It doesn't need to be a particular size denominator. I think it is just a question of do you feel like at the end of the day you have a reliable and valid estimate that will reflect the quality.

DR. SMITH-BINDMAN: But if you are looking at mammography quality, you need a large enough mammography facility. You know, Larry sort of slipped in there that this should be used to evaluate the physician level, which is not how we are using it. Then you are even talking more noise, but if only half of facilities in the U.S. would have sufficient volume to use this quality measure,

would that be okay or would that be a measure that is not okay, because it just doesn't find enough? You will have to come up with other measures.

DR. D'ORSI: Or can you grade them by size versus how often you are going to look at these numbers, so you have enough hits?

DR. BURSTIN: Sometimes a measure will be stratified. So, for example, there would be a facility that could only do procedure Y that is getting looked at. I think that is part of the issue here, is you may have a fairly specialized procedure that would be only be happening in a small proportion of facilities.

DR. SMITH-BINDMAN: No. This is happening everywhere. It is happening everywhere.

DR. D'ORSI: You have to reach a certain denominator count before the measure would have value.

DR. SMITH-BINDMAN: And if only

half the facilities could get to that count, would that --

DR. BURSTIN: I don't know. Small sample sizes -- you just can't get a sample size to make it something that is meaningful.

DR. CANTRILL: Steve Cantrill. Just a brief comment about CQI concept. Remember, those of us who work in training institutions, no matter if you have a static endpoint, that is always CQI, because we did the training, and then we graduated them. So we start over with a whole new dumb set.

DR. BURSTIN: That is --

CO-CHAIR GAZELLE: And in fact, even if it is not a new set of physicians, the same physicians having to achieve that performance on a new set of patients is still not entirely static. It is not like you have achieved it once, and then you automatically have it forever.

All right. Now would you like us to do the last mammo measure before we vote on

them?

DR. BURSTIN: I think that makes sense.

CO-CHAIR GAZELLE: So let's shift gears a little now to the one which is IEP-009-10, which is mammography follow-up rate among Medicare beneficiaries. Rebecca, you are the primary reviewer.

DR. SMITH-BINDMAN: I will be honest. When I read this measure, I was a little bit confused exactly what was trying to be measured. So the two possibilities are either it is looking at mammography recall rates, which is very similar to the measure that we discussed just before lunch, meaning of women who are sent for mammography, how many then are sent for additional tests, so recall rate; or if this is trying to measure, of women who are being sent for abnormal mammograms, how many actually come back.

So it is sort of -- it is the former? It is a little bit unclear, but okay.

So if it is the recall rate, then it is very similar to the discussion we had before lunch. I will go through it very quickly. That is sort of how I thought it was, but some of the text was a little bit confusing.

So in terms of how good and how important it is, I think it is a good measure and an important measure, the same as the discussion before lunch.

Opportunities for improvement is also a C.

If I move to 1(c), outcome, given the outcome for this consideration, is sufficiency. This is absolutely important for sufficiency, so it is a C.

If I move to 2, for the numerator versus defined, there are some questions I have with how it is defined, but in terms of in general defining it, I think it is very good. So 2(a) is a C.

In terms of 2(b), reliability

question, this metric is specifically made for use in Medicare data. So looking at the number of women who are insured by Medicare who have follow-up mammograms that our diagnostic defined by billing codes for diagnostic, I am not sure that the data are presented to let me know that the Medicare billing data is accurate for differentiating screening from diagnostic mammograms. So I think that is a significant problem.

The problem is twofold, whether things are captured and, in general, the follow-up rates are low in the Medicare data, and whether you can tell screening from diagnostic. So for 2(b), I gave it an M.

For 2(c), for the same reason, I gave it an M.

MS. STEPHENS: Excuse me. What did you say? I'm sorry.

DR. SMITH-BINDMAN: I am saying I don't have data to know whether the Medicare data are valid for assessing screening versus

diagnostic mammography in a relatively straightforward way.

There are new codes for it, CPT codes. I know a lot about the old codes, and they are not reliable, and the new codes I don't know very much about and I haven't seen the data to support that they are actually accurate.

So just to give people background, in the older codes most mammograms were billed as diagnostic, even though most mammograms were screening, for billing purposes they got higher reimbursement for diagnostic. So they were screened that way. Well, no, I take it back. I don't know why they were billed that way, in fact.

I have actually published on differentiating screening from diagnostic mammograms using the Medicaid data, and you can do it, and I argued you could do it. It just took a lot of work. It wouldn't be a reasonable thing to do. So again, it might be

okay.

For 2(d), there are no exclusions. For 2(e), risk adjustment, I think very strongly it does need to be stratified, but in the Medicare data it should be easy to do it.

Meaningful difference in performance is C. I think there are differences that could be improved upon.

2(g) is a C. There is great data on this from lots of different data sources.

Disparities in care, I gave it a C. There are some differences, not enough to waylay this measure.

3(a), I gave it a C. Harmonization, I gave it a Not Applicable. 3(c) also Not Applicable.

Feasibility, 4(a), is a C, assuming we can assure that the data are valid and reliable. My guess is we can, but then it would be an easy data to use electronic sources. C, exclusions, NA; 4(d), N.

Feasibility, I think, is a C; and

recommendation: I think the issue of validity needs to be established, but if they are, I guess it is risk adjusted or risk stratified, and I think it is a good measure overall.

CO-CHAIR GAZELLE: All right. Thank you. Other comments from the mammo review group before we throw it open to the whole group? Carl?

DR. D'ORSI: Let me just go down these, if that is okay with you, go down the numbers again, just on the ones that I had questions on.

CO-CHAIR GAZELLE: Sure. I'm sorry. Can you try and speak up a little?

DR. D'ORSI: I'm sorry. I am just going to go through some of these that I wanted to make some comments on, on this metric. I'm sorry. I will speak louder. Usually, I don't have that trouble, being from Brooklyn.

One of the things that Rebecca mentioned, which to me is problematic, is the

method that was developed to measure this recall rate. Remember, this is a recall rate attached to an event that happened previously, not an individual event.

So let's take this scenario, which is not uncommon. A woman comes in, has a screening mammogram. She has no symptoms. She hasn't seen her doctor for a year. She has her mammogram, and correctly is read as a 1. She goes away, and she says, oh, boy, I had better go have my exam now. She goes in, but two weeks later says, gee, I feel some thickening here: Go back and have your mammogram and an ultrasound.

Within 45 days, that gets tagged onto the normal mammogram as a recall, which it is not, and that is not an uncommon scenario. So I think that data is going to be corrupted by not a small amount. So I have a problem with measuring so called recall rate using that type of metric.

The other data that was used in

1b.2 to support the metric as a single event, one of the studies that was quoted was a 2005 study that says you should be within 4.9 to 5.5 percent as a good tradeoff between sensitive and positive predictive value.

If you look at that article, that was not the thrust of the article. Their basic conclusion was, when you compare performance metrics with other order programs, the time frame for a screen is important.

So those metrics can vary whether that woman comes in for a screen at 12 months, 18 months or 24 months. So that is an unfair statement to make regarding that article.

Another article, a retrospective study that was quoted -- this is also in 1b.2 -- was the lack of integrating what we discussed before the benchmarks, and I think we had enough discussion on that.

Let's see, what else do I have? The other thing is ethnicity. I think there is data coming out that not only is the breast

cancer different in African American women, but is more prevalent. You might want to consider that. No?

DR. SMITH-BINDMAN: No.

DR. D'ORSI: How no?

DR. SMITH-BINDMAN: Overall breast cancer rates are lower in African American women. The distribution of higher grade and higher stage tumors are higher. So they end up having worse outcomes, because the tumors tend to be in a higher grade, but in terms of the prevalence of disease, it is overall a little bit lower, which probably is just a reflection of screening.

So the true prevalence of disease is probably the same. Hispanics and Asians tend to have slightly lower breast cancer rates. Asians also have lower stage, but in terms of the pool of breast cancer in the U.S., it is remarkably stable by race and ethnicity.

DR. D'ORSI: That is all I really

had.

CO-CHAIR GAZELLE: Thank you, Carl. I have two -- yes, please?

DR. SMITH-BINDMAN: Can I say just one thing to agree with Carl. I think the measures, though -- the range of acceptables that is presented in that is not nearly specified enough, and I would expect -- you know, because I think it needs to be age stratified and screening cycle stratified, the numbers don't make a lot of sense, but those numbers that are cited, again, need to reflect more time limited.

CO-CHAIR GAZELLE: Thank you. I have two issues with this. The first is the general question, I suppose, of the -- I understand why it is valuable to CMS to have a measure that applies only to Medicare beneficiaries. I am not sure I understand why it is valuable to us or to NQF to have a measure that only applies to Medicare beneficiaries when the condition and procedure

of concern spans that.

It would be one thing if we were talking about a procedure that is only done in people over 65, but here we are talking about something from, say, 40 to 75.

DR. SMITH-BINDMAN: I'm sorry. Isn't this the same as measure 4?

CO-CHAIR GAZELLE: Except it only applies to Medicare beneficiaries, as specified. So my question is, you know, since they are similar, why would we choose this as opposed to one that applies to everybody?

DR. SPENCER: It makes the feasibility higher, doesn't it?

DR. ZERZAN: It is a huge payer, huge payer, in this category especially.

CO-CHAIR GAZELLE: I understand why it is important to measure, but I wouldn't support it personally as an NQF measure, because it is only 10 years of the, say, 35 years of mammo screening that is covered by this. So in my own opinion, I would rather

see measures that apply to the full spectrum of the condition.

The second issue is -- and I may be missing something here -- that it only applies to hospital claims, so hospital and it specifically excludes screening done in non-hospital facilities, and a lot of screening is done in non-hospital facilities.

So it is further narrowed in terms of its broad applicability. It does allow for the numerator hospital and non-hospital facilities to fully capture all of the events from the denominator patients, but the only way for someone to make it into the denominator is for the index screening exam to be done at a hospital facility, at least as worded. So I think that is a problem with the measure as well.

MR. BACKUS: This is Mike Backus. I agree with you that the hospital is too narrow. I think Medicare gives you two huge advantages, though.

One is the feasibility, because what you have taken out is the insurance question. So the ability to have the exam or the follow-on care paid for comes out of the equation. So I think you are probably more likely to have true follow-up or -- I mean, we talked before about the FQACs and how you could get a mammo, but then you can't get the biopsy paid for. That piece has been removed.

You know, the Medicare dataset -- it gives you the ability then actually to -- you know, if you are going to work in that dataset, you can head down the biopsy road as well, because you are going to get a path report, and it is all coming through one payer.

CO-CHAIR GAZELLE: But this is only about the follow-up. This is not about the biopsy.

MR. BACKUS: I understand. I am just saying that, as you -- if you think about where that measure might go over time, the

ability to have that dataset becomes --

CO-CHAIR GAZELLE: I see what you are saying with respect to biopsy, but I can't imagine a situation where the screening was covered, but the follow-up diagnostic was not covered.

MR. BACKUS: Right, in Medicare it is. In FQAC wasn't the exam --

CO-CHAIR GAZELLE: Only the biopsy was not covered. To the degree that you get -- you take out the insurance coverage question.

DR. ZERZAN: In Medicaid you fall off, and then maybe you have to reapply, and then it is another whatever period of time. So I think from that perspective, it does take out that insurance piece of the question, the access piece. You know it is covered. So it should be there, and this should be able to be sort of the best case scenario, because the extraneous factor has been taken out.

DR. SMITH-BINDMAN: This is

Rebecca Smith-Bindman. You are saying why start with Medicare. The answer might be this is the only place you can start, and maybe if you have this measure that is endorsed and you can see how it does, it might give you more insight into other data systems. Currently, with small groups, you don't have enough data, but maybe -- I don't know, but as a place to try it, it might be interesting.

MR. BACKUS: You would also address some of the stratification question, because now you are doing the 10 over the year, so to speak, instead of 30. So you have narrowed your stratification piece down.

CO-CHAIR GAZELLE: Clearly, you do. My issue is that, if we said for the other measure that recall rate wasn't valuable freestanding, by itself, and now we are saying this is essentially a recall rate. This is a slightly differently phrase recall rate measure, but the same problems exist. This is valuable as a stand-alone.

DR. SMITH-BINDMAN: But we could help them by suggesting that they could get at cancer detection rates, that they could identify breast cancers pretty accurately, about 80 percent in the dataset, maybe close to 90. So I agree --

CO-CHAIR GAZELLE: It doesn't exist.

DR. SMITH-BINDMAN: It doesn't work as it is.

CO-CHAIR GAZELLE: It is not a proposed measure.

DR. SMITH-BINDMAN: It is not stratified now. It is not adjusted now, but your concerns are completely valid, but as a measure they could also care.

CO-CHAIR GAZELLE: So the other question, though, that we haven't addressed is the why hospital only for the denominator event. I think it ought to -- and I am assuming it is because of some data feasibility problem.

MS. DaVANZO: No, no, not at all.

CO-CHAIR GAZELLE: Then what is it?

DR. BURSTIN: What is the logic?

MS. DaVANZO: -- hospital outpatient quality data reporting. We have to start with where our data is. We can look at it for IBPFs. We can look at it --

DR. SMITH-BINDMAN: This is Part B data we are talking about?

MS. DaVANZO: Part B. I can look at it for anything.

DR. SMITH-BINDMAN: How is it written? Is it written Part A or Part B?

DR. BURSTIN: Just a point of clarification. It is really important to -- obviously, we want to get at the best quality measure we can here. I think the Medicare-only issue, obviously, is we do routinely endorse measures for Medicare-only, because the data -- for example, the readmission rates, for example, for CHF pneumonia. But

the issue there is they are sometimes an older population, to start with.

I guess the real question would be I would like to find out what proportion of mammograms, in fact, that could have been at this rate are excluded because it is only Medicare.

The second question is what proportion of mammograms are excluded, because it is only hospital outpatient departments. I think my preference would be that, if possible, you would actually want to have the measure be broadest as possible, allow CMS to stratify it for their own payment rule issues. That is not our concern. NQF doesn't do payment. We do the quality measures.

So I think one recommendation of that might be, if the data is doable, why not do it for the entire population at facilities. You guys can stratify it for whoever you need to, for whatever payment rules you have, but the bottom line -- Scott is right. I'd like

to know what proportion of mammograms are done in hospital outpatient facilities versus not. Is that a known answer?

MS. DaVANZO: Sure --

DR. BURSTIN: It's got to be pretty small.

MS. DaVANZO: But the follow-up is in --

DR. BURSTIN: Exactly.

MS. DaVANZO: -- you can easily go through the initial screening mammography facilities --

DR. BURSTIN: I see.

MS. DaVANZO: -- as well.

CO-CHAIR GAZELLE: My point is that we have to separate what is important for NQF versus what is important for CMS, and it may be valuable for CMS to look at only hospital denominator events, but I don't think it is valuable for us. And as someone said, they could look at that on their own, if they wanted, but this is not a CMS committee. This

is an NQF committee.

DR. BURSTIN: But again, I think for Medicare only data issues are really quite reasonable.

CO-CHAIR GAZELLE: Right.

DR. BURSTIN: I do think the issue, though, of facility only versus hospital outpatient is one that I am not sure is justifiable.

CO-CHAIR GAZELLE: The only issue I have with Medicare only is if we are also proposing and supporting essentially the same event that is not limited to Medicare only --

DR. BURSTIN: Right, and this has come up repeatedly before as well.

CO-CHAIR GAZELLE: -- having two sort of competing same measures may be a problem.

DR. BURSTIN: This has come up repeatedly before as well. So at times NQF will endorse two measures when there are different data sources for the measures or

distinctly different populations.

So the question may be if there is -- this is logical on the Medicare side, given the data source. The key issue from our perspective is those measures have to be harmonized. They can't be different. You've got to be able to have apples and apples at the end of the day, accounting for the -- Obviously, there may be significant differences based on data source, but at least in terms of the way you are coming up with the recall rate, it has got to be defined here.

CO-CHAIR PETERSON: To clarify two things: One, we probably have an idea of the age breakdown of mammograms. Right? What is the percent 65-plus of all mammograms?

DR. SNOW: Percentage of all mammograms on people older than 65? I don't know.

MS. DaVANZO: We did a study at MCDS so we could combine the claims in the clinical and survey data that was in the

Medicare current issue survey, and we used the 2005 data, because that was the last one that had the claims in full over the period. We found that 22.7 percent of women, though it can be men as well -- but we found 22.7 actually got their screening mammogram in 2005, and then --

CO-CHAIR GAZELLE: Right. The question was what percentage of all screening mammograms are done in the Medicare population.

DR. BURSTIN: Right. So the MCDS is only Medicare. We are asking the broader question. So we are asking what proportion of screening mammograms are done for the Medicare versus the non-Medicare population.

CO-CHAIR PETERSON: Okay. So we are hearing somewhere in the 30 to 40 percent are, so a substantial minority.

Anyway, the second question is inpatient versus outpatient -- do we know that breakdown?

CO-CHAIR GAZELLE: No, it is not inpatient/outpatient. it is outpatient hospital versus outpatient other sites.

CO-CHAIR PETERSON: You have no idea? Do you guys have an idea?

DR. DEHN: Of all diagnostic imaging, 15 percent is now done individually. I would think that it would be far less than that for --

CO-CHAIR GAZELLE: No, that is not the question. The question is: Of all the mammos which are done as outpatients, what percentage of them are done in hospital associated outpatient facilities versus IVP or that are nonhospital facilities?

DR. DEHN: Well, it is apparent, obviously, on --

CO-CHAIR GAZELLE: The only question is average across the country, what the answer is.

DR. DEHN: Twenty percent, 25 percent at hospitals, and it is increasing,

because hospitals are buying practices. So those practices in which diagnostic imaging is performed is considered hospital.

CO-CHAIR GAZELLE: I understand that, but what we are trying to get at is mammography, not all diagnostic imaging.

CO-CHAIR PETERSON: So am I right in saying on the low end of -- the lowest extreme, this measure would account for 30 percent and then 20 percent of the 30 percent. So that would be six percent. That would be the low end.

MS. DaVANZO: No. The thing is, about 40 percent of women get mammograms in general.

CO-CHAIR GAZELLE: That is not the question.

MS. DaVANZO: In the Medicare surveys, we got a slice in time. So it was the people in Code 5 that got it, and there is a two-year -- you get it every two years.

CO-CHAIR PETERSON: All I am

saying is of the tests ordered, not of the people -- of the tests ordered, what percent are you capturing in this measure. You don't capture under 65. So that is 60 percent of the mammograms, approximately, or 70 percent.

Of the mammograms in 65-plus, you don't capture the outpatient nonhospital right now, and that was said to be 80 percent of the study. So if you took that --

DR. DEHN: I think there is a question before. If you choose that we include that, though that wasn't our mandate.

CO-CHAIR GAZELLE: So if we chose to approve it, we could choose to put the condition that it has to include in the denominator all mammography screening exams. That is what is in our purview.

MS. DaVANZO: Yes.

CO-CHAIR GAZELLE: And you could do that?

CO-CHAIR PETERSON: Okay. Then we are back to the question of what the measure

means by itself, which is where we are.

CO-CHAIR GAZELLE: Yes, which is where we are. So I think we have already -- before we turn it to formal comments from the measure developer, let us ask if there are any more questions from the committee or comments from the committee, either on the measure itself or on the merits of the measure -- a measure such as this in the absence of the other sort of balancing measures.

DR. FIESINGER: You are saying it is all the same recall rate. Do we need a similar measure, really, or can they be merged together, have one measure for everyone; because there a number of measure exploding every year to this group. We have measures on measures, and when I am practicing and seeing patients, it is very intimidating and costs a lot for practices to measure all this stuff.

So there if is a way to save a measure and achieve the goal, I would be in favor of that.

MR. BACKUS: This is Mike Backus. With this measure, we are suggesting, gets measured out of CMS data. Right? So essentially, there is no additional cost to the practice.

My question on the measure is: do we think that, because Medicare has a more stratified population -- right? You are only working 65 and over, excluding the disabled -- that you have taken out enough of the population bias that recall rate by itself is now substantially more meaningful and can stand on its own, or do you still need PPV2 to go behind it?

DR. SMITH-BINDMAN: Are you saying one measure is good enough in this population?

MR. BACKUS: I am always brought up, because I work in it -- it is like crawl, walk, run. Yes, there is a gold standard. I mean, there is a gold standard -- right? -- where you want to know the tumor size and -- but we will wait for the electronic health

record and being here in 20 years, but from CMS' perspective, if they are trying to get close, does this narrow it enough to be worthwhile? And I don't have a view.

