NATIONAL QUALITY FORUM

+ + + + +

SURGERY PHASE 3
STANDING COMMITTEE

+ + + + +

TUESDAY
AUGUST 16, 2016

+ + + + +

The Standing Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Lee Fleisher and William Gunnar, Co-Chairs, presiding.

PRESENT:
LEE FLEISHER, MD, Co-Chair; Professor and Chair of Anesthesiology, University of Pennsylvania/American Society of Anesthesiologists
WILLIAM GUNNAR, MD, JD, Co-Chair; Director, National Surgery Program Office, Veterans Health Administration
KARL BILIMORIA, MD, MS, Director, Surgical Outcomes & QI Center; Vice Chair for Quality, Northwestern University and Northwestern Medicine
ROBERT CIMA, MD, MA, Professor of Surgery, Mayo Clinic*
RICHARD DUTTON, MD, MBA, Chief Quality Officer, United States Anesthesia Partners
ELISABETH EREKSON, MD, MPH, Dartmouth Hitchcock Medical Center
FREDERICK GROVER, MD, Professor of Cardiothoracic Surgery, University of Colorado School of Medicine
JOHN HANDY, MD, Thoracic Surgeon, American College of Chest Physicians

CLIFFORD KO, MD, MS, MSHS, FACS, Director, Division of Research and Optimal Patient Care, American College of Surgeons/Professor of Surgery, Department of Surgery, UCLA School of Medicine, American College of Surgeons/UCLA School of Medicine
BARBARA LEVY, MD, FACOG, FACS, Vice President, Health Policy, American College of Obstetricians and Gynecologists
BARRY MARKMAN, MD, Senior Medical Director Medicaid, Aetna
KELSEY MCCARTY, MS, MBA, Senior Manager, Quality and Safety Program, Department of Anesthesia, Massachusetts General Hospital
LAWRENCE MOSS, MD, Surgeon-in-Chief, Nationwide Children's Hospital
AMY MOYER, Manager of Value Measurement, The Alliance
KEITH OLSEN, PharmD, FCCP, FCCM, Professor and Dean, College of Pharmacy, University of Arkansas for Medical Sciences
COLLETTE PITZEN, RN, BSN, CPHQ, Clinical Measure Development, Minnesota Community Measurement
LYNN REEDE, DNP, MBA, CRNA, Senior Director, Professional Practice, American Association of Nurse Anesthesiologists
CHRISTOPHER SAIGAL, MD, MPH, Professor, UCLA
SALVATORE T. SCALI, MD, Assistant Professor of Vascular Surgery, University of Florida-Gainesville
ALLAN SIPERSTEIN, MD, Chairman Endocrine Surgery, Cleveland Clinic
LARISSA TEMPLE, MD, Colorectal Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center
BARBEE WHITAKER, PhD, Director, American Association of Blood Banks*
A.J. YATES, MD, Associate Professor and Vice Chairman for Quality Management, Department of Orthopedic Surgery, University of Pittsburgh Medical Center

NQF STAFF:

HELEN BURSTIN, MD, MPH, Chief Scientific Officer
ANN HAMMERSMITH, JD, General Counsel
ELISA MUNTHALI, MPH, Vice President, Quality Measurement
MARCIA WILSON, PhD, MBA, Senior Vice President, Quality Measurement
KAREN JOHNSON, MS, Senior Director
MELINDA MURPHY, RN, MS, Senior Director
CHRISTY SKIPPER, MS, Project Manager
KATHRYN STREETER, Senior Project Manager

ALSO PRESENT:

SOPHIA CHEN, MD, Centers for Medicare & Medicaid Services*
TRACI CONNOLLY, American College of Cardiology
SHERYL M. DAVIES, MS, Stanford University*
KAREN DORSEY, MD, PhD, Yale-CORE
YAZAN DUWAYRI, MD, Society for Vascular Surgery
LILIANA GOUMNEROVA, MD, FRCSC, Boston Children's Hospital*
BRUCE GRAY, DO, FSVM, FSCAI, Greenville Health System
LEANNE HAHN, MD, Centers for Medicare & Medicaid Services*
JEPH HERRIN, PhD, Yale-CORE*
SARAH JERNIGAN, MD, MPH, University of Miami Health System*
BRAD JOHNSON, MD, FACS, Society for Vascular Surgery
PAMELA L. OWENS, PhD, Agency for Healthcare Research and Quality*
LISA G. SUTER, MD, Yale-CORE
GARTH H. UTTER, MD, MSc, UC Davis Health System*
PAT ZRELAK, PhD, CNRN, CNAA, BC, UC Davis Health System*

*** present by teleconference**

A G E N D A

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

8:38 a.m.

CO-CHAIR FLEISHER: Are we recording? I guess we'll get started. Welcome to the third iteration of the Surgery Standing Committee. As you may remember, I'm Lee Fleisher from the University of Pennsylvania.

As you may remember, one of the advantages of standing committees is our ability to actually monitor and create a model over time in a particular domain expertise. So it's great to have you all back. And that's where I'll end the brief welcome.

CO-CHAIR GUNNAR: So I'm Bill Gunnar. I'm National Director of Surgery for the Department of Veterans Affairs. Most of us are familiar faces. Do you want to do introductions around?

(Off the record comment.)

CO-CHAIR GUNNAR: Okay. That will be fine. Welcome. We like to be efficient and competent, right. That's our plan.

MS. SKIPPER: Good morning, everyone. My name is Christy Skipper. I'm the Project Manager for Surgery. And it's nice to see you all in person and not just read your words back over email.

MS. STREETER: Hi. Good morning. I'm Katie Streeter. I'm a Senior Project Manager here at NQF on the Surgery Team.

MS. MURPHY: And I'm Melinda Murphy. I'm a Senior Director at NQF. I've been here for about ten years. Many of you, I have seen in multiple different meetings and activities and so very much appreciate the fact that you continue with the high interest and high energy in working with this.

Today we've got three new members, Dr. Scali. Did I say it properly? Okay. Dr. Bilimoria, there, all the way at the very end. And Dr. Whitaker may or may not be on line today, but she will be here tomorrow.

MEMBER WHITAKER: I am online. Thank you.

MS. MURPHY: Oh great. Nice to have you, Barbee.

MEMBER WHITAKER: Thank you.

MS. MURPHY: So welcome all of you and thank you so much for continuing with the interest and the energy.

Did you want to say something?

DR. BURSTIN: I just want to say good morning to everybody and thank you for joining. Helen Burstin. I'm the Chief Scientific Officer. Nice to see so many of you again. I was just talking that I think this is your third rodeo for many of you and we appreciate you coming back.

It really does help when we have these committees give a sense of trust across the table. But also our process doesn't seem quite as opaque as it might to some newbees. I know the new folks will jump right in.

I'll be popping in and out. But again you're in great hands. Thank you so much again.

MS. JOHNSON: Good morning. I'm Karen Johnson. I'm one of the Senior Directors here at

NQF and I get to serve as our measure methodologist most of the time.

MS. MURPHY: So with that, I think we're ready to have introductions, disclosure of interest. We have Ann Hammersmith who is the counsel for NQF and always a pleasure to have Ann.

MS. HAMMERSMITH: Thank you, Melinda. That is rarely said about a lawyer. So I appreciate that.

I see some familiar faces, too. So I think many of you have done this before. For disclosures of interest, I'll give you a few reminders and then we'll go around the table and you can introduce yourselves and disclose.

This is a CDP committee. So we look at your professional activities in detail before we seat you on the Committee. We do the oral disclosures in the spirit of transparency so that you have the opportunity to let the public, let the staff, let your fellow Committee members know about your background.

I want to emphasize that just because you disclose does not mean that you have a conflict. You may have engaged in some activity that you believe is relevant, but it is not a conflict. And we do ask you to disclose that.

We're especially interested in research activities, grants and consulting. We want to remind you that you sit on the Committee as an individual. You don't represent your organization. You don't represent anyone who may have nominated you for service on this Committee.

And with that, we'll go around the table. Tell us who you are, who you are with and if you have anything you wish to disclose. And I would start with the Co-Chairs. It's the only perk they get as Co-Chairs. They get to start the disclosure.

CO-CHAIR GUNNAR: This is Bill Gunnar. I have no disclosures.

CO-CHAIR FLEISHER: Lee Fleisher. I work for the University of Pennsylvania. I am a member officially of the American Society of

Anesthesiologists' Committee on Performance and Outcome Measures, although I have not participated for approximately one year.

I am a consultant to Yale CORE in the development of the hospital-wide mortality measure. And I am a co-investigator on two grants from the NIH looking at performance measurement with Jeff Silber from Children's Hospital of Philadelphia.

MEMBER HANDY: My name is John Handy. I'm a thoracic surgeon from Portland, Oregon. And I'm the Director of Thoracic Surgery and Thoracic Oncology at the American College of Chest Physicians.

MEMBER GROVER: I'm Fred Grover. I'm a cardiothoracic surgeon from the University of Colorado, a past president of the STS. Currently, I'm not working on any of the database committees there. I serve as Vice Chair of the transcatheter valve registry. We work with the STS and the American College of Cardiology.