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. I was going to say a very similar point. I still think it needs to be stratified by age, but if extremes of poor quality were set in this measure, then I think you could identify those extremes with just this measure standing alone.

CO-CHAIR GAZELLE: Scott Gazelle. I assume this is a facility-level measure. Is that the intent? So basically, we are judging the facility and how it manages its Medicare patients. Okay. Right? If it is a facility-level in a Medicare setting -- so is that valuable?

DR. D'ORSI: Carl D'Orsi. This is a facility-based metric. You tie it to the woman. What happens if she goes to another facility for that diagnostic exam, the

screening?

DR. SMITH-BINDMAN: It should be in the range. She is billed.

CO-CHAIR GAZELLE: So it is the facility of the denominator, I would assume.

DR. D'ORSI: Got you. Okay.

CO-CHAIR GAZELLE: Eric?

CO-CHAIR PETERSON: Eric Peterson. Sorry, one more time. Clarification of what is good quality or bad quality? You said you could use it for that. How?

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. If a facility recalls more than 20 percent of their patients for additional mammography, that is a measure of poor quality and large cost. After a recall rate of about 10 percent, you are not getting much in the way of cancer detection. So we will give them from 10 to 20 to waste those resources, but above 20, whatever that cutoff is, that is poor quality.

CO-CHAIR PETERSON: And you would

argue for then some sort of binary?

DR. SMITH-BINDMAN: Think of it as a bubble in the window of a level, too much above, too much below.

CO-CHAIR PETERSON: Do we have or were we provided data that said what percent of institutions fall in that greater than 20?

DR. SMITH-BINDMAN: It turns out the way the data were presented were not age-stratified, were not first and subsequent, ended up being very misleading.

CO-CHAIR GAZELLE: They do present first and subsequent.

So let's finish comments from the committee, because I can't find it this moment, but there were data on first and subsequent.

DR. SMITH-BINDMAN: They come up with about -- recall rates of about 10 percent with a very narrow distribution. It was very low.

CO-CHAIR GAZELLE: Let's see.

Roger, then Mary, then Carl.

DR. SNOW: I may have missed it, but Carl earlier mentioned something that is important here, particularly if you are going to have an upset threshold for bad quality, that these data are at risk of being contaminated by independent events that send someone back for a mammogram, a second mammogram. I don't know the numbers. I have no idea, but it is not zero.

DR. SMITH-BINDMAN: The recall rate is driven by women who are normal. So of a thousand women, the recall might be 150. Those are normal. The concern that Carl raised is driven by cancers. So that is driven by a recall of one of those five women out of 1000 who have cancer.

So the recall rate of 150 could be contaminated by one of those a thousand with breast cancer. So instead of being 150 out of a thousand, it would be 151.

DR. SNOW: But she doesn't -- I

take the point, but she doesn't have breast cancer. She has a lump.

DR. SMITH-BINDMAN: She has a palpable lump.

DR. SNOW: She has got a lump. She's got a piece of fat there.

DR. SMITH-BINDMAN: But it is not -- it is an order of magnitude for prevalence.

DR. SNOW: So you are saying it --

DR. SMITH-BINDMAN: Not that it is not an issue. Carl's issue is absolutely real. It's just a small bit of noise.

DR. GEMIGNANI: This is Mary Gemignani. I favor this recall type of measure over the one previously, because it has a couple of things that are uniform about it. The population is more uniform. The payer is more uniform, and it is a small metric that we can start with.

The other one is much more broader, and it has so many variables about the institution, the population that you are

looking at. So if I were to pick one of those, I think I would favor this one.

DR. D'ORSI: I am confused, as usual. But let me ask this. What is the basic difference about the discussion we had with the other recall rate versus this as far as equating this to quality? Is there any difference in that discussion that I am missing?

CO-CHAIR GAZELLE: This group is age-stratified.

DR. D'ORSI: It is age-stratified and it is easy to get. But does it still give you a quality measure as a stand-alone?

DR. SMITH-BINDER: I am raising that. I am raising it as an extreme, not as a continuous metric where there is a lot of subtlety, but as a threshold.

DR. D'ORSI: You could do that with the regular recall rate, too, and as a matter of fact, you are stating only one edge of a group where you are saying above is not

good. What about one percent? Is that good?

DR. SMITH-BINDER: That is not good either.

DR. D'ORSI: So then you shouldn't say blank and above. If you are going to do it at all, you need a range.

CO-CHAIR GAZELLE: What are the ranges that is being proposed?

MS. DaVANZO: Ten to 14.

CO-CHAIR GAZELLE: Ten to 14?

DR. D'ORSI: So if you are under ten, you are no good?

MS. DaVANZO: No. If you were two, like you said, you would have to work --

DR. SMITH-BINDER: Ten percent involved half of the facilities not being good, because their recall rates are too low, which is an interesting state of affairs.

MS. DaVANZO: Older people -- I mean, the recount was eight and a half, different studies that we have done over the years.

DR. SMITH-BINDER: So you are saying ten is not -- lower than ten is not good.

DR. DEHN: I think we're in danger of rewriting it. I mean, the fact is that, as Rebecca said, there is a range, and we can identify those ranges, and if support from this group asks us to take a look again at what is too low versus what is too high, we can do that. I mean, it is not real complicated.

DR. D'ORSI: What will you use as -- this is Carl D'Orsi. What will you use as a gold standard to set those ranges besides just a recall rate? What would you say? Where would you pick, two, three, four, nine, ten, 11? Where would you pick it and why would you pick it?

CO-CHAIR GAZELLE: So why don't we finish our comments, and then we will ask for formal comments from the developer, and then we can have a back-and-forth.

DR. CANTRILL: If we are going to be setting a range, where does that data come from and has it been published? I mean, if this is proprietary information --

CO-CHAIR GAZELLE: We will ask them to address that in their comments.

DR. SPENCER: I mean, we have talked about it a lot. So if your recall rate is very low but your cancer detection rate is excellent, not only are you not not bad, you are excellent.

DR. D'ORSI: Supposing you are finding Stage 3. Are you still excellent?

DR. SMITH-BINDMAN: I think those cut-offs -- if the purpose is to identify really low quality, they have to be set at such extremes that that is unlikely to be the case. I would argue they would have to be very wide. The recall rate of two percent -- there are problems with it, but that is how the entire Danish mammography program operates.

DR. D'ORSI: In the UK, I think it is about five percent.

DR. SMITH-BINDMAN: Five percent. The recall rates in the UK are half what they are --

DR. D'ORSI: And they recommend below five. Five is the upper limit. The Dutch are 1.8, but their stages of cancer are much higher.

DR. SPENCER: Are Dutch women dying of breast cancer? Is that what you are saying?

DR. D'ORSI: Yes. That is exactly what we are getting at, that you need to know what you are finding.

CO-CHAIR GAZELLE: I am sensing that this is a good time to ask the measure developer to give their comments, and then we can ask them questions afterward.

DR. DEHN: This is Tom Dehn talking, and this is my second episode with Carl.

I think, for those of you who are not mammographers, I was somewhat, at least as a general radiologist. I have probably heard everything you could ever hear about mammography, and it was really very, very well done, in my estimation.

I want to thank the committee for looking at this, and especially thank Rebecca and Mary for your comments and support of it.

Let me just say that what we are really looking at, I think, as a radiologist -- what we are really looking at is indeterminate rates. That is kind of what you are looking at.

While we call them recall rates, what we are really talking about is, a radiologist has really three options when he or she looks at a study. It is either positive, negative, or I need more information.

There are some radiologists that always need a lot of information, and some

radiologists that don't need information and they are good. It doesn't get a lot more complicated than that, although it isn't anywhere near that simple.

When we look at data, yes, there is age stratification, but kind of the good news for the proposal that we mention is that, in and among the 65 and older age group, the results -- and we can certainly provide those for you -- the differences in those strata are relatively low.

What we do find when we compare it to private data -- and, certainly, Mike has access to that and we have access to that -- is the recall rate is very high in relatively young people for the reasons that you mentioned. Their breasts are denser, and the most important thing we have is the previous study and they aren't around in many cases.

I have the feeling that in transient populations that the same thing happens as young people, that you get more

recalls, because you can't find the previous studies that were done but we haven't really looked at that.

So what we have that is different than the earlier proposal that sounds kind of similar is that we have a fairly homogeneous group, and we are not dependent upon a voluntary BIRADS sort of participation. That is, that when we define an index study that is followed by a given number of studies, we can extrapolate that, that that was an indeterminate study because they asked for some more information or it was a positive, and the positives are pretty well going to be relatively stable.

So what did we find and what do we find? We find huge variations. Rebecca was very kind to our colleagues -- and I have worked with people like this and I think some of you have. They just can't -- they probably should not be reading mammograms, although they probably don't make a lot of mistakes.

It just takes them a long, long time to get there, and we see rates as high as 80 percent in some areas, and within communities that have -- nearly everybody has a nine to ten rate. I mean, I actually know some of your practices around here, and you are all doing just fine.

The thing is that -- but in that community where you are seeing the same kind of people in another radiology group or in another facility, you will have double or more the amount of additional information that is necessary for those radiologists or diagnostic imagers to reach their level of confidence.

So what we are really saying is that there are some radiologists that have a level of confidence that seems to be appropriate for reading and interpreting diagnostic imaging, and there are some that probably shouldn't be.

Now is there an -- and when you look at these high numbers, and we certainly

will look at the low numbers and report those out as well -- when we look at the high numbers, you begin to wonder whether asking to get lower will drive people into a behavior that they don't feel comfortable doing, and that certainly is a concern, or that when you start to see the data folks, you will find that small institutions with relatively low volumes have a very much higher additional imaging rate.

So what would that do to the rural areas that Roger talked about before, and others? I think that, in terms of policy, if we could make policy -- if it were my family, I would probably identify centers of excellence and with the digital imaging, teleradiology, send them in.

Radiologists in the middle of nowhere don't want to read mammograms anyway. So the fear of driving mammography from Chico, California to Sacramento is, at least in my estimation, not a realistic concern. It is a

concern, but not a realistic concern.

What we will give you is insight into the terrific variation between imaging providers. Now you say, well, wait a minute. That is kind of related, isn't it, to the amount of tumor discovery; and the next thing is we have the good Rebecca here who wrote the article, along with others, and they are really quite interesting.

MS. PETERSON: It is on Slide Three.

DR. DEHN: On Slide Three? Well, this is very interesting, because there is a point at which you can continue to add additional studies for call-backs or follow-ups, however you want to describe it but you really don't get anywhere, and this is somewhere around 14 percent.

So if, in fact, this committee or anyone on this committee would like to contribute a suggestion to us on what level we would like to set those thresholds, we can

certainly -- we can certainly do that. I think, if I were to do that, it would probably be back of the envelope. But when we know now that, after a given rate, you don't find any more cancers, they are in pretty good shape.

CO-CHAIR GAZELLE: Is this for the CMS population or is this all?

DR. DEHN: This is all.

CO-CHAIR GAZELLE: I thought we heard earlier that the numbers would be different in the CMS population.

DR. DEHN: The call-back numbers will be lower and, in fact, they are. The call-back numbers we looked at are somewhere in the seven to eight percent range. So we are operating down here.

So if you set -- if we are discussing where to set the threshold, I think that might be a discussion for another time. Should we set a threshold that experts suggest is realistic? Yes, of course, we should.

DR. SPENCER: I misunderstood. I

thought you said you had data that you could present from the Medicare population.

DR. DEHN: Now what we have here is kind of a peculiar -- I didn't do this. Radiologists don't do slides like this. But what you see here is, of the 2,800-some hospitals, there are some here, about half, that are below 8 1/2 percent national average, national average for Medicare, and there is about 50 percent that are over, and there are some that are really over -- really over.

DR. BASSETT: Please don't use that word, follow-up, because that refers to patients who are in a short term follow-up. As we go into this era of IT and all the electronic records, we don't want that overlay. So I just wanted to --

DR. DEHN: I agree, and we have all grappled with it. I noticed in yours it is called recall rate, and essentially, if you really looked at recall rate, that measures a whole different thing. I mean, you are

compliant and your enrollees or your patients that you take care of are relatively well educated and are compliant and you have a program.

That is a whole other issue is that, when you find something abnormal, are you able to get them back, and that is not what we are looking at. We are really looking at indeterminate rates. So when you look at a case, you need more information, some need a lot more than others, and that is what we are looking at.

So we think it is clean. We would like to take a look at it, get started on it, report it back to you, and let it change as time goes on. Please?

DR. SPENCER: This is Kirk Spencer. Two quick questions. So how does a Medicare database tell recalls from short term follow-ups?

DR. SMITH-BINDMAN: Short term follow-ups -- what exactly is that?

DR. SPENCER: Well, you are just going to find something done in less than six months.

DR. DEHN: That is correct.

MS. DaVANZO: The metric is 45 days.

DR. DEHN: And again, we are seeing less and less short term follow-up, and we are seeing more and more definitive imaging studies. It is either MR or it is biopsy or--

DR. SPENCER: I know it says -- from someone who remotely reads echoes, having anybody do the echo at the other end, and then they will send it to me to read, and the echo clearly doesn't work.

In mammography, is the technical aspects of it substantially less than the radiologist? I don't have a good sense for that?

DR. DEHN: Yes, and the good news is, as Dr. Forman indicated --

DR. SPENCER: And the reader is

the dominant variable?

CO-CHAIR GAZELLE: One other point is that we are including follow-up for mammo, diagnostic mammo or ultrasound, but not MRI in this measure.

DR. DEHN: And we intentionally left that out, because during the time of the study that we collected and intend to collect the data from, MRI is not real well defined, and I am not so sure it is yet well defined on one-use MRI in conjunction with an abnormal mammogram.

DR. FIESINGER: You know, on this curve -- it is a funny kind of curve, because it sort of suggests -- and maybe this is the fact in the real world, but it suggests there is a group of people who are just utterly lost in space.

I mean, you see a lot of variability in the clinical world. We can take their choice of that, but are there really a group of people who are lost in space

who are just doing all the MQSA stuff and figured out all of that, are getting paid, have all these other things, but somehow are just, as centers, congenitally unable to read mammograms?

DR. DEHN: Yes.

DR. FIESINGER: Because that is what is describing to me. That is what this curve is describing to me, and it seems like -- it just seems like who are these people?

DR. DEHN: What we will see, and when we did this in the private sector in that whole population that we were talking about -- that means the non-Medicare population -- I was totally surprised.

To answer your question, I thought you would have some variation, like you do. But I can only conjecture that there are folks out there that are either motivated by certain things, and then there are others that are so insecure that they always get additional films.

Now many of you have worked in radiology groups. I have. A couple of my partners had double what my call-back rate was.

DR. ZERZAN: Do you have any base for, like, numbers, because I could imagine the 100 percent one is somebody that reads three a year, and they are going to call back all three, because they don't know.

MS. STEPHENS: No. We have got minimum case counts on it.

CO-CHAIR GAZELLE: What do you mean by that?

MS. STEPHENS: It varies by the ratio level. We asked for the -- the case count asks them to count -- we had a lot of people at the low end and a lot of people at the high end. So the minimum case count actually varied by -- in this data, varied by ratio level, and we are working at a 90-percent confidence level.

DR. RUCKER: It doesn't look like

normal distribution to me. I understand you are doing some other funny graphing here, but it doesn't look like a normal distribution.

DR. DEHN: But what we do see, though, is you do see some outliers that are way out there. Unexpected to me as it might be to you -- I mean, how can you be that far off?

DR. RUCKER: Is that just fraud?

DR. FORMAN: Why aren't these -- I am still not clear.

CO-CHAIR GAZELLE: Hold on. Please give your name.

DR. FORMAN: Why aren't these tiny, tiny practices that are seeing three cases a week -- I mean, I am not trying to defend them, but --

MS. STEPHENS: I want to clarify. These do not include facilities who have a small case count. They have to have had a -- at the tail there, they have to have done at least 45.

DR. FORMAN: Forty-five what?

MS. STEPHENS: Screening mammograms.

DR. FORMAN: During what period?

MS. STEPHENS: During a year.

DR. FORMAN: That is nothing. You know, I once watched a resident in a practice that had -- you know, they did screening mammograms out of convenience, and you would see a patient once a week. So you are basically dealing with 140 practices at the tail, all of whom may account for less than .1 percent of the population.

So they are out there, but I wouldn't necessarily imply fraud. It is probably more likely that they -- people are saying that even pecuniary instincts are causing this. I have a feeling that most of the tail are probably radiologists who don't want to be doing mammography, and are just doing it because the set-up is in the office.

MS. STEPHENS: No, these are not

offices. These are hospital outpatient departments.

CO-CHAIR GAZELLE: Can I ask for a clarification. Is this a computer-generated curve or is this an actual curve? And the specific question I have is: are there really sites that are recalling 100 percent and zero percent or is this just --

MS. ARDAY: This is the real data. The maximum is 100 percent. The maximum between data where you know they started is 45 screening mammographs.

DR. DEHN: In the private sector, high-volume facilities have 80 percent. I have not seen any at 100 percent. There are some that have 80 percent with high-volume providers, and high-volume providers that have close to zero percent, in fact, I would worry about.

The deal is let's look. Now what I have produced that graph for is a different way, sure. But I think -- the radiologist put

a lot of work into this thing, but I am passionate about this. There are radiologists that have to have a lot more information than other radiologists, and they are out there in significant numbers, and we got to identify them.

DR. D'ORSI: I agree with that.

DR. DEHN: Okay. Carl.

DR. D'ORSI: I agree with you, John, but I am again confused. If the MQSA says an individual has to read 500, this would imply that somebody who is reading for 100 facilities to get that, do you know that or not? Where does 45 reconcile with the FDA minimum of 500?

DR. DEHN: You know, Carl, I had the same question, and I suspect that there are a fair number of radiologists that are not -- they are not qualified.

DR. SMITH-BINDMAN: As part of the Breast Cancer Surveillance Consortium, it seemed that there were a lot of low-volume

doctors, and there are a lot of low-volume doctors. But in fact, doctors read at many facilities, and so on a practical level you are only assessing the mammograms they are reading in the elderly, and you have no idea if they are making up their volume in other places. So I kind of agree with --

CO-CHAIR GAZELLE: Or with non-Medicare patients.

DR. SMITH-BINDMAN: Right. Those are the elderly.

CO-CHAIR GAZELLE: But it could have been in their same facility, just a lot of non-Medicare patients.

DR. SMITH-BINDMAN: Exactly. So it is very difficult to get.

DR. DEHN: From the back of the envelope you feel that Medicare is probably 30 to 40 percent or 30 percent of your mammography volume. That would be 90, you know, and there isn't a radiologist that I know that isn't terrified of someone coming

after you if you are reading two a week. I mean, basically, that is two a week. So they must be working at other facilities.

DR. SMITH-BINDMAN: But just looking at the distribution of your data, the 99th percentile distribution and the recall rate is 24.9 percent and I can give you six separate references that have gotten exactly that number: 25 percent.

So I think the one percent outlier which we are looking at is either a data issue or it is a -- I don't believe -- or represents a couple of doctors that are doing something odd. I think that is unlikely, and if your quality metric is only measuring that one doctor, it is not doing anything. It is doing nothing.

DR. DEHN: I understand that, and I would just say -- Offline I will share some of the blinded private information that we have and it really does happen.

DR. SMITH-BINDMAN: But that is

not the benefit of this measure. I mean, it might be a benefit to you to identify those few really, really extreme cases.

DR. DEHN: The thrust of this measure is to find --

DR. SMITH-BINDMAN: You don't need this measure to identify them. You can identify them in a lot of other ways without having an NQF measure.

CO-CHAIR GAZELLE: Do we have a proposed range, though, for this measure? Are we supposed to sign off on the measure or sign off on the measure with a range? Is there a range that is being proposed?

MS. DaVANZO: Yes. The literature supports ten. That is one of the benchmarks that you see a lot in the articles, and then 14 or 15.

DR. SMITH-BINDMAN: But you are saying you are applying a standard that half of your facilities would fail.

MS. ARDAY: No.

DR. SMITH-BINDMAN: Am I confused about that?

MS. ARDAY: We are not marking any of these as pass or fail. What we are really looking for is a more extreme rate of distribution --

DR. SMITH-BINDMAN: What was the ten to 14 percent you just cited?

MS. ARDAY: -- hospital outpatient departments establish a dialogue of what is going on with our patients here? What is going on with our clinicians?

CO-CHAIR GAZELLE: So that I understand, but the question is what are those numbers? Below what is not acceptable? Above what is not -- or is not good?

MS. ARDAY: We haven't done that piece. This is pay for reporting, not pay for performance.

CO-CHAIR GAZELLE: For CMS, but for NQF the question is whether or not we are going to approve a measure that doesn't have

a threshold. Right?

DR. BURSTIN: Just to be clear, not all measures need that threshold.

CO-CHAIR GAZELLE: No, I understand.

DR. BURSTIN: We would endorse, for example, an episiotomy rate. No one knows what the exact rate perhaps is, but the question is, is it useful for a bench purpose in reporting to begin to see where the --

CO-CHAIR GAZELLE: I completely understand that, but we have had a discussion back and forth about a lot of different ranges. What I am trying to get clarity on is, is there a range being proposed with this measure and, if so, what is it, or is there not a range.

I understand that there could be. That is not the question. The question is, is there one being proposed with this measure?

DR. DEHN: Let me speak, please. There is one that is proposed, and the

discussion today has prompted us to take another look at it.

CO-CHAIR GAZELLE: What is the one that has been proposed.

DR. DEHN: Ten percent and 14.

CO-CHAIR GAZELLE: Within 10 to 14 is the range that is being proposed?

CO-CHAIR GAZELLE: So eight and a half and nine would be outside of that range?

MS. ARDAY: No. No, because there is no cancer found. The 10 to 14 percent is on the general population. This is predominantly --

DR. SMITH-BINDMAN: So it has no relevance for our discussion. Is that correct? It has no relevance to the discussion.

CO-CHAIR GAZELLE: So there is no range.

DR. BRUETMAN: This is one of the issues that was brought up, which is we are talking about the stratification of data. I

mean, CMS stratifies their data into age: 65 and over and all that, but we have done that, and it is not significantly changed.

What it does indicate is that at a certain age, this sub-segment has at least a lower reach than average, a little bit lower, because of many clinical issues. So that is why you see it comes a little bit lower than expected, which the literature says ten to 14 percent is the expected recall rate that we see here. But CMS has stratified data, so a little bit lower level. Now we haven't defined do we think it should look at the low end and somewhere at the high end.

CO-CHAIR GAZELLE: So I think we need to be clear on this. First of all, I don't know what literature you are speaking of that says ten to 14 percent. So that would be the large study. The BCSC study was an average of 9.8 percent. Well, that is not ten to 14 percent. The European data is all single digits, and I have seen one study that

has a median of about ten, and I think at 75th percentile -- what was it, 16 percent?

So I don't know of a study, and it isn't cited here. So I don't think we should say the literature says ten to 14 percent unless the literature does say ten to 14 percent, which would mean that someone can cite that.

DR. SMITH-BINDMAN: AHRQ has old numbers, and I don't know if that is the number that you are citing.

MS. DaVANZO: I think the range five to 15, and it is an average of 12.3.

CO-CHAIR GAZELLE: The interquartile range was, I think, 6.4 to -- no, 4.6 to --

DR. D'ORSI: I have it right here. It is 6.4 to 13.3 is the 50 percent. Fifty percent of radiologists fall into that. If you use your numbers, 25 percent would fall into that.

DR. SMITH-BINDMAN: I think these

are for age-adjusted. Are these Ralph's data?

DR. D'ORSI: Yes.

DR. SMITH-BINDMAN: I think they are age-adjusted.

CO-CHAIR GAZELLE: So is it fair to say that the ten to 14 percent is not relevant to this measure and not relevant to this discussion? Right? So we can leave that behind? All right.

Okay. Other questions of the committee to the measure developer? Ray?

DR. GIBBONS: Ray Gibbons. Just two broad comments. One point has already been made, but in terms of the potential impact of this, you would have to know the volumes of these studies being performed and the extremes to know how useful this measure would be to CMS for overall quality.

The second observation I would make is that using different kinds of datasets in far larger populations -- if I look at published data for cardiac procedures, these

extremes don't look bad at all.

DR. DEHN: Believe me, I know.

DR. SMITH-BINDMAN: From zero to 100?

(Laughter.)

DR. GIBBONS: For example, the published data on cardiac procedures based on Medicare markets -- so these are hundreds of thousands of patients -- show customarily five- to ten-fold differences in non-zero rates, and a well known, published example of one referral region that is three times higher than a referral region 60 miles away.

So I am surprised. These look pretty good.

DR. D'ORSI: So, John, these are facility numbers; right?

DR. DEHN: Yes, and they can be broken down into individuals, but were are instructed not to do that. When you look at it, however -- when we look at it in the private sector --

DR. D'ORSI: But this metric you are presenting is relatively unfair, because there is no facility standard with a recall. It is an individual metric. So it is a little bit unfair to say those people are really -- those facilities are stupid, because they may be going to somebody who is very good at reading, but only doing 45 a year.

DR. DEHN: By extrapolation, we simply say that there is a quality issue if you know that your partners are not reading or, in the aggregate, you are doing well. So you can blame it on something systemic within the facility.

DR. D'ORSI: But it is a little misleading.

CO-CHAIR GAZELLE: All right.

DR. GIBBONS: The thing I take from these, and I saw another similar curve, is that you can't draw final conclusions from these data, but you can say, well, there is a sector of interest out there at that far end,

whether it is because they are very conscientious, whether it is because they are this or that or the other.

So you know, the ones in the middle maybe you don't have to worry so much about, and use your resources the same way you have used resources on the people at the end. That is as far as you can go with those data, I think.

CO-CHAIR GAZELLE: Okay. Other comments on this specific measure?

DR. SMITH-BINDMAN: I raised something when I reviewed it. Rebecca -- sorry. Are you guys -- do you have some measure of the ability of these new CPT codes to differentiate screening from diagnostic exams?

MS. DaVANZO: They are separate codes. They are separate CPT codes.

DR. SMITH-BINDMAN: Right. Do you know the ability to differentiate screening from diagnostic using those codes to get some

reference standard like data in the Breast Cancer Surveillance Consortium data, in self-reported mammography, that sort of thing?

We did a paper that was published in Medicare a couple of years ago that looked at the classifications of mammograms using CPT codes, using CMS data compared to Breast Cancer Surveillance Consortium. So, certainly, it would be something that you guys could repeat using your new codes.

MS. DaVANZO: Right. We have it from the old codes.

DR. SMITH-BINDMAN: If we were using the old codes.

MS. DaVANZO: If we used the old codes.

DR. SMITH-BINDMAN: Then I would say you don't have a measure here.

CO-CHAIR GAZELLE: Because you don't believe in the validity of the reporting of the total.

DR. SMITH-BINDMAN: The ability to

frame it from screening to diagnostic. So I am assuming your new codes are going to be better. I am asking you if you have looked at that. I am suggesting it might be useful. It requires some chart abstraction, or the simplest thing to do -- the simple thing that you could do is in states that have a SEER tumor registry or Breast Cancer Surveillance Consortium registries -- so you can do it in New Mexico; you can do it San Francisco; you can do it in Washington -- they currently have done the linkage for you.

So the linkage is done between the Breast Cancer Surveillance Consortium and the Medicare data. So you just have to put in this request, and if you speak to me after, I will tell you how to do it, and then you can find out the rest.

MS. DaVANZO: And it is very possible that CMS is also researching demonstrations for these, probably after looking back at the SEER registry. So it

might be as simple as having a talk with Jerry Riley or somebody and say, hey Jerry, have you looked at this number lately.

DR. SMITH-BINDMAN: It's the Breast Cancer Surveillance Consortium. So it is Rachel Ballard Barbash. It is under her. Diana is the Coordinating Center person in Seattle.

CO-CHAIR GAZELLE: So that would be an important point of clarification, if we would decide to --

DR. SMITH-BINDMAN: If we would decide they haven't shown us which measures can be used.

CO-CHAIR GAZELLE: Okay. Carl?

DR. D'ORSI: Just one quick one, John, and you can answer this yes, no, I don't know. So you have information at that end of readers who are MQSA-certified with these recalls and if somebody is reading who is not MQSA-certified with these readings.

DR. DEHN: That is correct.

DR. D'ORSI: Okay, thank you.

DR. SMITH-BINDMAN: Do you know their personal individual MQSA?

DR. DEHN: That is correct.

CO-CHAIR GAZELLE: Mike?

MR. BACKUS: I just look at the curve, and 2801 is there, and the total sample size is 2957. So this tail that we are spending all this time talking about -- this is like 30 guys.

MS. DaVANZO: That curve there, Mike, represents 2.7 million mammograms --

MR. BACKUS: Well, no. I am talking about number of facilities. So you look at that list of facilities -- I mean this is what we do from the plan perspective all the time.

You know, I have said just back of the envelope -- you set that line at a standard deviation or a standard deviation and a half off, and you go, okay, I want to look at the guys that are sub-three, and I want to

look at the guys that are over 20, and I am going to end up with 200 facilities to look at. That is what is going to tell you, because CMS or any organization -- we can't be in the position Rebecca has talked about where half of the facilities in America don't meet the measure. That doesn't serve anybody any good. Just look at the tails and -- you know.

CO-CHAIR GAZELLE: I would like to raise one issue for discussion that we talked about this morning with the recall rate measure. The question is, is there value, if we were to approve this, in having essentially a recall rate measure that doesn't include a cancer detection rate or possible prediction values in the measure?

Should we go back and ask CMS if they wanted a Medicare population measure for recall rate to also have a cancer detection rate?

DR. CANTRILL: Steve Cantrill. I

was impressed this morning in the discussion for each of the first four measures where we were saying this alone is not good; you've got to take it in conjunction with other measures, and this alone is what we are talking about.

CO-CHAIR GAZELLE: Yes.

DR. CANTRILL: So I don't understand how we can strive to have a, quote combined measure or call it what you will, firstly, and then say, oh, but in this case, because the data is easy to get, we are just going to do this alone.

So I would say we are obligated to go back to the makers of this measure and say, do you have the data. Can you do what we were talking about in that set of first four measures as well as this single measure?

CO-CHAIR GAZELLE: Thank you. Other comments on that topic or other topics?

MR. BACKUS: Look, CMS doesn't hold the BIRADS information, though. Right?

CO-CHAIR GAZELLE: No, it doesn't.

MR. BACKUS: So that becomes a drop-off question. If you assume everybody with Medicare that comes zero, four, five has an insurance coverage -- or does not have an insurance coverage issue, then you would expect a dropoff of zero, four, fives that don't get follow-on care, assuming they are continuously enrolled or whatever, you know, the drop-off should be trivial, you would hope. So you would end up with cancer detection down the stream, because you have the path data.

CO-CHAIR GAZELLE: Don?

DR. RUCKER: Maybe for the measure developers -- Don Rucker. I think some of our requirements for the NQF process -- I think the first one on importance -- did we have a sense of the area under the tail here in terms of the requirement for the importance?

We are asking a lot of people to do a lot of reporting, as far as I can understand here, that has a cost to it.

DR. SMITH-BINDMAN: No. It is all paid for.

DR. RUCKER: So it is sort of, quote/unquote, "free"?

DR. BURSTIN: Right.

DR. RUCKER: Then maybe just on the importance question -- I don't have it; is that 1(a)? It is pretty high here. -- just a raw importance metric, if we could understand that, because --

DR. BURSTIN: I think that was referring to 1(b), which is the demonstration of quality and opportunity for improvements. If you are making the argument, the tail is fairly small here. It is a facility level measure. So the question is how many facilities does that 1000 cases represent.

MR. BACKUS: Well, the 95th percentile is 17-1. So you would have 200, right? You would have five percent on top out of 2000. It is 150 on the top, 150 on the bottom. Right? If you went fifth percentile

and 95th percentile? So 300 facilities -- that is five percent of the hospitals in America. That is pretty substantial.

CO-CHAIR GAZELLE: Correct.

DR. SMITH-BINDMAN: It is the 99thh percent.

MR. BACKUS: I'm sorry. Right. That was at 25, right. At the 95th percentile, if you cut it at 17. If you cut it at 17 and 5, you are up to 300 hospitals, fifth percentile and 95th percentile.

DR. SMITH-BINDMAN: I like that.

MR. BACKUS: Three hundred hospitals to go look at.

CO-CHAIR GAZELLE: Troy, you look like you are about to raise your hand. No?

DR. FIESINGER: No. I was just pointing to the data. I am fine.

CO-CHAIR GAZELLE: All right. Now we need to move toward decisions, voting, and it is complicated. I am not sure how best to approach it, because seems like we have the

four measures this morning that we want to consider as a group, and then our decision on that might affect our decision on this afternoon's measure.

So what I propose is we have a brief discussion, try and limit it to about 10 minutes or so, on which of the four we would -- on the merits of approving them individually this morning or of grouping them.

I will throw out a straw man proposal based on what I thought we heard this morning, is that the measure developer wants to see them approved or presumably not approved, but approved as a group, and I think from our discussion, the consensus was from the four this morning, the three that would make sense to bring together or consider together would be the recall rate, the cancer detection rate, and PPV2.

DR. SNOW: The PPV2 on the diagnostic?

CO-CHAIR GAZELLE: Yes. So I

think that was -- If I am off, speak up, please, but I think that was kind of where we were thinking based on the morning's discussion.

So I don't know then how we go about voting for that without voting for the individual measures.

DR. BURSTIN: You still need to look at each of the individual measures, make recommendations, recommendations for conditions, whatever the case may be.

CO-CHAIR GAZELLE: And the condition could be only with the other two?

DR. BURSTIN: Yes, although I think -- Again, it is really the question of how the three at the end of the day get presented together, but I still don't think we have clarity since they are not a composite.

CO-CHAIR GAZELLE: Right.

DR. ZERZAN: This is Judy, I have a quick question. The one thing that I do like about the first PPV2 or 1 is that it is

based on tissue diagnosis. So it is a real outcome rather than asking for follow-up. So I don't know if there is a way to change or recommend that the second one move to tissue diagnosis or -- I guess I still don't know.

CO-CHAIR GAZELLE: Is diagnosis recommended?

DR. ZERZAN: It says it is recommended to get a tissue diagnosis rather than the actual tissue itself, which to me is a difference in terms of, I think, my philosophy of quality measures in general is that we should be pushing toward more outcome based things and measuring more things that really change health rather than the indeterminate process-ey things that we sometimes focus on.

So, to me, tissue sounds more definitive than, oh, I recommend that you go there by --

CO-CHAIR GAZELLE: You speaking for what? You are speaking for the

denominator?

DR. ZERZAN: I like the second one, but the part I don't like about it is that it just recommends. It doesn't say get the tissue.

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. Your point has been raised by others, and the argument -- not that I endorse it or not -- is that from the quality point of view, all the radiologist can do is recommend that something else happen, and there are a lot of factors outside that doctor's control in terms of whether the person chooses to follow up at that facility or any facility, and that it would be better to separate -- Your point is that the doctor can take responsibility. The doctor can't, and that is why it is adopted as a recommendation rather than what actually happens.

CO-CHAIR GAZELLE: It is also the issue of --

DR. SMITH-BINDMAN: Impractical.

CO-CHAIR GAZELLE: -- you know, in terms of positive predictive value, it is the positive predictive value of a positive mammogram. Right? So that is why a positive mammogram is a 4 or 5, which is the recommendation for biopsy, and what percentage of those positive mammograms are actually positive.

I think we can't redefine a commonly used measure.

DR. ZERZAN: But why not push for -- I mean, I understand that the doctor doesn't necessarily have control over that, but that is also a reason why doctors say they can't address obesity, you know. They are still -- Did it help push the system, health system, the payers, as well as the providers to a higher standard than what is already there? Maybe we are not there yet in terms of data, but if we are close, I guess I would argue for getting the tissue rather than just

the recommendation, to push that a little further.

DR. SNOW: Roger Snow. I am very sympathetic with what you say, but I think that that is an argument for another table, because what is being done here is a measure that works on what radiologists do and can do. The point has been made that they can't get the biopsy. The interventional guys may, but that aside, the actual thing, the step of getting the outcome, would be a separate measure. That would use PPV3, I think, and maybe we all come back in a year and go after the primary care guys.

I think it really is a measure of quality at the care delivery level rather than at the diagnostic level. It is a different measure.

CO-CHAIR GAZELLE: So what I am looking at --

CO-CHAIR PETERSON: Can I just ask for a clarification? So let's take one

assumption. When could this come back to this, if it were not passed today? When would it be potentially re-eligible to come up again?

DR. BURSTIN: It is not clear. When we have another project with the right expertise, we could review it. So I don't see any --

CO-CHAIR PETERSON: But we don't know when the next imaging efficiency group will --

DR. BURSTIN: I suspect, given how important this area is, it is probably within the next two years, but I wouldn't say it is less than that. Since this is a starting point --

CO-CHAIR GAZELLE: Maybe what we should do it vote on this measure in isolation first, because if it passes in isolation, we are done -- each of them.

DR. BURSTIN: Each of them.

CO-CHAIR GAZELLE: The first four.

Then if they don't pass in isolation, come back and vote again with the grouping; and if they don't pass there, then they haven't passed. I don't think I can think of another way to do it. Voting is endorse/not endorse.

CO-CHAIR PETERSON: We are looking for simple majority here?

CO-CHAIR GAZELLE: Yes.

DR. BURSTIN: Although, again, if it is a split vote, we will just present it to the public as such.

CO-CHAIR GAZELLE: So let's -- Carl?

DR. D'ORSI: I just want to make one quick statement. In this country, 2 and 3 are almost the same. So the vast majority of PPV2s will have tissue, the vast majority. So it is not like --

DR. ZERZAN: Well, then why not go for tissue?

CO-CHAIR GAZELLE: Well, because tissue hasn't been proposed. So we can't vote

on it.

DR. D'ORSI: What Rebecca said is correct. The 2 is the cognitive part of the radiologist and the surgeon to say, out of here. So nobody is talking about it. Go away from me. So she doesn't get it. No, but this is -- I am hyperbolic, but this is a scenario. So you are really judging the cognitive thinker on doing the 4 or 5. After that, they can't really control what happens, but it is very close.

CO-CHAIR GAZELLE: But also we don't have a PPV3 measure to discuss or vote on.

DR. BURSTIN: And it may wind up being that is a research recommendation. Just to follow up on Judy's point, there is a strong interest in measures that get at shared accountability. It doesn't need to just reflect the facility, if the end game really is to zoom in with positive mammograms, get the outcome we expected, and that is, I think,

a very reasonable expectation. I just don't know that the measures in front of us today offer us that option.

CO-CHAIR PETERSON: So to clarify one more time, we are going to go and vote on these individually. If they are voted up, then they are in. If they are voted down, then we will take them as a group.

CO-CHAIR GAZELLE: As a group, with the condition that we would approve them if they were a group. Then they may or may not pass.

Okay. So do you want to call for the voting or should I call for a vote?

MR. CORBRIDGE: I just want to bring something to the screen. We do have an NQF just kind of form to capture the process that you are going through. Sarah has been working on getting the Steering Committee comments and recommendations, covering the black discussion points, response of sponsor measure developers or response from the

public, which at the end of discussing mammography measures we will open it up to the public to see if there is any responses.

On the lefthand side, we have NQF's criteria for looking at measures. So you have importance, scientific acceptability, usability and feasibility. Our plan is, as we are going through, I will collect the Steering Committee's votes on that.

So we are looking at how many people are voting on each.

CO-CHAIR GAZELLE: And an overall?

MR. CORBRIDGE: Well, taking -- I guess taking -- For the four main criteria.

CO-CHAIR GAZELLE: Right. So there's five votes on each one. Okay.

MR. CORBRIDGE: Then, I guess, depending on how things lay out, if there are comments that are needed to justify some of the recommendations that the Steering Committee puts forward, we will put those comments in.

CO-CHAIR GAZELLE: And these are binary votes on each of these five measures?

DR. BURSTIN: You mean yes/no?

CO-CHAIR GAZELLE: Yes/no.

DR. BURSTIN: I'm sorry. It is recommendations specifically on a criteria are high, medium, low.

CO-CHAIR GAZELLE: Okay. So do you have a matrix to capture these four by four, and then the one by two?

MR. CORBRIDGE: Yes. We are just going to take this down.

CO-CHAIR GAZELLE: All right. So now here we are. We are voting on measure number 1, cancer detection rate. We have discussed it this morning. We are voting on it in isolation, and we need people to raise their hands. This is Steering Committee only members. We need you to raise your hands under the importance.