MEMBER KO: Hi. I'm Clifford Ko. I

work at UCLA in the Department of Surgery. I'm a colorectal surgeon. I also am the Director of the Division of Research and Optimal Patient Care at the American College of Surgeons which houses all the quality programs including NSQIP, the Bariatric Program, Cancer, Trauma and Peds. So the disclosure for this meeting is that there are a number of measures that are either developed with some of these registries or programs and another one that uses the Peds NSQIP Program.

MEMBER SIPERSTEIN: Allan Siperstein. I'm with the Cleveland Clinic. I do endocrine surgery. I serve on the American College of Surgeons Performance Measure Committee.

MEMBER YATES: Adolph Yates. I'm in the Department of Orthopedic Surgery at the University of Pittsburgh where I'm the Vice Chairman for Quality Management and I'm also serving as the chief of our hospital in orthopedics.

I'm also the Evidence-Based Medicine Chairman for the American Association of Hip and

Knee Surgeons. And I have served unpaid on technical expert panels with both Acumen and Yale. But none of the measures today are of note. And I will also say that those are unrelated. Thank you.

MEMBER MOSS: Morning. My name is Larry Moss. I'm a Pediatric Surgeon at Nationwide Children's Hospital in the Ohio State University. I am on the steering committee for Pediatric NSQIP as well as the Children's Surgery Verification Program in the American College of Surgeons.

I serve for the Children's Hospital Association on the Quality Measure and Standards Committee that's developing measures, but none are under submission at this time. And also I'm President-Elect and Chair of a group in the organization, the Children's Hospital Surgeons in Chief that's developing quality measures for submission but none currently submitted.

MEMBER TEMPLE: I'm Larissa Temple. I'm a colorectal surgeon in Memorial Sloan-

Kettering Cancer Center in New York. I'm the Vice Deputy Physician and Chief for Quality there.

My research is in patient-reported outcomes and I'm doing some work with the college on working towards building that out. But I don't have any conflicts.

MEMBER MCCARTY: My name is Kelsey McCarty. I'm the Director of Operations and Strategy for Cardiology and Endocrinology at Boston Medical Center. And I have nothing to disclose.

MEMBER OLSEN: I'm Keith Olsen, Dean and Professor at the University of Arkansas for Medical Sciences, College of Pharmacy representing the American Society of Health System Pharmacists.

I do sit on the Board of Regents for the Society of Critical Care Medicine, American College of Critical Care Medicine. And we do review all the guidelines for the Society.

MEMBER SAIGAL: I'm Chris Saigal. I'm

a urologist at UCLA. To disclose, I worked with AUA in a variety of ways in the past. I've been on the Quality Improvement Patient Safety Committee. I'm on the Data Committee right now which looks at how to use data to improve care. I'm also, in our department, one of our QA participants.

MEMBER BILIMORIA: Hi. Karl Bilimoria. I'm a Surgical Oncologist in Northwestern. I run a research center that's funding in quality measurement and quality improvement programs. And I do a variety of projects with the American College of Surgeons.

MEMBER DUTTON: Rick Dutton. I'm a former Chief Quality Officer for the American Society of Anesthesiologists. Now Chief Quality Officer for USAP, a large group practice. I do work with Yale and CMS on a number of technical expert panels, but I have no measures here today.

MEMBER MOYER: I'm Amy Moyer. I'm the Manager of Value Measurement for the Alliance. We are a healthcare purchasing cooperative, a

not-for-profit, largely in Wisconsin, Illinois and Iowa.

We use several of the measures that are under consideration today in our programs, but I have no other conflicts to declare.

MEMBER PITZEN: Collette Pitzen, Minnesota Community Measurement. I'm a measure developer. We have no measures in the surgery portfolio. I also serve as the consultant to ANA Yale CORE TEP group and the PCORI grant with the American Society of Clinical Oncology.

MEMBER MARKMAN: Barry Markman. I'm a Senior Corporate Medical Director for Medicaid for Aetna. I do have some research projects with MTF, the Musculoskeletal Transplant Foundation. Otherwise I don't see any conflicts with any of the measures today.

MEMBER EREKSON: Hi, Liz Erekson. I'm a gynecologist. I work at the Geisel School of Medicine at Dartmouth and also with the Dartmouth Institute for Health Care Delivery Science.

I have done a lot of work with the

American Urogyn Society and currently serve as the National Advisor to their Outcomes Research Network. But I have no conflicts of interest for today.

MEMBER LEVY: Good morning. I'm Barbara Levy. I'm the Vice President for Health Policy at the American College of Obstetricians and Gynecologists. I do serve on the PCPI Executive Board, but have no conflicts in that I don't participate in any of the measures or measure development.

MEMBER SCALI: Good morning. My name is Sal Scali. I'm a vascular surgeon at the University of Florida. I also serve as the Chair of the Endovascular Aortic Aneurysm Committee for the Vascular Quality Initiative.

And I also serve on several other standing committees through the VQI and SVS on quality. But I have no other disclosures or have not been a participant in any of the measures that are being discussed today.

MS. HAMMERSMITH: Okay. Thank you.

Now I'll call on the people on the phone. Robert Cima?

MEMBER CIMA: Yes, this is Bob Cima. I'm at the Mayo Clinic in Rochester. I serve as the Surgical Quality Officer and I serve on the ACS measurement committee, much like Allan. I'm the American Society of Colorectal Surgeons Quality group, but we have no measures on this. I'm not a very big fan of Delta right now.

(Laughter.)

MS. HAMMERSMITH: Sorry to hear that. Mark Jarrett. Is Mark Jarrett on the line?

(No response.)

Melissa Thomason.

(No response.)

Barbee Whitaker.

MEMBER WHITAKER: Hi. I'm Barbee Whitaker. I'm Senior Director of Research and the AABB Center for Patient Safety at the American Association for Blood Banks. And I have no conflicts of interest.

MS. HAMMERSMITH: Okay. Thank you,

everyone, for those disclosures. Before I leave, I want to remind you of a few things, and then I'll ask you if you have any comments or questions.

The big reminder is that in order to make a conflict of interest process work, we rely on the Committee members. Everybody has a part in this. NQF does. The public does. And so do the Committee members.

So if you're sitting in a meeting and you think that you may have a conflict or that one of your fellow Committee members may have a conflict, please speak up in real-time. What we don't want is to get six months down the road and have someone say well, you know. Actually, I think I may have had a conflict. We want you to tell us now.

If you'd like to bring it up in the meeting, you're welcome to do that. If you'd like to approach your Co-Chairs, you can do that, or you can connect with NQF staff and they will talk to the Co-Chairs.

Any questions or comments regarding anything that's been disclosed today?

(No response.)

Okay. Thank you.

MS. MURPHY: Unless anyone has anything they would like to comment about right now, we'll get started with a little bit of an overview of the process and an introduction to the project which won't be new to any of you. So we'll move through it pretty rapidly.

As you know and as many of you have participated in multiple rounds of the surgery project, there are over 100 measures of which about 64 of them are assigned to this Committee. At this point, you have seen, or will in these two days, be seeing the last of all of them. So you can see there you're going to be looking obviously at all kinds of surgery care, perioperative, adverse outcomes, other related surgical topics.

And what you have there is just a list of the maintenance measures that are under

review. You have those. We'll move on.

And then there are in addition to those 14 maintenance measures, there are ten new measures that you'll be considering. One is an orthopedic trauma measure, the first one you see there. There are five blood-related measures from The Joint Commission that you'll be looking at.

There's a new one that actually the group has seen an earlier iteration of it on carotid endarterectomy. And there are three new composite measures from STS that you will be looking at today.

At the last phase, you reviewed the portfolio of measures and made recommendations about where you felt gaps were still occurring. And what you see on this slide are some areas in which you felt there were gaps.

The next slide has information about the kinds of measures that you indicated that you were interested in seeing being brought forward. So what you will be doing, the next slide just

has a little bit of very difficult to read information. But what this does is to show you that in that last report, you had identified gaps. We provided the list of measures that were in place. You took that and you identified gaps where you felt that measures were needed.

What we will ask you to do after this meeting -- you'll get it at your home site -- is that list of the measures that are currently endorsed plus those areas where you felt there were gaps and ask you to provide us update on where you see gaps now occurring. But we'll send that to you after this meeting after you're back at home.

What you have in starting off with your measure evaluation is that you have among your souvenirs a measure worksheet that has a preliminary analysis that NQF staff has put together. I want you to know that in the course of that preliminary analysis, there were multiple NQF staff who reviewed those documents and who provided input directly or to staff who were

working at those.

In no case is what you see the work of an individual. We do this as a collaboration to try to do the best we can do to give you information in advance.

What you also will have -- Christy, I think this is right in PDF. Is that right?

CO-CHAIR FLEISHER: Can I just comment?

MS. MURPHY: Absolutely.

CO-CHAIR FLEISHER: I mean we really need to give kudos to the staff. This is something that the CSAC really felt was a great advance in that there's consistency across all the measures. We can take exception to what the staff did and that's part of our role, but I really need to give kudos because they did a phenomenal job as they always do. It really provides a uniform foundation across the entire NQF portfolio. So kudos to the team, to Helen and her team. But she's gone.

MS. MURPHY: A large team. And you

should know that the team members don't always agree. But we do come to a point that we are reasonably comfortable with going forward.