So how many people want to rate the importance as high? C? High up here? So

this is all of the different subparts of High together.

DR. BURSTIN: Yes.

CO-CHAIR GAZELLE: The options are High, Middle or Low?

DR. D'ORSI: Can you read the evaluation criteria, the main ones, before you ask for a vote?

CO-CHAIR GAZELLE: I will, once we count. The importance, everybody knows. I will read it again while we are counting.

"Importance: Extent to which the specific measure" -- Hands down. It is, "extent to which the specific measure focus is important for making significant gains in health care quality, defined by the six dimensions of the IOM, and improving health outcomes for a specific high impact aspect of health care where there is variation in or overall poor performance."

So that is the importance. Now we've got -- How many people would like to

rate that M for Middle rating? Four?

How many people would like to rate that Low for low? I figure we need to say it.

Okay. So next we are going on to criterion number 2, scientific acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

Remember, we are voting on this measure now in isolation. How many people want to give it a High rating? None.

How many people want to give it a Middle rating? All right. And how many people would like to give it a Low rating? We should have an easy way to calculate that.

Now the next is -- I am not going to read these definitions with every measure, but there was a request to read them.

Next is usability, which is the extent to which intended audiences can understand the results of the measure and are

likely to find them useful for decision making.

Again, we are voting on measure number 1 in isolation at this point. High? It looks like three. Middle? Looks like six. And Low?

DR. SMITH-BINDMAN: Can I just clarify. When you read the second one, you said as written.

CO-CHAIR GAZELLE: As written.

DR. SMITH-BINDMAN: But you didn't say for this usability as written.

CO-CHAIR GAZELLE: Oh, I thought I did, but we are voting on this thing as written.

DR. SMITH-BINDMAN: Only as written?

CO-CHAIR GAZELLE: Only as written now, because we agreed we would just vote on them as written first, and then talk about the modifications.

DR. SMITH-BINDMAN: I want to

change my vote.

CO-CHAIR GAZELLE: We can do that just by counting. What do you want to shift from what to what?

DR. SMITH-BINDMAN: High and Middle.

CO-CHAIR GAZELLE: Okay. So that would be two High and Seven middle then.

Okay, the last category is for feasibility, extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

Again, this is measure number 1 in isolation. How many votes for High? Five.

How many votes for Middle or moderate?

MR. CORBRIDGE: Is it 15? Yes.

CO-CHAIR GAZELLE: Okay. And now we have an overall -- Oh, Low, sorry. How many Low? Who wants to vote Low? Should be a couple. You could abstain. Okay.

The important thing is the NQF will report the numbers of the votes. They are not going to come to a binary decision.

So now we want to have an overall recommendation, and that is either Yes or No. So you vote either to approve to recommend this for endorsement or not.

So who would like to recommend this for endorsement as is, as written, in isolation? Okay, who would vote not to recommend this? Okay. So that is this measure.

So we will go through. We are going to do the same process now for measures 2, 3 and 4, and then we can come back and talk about a proposed either conditional approval and what the condition might be as a group.

Let's go to measure 2, which is screening mammography, positive predictive values, PPV2, which as a footnote should really be PPV1, but as long as we are voting on it as it is written and defined in the

measure. Okay.

We are on the first category, which again is the importance. Who wants to give it a High? Is it eight?

Who would like to give it a Middle or Moderate? Eleven. And who would like to give it a Low? None? Okay.

So now we are going to move on to the second category, which is scientific acceptability of the measure property. Who would like to give it a High? Zero. Who would like to give it a Middle? Seventeen.

Who would like to give it a Low?

MR. CORBRIDGE: Is it Four? Five, sorry.

CO-CHAIR GAZELLE: We keep getting different totals. Are there 22 people? How many people are there?

MR. CORBRIDGE: Are individuals abstaining?

CO-CHAIR GAZELLE: There are 22 people.

DR. SNOW: Vote early, vote often.

MR. CORBRIDGE: The problem with the 17, I can't see -- I don't know if you would like to be in the middle?

CO-CHAIR GAZELLE: So raise your hand if you want to give this a Middle.

DR. D'ORSI: This is a lesson in statistics.

CO-CHAIR GAZELLE: I got 14. Who would give it a Low?

MR. CORBRIDGE: I saw 14, yes.

CO-CHAIR GAZELLE: Who would give it a Low?

MR. CORBRIDGE: One, two, three, four, five. So that gives the right number.

CO-CHAIR GAZELLE: Thank you. Okay. So the next category is category 3, which is usability. Who would like to give it a High? High for usability? No? One high.

Who would like to give it a Middle?

MR. CORBRIDGE: I count 15.

CO-CHAIR GAZELLE: I got 14. Who would like to give it a Low? Three?

MR. CORBRIDGE: Three, yes.

CO-CHAIR GAZELLE: I think we need to ask everybody to vote. You have to make a decision. You can't really abstain.

DR. BURSTIN: You can abstain. You just have to let us know you are abstaining.

CO-CHAIR GAZELLE: I can understand how you could abstain on the for or against it, but how can you abstain on the high, medium or low?

The next one -- The last one is feasibility. How many people would like to give this a High on feasibility. Raise your hands high.

MR. CORBRIDGE: Looks like we have three.

CO-CHAIR GAZELLE: How many people would like to give it a Middle for feasibility?

MR. CORBRIDGE: Seventeen.

CO-CHAIR GAZELLE: So that should be zero Lows. Okay, good.

So now you have the option of either voting to recommend for endorsement or not to recommend for endorsement. How many people would like to vote to recommend for endorsement, again single measure in isolation? All right.

How many people would not recommend for endorsement? Looks like everybody. All right, we are making -- Oh, that is not progress.

Okay. So now we have measure number 3, which is diagnostic mammography PPV2, which is the percentage of positive mammograms that lead to a diagnosis of cancer.

Again, we are voting for importance. How many people would like to give it a High?

MR. CORBRIDGE: Eighteen.

CO-CHAIR GAZELLE: How many people

2 is compleatly out of the running

would give it a Middle? Two?

How many a Low? Zero.

Next is for scientific acceptability. How many people would give it a High?

MR. CORBRIDGE: Seven.

CO-CHAIR GAZELLE: Middle?

MR. CORBRIDGE: Thirteen.

CO-CHAIR GAZELLE: And a Low?

MR. CORBRIDGE: It would be zero.

CO-CHAIR GAZELLE: And next is for usability. How many people would like to give it a High?

MR. CORBRIDGE: Four.

CO-CHAIR GAZELLE: Middle?

MR. CORBRIDGE: Sixteen.

CO-CHAIR GAZELLE: And Low? It should be zero. Okay. I am not trying to influence your vote.

And for feasibility, how many people would like to give it a High?

MR. CORBRIDGE: Six.

CO-CHAIR GAZELLE: Middle?

MR. CORBRIDGE: Thirteen.

CO-CHAIR GAZELLE: And Low? It would be one -- No? One abstention. So should we ask for abstentions, just to check our math, Helen?

DR. BURSTIN: Did somebody abstain?

CO-CHAIR GAZELLE: Did somebody abstain on that one? It was six, 13 and zero, but no one is claiming an abstention. So we must have counted wrong. Could we count again, please? Highs? How many Highs?

MR. CORBRIDGE: It looks like there is six. Should be 14 middle.

CO-CHAIR GAZELLE: All right. Who would like to vote to recommend endorsement of this measure? One. One for.

Final voting outcome is only one for recommended

Who would vote against endorsement? That looks like 19 to me. Any abstentions? That is 19.

Okay. Now let's go on to measure

4, which is recall rate, and we are back to importance. How many people will give this a High importance?

MR. CORBRIDGE: Thirteen.

CO-CHAIR GAZELLE: Okay. How many people will give it a Middle?

MR. CORBRIDGE: Seven.

CO-CHAIR GAZELLE: Should be no Lows. Any Lows? All right.

Now we are on to the next measure, which is scientific acceptability. How many people will give it a High?

MR. CORBRIDGE: Five.

CO-CHAIR GAZELLE: How many people would like to give it a Middle?

MR. CORBRIDGE: Fifteen.

CO-CHAIR GAZELLE: How many Lows? We must have counted wrong.

MR. CORBRIDGE: Fourteen.

CO-CHAIR GAZELLE: All right. Next is usability. How many people would like to give this a High? Middle?

MR. CORBRIDGE: Nine.

CO-CHAIR GAZELLE: And how many people would like to give it a Low? One.

Feasibility: High?

MR. CORBRIDGE: Six.

CO-CHAIR GAZELLE: Middle?

MR. CORBRIDGE: Thirteen.

CO-CHAIR GAZELLE: And Low? So could we recount the Highs. I think there were seven High. High? Okay.

MR. CORBRIDGE: Eight.

CO-CHAIR GAZELLE: Okay, let's recount the Middles then. This is Middle. Raise your hand for Middle, please. And Low? Okay, we are at 19. Did anyone abstain?

DR. CANTRILL: I don't think I voted on that one. I vote Middle.

CO-CHAIR GAZELLE: Add one more to Middle. So that is 12.

All right. Now we need to vote either for or against recommending for endorsement. Who would like to vote for

recommending for endorsement? All right, one. Again? Okay. One for, 19 against.

CO-CHAIR GAZELLE: So now what we will do is we will take a 10-minute break, and over the break I want to think about what we are going to do next.

What we are going to do is come back and think about something that we could vote on -- I don't think we need to vote for the individual characteristics so much as approval or not approval, if they were proposed as a package. So think in your mind about what that might be.

DR. SNOW: Roger Snow. Are we going to be taking a single vote to approve the concept of a package?

CO-CHAIR GAZELLE: No. I think we will take -- We will start by taking one vote of a proposed package, and we can vote on a couple of proposed packages, if we need to, because there are a couple of combinations. The logical one is recall rate, PPV2 and

What should be in the package! (1, 3, 4)

cancer detection rate.

CO-CHAIR PETERSON: I am not so sure I could -- Given the fact that -- I am not so sure that this is beyond our task here. I will come back pretty strongly and say that we don't have a set -- We don't know what that package would look like. So it is very hard for us to vote intelligently about that.

I am not so sure that they can come up with a package in that short order. This is writing a new measure that we don't have.

DR. BURSTIN: The only thing that, I think, would be appropriate to specifically vote on, if you wanted to, is the fact that they proposed them as measures to always be presented together, not as a composite, not in some combined way.

CO-CHAIR PETERSON: Okay. So would this be meaningful for the public, had you gotten the three scores together? Would you like that?

DR. BURSTIN: That's all. I actually think you might just want to take care of it now, so long as everybody is thinking about it.

CO-CHAIR GAZELLE: Do you want to do it now before the break? Okay. So here is the vote. Pay attention.

The vote is -- We are going to ask you to vote in favor of recommending for endorsement or not the combination of recall rate as written, PPV2, the second one of the ones, the true PPV2, and cancer detection rate.

DR. D'ORSI: Can you give us the numbers, please?

CO-CHAIR GAZELLE: Yes. One, three and four, as written.

DR. SMITH-BINDMAN: As written.

CO-CHAIR GAZELLE: So note, as written there are no specific ranges being proposed. The question is --

DR. D'ORSI: And no risk

adjustments.

CO-CHAIR GAZELLE: And there is no risk adjustments being proposed, and after the break we can come back and talk about possible conditions or modifications.

DR. BURSTIN: Usually, you would vote on what you actually want the package of true measures to be. So I think it may make sense to say are there truly conditions on these.

CO-CHAIR GAZELLE: What if we approve it as written without, the three as written? I was thinking we could see if we would do that.

DR. SMITH-BINDMAN: Hypothetical. You would, as written?

CO-CHAIR GAZELLE: So again, we are talking about one, three and four, as suggested by the measure developers that they be endorsed as a group, without further conditions. We will vote on this, and then we will have a break. So we can have discussion

during the break, if we want, and come back refreshed.

CO-CHAIR PETERSON: Just to be clear, while we might prefer the conditions, if we say we don't want it unless there is a condition, essentially we are pushing -- we are going to end up pushing it off for some number of cycles or it can come back within this cycle with conditions?

DR. BURSTIN: No. If there are really reasonable conditions, they could pass them now, which is why I think --

CO-CHAIR GAZELLE: So let's take this vote, and then we will talk about it, because I was thinking that was sort of a natural break point.

How many people would vote for recommending for endorsement the package of one, three and four, as stated, without ranges and without any modifications? You got a number there?

MR. CORBRIDGE: There were nine.

I'm sorry.

CO-CHAIR GAZELLE: How many people would vote against endorsement?

MR. CORBRIDGE: I get 11.

CO-CHAIR GAZELLE: So no abstentions. So let's take a 10-minute break, come back ready to discuss possible conditions that we would like to request the developers.

(Whereupon, the foregoing matter went off the record at 2:58 p.m. and resumed at 3:11 p.m.)

CO-CHAIR GAZELLE: All right. Here is the plan for the rest of the afternoon. We are going to try and get through the remaining discussion and voting on the mammo measures, and then if we have time to move on to some of the measures that we are slated for tomorrow.

So we will finish by five. No need to worry, and if we get through some of tomorrow's work before five, then we will have a better chance of finishing easily tomorrow.

Before the break, here is what happened. We all voted, I think on balance, favorably for the individual aspects of the four ACR proposed measures, though we had a lot of High and Middle for individual characteristics, but we voted against recommendation, almost unanimously, for all four of them individually.

Then we had a nearly split vote, 11 versus 9, against for the combination of 1, 3 and 4 unmodified.

So now what we want to talk about briefly, because there is an unlimited number of potential modifications -- The question is: Is there an easily described and voted on combination of conditions that we would propose to that one, three, four combination that would get people who voted no to vote yes without taking people who voted yes and making them vote no? Right?

What I heard is the conditions that some people would like to see added to

these measures are stratification -- so it is probably stratification in reporting, since we are not proposing thresholds anyways -- for some or all, and that could be both by age -- It could be by age and/or by first versus repeat mammogram.

So that is what I heard, but I would like somebody to propose, because I voted for approval without modifications. So I would like for someone who voted no to that combined group of three to propose conditions that they would find acceptable enough to vote yes.

So if there is no response to this request, that means that all of the people who voted no, the 11 people who voted no, there is nothing that could get you to vote for these measures. Then we can move on, if that is the case. Is that correct?

DR. GEMIGNANI: My vote could be moved. So how many of us would have to move for you to --

DR. BURSTIN: It doesn't really mean -- Either way, this is going to go out to the public and membership as a split vote. So I think, unless there is truly a huge -- everybody just says stratify it, and we are good, we will present it as is. This is not Congress so don't feel like you've got to go peddle for the vote.

CO-CHAIR GAZELLE: Right, but if there was something lurking below the surface that kept -- that you felt, ah, geez, if it was only for that condition or set of conditions, I would have voted for it, this is the time to speak up.

Look for smiths stratification

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. If these measures were age stratified, I would be willing to accept them as a group. I would like them to also be stratified by whether mammograms are first or subsequent, but that makes it more tricky in the feasibility category; whereas, the age doesn't seem to add complexity to doing it,

and feels it is imperative to making the numbers remain the same.

CO-CHAIR PETERSON: To be clear, how many strata do you --

DR. SMITH-BINDMAN: By decade.

CO-CHAIR PETERSON: By decade. So you are going to have three measures times X number of decades.

DR. SMITH-BINDMAN: Forties, fifties, sixty, seventy. So four strata. Four times three is -- It is not bad.

CO-CHAIR PETERSON: Twelve numbers.

CO-CHAIR GAZELLE: Okay. So now are there other people who voted against the combination for whom that would make it appealing enough to vote for it? So we got three others. So that would -- four others. So that is good information.

Are there people who voted for the combined measures unmodified that would be opposed to the reporting of stratified? Carl,

did you vote for them?

DR. D'ORSI: I voted for them. I am just a little bit worried about the number of events you need when you put that decade in, and I don't know if we can get that much data on decades.

CO-CHAIR GAZELLE: So that could be a condition that we asked the measure developer to come back to us with, if they had data about the statistical effect of stratification.

DR. SMITH-BINDMAN: Can I add one more thing as well?

CO-CHAIR GAZELLE: Yes, please.

DR. SMITH-BINDMAN: If the measure developer can give us a sense of what sample size they would want for each of these measures. So how small a facility could they go down to reliably?

CO-CHAIR GAZELLE: So let's do this vote. Again, we are going to be asking you to vote for or against, for or against

recommending for endorsement, and what it is for, measures 1, 3 and 4 with the two modifications, that they would be reported by decade age strata, and we would ask the measure developer to come back and present information about sample size, the likely sample size, and statistical considerations, if stratified by decade.

DR. BURSTIN: Just one other possibility that might be perhaps not as messy would be to actually ask a series of questions to the measure developer we can feed back to you and allow you to re-vote, and see if, in fact -- I mean, you are sort of voting without complete information.

CO-CHAIR GAZELLE: You think they could answer those questions now?

DR. BURSTIN: No, not today. We will give them a week or so to get back to us, and the committee can easily do it on the phone or e-mail. I am not sure you are going to have enough information today to make an

informed decision, unless you feel strongly they already know that information.

DR. D'ORSI: I agree.

CO-CHAIR GAZELLE: So should we go take the vote first without the additional information, since we had four people, five people that switched over, and at least we know how many people we are losing?

DR. RUCKER: But it will be faster if you have the information. We can all vote in a week.

CO-CHAIR GAZELLE: So this is what -- Just so we can have this clear since it will be coming by e-mail, what we are going to do is we are going to propose -- We are going to ask for the measure developer to give us information on some likely sample size in the cells, each strata, and then we would be voting on the combination of the three, 1, 3 and 4, reported by decade age strata, and we would be able to make that vote after we had some indication of the effect that that would

have on statistical --

DR. BURSTIN: And how many strata. There is a lot going on here.

CO-CHAIR GAZELLE: It would be 12 strata, four per measure -- four decades.

DR. BURSTIN: So from 50 -- I am just trying to -- So 40 to 50 -- You need to define that.

CO-CHAIR GAZELLE: Forty to 50, 50 to 60, 60 to 70, and 70 to 80. So one decade -- So those would be the four strata. So what we would like to know from the ACR is an estimate of over, say, if we had it a year reporting period, how many -- what would be the precision of the estimates.

DR. SMITH-BINDMAN: And how many facilities would or would not have sufficient data?

MR. BACKUS: Is it data to stratify 60 to 70 or are you really talking about for usefulness of data? How many stratifications do you need, and does it make

sense to break the line at 65 or 66, since essentially that is where the Medicare data comes into play.

My only concern with the stratification is that, all of a sudden, so now you are a 53-year-old woman, and you are looking at where I should go to get a mammogram, and now I am trying to look at that center's data, and then, well, they are better at 50-year-olds, but worse at forty-year-olds, but good at 60-year-olds.

I just wonder to what degree you start creating confusion in the general public.

CO-CHAIR GAZELLE: Yes. My argument against stratification would be partly that a few of us in the room, and maybe a number of people outside of the room having discussed it, might understand why it is valuable to do, but I think most people would find it confusing.

I think, besides that, even though

there probably is a difference between the numbers you would obtain in a pure, say, 40-50-year age population and a pure, say, 70-80 age population, most practices are blended populations. So that the true range of variability is going to be a lot less than comparing the two extremes.

So those would be the arguments against stratification.

CO-CHAIR PETERSON: I would agree, but if we are going to get numbers, we are going to get -- Within each of these 12 strata, we are going to get the number per hospital that would qualify for that measure. Correct? That is what you were asking for.

So how many 50 to 60-year-olds across the data they have -- how many? -- n is that per 100? So that would be the range, and they would give us 1000 to five cases within each strata.

The other number that would be somewhat valuable to see would be to see what

is actually the range and performance for that measure for that metric, because that would, in fact, inform the issue of do you need the strata at all, because there isn't varying from 50 to 60-year-olds.

CO-CHAIR GAZELLE: How much of that would you be able to give us, do you think? Well, one to two weeks, right, Helen? Re-vote would need to be then.

DR. SMITH-BINDMAN: Do you have data on performance for these facilities?

MS. BURLESON: So the issue is it involves new. So the amount of facilities that we have a full year just started this year, and have a full year of outcome data for some of this. But we won't have a full year of outcome data until next year, even the year following.

DR. SMITH-BINDMAN: So the data that you are asking for from this source is not available.

MR. BACKUS: So I guess the

question, to me, that comes back to the committee then is are we comfortable in an issue like breast cancer saying that, if we don't have strata or the set of performance measures, that we are willing to just let the core combination of the three, which is essentially good enough for a lot of Europe and stuff to use as a basis for at least some measure of reporting -- Are we willing to let that measure die out until whatever the next cycle is, two years, three years, four years.

DR. SMITH-BINDMAN: Versus using a measure that we don't know the association of quality.

MR. BACKUS: You know it is directionally correct.

DR. D'ORSI: And we won't know that even with stratification. Do you know that with stratification, what the cancer detection rate should be at 40 to 50?

DR. SMITH-BINDMAN: Yes.

DR. D'ORSI: Then you should know

it from 40 to 60.

DR. SMITH-BINDMAN: I do know it from 40 to 60.

DR. D'ORSI: Then you should know it from 40 to 90. You should know the whole range.

DR. SMITH-BINDMAN: If you find two cancers per thousand in a 40-year-old, you are doing just fine. If you find one cancer per thousand in a 28-year-old, you are doing fine. If you find one cancer per thousand in a 70-year-old, you are doing horrifically, and I think averaging these measures gives you a very meaningless summary.

DR. D'ORSI: Well, I agree with you that, statistically speaking, you are absolutely correct. Clinically speaking, I don't think it is meaningless. It is often meaningless, but I think you can group these together in a reasonable range and still get some performance metrics, but I understand what you are saying. It is a much stricter

criteria, and you get some more information. But I don't know if it is necessary for what we are aiming at, at the NQF.

MR. BACKUS: This is Mike Backus. See, your are hypothesizing, though, then that, first, sites -- let's say they are doing 2000 exams, so that we are in the realm of reasonable -- that there is significant enough differential in the age of the patient population to swing that data.

You think that -- I mean, I am just hypothesizing, but I would guess that the average center that is doing mammos, the distribution of ages of the patients that they see is very similar. Maybe that is an easy piece of data.

If age is in the stratification, maybe the easy piece of data that you can get in one week or two weeks out of that MQSA or whatever is look at the age distribution of centers and see whether or not there is statistically meaningful differentiation in

that age band.

CO-CHAIR GAZELLE: That would answer the question as to whether or not stratification is out there.

MR. BACKUS: Right. If there is not --

DR. RUCKER: Don Rucker. There is a lot of reason to believe it might be right. If you are in someplace like Scranton, Pennsylvania, where people are moving out on a continuous basis versus Scottsdale, Arizona, where that may have retirees in Phoenix that is booming, you are going to have quite different populations.

In places where there is more Medicaid or more Medicare or something, you are going to have very selective age mixes.

CO-CHAIR GAZELLE: It is an answerable question. Right?

DR. RUCKER: Yes.

DR. GIBBONS: I will just offer the thought that from Cleveland to Rochester,

Minnesota, to Jacksonville, Florida, Mayo to Scottsdale, Arizona, Mayo, very different age distributions.

DR. SMITH-BINDMAN: Give us some magnitude to understand.

DR. GIBBONS: Oh, percentage of people over Medicare is 30, 38; Scottsdale, 61; Jacksonville, 58.

MR. BACKUS: So you can give me the outliers, but if I am the consumer, again, or the public trying to interpret --

DR. SMITH-BINDMAN: No, but 30 versus 60 percent being old versus young.

MR. BACKUS: But if I am the public trying to interpret this measure for quality, I am not picking my mammo, should I go to Scottsdale or should I go to Rochester. I am like should I go to Sloan Kettering or should I go to NYU.

DR. SMITH-BINDMAN: I think your point is completely -- This is Rebecca Smith-Bindman. I think you are raising a really

valid point. I think that, before we put it out there as a measure, it would be nice to have some sense of how much difference it would make it. I think the narrower the allowable that they decide the criteria should be, the more important it is, and the broader it is.

Your point is you want one measure. So the ideal metric would be some relationship within each age category combined, but it would be nice to know that from the data. Is there a big difference based on the distribution of age?

DR. STILLMAN: This is Art Stillman. Scott, you raise an issue about how confusing it might be for patients having risk stratified data. But I think, even more confusing, at least for me -- I am confused -- is how we are going to be using three different metrics that are coupled and use that to rate different facilities, so that patients know that they would rather go to

this facility rather than that one.

CO-CHAIR GAZELLE: Well, as I understand it, we are not proposing a rating mechanism. We are just proposing public reporting.

DR. STILLMAN: But public reporting doesn't happen in a vacuum. It is going to be used for something.

CO-CHAIR GAZELLE: I would assume that patients would do it and --

DR. STILLMAN: Well, but then it needs to be something that is understandable to a patient. It is not understandable to me.

CO-CHAIR GAZELLE: That would be the basis on which you would vote then, I suppose.

CO-CHAIR PETERSON: Okay. So we have clarified what the request is. I think at least we put in our request, and we say we would want the Ns, range in hospital Ns, and we would want -- secondly, would be the average or mean age distribution for those

hospitals, how much variance there is among hospitals.

DR. SMITH-BINDMAN: The mean or median age?

CO-CHAIR GAZELLE: You would get both.

CO-CHAIR PETERSON: Range and mean.

CO-CHAIR GAZELLE: I mean, the real question is within a given region.

DR. SMITH-BINDMAN: No. No, it isn't.

CO-CHAIR GAZELLE: It isn't, because again you want everybody in Florida to go bad, because they are all on the bad side of the score. So it is not going to be popular.

DR. RUCKER: Don Rucker. It also varies by practice within a city. Honestly, within a city --

CO-CHAIR GAZELLE: Well, that is the question.

DR. RUCKER: -- it is surreal. Somebody made the point -- I think, Mike -- about, you know, you are not going to go to Scottsdale or Rochester, but I think within a city, you know, if you are in a clinic situation or something that has some sort of catchment mix, I think these things vary a lot; and if we are asking people, even before the confusion, which I am sort of also quite confused, but even before the confusion, I think it has to have just an intellectual honesty about, if you made the effort of understanding it, that this represents reality, that this represents sort of total stand-alone data.

MR. BACKUS: As you get down in the city -- This is what I do all the time -- you know, the acuity of a practice is always something -- For any practice that is an outlier in utilization, the first discussion is about the acuity of that practice's patients.

I am just a fan of even getting some version of a measure out there, and if you say that your practice is different and you can document it and things -- Remember, you know, we have talked about there is a range, and we are trying to look at the outliers.

If you are really, truly that outlier and you can really, truly document that acuity or whatever that argument is, then I think you've got a very valid explanation, and there are things that make that practice unique and understandable. But I think, until we get at least some version of measures even under discussion, we will just forever be in conjecture.

DR. D'ORSI: Carl D'Orsi. One other thing is feasibility. There are people now who are on the edge of not doing mammograms. So there is a possibility of an access issue if we add more, which is the three general measures. If we then ask for 12

strata, you are going to drive a lot of people out, maybe for no good reason.

Even the three general conditions are going to be difficult to get, even with an electronic model or module, unless you go to some organized database where you can get feedback. If you have to do that by hand, there is no way you are going to do it.

So this, on the feasibility side, may be an impetus to drop access. I just think we should keep this in the back f our heads.

CO-CHAIR GAZELLE: So we have time for maybe one or two, three more comments, and then we are going to need to move on. So, Troy, and then Judy.

DR. FIESINGER: I will be brief. I agree. I think some measures would be better than nothing. I think the stratification will matter a lot if I am the medical director, depending on my practice.

To me, as a physician, is it

important? Is it close enough to the patient they can get there? Where was the patient's last mammogram? That is really what I am going by.

Kaiser Foundation did a great study five years ago on whether patients use quality measures to choose surgery and physicians and hospitals. No. They ask their neighbors and their friends, and I have seen that true in five years of practice, which is frustrating to NQF, but that is the reality.

DR. BURSTIN: The end user is not just consumers. It is those who purchase care on their behalf. It goes beyond just whether an individual consumer can figure it out. So just keep it really broad, and again, lots of people -- The number one consumer of a lot of the information on these various compare sites are actually clinicians looking for stuff for their patients. So don't limit ourselves to thinking it would --

DR. GEMIGNANI: A brief comment

about the age stuff. I think that it would make sense from my view to stratify it into two age groups, under 65 and 65 and older, because of the Medicare payer issue, and then it is not too many different age categories.

I recognize that it is not perfect in terms of where cancer is diagnosed, but in terms of access it makes sense in that way. I would absolutely second that I think these measures are more used on the facility level to say why are we a total outlier.

No one wants to look bad, and in terms of payers and system issues, I think that this moves quality that way, although it is less understandable to an individual patient.

CO-CHAIR GAZELLE: Thank you.

DR. SPENCER: Just to answer Mike's question -- So I voted no, but if this data is not available, I am not in favor of seeing the measures die.

CO-CHAIR GAZELLE: You would vote

for it?

DR. SPENCER: Yes, if this data is not available.

CO-CHAIR GAZELLE: If we couldn't stratify it. Okay.

All right. I think, as hard as it is to vote by e-mail because there is really no opportunity for a dialogue that we can sit and look at each other -- I think we have probably had all the dialogue we can have about this measure.

Clearly, there is a lot of sentiment for this combination, and also a lot of concerns about -- you know, the devil's in the details sort of thing -- about how they would be used and understood.

I think it is time now to move on to the remaining mammo measure. So we are going to go through the voting again, all four levels plus an overall. Luckily, I don't think we are going to propose to combine it with others. So that part should be shorter.

So we are now voting on measure 009-10 mammography.

MR. CORBRIDGE: Scott, I hate to just interrupt. Quickly, I forgot on the last measure set, is anyone on the public line who would like to make a comment?

CO-CHAIR GAZELLE: Is anyone still on? Or anyone from the public, and I know we have measure developers, but anyone from the public that would like to make a comment before we proceed to voting?

So we are going to go to 009-10, mammography follow rate in the Medicare population. I think, before we vote, we should -- My sense was all agreed that it should not be limited only to hospital outpatients, that it should be -- So that would be a condition we would propose.

We, I think, all agreed that there wasn't a specific range that was going to be part of this measure. So we are not voting on a specific range so much as publicly reporting

all Medicare beneficiary hospital outpatient and other facilities.

So we need to go by the four categories ago. Importance: Who would --

DR. SPENCER: I'm sorry. With the change we are voting, or without?

CO-CHAIR GAZELLE: With the changes.

CO-CHAIR PETERSON: The changes that we are going to do outpatient -- hospital and outpatient.

DR. BURSTIN: And the developer has already agreed.

CO-CHAIR GAZELLE: Okay. So we are voting on the importance of the measure and report. We all have it. Who would give it a High? Nine? Middle?

MR. CORBRIDGE: Ten.

CO-CHAIR GAZELLE: And Low?

MR. CORBRIDGE: One.

DR. FIESINGER: I voted High.

CO-CHAIR GAZELLE: Do we have an

abstention? Did somebody Abstain? Let's have it again. High? How many Highs?

MR. CORBRIDGE: Still nine.

CO-CHAIR GAZELLE: How many Middle?

MR. CORBRIDGE: Eleven.

CO-CHAIR GAZELLE: Good. Lows? Good. Okay, now we are moving to the second category, which is scientific acceptability of the measure properties. High?

MR. CORBRIDGE: Six.

CO-CHAIR GAZELLE: Medium? Middle?

MR. CORBRIDGE: Thirteen.

CO-CHAIR GAZELLE: And Low?

MR. CORBRIDGE: One.

CO-CHAIR GAZELLE: Next is usability. High? We are talking about usability. Feasibility is the next one. How many want to vote High for usability.

MR. CORBRIDGE: Eight.

CO-CHAIR GAZELLE: Now Medium for

usability?

MR. CORBRIDGE: Twelve.

CO-CHAIR GAZELLE: Okay, no Lows. And now feasibility. High for feasibility?

DR. RUCKER: This is just getting it from Medicare data themselves. Right?

CO-CHAIR GAZELLE: Should be 20. Okay. Now we are voting either to recommend for endorsement or not to recommend for endorsement.

DR. SMITH-BINDMAN: With the condition.

CO-CHAIR GAZELLE: With the condition which we talked about. Who would like to vote for -- to recommend for endorsement, with the condition meaning all instead of just hospital? Four. No range, yes.

MR. CORBRIDGE: Looks like nine.

CO-CHAIR GAZELLE: And who would like to vote against recommending for endorsement.

MR. CORBRIDGE: Eleven.

CO-CHAIR GAZELLE: No abstentions? All right. We have finished the mammo.

DR. BURSTIN: Identical.

CO-CHAIR GAZELLE: Yes. Okay. Yes, Don?

DR. RUCKER: Do we want to do anything -- Some of this, I could imagine, is on what we do with the other mammo in terms of the overlap, or are we sort of saying there is just no real overlap. I would be curious to see, because the group of four, or group of three mammo things -- I am just still --

DR. DEHN: I think we can certainly do combinations, but I would just ask on the last three, you would ask if there was anything on their mind that we could include that would change their mind. I would ask, and we are entitled to that.

CO-CHAIR GAZELLE: Sure.

DR. SMITH-BINDMAN: The same as the prior.

CO-CHAIR GAZELLE: The same as the prior, yes. We talked about that with conditions, but are there other conditions?

DR. SMITH-BINDMAN: Rebecca Smith Bindman again. It would be nice -- I would be more favorable to the measure if the results were age stratified, and if there were some validity data provided on the new Medicare CPT codes.

CO-CHAIR GAZELLE: So the request would be age stratification, and it would be--

DR. SMITH-BINDMAN: And it is less than 40 in this measure, but there is no reason not to have a 65 to 70 in this population. It is less important than the other one.

MR. BACKUS: How much do those ranges change, the 65-70, 70 and 75, 75 above. How much is that?

DR. SMITH-BINDMAN: There are two reasons the recall rate changes. Partly, the incidence of cancer, but that is a trivial

amount. Most of it is breast density continues to decline, and so the false positives just happen to go down a lot, not the same rate as 40 to 80, but --

MR. BACKUS: What was the discrepancies in screening and diagnostic? What was the range in the code? The issue of the accuracy of the code?

DR. SMITH-BINDMAN: For the old code? About half of the screening exams were coded as diagnostic. So my guess is the purpose of these codes was to fix that problem, but it was an enormous issue.

CO-CHAIR GAZELLE: So then, just to be clear, I think we can all -- I am presuming we can all agree that that is an important piece of information we would like.

Let's take a quick look to ask for how many people is stratification for the CMS measure important? How many people feel that that should be done? One, okay.

How many people feel that it

shouldn't be done. Then I think I am going to ask how many people are neutral and how many people feel that it shouldn't be done.

So how many people are neutral, don't care one way or the other? And how many people would prefer that it not be stratified?

MS. DaVANZO: I think Medicare patients include -- presumably dominated by the Medicare 65 and older. The disability population doesn't consider it at all.

MR. GIBBONS: Mr. Chairman, just to clarify. You said this condition of the CPT codes was something everyone would accept. I didn't accept it. That is why I was the single low vote on scientific acceptability.

CO-CHAIR GAZELLE: No, the question was whether or not we want to ask them to provide that information.

MR. GIBBONS: Okay, but in terms of the previous vote, that was the basis for my low scientific acceptability vote.

CO-CHAIR GAZELLE: We are only

voting now on approve or no. So the question is whether or not we would all like to have that information, and I was just presuming we would all like to have that information.

DR. SMITH-BINDMAN: In fact -- This is Rebecca Smith-Bindman. So the information -- it doesn't have to be a perfect reference standard. If you can show that the distribution of current mammograms is about 90 percent with your screening code and 10 percent or 15 percent of your diagnostic code, that would be consistent with the distribution that I have --

DR. BURSTIN: The problem is you just let that information flow back to the committee. Again, it was equally split vote.

CO-CHAIR GAZELLE: Carl. Then Don.

DR. D'ORSI: I just want to make a point. We don't have to discuss it. Since this metric is very close to what we think of as follow-up rate or recall rate, I would

think we need the same kind of information that we requested on the other recall rate; and if CMS has a valid way to produce that information, I think that would be nice, but I am just saying that I know we are not thinking of this with other metrics, but just as a point of discussion, I think it becomes not as relevant when you don't have that information. It is very similar to recall rate.

DR. GEMIGNANI: My only point, I guess -- This is Mary Gemignani -- is that this group is so uniform that you probably have data on cancer detection rates already. So you don't really need to collect it, as you would in the other three measures, and this is separate.

So I think that, when you have got a recall rate within whatever center and you wanted to evaluate it, you could get the cancer detection rate, because of where the data is coming from and the population that is

cited.

DR. D'ORSI: But I would like that bundled in automatically, not that somebody has to -- I would like it as a package, not that this goes out and that somebody says, okay, what is the cancer detection rate.

DR. DEHN: Carl, you would like us to report out not only the indeterminate rate, but also whether that indeterminate rate seems to be generating more cancer.

DR. D'ORSI: And if you can -- I don't know if you can get the type of cancer.

CO-CHAIR GAZELLE: Carl, I think what you are proposing is another measure.

DR. D'ORSI: That is true. I said it is not for discussion. I am just pointing it out as a point of information that, to me, it becomes not as relevant as when we discuss recall rate. That is all.

CO-CHAIR GAZELLE: Okay.

DR. SMITH-BINDMAN: Can I just give you numbers for the recall rate by age,

just because we talked about it.

The recall rates for women less than 40 goes from nine, and it drops to 8 for women in their fifties, 7 1/2 for women in their sixties, and 6 1/2 for women in their seventies. Those are the average.

MR. BACKUS: So the stratification, though -- what you are saying, if those are the recall rates -- I mean, the stratification that you are talking about is -- I mean, you are only going to move -- You moving such a trivial --

DR. SMITH-BINDMAN: For the older women, it is much smaller. For the young woman, I think it is a much --

MR. BACKUS: Well, no, you said it goes from like 9 to 8 to 7, 7.

DR. SMITH-BINDMAN: Six to nine is a 50 percent difference based on --

MR. BACKUS: Understood. But so if you think of a distribution of age of people in the practice, now for that

stratification I would have to have -- A 10 percent or 15 percent change of old people to young people within a practice will get ground out in there, because I am looking at 15 percent on four. So I am looking at a half a percent of recall rate.

DR. SMITH-BINDMAN: I think that is why it matters whether you are talking about coming up with really narrow ranges of quality or really broad. At the really broad ones, I completely agree with you. If you are getting a narrow, we are talking about 10 to fourteen.

DR. GEMIGNANI: We eliminated the rate.

CO-CHAIR GAZELLE: We eliminated the rate.

DR. GEMIGNANI: We weren't thinking a rate. We were just going to report.

CO-CHAIR GAZELLE: We are not thinking a rate. All right. Just to tie up

the discussion on this, we had a split vote. We are asking the measure developer to come back to us with information on the accuracy of coding screening versus diagnostic, and I think we are of a mixed mind on stratification, one person strongly in favor of reporting the stratification, a handful of people against it, and most people neutral.

So we will vote again on this as well, Helen? Is that -- We will vote again with the additional information on this, but cement it in your memory.

We are going to now change direction, and I am going to pass the gavel to my colleague, and we are going to move to measures number --

CO-CHAIR PETERSON: Measures number 7 and 8. For those who are not aware, one of our members is going to be leaving tomorrow and will not be around in the afternoon. So we might do these two measures, and get through the day without him.

Some people didn't get 7 and 8.

DR. BRUETMAN: Based on the discussion we had previously, I would like to know from the committee if that information that was requested, the stratification work to be done and the new CPT codes were in the range and would be accessible, would the committee endorse it or not? The other --

CO-CHAIR GAZELLE: We are going to vote again. We are going to vote again. We are not going to make a commitment based on information we don't have.

DR. BRUETMAN: I ask because the other one, the age based, all those things were endorsed.

CO-CHAIR GAZELLE: No, we didn't. We didn't vote on either of them. We are just asking for information, and going to vote again by e-mail.

Okay, now we will move on to seven and eight.

CO-CHAIR PETERSON: Seven and

eight. The measures are appropriate head CT imaging in adults with mild to traumatic brain injury.

So EP-007-10. Numerator is the number of denominator patients who have a documented indication consistent with the clinical quality for mild traumatic brain injury prior to imaging.

The denominator is the number of adult patients undergoing head CT for trauma and presenting within 24 hours of a non-integrating head injury, which is Glasgow Coma Scale.