And I would want to say, certainly repeat what Lee has said. Please do challenge where you feel it should be. What we've tried to do is give you preliminary information. Yours is the opinion and the consideration that counts.

In those measure worksheets, where we have received them, you will have pre-evaluation comments that those of you who reviewed the measures have provided. There is only one set of pre-meeting public and member comments and that's on Measure 2998. Typically we would receive more than that, but this time one set of comments.

And then among the items that you would have seen in the materials available to you on the SharePoint site, in addition to the measure worksheet, the measure information form submitted by the developer, there are attachments speaking to the evidence related to each individual measure. There are attachments that

speak to the testing that's been done by the developer in putting together and bringing the measure forward.

And then there are other attachments. You will see attachments with codes. You will see attachments with definitions. So those things that the developer provided in order to be helpful to you as you were evaluating their measures you will already have seen.

And probably enough said on challenging, challenging what NQF has brought to you and in your discussions.

CO-CHAIR FLEISHER: Just a couple of things and then Barbara and Marcia will keep me honest in this space. The CSAC, one of the things we're supposed to do is bring back some of the general consensus of reserve status. If you remember, this Committee was one of the first to really utilize reserve status.

There have been some areas in which we feel that the measure may have little gaps for improvement. But it's still important to keep on

the books because it's still a valid and reliable measure. It can be used as part of a composite. However, if the gaps are small -- for example, it's a never measure and there are rare events -- but it remains important to keep it in active status, then we have the right to keep it in active status even though the gap may be small. Correct? That's the most recent determination.

MEMBER DUTTON: On that point, we've discussed this many times here before, the concept of topped out measures and particularly the very important ones like mortality after surgery that seem to have a lot of face validity or public value, I noticed in several of the measures there was a comment from the NQF analyst this time and I'll quote it. Generally rare-event outcomes do not provide adequate information for improvement or discrimination. However, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

Is that our policy, and are we the

ones who decide whether a measure fits that bill?

CO-CHAIR FLEISHER: Marcia, if you have comments you can please add them. But I think that's essentially where the CSAC, we feel it's important to keep that measure active rather than reserve status. The answer is yes. But, Marcia, if you could.

DR. WILSON: Yes, it was very much CSAC wanted the standing committee to have the leeway to make that decision. What they recognized is a single, across-the-board policy would not be appropriate in these instances. And they felt that the standing committee would have the expertise to understand on a measure-by-measure basis because obviously this is a complicated topic. They wanted the standing committee to have the ability to make the decisions themselves.

And the guidance from the CSAC was that the Committee was allowed to decide when it was appropriate to leave a measure active, meaning not move it to reserve status especially

in the case of patient safety, mortality, a number of different issues. So the authority rests with you. But CSAC would want to hear the explanation like in the report when a measure goes back to CSAC. They would want to hear your rationale for making the decision.

Does that help, Lee?

CO-CHAIR FLEISHER: One of the important things is that the Board -- I don't know if it's implemented yet -- CSAC is the final arbitrator now.

DR. WILSON: Right.

CO-CHAIR FLEISHER: And the Board is getting out of the business of readjudicating. And CSAC is trying to get out of the business of readjudicating what the standing committee did unless the CDP process was not appropriately followed or it's not consistent with the overall theme of the direction.

So being clearer about why we have chosen a particular route will actually help not readjudicate and leaving a lot of the decisions

in this Committee and therefore the place where it's readjudicated if it's not consistent with the theme of the entire body of measures.

Bill, I'm sorry.

MEMBER DUTTON: I thought that was a huge step forward.

CO-CHAIR GUNNAR: So help me understand a little bit. I think the umbrella of focus for us is quality improvement. But there are certain situations or measures where the incidence has declined or the measure is topped out. And public reporting of that or the ability of that facility to publicly report is the right thing to do. Is that the perspective?

MEMBER DUTTON: Yes. So if we take the classic death after CABG that is 30 years old and the best measure out there -- that's obviously very important to the public -- it has huge face validity. But look at 800 hospitals and there are three low outliers and three high outliers. It's not incredibly useful for quality improvement as written. But it is a very

important measure to collect and report. And I guess you could phrase it that you want to keep reporting that because you wouldn't want to see backsliding or regression on something that important.

I'm glad that we have language that we can use to say, this measure is important essentially regardless of the variability in it, regardless of the fact that performance is very good. There are other process measures that are exactly the opposite.

From my own specialty, did we give the antibiotics on time? We sent that one to the Measure Hall of Fame a couple of years ago because it was clearly topped out. And as a process measure, it's not in itself critical. So I thought that was a very good decision.

But something like the mortality measures that we have in here -- and there's a bunch of them -- I think regardless of how they perform as quality indicators, they're going to be very important public measures.