DR. FORMAN: So just as background for this --

DR. BURSTIN: Is the measure developer here or available? The only issue in us reviewing the measure in their absence is they are having to be here tomorrow.

CO-CHAIR GAZELLE: And is somebody from Brigham coming tomorrow? Do we know?

DR. BURSTIN: I don't know.

MR. CORBRIDGE: I haven't heard, actually, if anyone is coming in person. They may be on the phone, but I don't --

CO-CHAIR GAZELLE: Is there a way to find out, because if they are not going to be here anyway, then there is no reason to do it today versus tomorrow.

DR. BURSTIN: Well, they would at least be on the telephone.

CO-CHAIR GAZELLE: Can we do the cardiac, start off with the cardiac?

CO-CHAIR PETERSON: The cardiac? Well, the cardiac -- they are not here either. What is the other? The third one is fine?

DR. SPENCER: Well, there are two cardiac studies here now.

MR. CORBRIDGE: I can go place a call with them to see if they are going to be on the line early in the morning, and we could run through this maybe right in the beginning.

DR. BURSTIN: We could do them right now, if they could call us.

CO-CHAIR PETERSON: Shall we start then?

DR. BURSTIN: Just call them, just so that we would hate to have to rehash it if they are not here.

(Whereupon, the foregoing matter went off the record at 3:52 p.m. and resumed at 3:57 p.m.)

CO-CHAIR PETERSON: Could we do some very quickly?

DR. CANTRILL: It won't be so fast. There is a lot of good stuff.

DR. ZERZAN: How about the applicability of their ratings?

DR. CANTRILL: Applicability is an issue, but I think, especially now with the number of denials that people are seeing, they are learning that they have to have an ordering system that gives you not a process, not rule-out, but an indication. That is where this falls in with that very nicely.

All I need to do is give you one

reason, one thing that the patient has that is consistent with that guideline, and then that is success.

DR. ZERZAN: One is a guideline, not quality. Is that linkage hard to find? Everyone who knows computer order entry will game, once they learn the right thing. So proving that they really have that condition is much harder.

DR. CANTRILL: That is true with anything, without question, and they can be gaming and can game almost anything, as we have seen.

DR. ZERZAN: Absolutely.

DR. CANTRILL: Certainly, with a lot of the guidelines.

DR. ZERZAN: With me, in my world, people do it all the time. Then we change the rule.

DR. CANTRILL: Are we just going to give up and go home? I think that the issue is overuse. There clearly is overuse in

head CTs. The question is how do we go about addressing that issue. Do we just say order less? What the hell does that mean?

Does that mean on Thursdays I don't order head CTs or do I try to about it in an organized fashion, looking at what we have in the literature based on clinical guidelines.

So they are guidelines that address the patient population that we want to address in terms of the emergency department, and we look at graded literature, not to someone's notions, not a consensus panel. So this is done based on a guideline that is pretty rigorous in the way it is put together.

Now I will also divulge, I was part of the panel that put that together. I have the scars to show for it, but I think that this is a reasonable approach.

The CMS guideline -- all that is, is a count. You know, how many head CTs did you do per head. That doesn't get at the

issue. The issue how many appropriate or inappropriate head CTs did you use.

That is where this, although, yes, there are some difficulties with applicability, I think that this really does get to clinical medicine, not just someone with a dull sword trying to cut down the number of studies.

Other than that, I don't have anything.

DR. FORMAN: He is calling in. So I can give a preamble. I don't think he will miss the preamble.

CO-CHAIR PETERSON: Okay, good.

DR. FORMAN: I think the preamble about both of these are -- and I will state for both of them first, both the CT and the cervical spine CT in the setting of trauma, is that there are good evidence based guidelines in both cases.

There is evidence in the literature, to begin with, that - both

evidence based guidelines -- that current imaging far exceeds the evidence based guidelines, and that there is evidence of overuse, and perhaps the only limitation -- and we will go through it point by point, but the only limitation for all of this is that much of the evidence based guidelines were first predicated on cervical spine radiographic imaging, not necessarily cervical spine computed tomographic imaging.

Cervical spine computed tomographic imaging has been available for both head and cervical spine for over 20 years, has been used. So we have very good evidence that it is more sensitive than radiographs in the detection of injury.

There is no evidence existing to date, even anecdotally, that the incremental cases that are picked up are actually -- that affect outcome in a meaningful way, although they are more sensitive, and they are useful in the guidelines that have been presented.

So then starting with the appropriate head CT in adults for mild traumatic brain injury -- So the main reason why I am just using that whole preamble is it is not that CT has not been available 10 years ago when many of these evidence based guidelines were used. It is just that we were still under the paradigm of using cervical spine radiography.

Now in most practices, a lot of the radiography has just migrated right over to CT imaging. So it is just something to consider in terms of judging the evidence.

Just starting with the importance of the measure and, of course, in looking at the demonstrated high impact aspect of health care, it is an enormous part of both the radiology practice as well as the emergency room/trauma practice in head CTs and cervical spine imaging in the setting of trauma.

So following down, I don't know if we give the rating as I go alone. So as far

as high impact, I think you meet completely the standards.

Opportunity for improvement: There is also substantial evidence in the literature of both the use of the CT head rules in the setting of trauma and the fact that, despite the fact that these rules have existed for quite sometime, that there is still excess use and considerable variability in the use of head CT in the setting of trauma.

So again, I would argue that for this, more so than on the cervical spine, there are still some questions. It meets completely the opportunity for improvement standard.

Under outcome our evidence to support the measure, there is considerable purity in the literature that goes way back. Like I said, CT in the setting of trauma has been used for well over -- probably into 30 years now, but really in broad usage for at

least 20 years, and really considerably bigger usage over the last couple of decades as CT imaging has been a lot quicker and easier to do.

So there is the Canadian CT head rule CTOHR, which has both been -- you know, initially validated and then subsequent studies were applied, and in the subsequent studies, they compared that rule to the New Orleans Criteria, and so that the Canadian CT head rule was more specific overall, and that both rules were 100 percent sensitive to patients with injuries requiring intervention.

So overall, on that basis, again I think it meets completely the standard of outcome or evidence to support the measure focus.

Then subsequently, the strength and the quality evidence: Like I said, there is considerable evidence, particularly on the CT standard, and there really is no quarreling about the previous applications, since

radiography for the head CTs has just been doing head CTs throughout this entire period of time.

Let me see what we are down to then. I think we are up to number 2 now, scientific acceptability of the measure properties, bench specifications.

The numerator statement is basically the number of denominator patients who have had trauma, as we will define, who meet the criteria for imaging prior to imaging. It is basically affecting just the initial visit, does not really include cases of follow-up imaging in the setting of trauma where either there is a known finding or a questionable finding.

Then the listed indications that you see below are from the evidence based criteria, which either include loss of consciousness or post-traumatic amnesia and at least one of the following findings, as you see below, and again I am on page 70 of this

guideline, patients without loss of consciousness or post-traumatic amnesia, and either severe headache or vomiting -- and it goes on, age over 65, etcetera.

We said the denominator is all those that present in the setting of trauma.

DR. CANTRILL: I think there is a typo there. I think the denominator is supposed to be people with GCS greater than or equal to 14.

DR. FORMAN: Oh, okay. I didn't know that.

DR. CANTRILL: Right. By reading it very carefully --

DR. BELLO: Comparing it with the one at the top.

DR. FORMAN: Yes. There is a definite little typo in line 1.

Okay. So what are we up to now. And the denominator exclusions are listed here. And I think that is it for 2(a).

I think we are on 2(b). So

reliability testing: There is evidence on all this, and it has been validated, although I believe that they are -- well maybe it is just the c. spine one that they are actually undergoing validity testing right now as well.

So I think, actually, on the reliability testing you do have -- it does meet completely the standard for reliability testing. Right?

DR. RUCKER: Are you talking about 7 or 8?

DR. FORMAN: I am on 7. Yes. Same thing for validity testing. They are not presenting validity testing. So I don't know what -- I guess I need some guidance on that. They have -- These measures have been tested over and over. I mean, we have the 2005 paper, a comparison of the Canadian CT head rule and the New Orleans Criteria.

So what level do you need to actually judge something that is being ruled as valid when you have already done a

validation study?

DR. BURSTIN: Those are research studies, and the difference would be this would be in real practice. Can you reliably collect these data elements, they are saying here, either in terms of paperwork or electronically.

DR. CANTRILL: Several of those studies, in fact, are from their practice.

DR. BURSTIN: Oh, good. That is good to know. It is not clear. This would be the kind of thing we would love to have --

DR. SMITH-BINDMAN: The data weren't collected for the research project. They were collected from routine clinical practice?

DR. CANTRILL: Some were, especially if you look at some of the Dutch. They have a very good registry, and they did everybody for a period of time.

DR. RUCKER: This was a prospective research study? It is not?

DR. CANTRILL: This? Well, this is the culmination of a lot of - multiple sites in terms of the setting of the criteria. Now I don't know if Jay in terms of his work -- I don't know if he did a study on this or not.

DR. GEMIGNANI: This is Judy. What is the range? You know, if people measure it, what do you get out of that, which wasn't clear from this measure. What are you measuring? What is an appropriate -- You know, presumably they have applied this to their practice, and so they have a range of 10 percent or --

DR. FORMAN: Ten percent that are outside the guidelines?

DR. GEMIGNANI: Right.

DR. FORMAN: Okay.

DR. GEMIGNANI: You know, there is no -- It is hard to figure out what they mean by their ratio and what gives you.

DR. RAKSIN: This is Patti. This

is going to come up tomorrow. It came from the Brigham. It is the same issue of what you are really assessing here is adherence to a single clinical guideline, and what kind of QI initiative is that, really.

DR. BELLO: My interpretation -- This is Jacqueline Bello. My interpretation of it was that range in the sphere of overuse and efficiency, that the ratio would tell us what percentage of the gazillion CT scans that you are doing from that ER are actually meeting some criteria.

So, back-pedaling, they go and they evaluated the Canadian head criteria, the New Orleans Criteria, and then came up with this nice little A set list which they published, which is a collaboration of radiologists, ER physicians, and others.

So once we know how many of your gazillion head CTs would really meet these criteria, and they are trying to balance it with "and, no, we are not being dangerous,

because you have to have a Glasgow Coma Scale of 14 or better," so we are not talking about not scanning the comatose -- No, their implication is -- Well, that is another issue, I guess. But anyway, their implication is that may be somewhere between -- they say 37 percent scans could be deemed as overuse.

So the measure is to get a handle on, institution by institution, ideally, whether the number of scans you are actually doing meet any criteria at all. In today's operations, it has got the balance of the radiation use and, other than the dollar, attached to it.

DR. CANTRILL: What is really going to happen -- you all know this; anyone who practices clinical medicine. It is the Hawthorne effect. We start looking at this, and the numbers are going to drop dramatically.

When I am told, well, they are going to be looking to see for every head game

that they have at least got something -- you know, show me something in this guideline. Then suddenly you are going to start seeing adherence, and your number of head CTs is going to drop or at least the rate of climb is going to slow.

So that really -- So it is going to be very hard to say, well, look at the quality that we have given here. We don't have a baseline. If we could sneak in there right now and get a baseline across different institutions and then put this in place, then we could say look at what we have done.

DR. RAKSIN: Patti again. I think this is going to come up again tomorrow as well. The other thing that is missing here is we don't know how many positives show up out of the ones that don't have indications. That is part of you need to really understand overutilization.

DR. SMITH-BINDMAN: Although -- This is Rebecca Smith-Bindman. What the

writers have said is they have cited guidelines that have 100 percent -- I am not defending this, but I am saying in application we have a guideline that you know are not going to miss anything significant. Then you can just start looking at adherence to the guideline. You don't need to worry about the primary misses that you are asking about.

DR. CANTRILL: If you really want to understand that -- Steve Cantrill -- you need to understand the evidentiary table that goes along with this guideline, which is about 16 or 17 pages long. It goes into detail of the evaluation of all the different papers, and that is how -- We agonized over that. We really did, in terms of -- because no one wants to miss a -- But you can't, by the same token, head buzz everyone who walks in the door. So you use random criteria or no criteria or you try to be somewhat scientific.

DR. FORMAN: Can I just finish up a couple of other points, just to add on there

as somebody who practices in the environment of trauma imaging for 15 years right now.

I agree with you fully, but I actually think that a guideline put into place appropriately will influence practice. It will influence the adoption of computerized physician order entry. It will have so many external effects that will be favorable to the overall system that, without overdoing the pun, this is a no-brainer to me.

I think you really -- You know, the opportunity here is to take something -- This is, to me, like aspirin after MI. It is something where you try to find institutions that come very close to 100 percent compliance with the guidelines.

Now there is no question, we will find a certain degree of gaming by physicians that are ordering. They are going to remember a few symptoms that they have to put in there. That is the only way they are going to get it, and they are going to improvise about whether

it was really a high impact collision with, you know, intrusion of more than 18 inches or whatever the criteria are to make a major high impact accident. But I do think that you will actually -- because they have these very specific criteria.

I do think that you will have an opportunity to really impact and improve care, just by a relatively simple guideline. I would say you go in academic institutions; you find very -- Well, I won't say important clients -- you have some people with excellent clients who are telling you precisely why they are ordering a head CT on everyone, and as we have joked since I was trained at Wash U 20 years ago, that the indications for a head CT is if you have a head.

DR. CANTRILL: And we prefer a pulse as well.

DR. RAKSIN: Two other things. Having said what I said earlier, there are indications for ordering a head Ct are pretty

loose and far encompassing. So virtually anyone who has a headache, who has a head, would qualify for a head CT scan criteria.

DR. FORMAN: I am not sure about the -- I mean, they show applications --

DR. RAKSIN: Right. The other thing was that I think we have to ask the developers has to do with the definition of mild traumatic brain injury and who they are actually including, because traditionally, the GCS is 13 or 14 or 15, and they seem to have excluded the 13s.

CO-CHAIR PETERSON: So can we get back? I am just going to keep a little -- We have got a lot of discussion going on. I believe you are at -- You have gone down through reliability. Are you at reliability?

DR. FORMAN: I was, and then I backed up. So let me get back to that.

DR. BURSTIN: The measure came in as non-tested. So it will be time-limited.

DR. FORMAN: Okay. So let's go to

-- We can skip over SC analysis, and there is some degree of evidence supporting exclusions. They mainly point out the populations that weren't included in the previous studies, because they were either perceived to be a virus with serious injury or indicates a pregnancy, either concerns with radiation exposure to the fetus. So I felt those were at least either partially or completely supportive based on the evidence that we have.

No risk and non-applicable for risk adjusted for outcomes in equal difference in performance, I think, we are not evaluating.

Overall, to what extent is the criteria of scientific acceptability of the measure properties met? I would say completely, notwithstanding the small groups.

Then on the usability, whether it is meaningful, understandable, and useful information, still undergoing current testing. So we don't really know what the findings will

be from various institutions, but we would imagine that it would be along the spectrum of like it did with aspirin where you have a percent compliant with the guidelines, and that it would probably be less than 100 -- obviously, be less than 100 percent.

These institutions will have some latitude within the guidelines where other measures may be taken, but in general, it would be that type of measure.

No harmonization, because there is no prior guidelines at NQF.

So to what extent was the criteria usability met? You know, I would say at least partially in the absence of actual applicability and data.

Under feasibility, this is probably the most contentious issue, and this is, I think, the challenge. I don't know where the group comes down on this, but I will tell you, feasibility-wise these are not easy to institute in terms of capturing the

information.

This is not dissimilar in terms of getting the information from PQRI and the Physicians Quality Reporting Initiative, and I can tell you that, even a huge practice like we have at Yale, if you don't have well coordinated, computerized physician order entry and coordinated with data collection, it is an administrative burden.

It is possible, and I think it is possible for everybody to use, but how you define usability is an open question. I would say that, on this count at least, one would have to say partially.

You know, how are the data measures generated? I think it is a by-product of care processes, but it is not easily generated. It is not necessarily captured automatically, and you will find, I think, that at smaller institutions, which is where the majority of patients are cared for, it may be more difficult to capture that

information.

They mention computerized physician order entry, and I think that that is the way to do the validation studies, and it certainly is the future of being able to use a measure like this, but I think this is the only limitation around the measure itself.

DR. SMITH-BINDMAN: Can I ask you a question. this is Rebecca Smith-Bindman. When you say the feasibility, I think what they are saying is that, if you have ordered a head CT and you have ordered it for mild traumatic brain injury, then you need one of these indications.

So you need two steps. You need defining the patient population, and within that population defining the category.

DR. FORMAN: Right.

DR. SMITH-BINDMAN: Is that feasible within the data order entry? The specific category, I get, so vomiting or not vomiting.

DR. FORMAN: Right.

DR. SMITH-BINDMAN: But the denominator -- is that possible at Yale?

DR. FORMAN: The denominator is stated as a positive finding of --

DR. SMITH-BINDMAN: No, mild traumatic brain injury.

DR. FORMAN: That is a clinical finding, mild traumatic brain injury.

DR. SMITH-BINDMAN: Right. So I don't know if this is defined from the radiology point of view, from the data that the radiologist could have had access to, or--

DR. RAKSIN: It is probably -- What happens at our institution is that, especially in trauma or in the emergency department, it is the emergency room physician who is ordering the study who has to list an indication for the study.

Now sometimes they will, in their indications, put mild TBI rather than headache or nausea and vomiting. So that is an

education issue, but I know that we certainly do our share of trauma head CTs, and for us data collection in the trauma unit -- we are not computerized in the emergency department.

CO-CHAIR GAZELLE: Even in the measure as submitted, 4(b).2, it says "All data elements are not likely to be available electronically to most providers currently. Although many electronic health records include CPOE, most are not programmed to have" -- and they go on to say how they are doing it at the Brigham and this and that.

They say it would be technically feasible to reprogram the system to do this. Then they go on to say that it would also be possible to do chart review, but that is not likely to be useful, since a lot of times the information isn't in the chart at the time, and it is not feasible.

So I think this is the Achilles heel of this measure, if it can only be done at a small handful of institutions.

DR. RUCKER: It is not that they actually use their core HIS system to do this. Right? This is a stand-alone separate order entry system, is my understanding of it, that was custom built for this. So this is not --

DR. FORMAN: But integrates with their --

DR. RUCKER: It may integrate, but it is not like they used a commercial CPOE system and quote/unquote "reprogrammed it." This is a hand-built custom system.

DR. GRIFFEY: Actually, no. I work there. So I know that they use a Precipio proprietary system for CPO. It sits on top.

CO-CHAIR GAZELLE: No, no. They built the interface between that and the electronic medical records system. That is what they built.

DR. GRIFFEY: I think that is right. No, I agree with you. I think it is a great measure, but the difficult piece of it

is this piece, and they talk about putting together a template to try to collect this data, and your concern, I think was how do you define the denominator. Is that right?

DR. SMITH-BINDMAN: Right. If it is two separate populations, one is an ED defined variable. The other is a radiology defined variable. I am not sure if --

DR. GRIFFEY: You would have to use the ED defined variable, I would think, and it would have to -- Typically, the indication almost never is going to say, you know, TBI. It is going to say evaluate for intracranial hemorrhage or --

DR. SMITH-BINDMAN: Right.

CO-CHAIR PETERSON: So this system, a proprietary system that does measure this, is proprietary to? Who owns that?

CO-CHAIR GAZELLE: It was developed at the Brigham, and it is now licensed to a company that you can buy. That is the order entry system, but the interface

between the order entry system and the electronic medical record is a Brigham system.

CO-CHAIR PETERSON: Okay. So is there other proprietary systems out in the market, other than this one, that would allow you to measure this measure?

CO-CHAIR GAZELLE: There is one other one, but again you would need to develop the interface between that one and the medical record system.

DR. CANTRILL: You don't need a computerized system. You can do this manually. It might require some work, but you can get it. We don't need to worry about proprietary systems.

I think the other issue is what direction do we want to push American medicine in? This is the direction. We would like to have studies done for a valid indication, and we would like to have the appropriate information conveyed to the radiologist. Does this push us in that direction?

DR. SMITH-BINDMAN: What would this system be, just to understand this. someone has to define it.

DR. CANTRILL: We are working on a paper system right now that we would be able to use.

DR. SMITH-BINDMAN: So just walk me through how you would do this with paper.

DR. CANTRILL: Sure. Well, it is partially computerized, but I click on the patient's name, and I said I want to order a CT, and then it says what are the indications, altered mental status, whatever and listing the mechanism, and what study do I want. I want a head CT. And what am I trying to rule out? I am trying to rule out intracranial hemorrhage.