CO-CHAIR FLEISHER: I think that's well said. Just a refresher because I forgot for a second, we should raise our placards when we want to speak. I think Barbara's hand is on it signifying.

MEMBER LEVY: Yes, I just want to give an analogy that maternal mortality for example is exquisitely rare, and it's very difficult to use as a measure for an individual institution because they may not see one for two or three years we hope.

But it's really useful for quality improvement. Maternal mortality reviews are things we're trying to get passed throughout the country.

Let's remember that even rare events can be very useful for quality improvement, and those kinds of events have a great deal of learning for us.

CO-CHAIR FLEISHER: Great. Others? Yes.

MEMBER MOSS: Yes, I just wanted to

put on the record that across-the-board event rates in children's surgery tend to be in the order of magnitude below adult surgery. So this concept is relevant across the board in children's surgery. And I support the decision of the group.

CO-CHAIR FLEISHER: Great. I mean I think what we were hearing is outcomes could be kept and not put into reserve status. The question will really be around process measures rather than outcomes. So that's invested back in this Committee.

Ready to start?

MS. MURPHY: All right. So I want to just tell you as you're thinking about the measures is that you're going to be thinking about new measures a bit differently than you're thinking about measures that are undergoing maintenance, in terms of the criteria.

So in looking at new measures, and you're evaluating evidence, you're evaluating gap as you always have. When you're looking at the

maintenance measures, you can place less emphasis on the evidence and gap -- well, not with gap, that's in fact increased emphasis -- but on the evidence piece. Because if they will have been through the process, they will have been endorsed. If there are changes in the evidence, then you will want to look at that carefully.

In many, if not most cases for the ones you'll look at in these two days where there is new evidence, it is directionally the same as evidence has been in the past, so maybe additive as opposed to in conflict with. If there is new evidence, you will consider that.

One of the other things that you will be looking at is with the maintenance measures if whatever was done before and was endorsed before raises a question in your mind in terms of the evolution of measurement through the NQF process. You may want to revisit that. But in general there will be decreased emphasis on the evidence piece.

There's actually greater evidence on

the gap because remember the maintenance measures will have been in use for a while. You would expect to see some improvement. If there hasn't been improvement you're going to be interested in wanting to know why there has not been any improvement. What are the issues? What's the variation? And what's contributed to those?

With respect to the criteria for scientific acceptability, you will be considering for -- again, for the new measures, you're looking at them in the same way you have looked at the measures in the past: complete. In terms of maintenance measures with scientific acceptability, there will be no difference in the way in which you look at the specifications of the measure other than you're looking for them to be updated where appropriate.

In terms of reliability and validity, if prior testing was adequate -- and that's a key question for you -- then there's not any need for additional testing for the maintenance process. For those measures to which it applies, the

effective sociodemographic factors should be addressed. You will be looking for that to have been addressed with the outcome measures.

Going ahead then to feasibility, there will be no difference in the way you're looking at feasibility whether it's new or a maintenance measure. And in terms of use and usability, you actually will be having an increased emphasis on the way in which you're looking at that. Much as you do with gap, you're going to be looking at is the measure in use? Is it useful? In terms of its impact, are there any unintended consequences? And if there are unintended consequences, what's being done to address those or what should be done to address those?

That's the CliffsNotes version of what you're going to be looking at differently in terms of maintenance measures as opposed to new measures.

We've talked about endorsement with reserve status. I think we've covered that adequately, Lee.

CO-CHAIR FLEISHER: Yes.

MS. MURPHY: And, Bill, are you okay with that? Okay.

MS. SKIPPER: Thank you, Melinda. Good morning, everyone. Just a couple of housekeeping items I meant to mention. Our restrooms are past the elevators and to the right.

And microphones, please be sure to speak into your microphone. You may have to lean a little bit closer. The meeting is being transcribed. And only three microphones can be on at one time. So once you're finished speaking just be sure to click off.

And we'll continue with the role of the Standing Committee. Some of you have been through process before and to our new members, welcome. So your role on the Standing Committee is to act as a proxy for the NQF membership. Each of you will serve a two to three year term.

Overall, you all are tasked with overseeing the surgery portfolio of measures and

working with the project team to achieve the goals of the project. Additionally your role is to complete the measure evaluations as you all have done -- thank you so much -- and to make recommendations to NQF regarding endorsement of measures.

Here are a couple of ground rules for today's meeting. I will not read them, but if there are any questions, please let us know. But I'm sure that you all will agree to adhere to the ground rules. We'll be sure to keep you on task. But I don't believe we'll have an issue with that.

Okay. When it comes to discussing our measures today, we will be asking the measure developer to introduce the measure. They have about two to three minutes to do so. Some of our developers are on the phone. Those that are in person, we'll ask them to come to the table and they'll sit right next to Karen Johnson. And once we begin the discussion of the measures, should the developer have a question, they'll be

asked to raise their placard to respond.

And then also the lead discussants will begin the discussion of the measure providing a summary of their pre-meeting evaluation comments or emphasizing any areas of concern or differences of opinion. Within your packet, you should have a measure discussant script that can help you start off as we discuss each of the measures.

Then again developers will be able to respond to any questions or clarify anything that perchance is misspoken. Then also the Committee will vote on each criteria and subcriteria.

When we move to vote, there are four criteria that we will be voting on. The first one is importance to measure and report. And there are two subcriteria: evidence and performance gap.

For some of our composite measures, we do have three composite measures within the portfolio. There is an additional criteria that we'll be voting on. That's quality, construct

and rationale.

We'll then move to vote on scientific acceptability, so whether or not the measure is reliable and valid, credible and consistent. And then in addition for our composite measures we'll be taking an additional vote there as well. Then we'll also move to vote on feasibility and usability and use.

And I want to note that the first two criteria -- importance to measure and report and scientific acceptability -- are must passass criteria. If the Committee does not pass a measure on either of those two criteria then the discussion of the measure ends there and we will move to discussion on the next measure.

Achieving consensus. In order for a measure to be passed or recommended, greater than 60 percent. So 60.1 percent of you must vote yes, high or moderate, on any criteria in order for the measure to pass. If at any time the vote falls between 40 and 60 percent inclusive of both of those ends, it's considered consensus not

reached.

If we vote consensus not reached on any of the two must pass criteria, then we will continue to vote on the remaining criteria, but we will not take an overall vote for suitability for endorsement. A measure is not passed if less than 40 percent of the Committee votes low or insufficient.

And just in general, quorum is 66 percent of the Committee. And I believe we definitely have that today. And are there any questions at this point?

(No response.)

All right. So we will begin the consideration of candidate measures. I'll turn it over to Lee.

CO-CHAIR FLEISHER: Great. And if you remember and for those who are new, if there are measures that have the same validity and reliability, we may ask if anyone has any objection to using the previous vote and carry that forward. There are certain criteria that

it's essentially the same dataset and the same construct.

So we are starting with PSI 4. And I believe we have AHRQ. Pam and Garth are on the phone.

DR. OWENS: That's correct.

CO-CHAIR FLEISHER: That's correct.

DR. OWENS: Thank you. Should I go ahead and start?

CO-CHAIR FLEISHER: Great. So Christopher and Amy are lead discussants. We don't need to repeat the same thing for whoever is going first. And then whatever you add, Amy can add.

Christopher, do you want to start with this?

MEMBER SAIGAL: Sure.

CO-CHAIR FLEISHER: Or do you want to let the developer? Okay. Great. Why don't we have Pam or Garth? Do you want to give us an overview?

DR. OWENS: Sure. Thank you very

much. My name is Pam Owens. I'm the Scientific Lead of the AHRQ Quality Indicators. And first I want to apologize for not being able to be there in person or have any AHRQ staff member there in person. I also want to apologize that our clinical representatives aren't able to be there in person.

But as was mentioned, Dr. Garth Utter and Pat Zrelak are on the phone from UC Davis. They can answer any additional questions as I can as well.

AHRQ very much appreciates the opportunity to have PSI 4 or the death rate among surgical inpatients with serious treatable complications reviewed today. Before Garth gives a broad overview of the measure, I want to take a few seconds to tell you about some of the core principles of the AHRQ QI program as I think these are critical aspects to keep in mind during the review.

The hallmark of the AHRQ Quality Indicator development process is the continuous

enhancement and refinement of all indicators based on user feedback, review of clinical practice changes, validation studies, empirical testing for validity and reliability and input from expert panels such as yourselves like the NQF Surgery Committee or experts from the AHRQ Standing Work Group.

In addition, I want to highlight another key component of the AHRQ QI program and that is the transparency and usability of the indicators. Not only does the AHRQ QI program publicly post all of the technical specifications, but AHRQ also provides users with SAS and Windows-based software to be able to calculate thorough numerators, denominators, observed and risk-adjusted rates using their own administrative data.

Users are in fact a critical component of the QI program at AHRQ. For example, this month we released an updated AHRQ QI Toolkit that can be used by hospitals as a general guide to applying improvement methods in a hospital

setting as well as guidance to improving performance specifically to the PSI such as PSI 4.

Garth, would you like to continue with an overview of the indicator?

DR. UTTER: Yes. Again, my name is Garth Utter. I'm a general surgeon, a trauma and acute care surgeon at UC Davis and a clinical lead on the QI project.