Then that has all the necessary information on it, and that goes to our radiologist.

CO-CHAIR PETERSON: So how could you get it out of that to somebody to do NQF

reporting?

DR. CANTRILL: Well, as was pointed out, that is the tough question here.

CO-CHAIR PETERSON: So, currently, the only way that that could be done -- that is what I am getting back to, using one or two proprietary systems, one developed by the persons putting forth this measure -- just bringing this out. That is pretty clear. This would generate a large market.

DR. BURSTIN: Just to be fair, what they are actually putting forward is -- There was an attachment as well and a link to their website. It is actually really a paper based chart reporting.

They are indicating they can collect this electronically using their system, but they are putting it forward as any other process measure which you need to go to the chart to collect the data, and currently we don't have reliability capabilities to this measure. It could only go forward as for time

limited endorsement, since the measure has not been tested.

We don't know, for example, how well that paper form performs. How often can you -- Just looking at the extra data here, how often can you find evidence of a sticky one, short term memory deficit, clearly indicated in the chart?

That is what I think the time-limited endorsement period is for, is to look toward that.

CO-CHAIR GAZELLE: I am with you, Steve, on the importance of pushing American medicine to get to this. I just think that, for us to vote to recommend for endorsement a measure where it can't be done now, is too early.

DR. FORMAN: It is relative. We have been dong PQRI, which is not dissimilar to this. For radiology PQRI has been a paper, completely paper based --

DR. SMITH-BINDMAN: Can you give

us an example?

DR. FORMAN: On our head CTs -- and I can get the exact measure; you all may remember it -- we have to put down the time the patient hit the emergency room and the time they did the study, and whether we documented it as an intracranial mass, hemorrhage or shift.

DR. D'ORSI: But what percentages of practices are participating in PQRI?

DR. FORMAN: Not a lot. I don't know. A minority.

DR. BURSTIN: But we can. Fifteen to 18 percent.

DR. CANTRILL: How about the concept of sampling? We haven't discussed that. Is that an acceptable approach here? So you are not doing 100 percent, but you are doing a specific sampling, and that gets away from some of your concerns.

I hate to see a good idea really turned off, because we don't think we can do

it. How can we maybe get this thing so it might be acceptable?

DR. RUCKER: Don Rucker. I think one of the challenges with this, and sort of follow the stuff above the neck, the neck and above as opposed to the knee and ankle and maybe heart. You know, it is sort of in the definition.

So, for example, a Glasgow Coma Scale of 14 is something where the person is potentially messed up and can't hold a job again. I understand it could go away tomorrow or later in the day or when they are sober, but if you came to me with a Glasgow Coma Scale of 14, there is some potential serious, life altering deficit there that, I think, in this particular thing -- again, this could be in the comment -- that needs to be shown.

Then when you get to --

DR. GRIFFEY: But, Don, if someone had a life altering injury like that, you would hope to have seen one of these other

elements there, and that is what those other studies addressed and bore out.

DR. RAKSIN: There are so many reasons that someone might be a 14, I mean, he might be an adult football player.

DR. RUCKER: I understand you can find counter-examples, but I am just saying that, when you have somebody who has a neuro-deficit, for whatever reason, I think -- and certainly in the emergency department setting, that is something you have to give some significant benefit of doubt to.

I think the other issue is severe headache without loss of consciousness or post-traumatic amnesia and severe headache. I mean, many of these people come in with severe headache, the number of worst headaches in their life. I mean, we all do.

DR. RAKSIN: That is a different measure.

DR. SETZEN: What about the person who hit his head walking down the street, hit

it on the side, didn't have loss of consciousness, and GCS is over -- you know, is normal. Those are the ones you are trying to get rid of. Right? All the BS. Right? So that is the value.

DR. RUCKER: I understand that, but I am just saying, if you have severe headache, this is a very judgmental -- It is a very judgmental standard.

DR. CANTRILL: But, Don, you know, we are not worrying about those. We are not even into the gray zone. They are the stuff that, you know, this shouldn't even see the inside of a department of radiology. You guys never seen those, right? Every day.

DR. FORMAN: No.

DR. GEMIGNANI: Actually, I would say that, contrary to the one that will be about the headaches tomorrow, this one at least has evidence, and it has got really good studies, better than others. It is hard to measure, which is the hard part of this, but

I would say that out of our options, this is one very obvious place that there is overuse, and that there is good evidence that there is overuse.

CO-CHAIR PETERSON: Any other comments?

DR. STILLMAN: I have a question which reflects my ignorance perhaps. How reliable do we think the Glasgow Score is in the medical record to be extracted or is it going to be in there in some other form? So if we have a cutoff for a metric, then we should be able to pull out a score and make sure that it is there.

DR. FORMAN: There are institutions who reliably document anybody below 15. So it is pretty reliable.

DR. GRIFFEY: If it is not there now, it would be when you went to get the measure or else it would probably fail the measure.

DR. RUCKER: It should be. It is

not like -- Take an Apgar score. It is really something that --

DR. STILLMAN: So anybody who walks in the emergency department with mild head trauma will have a Glasgow Score in the record?

DR. RUCKER: Yes.

DR. CANTRILL: As soon as this becomes part of a measure, it will be in the record.

DR. RUCKER: I think -- I am not sure about that, because I think a lot of times what is in the chart is the actual lesion, depending on how severe the thing is. You have an XYZ in the scan or you don't.

If you look at people who are hand scanned in these traumas now, all that -- I mean, there is sort of a crowd that is getting the major trauma. This is what I was getting at, the walkie-talkie crowd. I am not sure these people have Glasgow Coma Scores.

DR. FORMAN: They should. Look at

the nurse's notes.

DR. RAKSIN: They do. It is part of a primary trauma survey where a patient comes to the resuscitation -- Granted, if they are a walkie-talkie, they are a 14 or a 15, but that is part of what is documented for every patient that comes through the trauma center.

DR. RUCKER: Well, we are trying to improve the trauma center per se.

CO-CHAIR PETERSON: Mike, you had a comment.

MR. BACKUS: Yes. The only thing -- You know, we are in the radiology benefit management area, and we do outpatient preop, and every insurance plan comes to us and says, well, what are you going to do about the ED. What can you do about the ED?

You know, we have looked at it a lot, and from a straight preop perspective, there is not a ton that you can do. I completely agree that you will generate

Hawthorne Effect here by saying that you are going to look at it, and I agree with that completely as well.

I think that the really tough piece is, if I compare it to the breast stuff that we just talked about where you have kind of this mandatory BIRAD and the data is easily extractable -- you know, CMS's stuff is easily extractable out of the claims and everything.

I think I completely agree with the measure, and I have no issues or basis to have issues with the scientific judgment of them. The data collection is just so, so tough for me on this one.

If you are running a Medicalis or a Precipio or whatever, you can get it. I think, as a national body, that becomes very tough. To me, it is like an unfunded mandate.

You know, we want to be taken seriously in the provider community, and accepted; and to say, oh, we want you to do this and, by the way, all the ED physicians

got to work with a piece of paper now, and you got to fill this thing out; you are going to send it in, and we somehow going to get the stuff in Excel and pull it together, it becomes very expensive.

All that said, I would love to see progress made on the measure in some method, because what you are getting at -- and we have all made jokes about the ED -- I mean, the running one in our shop is that the door to the ED is not a set of bifolds; it is a tube.

So I am hugely in favor of the -- I am huge in favor of doing something down the road.

DR. MECHTLER: Without being selfish, I am very pleased we are not talking about mammograms.

My issue at this stage is that the Glasgow Coma Scale, among neurologists, is really a poor -- poorly associated with mild -- moderate and severe maybe more, but mild head trauma.

A couple of issues that I have is: (a) in the previous discussions we have looked at EDs, but then a comment was made that we may look at outpatient facilities. Let's be fair. Nobody does CT in outpatient facilities for mild head trauma. So the science has gone in a different direction.

We are looking at ERs or EDs that have 24/7 MRI right now. I am very interested, and I agree there is over-utilization of imaging in EDs and outside of EDs. The real question in my mind is, if we put these rules for CT, would you think, with mild head trauma, that the frequency of MRI may increase in emergency room 24/7 coverage?

The other issue may be that it has in outpatient. If this discussion here is going to not only represent for ED but will be at freestanding centers, hospital imaging centers off-campus, then I promise you that in our practice we actually have the largest neuroscience center in the country. We see

130,000 patients a year, and our CT numbers are decreasing with MR increasing, and we have both modalities within the facility.

So the reality is MRI in tomorrow's discussion for headache and mild head trauma -- I mean, that has to be on the table also, the evaluation and utilization of MRI and CT.

DR. CANTRILL: Steve Cantrill. I think you bring up a valid point, but I don't think it is our concern in the immediate future. I can get a head CT in 18 seconds. I can get a head MRI in 45 minutes. That is after I go through 27 different hoops.

So that is not going to happen very soon.

DR. MECHTLER: We have trauma protocols less than 15 minutes. We do.

DR. CANTRILL: Say 15 minutes, 15 seconds.

DR. MECHTLER: Of mild head trauma.

DR. BELLO: I think the other issue is -- Jacqueline Bello. I think the other issue is the monitoring through the study and the other CT scans in a trauma setting that that same patient is getting.

So we are here to discuss efficiency. Way before you start sending the patient to four different ZIP Codes, they are going to see CAT anyway for the chest. They get a CT of the head.

So I really think that we are stuck, like it or not, with a CT. I also really think that we bear the burden of having some sense of responsibility when it comes to the repeated radiation dose. Yes, this starts at 16; so we are not going to say the 10-year-olds, but I take an ER shift every month, and there are people who come in from nursing homes once a month, because they have fallen at the nursing home -- instant CT of the head and C-5, and these are patients who -- They are unchanged over 12 months, and hello,

Medicare, you know. I mean, these are there. They are not going to die of the radiation dose, but they are going to kill our medical system.

DR. SPENCER: But they get a scan if they are over 60.

DR. FIESINGER: Troy Fiesinger. Just a technical question. In the numerator it says mild traumatic brain injury, in the denominator nonpenetrating head injury. Are those equivalent terms or synonymous terms?

DR. BELLO: No.

DR. FIESINGER: Because it is a technical problem. It may be a minor one, but using two different terms -- We are arguing about definitions.

DR. BELLOW: No. It is an additional requirement. Once it is penetrating, it doesn't matter --

DR. FIESINGER: Right, but the language should be the same in the numerator and denominator and not different between the

two.

DR. GIBBONS: I think I feel totally ignorant in terms of this discussion of feasibility with respect to a couple of things, and maybe some of the people in the room can clarify this, which is: (1) the actual current level of penetration of electronic medical records into emergency rooms which, at least in our area of the country, is clearly lower than the rest of the medical system; (2) whether insurers have already tried to do something about this with respect to indications, and that might include CMS, which at least as I have asked questions over the years regarding chest pain, some of the things that are done in the outpatient sphere seem to be handled so differently administratively within emergency care that it is like a mystery to me.

So maybe other people in the room could shed light on that.

CO-CHAIR PETERSON: Clarify the --

What you are asking for the EHR is how many could do this measure?

DR. GIBBONS: Yes, or how many even have an EHR currently in --

CO-CHAIR PETERSON: And a CPOE system that has indications.

DR. GIBBONS: Yes. In an emergency room setting.

CO-CHAIR PETERSON: Less than 15 percent.

DR. FIESINGER: I think maybe 25 percent or something, but it is in that range, certainly not the vast majority.

CO-CHAIR PETERSON: Okay. That help?

DR. GIBBONS: Yes, that helps, but how about this issue of handling it from an insurer standpoint, and indications, because certainly, CMS tries to regulate indications for procedures in the outpatient sphere and denies payment. Is this something that insurers have tried to do already and, if so,

what happened?

CO-CHAIR GAZELLE: I can tell you our experience in the northeast is that, for the most part, they don't get into ED image.

DR. ZERZAN: And especially -- This is Judy from Medicaid -- there is no way to narrow with administrative data. There is certainly no way to narrow at point of contact.

The best we could do, I think, is similar to one of those CMS measures that is proposed to sort of find out what the rate of things are, and maybe in that way encourage people to change their rates, if they are an outlier. But that is super-blunt tool.

This is much more specific and evidence based, but there would be no way that we could collect that data, and if we asked our managed care providers to give us that data, what percent, they would run screaming and yelling at us, and say no.

You know, honestly, we pay crappy,

and we are certainly not paying for this additional thing that they would feel was burdensome, even though this is a huge problem of overuse.

CO-CHAIR PETERSON: Carl?

DR. D'ORSI: I just wanted to back up a little. We are creating a metric. What is a good event metric? One is ideal. So what is acceptable --

CO-CHAIR PETERSON: Can't hear you, Carl.

DR. D'ORSI: I'm sorry. We are creating a metric which, to me, means that it is a measure of something that is going to tell whether you are abusing it or not. So what is an abuse, and attached to that, what is the false negative rate or the true positive rate of doing a CT without these criteria?

Also, related to something a radiologist stated before, are we thinking of malpractice issues in this at all, or is that

excluded?

DR. CANTRILL: The guidelines -- Practice guidelines are practice guidelines, and malpractice is always a concern. I think the tort issue is less of an issue here than it is for some of the other measures that will come before us while we are here.

DR. SMITH-BINDMAN: I am not sure that I would agree with that. We have a paper on this topic exactly looking at mild traumatic brain injury in the Medicare population over time, and imaging is basically approaching 100 percent across the board.

DR. CANTRILL: I am not saying it is not an issue, but what I am saying is here you are trying to give guidance to decrease overuse, as opposed to just saying decrease over use with no guidance. So I think that is the difference.

I think the whole issue of tort concerns is something that this committee should think long and hard about, because why

do we overuse? Because we don't want to make a mistake or because we are lazy. There are a couple of reasons for that.

DR. SMITH-BINDMAN: Any other reason?

DR. CANTRILL: There's several, but we don't want to make a mistake in terms of our patients. So if we are going to be put in the position where the chance of making a mistake goes up, then we do need to worry about the tort issues. I think every practicing clinician is worried.

DR. D'ORSI: So what is a good and bad metric in this?

DR. CANTRILL: Well, here -- I don't know what -- I can't tell you what a good would be. Good would be probably close to, you know, above 90 percent, 95 percent. Who knows?

DR. SMITH-BINDMAN: This is Rebecca Smith-Bindman. I can't remember from the papers, but they are close in numbers.

What would the impact of this be on utilization in the setting of mild brain trauma? How much would this decrease imaging? So you would reduce a pretty common indication imaging by 40 percent, potentially.

DR. FIESINGER: We talk about demand side changing practice.

DR. SMITH-BINDMAN: This is big.

DR. FIESINGER: This is huge.

CO-CHAIR PETERSON: So in the interest of our developer, are there questions? We have our developer on the line. Dr. Schuur, are you on the line?

DR. SCHUUR: Yes. Jay Schuur calling from Boston. I am joined in the room by Ali Raja who is an emergency physician and works on evidence based imaging. Good afternoon.

CO-CHAIR PETERSON: Good afternoon. Were any things that you wanted to specifically address to us relative to the comments you have heard, and then afterwards

we will have a short Q&A for you from anybody on the panel who might other questions.

DR. SCHUUR: Sure. I think I will take just one minute and give you a brief background on the measure development process, and that should sort of apply to all four measures. Then we can both try to address a couple of the questions.

These four measures were developed primarily by four emergency physicians, none of whom have any financial interest in the Precipio system or any other decision support system, and have been vetted through providers in multiple fields at the Brigham and other Harvard hospitals.

We are practicing emergency physicians, and know that the evidence shows that there is widespread variation in the use of CT, that there is evidence that CT radiation exposure is high, driving high Medicare costs, and the use has gone up in the last 10 years.

So we looked for clinical indications where there were consensus evidence based guidelines primarily applicable to the emergency department, and then we developed measures for those indications. That is why we focused on these four areas.

All of the measures were set up with the same general construct, which is that the denominator would be the population getting a CT, and the numerator would be the patients who had received a CT who had an appropriate indication.

An alternate approach might be to define the population that had a traumatic brain injury, but as published literature has shown, ICD-9 codes and other administrative data are not reliable to define these populations.

So we set up the measures in that structure. We have also submitted them to be reported at the emergency department or facility level, not at the individual level,

because as we all know, guidelines are developed for populations, and we didn't want to put pressure on any individual clinician. We didn't think the evidence was strong enough to not order that one individual test.

We did think it would be very useful to know if one emergency department -- 80 percent of their scans were consistent with evidence based guidelines, and another ED 20 percent of their scans were in that form.

So let me just turn it over to Dr. Raja for a second, who works with the Center for Evidence Based Imaging, and he can describe the work that they have done from the published research.

DR. RAJA: I know that at least two or three of you are very familiar with our system here at the Brigham, since you guys have worked here in the past or you were with one of our partner institutions. So I won't belabor the point here. I have heard your discussions. I think they are right on.

It is very easy to do this kind of data gathering with our Precipio and Medicalis systems that we have here, but what we have been doing is we have been actually looking at how many of our CT scans have evidence based indications for them.

One of the most amazing things we have found is that there is such broad variation. Among the traumatic head CTs, we found variation, everywhere from five to 17 percent of patients specifically by emergency physicians.

So there is some sort of a need for some sort of a better practice to see if we can diminish this variation. I know you guys all agree with that in general concept. Now as far as making this happen in feasibility, what we are envisioning for emergency departments that weren't able to -- for the vast majority of emergency departments who aren't currently able to do this on a complete computerized fashion, a simple paper

form.

Dr. Schuur and I just e-mailed a paper form to you guys as well, but you can, I am sure, envision with, for example, a head CT for trauma a simple paper form with the indications that were outlined here requiring only a checkbox if they applied to that patient, which would then meet the criteria for the imaging efficiency guideline.

It wouldn't take that much more work for the emergency physician. It would allow for pretty good review of those scans that did actually meet these guidelines.

That is what we were actually going with this, but we would love to hear whatever other questions you guys have for us.

DR. SCHUUR: And just to address a couple of specific questions, I think there was a discussion around the GCS and some other questions on -- I think the discussion was around the traumatic brain injury measure.

The traumatic brain injury measure

is based on a consensus guideline that was developed by the American College of Emergency Physicians, and included a representation from multiple specialties and include both the evidence behind the Canadian head CT rules and what are called the New Orleans head CT rules, and a long discussion about which one of those is preferable, and there actually have been comparison studies. But in order to be inclusive, our measure would allow any indication from either of those two measures.

So this is really the broadest inclusion of accepted consensus evidence based standards that have been promulgated by the larger specialty society for emergency medicine.

CO-CHAIR PETERSON: Perfect. Questions at all for the measure developers?

DR. D'ORSI: Just one -- Oh, I'm sorry.

DR. SMITH-BINDMAN: No, no. Go ahead.

DR. D'ORSI: What was the gold standard for these ACEP finding? What did they find to say, wow, okay, it is worthwhile to do this to find hemorrhage trauma, and how often did they find hemorrhage trauma, and how often did they find it to say this was a valid indication?

DR. SCHUUR: Let me make sure I understood the question. What was the gold standard in these clinical studies for comparing to the CT?

DR. D'ORSI: In other words -- Yes, what did they find to say, yes, these are great --

DR. SCHUUR: So both of these studies used follow-up with either direct contact by telephone and/or review of medical record. Both were -- One was published in JAMA, the other one in the New England Journal, or actually in Lancet and the New England Journal, and they have been -- The Canadian study has been replicated with over

95 percent follow-up.

They are considered the gold standard of diagnostic test studies. So the difference between the two measures -- the New Orleans criteria, which were developed at Charity Hospital, used many CT significant findings on radiology; whereas, the Canadian gold standard outcome was any finding that would require a neurosurgical intervention.

Since there are things you will find on a CT, say a small subarachnoid hemorrhage, which do not end up requiring neurosurgical intervention, by definition the Canadian rules will use less scan -- will require less scan.

They have studied them head to head, and in the head to head study, actually,. the Canadian rule was as sensitive and more specific, but a lot of doctors in the United States use the New Orleans criteria because of their concern about medical legal liability associated with missing a

craniographically visible hemorrhage, such as small subarachnoid, even if it doesn't require any specific treatment.

DR. D'ORSI: thank you.

CO-CHAIR PETERSON: One other question?

DR. BELLO: Yes. This is Jacqueline Bellow. One of the points that came up in discussion earlier was wouldn't it be great to be able to sneak in there and see what is going on now in terms of this being -- these criteria being met and, therefore, you would have something to compare the measure to.

Did you do any preliminary snooping around before you instituted this that you could answer that question for us?

DR. SCHUUR: So I am going to turn it over to Dr. Raja, and he can address that. There is data on what the current variation is and they are now implementing these.

DR. RAJA: So right now we are

actually implementing these rules, and that is, obviously, ongoing.

What we have found is that at this point -- and again, we only have a few months worth of data where we have implemented this rule, but at this point we are looking at somewhere between a 60 to 80 percent compliance with one of these rules.

Now, obviously, as you know, as you guys have already discussed, there is the Hawthorne effect where, now we are asking people to click on a box, they may be clicking on a box that they wouldn't have necessarily have clicked on otherwise, but there seems to be somewhere 60 and 80 percent compliance with these rules.

DR. SCHUUR: But multiple published studies that are referenced in our application and also in the Canadian head CT rules in the literature show that in sharper views of current practice, there is a large gap between what is the number of scans --

around the country, the number of scans that are done without evidence based indication.

DR. RUCKER: Don Rucker. Three definitional questions. One, what is your operational question of loss of consciousness, because patients are often goofy on that.

The second is how do you distinguish severe headache from non-severe headache, because it was my experience patients sort of tend to say their headache is severe.

The third one on the numerator and on the denominator, I was wondering why choose the Glasgow Coma Score of under 14 as opposed to under 15?

DR. SCHUUR: Going by our standards, we are basing this on a consensus of a published evidence based guideline based on multiple, well done follow-up studies through the Canadian and the New Orleans Criteria, and those studies use clinicians' decision about loss of consciousness and

clinicians' decision about severe headache.

Although I agree that one could say that those are subjective, when actually studied with tens of thousands of patients, they have been shown to be highly sensitive.

DR. SMITH-BINDMAN: I have one question. This is Rebecca Smith-Bindman. The way you described just minutes ago this would be applied, you talked about all CTs, how many fit within some appropriateness criteria.

I want to understand it. Is this measure limited to a patient population defined at the point of referral from the emergency department as having mild traumatic brain injury or is it meant to be applied from a point of view of all CTs that are done, and how many fall within an appropriateness criteria?

So one of those you could use decision support software or entry from the radiology point of view to get at. The other, you would have to do from the ED point of

view.

DR. SCHUUR: It is my understanding that all the CT scans that get reimbursed require a physician's order. so that would be the way that we implement -- constructed the measure to occur for all CTs. So it is based on -- If you look at the documentation, the denominator statement, the number of adult patients undergoing head CT for trauma who present within 24 hours of a nonpenetrating head injury with a Glasgow Coma Score greater than or equal to 14.

There are then five denominator inclusion criteria, and there are a set of exclusions that define who would not be included in the measure.

DR. SMITH-BINDMAN: So my question is: The data form that you have provided to us or that we just got by e-mail is creating a cohort and denominator from the point of view of the emergency room, and creating that cohort based on mild traumatic brain injury.

The way you have described the measure right now is defined from the radiology database point of view, where I am not sure if that information on trauma, mild traumatic injury would necessarily be included in those data.

So you might understand vomiting or severe headache, but you wouldn't know if that was a patient who was post-stroke or post-trauma. You are describing it from a CT point of view. The data that we have just been sent is from the ED point of view. How is the cohort defined, and how do you define it?

I can easily imagine applying it from the radiology point of view, but we couldn't get the cohort on trauma defined.

DR. SCHUUR: Well, I think there are two questions. One is how do you define the cohort, which is think is very explicitly defined in the measure. The second is how do you collect the data.

That would depend on what hospital, what system the hospital would have and would want to implement. If a hospital has an EMR with physician entry, this could be programmed into the radiology ordering platform.

If they did not have that or they did not want to use that, they could make up a paper form that applied every time a head CT was ordered and have the exclusions and then the inclusions, and it would be a simple check process.

CO-CHAIR PETERSON: I think that answers it. Other questions?

DR. MECHTLER: I have a question. Laszlo Mechtler, neurologist. Your category of patients with head injury, no loss of consciousness, no post-traumatic amnesia who have a severe headache and nausea, you have just described a post-traumatic migraine.

So are you saying that every post-traumatic migraine should have a CT? These

type of headaches are very common, especially if you have a previous migraine history. I think Donald alluded to that, and many of these patients of head injury have also whiplash injuries. So many of them come with cervicogenic headaches.

Are you presuming that a cervicogenic headache or so called acute or episodic tension type headache or a post-traumatic migraine -- are these individuals, by your measures, your numerators, these individuals will be getting CT scans, and are you concerned that the frequency of CT scan, in fact, may increase in that subset of that population, and should you define headache somewhat more specifically than just saying severe headache?

DR. SCHUUR: I think these are good questions. Again, the numerator details are not based on something that we sat around and made up. This comes from the evidence based consensus guideline published by the

American College of Emergency Physicians, and their evidence based consensus guidelines were based on those two large studies, and all of those terms were what were used in those studies.

It is possible that someone with a post-traumatic migraine would meet these criteria. The clinical question that is presented to the emergency physician is does this patient in front of me who has a mild traumatic brain injury and a headache require scanning?

That is the question that the guidelines attempt to address. So whether -- They may ultimately have a migraine, but that is the clinical question people are addressing and what the clinical decision rules have been addressed for.

It is very unlikely that these measures would increase imaging, because what they are going to do is they are going to measure patients who received an image and

said whether or not it was appropriate. They are not setting up a population with a diagnosis and saying you didn't get an appropriate scan.

So everyone who is in this population already has had a scan. The only way you will look worse is by ordering scans on patients -- or your institution ordering scans on patients without indications.

DR. GRIFFEY: This is Rich Griffey. Jay, you may have heard Howard say that I like this measure. I think it is a good measure, and the Achilles heel of this measure may be the feasibility component in terms of reporting.

It is great to have a paper form, but a number of people have brought up that, well, then we've got to do something with those forms or you have to have someone to enter that data, and it is sort of an unfunded mandate, a lot like the pneumonia measures, for example. That is all chart extraction in

a similar way.

Do you have any thought about how to get around that or how to make that simpler? I know you talked about sampling. If you did that, you would want to make sure you had a denominator, so that not just the good papers or the compliant studies were filled out. Do you have any thoughts about that?

DR. SCHUUR: I may refer to Dr. Raja the technical aspect.

DR. RAJA: Dr. Griffey, that is a great point. This is, obviously, an unfunded mandate. It would take a lot -- It would take some time. It would take somebody to actually collect the data. It would take somebody to actually go through and measure it.

I guess our biggest overarching point is simply that this is somewhere that we need to move toward, and I think this is a first step. If we can figure out a better way to do this that would take less man-hours or

if we more widely implement electronic physician order entry, that would be great, and it would make this a lot easier. But to get things started, it takes a paper form, and that actually pushes people to spend money on electronic order entry systems rather than having to fund somebody to go through and collect forms, great, because that is where we want to go.

Unfortunately, you are absolutely right. We don't know how to get this funded, but I think we all agree that this is where we want to go.

DR. SCHUUR: The second point I would make is that I don't think the term unfunded mandate is correct, because the facility and the reviewing physician are both getting well compensated for each of these scans. So the time and effort to properly document indications doesn't seem onerous.

The second comment is that, like the pneumonia measures and other core

measures, I think sampling would be very appropriate for facilities that could not easily collect data on all of them, and CMS has well validated sampling numbers and what would be appropriate.

CO-CHAIR PETERSON: Helen.

DR. BURSTIN: Just a couple of points of information. This is Helen Burstin. Hi, Jay.

So I just want to point out that this measure would only go forward for time limited endorsement. I just want to emphasize that again. NQF has endorsed numerous measures based on medical records. I don't want this to seem as if it is a real aberration.

Oftentimes in new areas, the first thing that happens is a medical record based measure. It gets tested. There may be other feasible ways to follow it, but I just don't want it to seem like this is actually all that different than the majority of core measures

we require hospitals to do, which are all paper based at the moment.

So I guess a major question for Jay is I just want to understand that. If it is time limited, do you have a plan and the capacity to test it within 2 months and report back to NQF?

DR. SCHUUR: Absolutely. We are actually doing that right now.

DR. BURSTIN: Just one last comment. You know, if there is anything we hear a cry for, particularly -- and this committee doesn't have as many consumers and purchasers on it; one is out sick, and we have a limited number at the table on Medicaid. It is for overuse measures.

So I think this is where those four criteria are intended. They are not weighted. They are not do one versus another. You have to make an overall assessment of how you think those four play out.

Feasibility is a concern, but you

have to weigh it against the other things.

CO-CHAIR PETERSON: Any final comments?

DR. RUCKER: Is there a worry that the studies -- you know, about the gaming in terms of the severe headache versus headache, because I think it is a different crowd when the study researchers who are motivated in these big studies to prove the point that we don't need the image is sort of a very different dynamic than ER docs who are ordering these studies for some intrinsic reason, presumably since they are actually not paying to get radiology studies, contrary to what was mentioned, who might just say, well, it is a severe headache; because that is sort of what the patients typically say in this.

You know, I hate to harp on this, but that is -- It is the severity of this nebulous symptom that is the big clinical concern when you are seeing these people. It is that sort of subtle judgment, I think.

DR. SCHUUR: I would strongly recommend that, if people have questions about this measure, that they review the original studies from the Canadian and/or the New Orleans Head CT rule.

The way that those studies and well designed diagnostic tests on decision rules are designed, the clinicians were not pressured to do anything.

They just had an order form, and they implemented this in a number of emergency departments and basically said do what you would normally do, and then after a period of time, they compared what was on order forms to patients' eventual outcomes, and using regression and sorting statistical techniques, they figure out which indications have the most association with the outcome.

CO-CHAIR PETERSON: Okay. I think, in the interest of time, we are going to -- Thank you very much for your effort of answering our questions and for putting forth

this measure.

Helen, I think in the interest of time -- we are beyond the hour. I assume we will hold votes until tomorrow. Do you want to vote tonight?

DR. BURSTIN: Let's finish up.

CO-CHAIR PETERSON: I am all for voting. I don't want to short-change, if thee are questions.

DR. SMITH-BINDMAN: Before we vote, can the people who read it carefully sort of give us a summary of their review?

DR. FORMAN: I would just say, from my point of view, the only issue that is really a question -- I am not that concerned about people dealing with this anymore than anything else, and I think that goes on.

The fact that you might have five percent gaming and still get rid of 25 percent of excessive imaging, I think, is acceptable to me. So that doesn't concern me much.

The only part of this that I think

raises any real concern is the feasibility. You know, I am speaking from an institution not dissimilar from the Brigham, but without the computerized physician order entry piece in place, and I think it will be difficult to implement for even us. I think it becomes that much more difficult at other levels.

I do agree that the form that they are presenting is so simple that you could plot this data, and it is such a high dollar item that it should motivate practice change.

CO-CHAIR GAZELLE: If you look at that paper form, I can't imagine ordering a head CT for mild traumatic brain injury and not circling at least one of those indications. You are getting it 100 percent.

DR. FORMAN: You know, I disagree.

MR. BACKUS: No. You might get 100 percent of people that, when they say -- You say you can't imagine ordering it and not circling one of those. But the question is: Can you not order it, because then you really

look down on that list and go like, ah, there is nothing really here for me.

DR. GRIFFEY: That is why it is time limited, and that is why you will learn what you learn, I would think.

DR. ZERZAN: Prior authorization, you don't really -- Especially in Medicaid, if you fill out a prior authorization form, we pretty much approve it, but the part where you say is that barrier to get there, and I think that this is exactly that same thing.

You will probably approve everyone that fills out the form, but there will be some statement that you have avoided, and that is what you are looking for.

MR. BACKUS: You are just bringing that thought to top of mind. That is all that form does. It just brings that score to top of mind before you order the CT, and that is all you can hope for.

CO-CHAIR PETERSON: So, Helen, just a point of clarification. Time limited

data that you would require -- Clarify for the committee here what that really means.

DR. BURSTIN: Right, and it is spelled out in the form. Essentially, what it means is you guys agree this measure would pass all the other NQF evaluation criteria with the exception of the fact that it has not been tested.

They would need to go back and test whatever form the measure is going to be used in, in this instance the paper form, maybe to look to see how reliably they could collect the individual data elements, whether the reliability is tested, probably in this instance whether they have an electronic system. It would be particularly interesting to understand if, in fact, the results are similar between the electronic system and paper record.

That should, at the end of the day, allow enough to say can you validly and reliably collect this data; and given the

feasibility concerns, I would hope they would also give us some information about how difficult it is to collect.

CO-CHAIR PETERSON: I am just curious about what would be considered a reasonable test of this? Can this be one institution?

DR. BURSTIN: No. It cannot be one institution. There is actually specific guidance. It depends on the kind of measure. It is probably at least five to 10 institutions or a certain number of patients.

It really depends on the level of analysis of the measure. So we will need to take a look.

DR. SMITH-BINDMAN: So they have to test it?

DR. BURSTIN: They have to test it.

DR. SMITH-BINDMAN: They have to test this measure on 10 institutions?

DR. BURSTIN: I can't remember the

exact protocol, but whatever the protocol is they need to undertake efforts to test the measure, provide information back on reliability and validity, or the measure isn't endorsed. So that is the issue.

That is the fail safe for measures like this, if you think it otherwise meets all the criteria. We just don't know how well it is going to perform in the real world on paper, since not everybody is like The Brigham or other places like that.

DR. RUCKER: I had a question. It wasn't clear to me that they were actually going to do a multi-site study on that. I don't know if that is a question to them or somebody else.

DR. BURSTIN: They understand the requirements for time limited.

DR. RUCKER: So they know that that is sort of part and parcel of what --

DR. BURSTIN: We will give them further --

DR. RUCKER: It would need to be a place that don't have computerized ordering. Right?

DR. BURSTIN: They are going to need to test the paper form, if that is what they are arguing is the dominant mode of collection.

CO-CHAIR PETERSON: Roger.

DR. SNOW: Yes. Multiple sites, but do they have to be outside of the same network or could they be within the network?

DR. BURSTIN: They could be within the network. Again --

DR. SNOW: I know it is a detail, but I just raise the question.

DR. D'ORSI: do they have to have any discussions about what they produce, what it does, that number? Is there any discussion that it is useful in any way or just proving that it can be done?

DR. BURSTIN: At this point, you should be making the assessment that you think

it is already useful, usable. I think --

DR. D'ORSI: Okay. So my number is 8 -- .8. Why would it not be like aspirin?

DR. BURSTIN: This is proportion of CTs for mild traumatic brain injury that meets some guideline. You would like it to be fifty.

DR. D'ORSI: Oh, no.

CO-CHAIR GAZELLE: Can I just say quickly, I think -- You know, Helen, you said that some measures rely on chart abstraction. I think there is a very big difference between going through the medical record to see whether or not these criteria are met, versus forcing someone to fill out a form where the only things they can check off are the criteria that is needed.

I think, for this to be a useful measure, the paper form is not enough. You have to do the review of the medical record, either manually or using the MR, because in my opinion this form is just not acceptable.

DR. GRIFFEY: Why? Because you think that it is going to be garbage in, garbage out?

CO-CHAIR GAZELLE: Yes. You are asking some intern in the emergency department, while the patient is on the way to the head CT, to fill this thing out. They are just going to check the --

DR. SMITH-BINDMAN: These are your choices of why you ordered that scan.

DR. GRIFFEY: Well, that may be the case. The proof is in the pudding with the utilization data.

CO-CHAIR GAZELLE: But we are not going to be tracking utilization. We are only

DR. GRIFFEY: But computerized tracking the percentage of the head CTs that have the ACS criteria. So I would argue that you either have to do it by looking at the medical record to show that it has been documented as opposed to a paper form filled out, or EHR. I just think this is absolutely

not acceptable.

DR. CANTRILL: To mis-fill out this form, we call that lying. No, but his question is how do you get the denominator?

CO-CHAIR PETERSON: They have yet to produce evidence. They are getting it now, but they have yet to produce evidence to say we influenced the system and utilization of this test goes down.

DR. BURSTIN: And that is why I am just trying to get at the denominator.

CO-CHAIR PETERSON: There is not multiple studies that say that, if we have a system that has to check this box, it will reduce the number of ordered tests. There is 30 percent of tests that don't meet this criteria under current --

DR. GRIFFEY: But computerized decision support tools outperform education or Physician Champion or CME or any other intervention you have. This is the best thing you have. Now they won't all be computerized.

There will be a paper, a piece of paper, but it is --

CO-CHAIR PETERSON: Clarifying where we are. Okay.

DR. GIBBONS: I understand the points that have been made, but I would just point out that there is a fair literature that just -- as we have pointed out, if you audit something, it will get better.

DR. SPENCER: the one thing that is going to come up again tomorrow and tomorrow about the NQF stuff is the feasibility stuff. Again, what I don't understand is we don't make people follow these things, and there are several things we are going to look at that are just exceptionally clear that are overused in the scientific literature, that there are exceptionally clear criteria for what these should be.

We are going to see lots of those type of things, another easy-easy, and then we

hit this feasibility thing, and we get stuck. So what is wrong with saying that these are just good and right things, and -- The accreditation on a payer say, hey and, you know, NQF says these are important; you start reporting these or we are not going to credit your ER or we are not going to pay for these. Then people have to do them.

What is our obligation to say that it is a really easy thing to do or not? That is what I am struggling with, because this is -- No one argues about this. These are exceptionally well ordered, and they are unbelievably good criteria for when they should be ordered.

This is like one of the best things of all the things we have done here that is supported with literature, but we are stuck on what a pain in the rear it is to do. But nobody has to do it. Right? There is no Federal thing that says everybody must follow the NQF or CMS does it or Wellpoint does it or

somebody says we got to do it.

DR. BURSTIN: NQF does not implement the measure.

DR. SPENCER: Right. That is why I am stuck on feasibility with a lot of our measures.

DR. BURSTIN: If it is appropriate, the public supports it.

CO-CHAIR PETERSON: Shall we get to the vote?

DR. SPENCER: So does feasibility kill the deal? Well, we will find out. We will find out in a few minutes here.

CO-CHAIR PETERSON: Any other -- I think people have stated pretty clearly where they stand. Okay. Can we call for the vote? We'll go through the criteria. I know how the first scores will go.

I guess there will be 19 voting. Right?

DR. BURSTIN: Yes. We lost one.

CO-CHAIR GAZELLE: Oh, she gave me

her proxy vote.

DR. SPENCER: No. She gave it to me.

CO-CHAIR PETERSON: Okay. How many think the importance rating is High?

MR. CORBRIDGE: I've got two laptops. I can't really stand up.

DR. BURSTIN: I can do that. Eighteen.

CO-CHAIR PETERSON: Moderate? Okay. Low?

MR. CORBRIDGE: Moderate was one?

CO-CHAIR PETERSON: Yes. Okay, now we are to scientific acceptability. Okay, High? Moderate? Three. Low?

Okay, usability: How many say High? That would be a zero. Moderate? And Low? One.

Okay, feasibility: High? Moderate? Low? Okay.

We have the yes or no. So let's do Yes?

MR. CORBRIDGE: Before we do that, we have to open up -- Sorry. Just to make sure, is anyone on the line for public comment? Okay.

DR. SCHUUR: Yes. Record my vote.

CO-CHAIR PETERSON: The vote on this is Yes?

DR. BURSTIN: Sixteen.

CO-CHAIR PETERSON: No? Two, three.

DR. BURSTIN: Three, okay.

MR. CORBRIDGE: It is 15.

DR. BURSTIN: Sixteen and three.

DR. MECHTLER: Could we add comments, too, that can be added even to the vote?

DR. BURSTIN: Sure. Anything you want to recommend.

DR. MECHTLER: As I mentioned, I think this should be -- I would not like to see this presented for headache centers around the country. It would not make sense for

urgent care centers and even probably a fusion labs that deal with headache.

So if this is ED, that will probably be --

DR. BURSTIN: This is just ED.

DR. D'ORSI: This is acute trauma. This is for time limited.

DR. BURSTIN: It is time limited.

CO-CHAIR PETERSON: Any other comments?

DR. FIESINGER: I like the comments that at least we would have a paper system. What about testing that?

CO-CHAIR PETERSON: For the morning, everybody okay starting at nine?

(Whereupon, the foregoing matter went off the record at 5:23 p.m.)

G

O