PSI 4 is fundamentally a measure of risk-adjusted, post-operative inpatient mortality across the --

MS. MURPHY: Sorry I'm interrupting. We can hardly hear the speaker. So you may need to get a little closer to the microphone.

DR. UTTER: Sure. I'll try again. Can you hear me better now?

MS. MURPHY: Yes, we can.

MEMBER SAIGAL: Yes.

DR. UTTER: Okay, great. So again I'm Garth Utter at UC Davis and a clinical lead on the QI project. I just wanted to emphasize that

PSI 4 is fundamentally a measure of risk-adjusted, post-operative inpatient mortality. It involves really the full spectrum of major operations.

Its denominator is limited to the subset of patients who experience one of a few common life-threatening complications. And the rationale for this is that it has been shown to support fairer comparisons across different types of procedures as well as it appears to be especially sensitive to the role of nursing care, such factors as staffing, skill mix, retention and turnover and the ability of hospitals to respond rapidly through effective teamwork as a patient's condition deteriorates.

It is, I should note, similar to another NQF-endorsed measure that was reviewed last year by the Patient Safety Standing Committee. That measure which is called Failure to Rescue differs slightly in that it captures all post-operative inpatient deaths, not just those that occur after one of these common life-

threatening complications. That measure was stewarded by the Children's Hospital of Philadelphia or rather Pennsylvania rather than AHRQ.

PSI 4 is built as a stratified measure combining deaths that occur after five different types of complications -- sepsis, pneumonia, shock, DVT/TE, or GI hemorrhage -- using different risk adjustment models tailored to predict the mortality related to each of these types of complications.

It's important to recognize that some of the deaths captured by PSI 4 are preventable while others most certainly are not. But these deaths appear to be worthy of investigation. The stratified risk adjustment models are designed to reduce bias in comparison across the hospitals. Thanks for the chance to summarize the measure.

CO-CHAIR FLEISHER: Great. Thank you. Any other comments from the developer?

(No response.)

Perfect. Chris.

MEMBER SAIGAL: Okay. First we go over the importance to measure and report criteria.

CO-CHAIR FLEISHER: Yes. And then if they need, any additional comments.

MEMBER SAIGAL: Okay. As we have just heard, this is a measure looking at the number of deaths in surgical patients with serious, treatable conditions post-operatively. And the developer submitted evidence that it's into the quality of nursing care. So there's a process that supports that.

In terms of performance gap, the performance has improved nationally by about six percent per year. But there is still a pretty significant gap in terms of there being 43,000 deaths per year in 34 states measured in all-payer datasets. That's a pretty significant number of perhaps actionable deaths.

There are variations in those deaths by age, insurance status and other subgroups. In terms of -- anything else that my co-discussant

wants to add on that?

(No response.)

So then in terms of how it's specified, do we move on to that? Or should we vote?

MS. SKIPPER: Yes. So if there are no other comments or discussion on evidence and performance gap, then we would stop here and vote on this criteria.

CO-CHAIR FLEISHER: Questions or comments? Yes, A.J.

MEMBER YATES: Yes, I just have a question for the developers. Of the five complications that you list, at least four of them could be indications for emergency surgery at a hospital to where they've been transferred. An example would be sepsis from a surgically correctable source of sepsis.

And I don't see an exclusion criteria for present on admission. And I don't see any sort of mechanism for deciding how those patients arrived and are they excluded when the actual so-

called complication is the indication for surgery.

DR. UTTER: Sure. I can help address that concern. So you're correct. The indicator is specified such that present-on-admission conditions are not excluded, however principal diagnoses are.

So the rationale here is that there were some prior analyses done by Needleman's group that pretty clearly established that excluding these cases that involve one of these complications having had an origin prior to arrival of the patient at the hospital didn't really improve the validity of the measure. It did detract pretty substantially from the number of events that could be ascertained with the measure.

MEMBER CIMA: I have a comment. But the assumption was initially that this was a measure of nursing care and quality of care in an institution. But as an institution like ours where we're surrounded by very small community

hospitals for a region about 400-500 miles, we get a lot of these issues. And it has nothing to do with our system of care.

For centers that are regional centers that receive this, it may make a very big difference because their volumes are so high of doing that.

MEMBER SAIGAL: So this is a fact of validity we're talking about I think. This is the next session. We should probably get into that there. It sounds like we're talking about validity questions.

CO-CHAIR FLEISHER: Okay. Why don't we get back to that? Rick, do you have a comment on importance?

MEMBER DUTTON: I think failure to rescue is an important thing to measure. I worry or maybe in the next discussion there's an exclusion for patients transferred out of the hospitals. So if my patient develops sepsis and I quickly transfer them out, I'm good. But isn't that creating a loophole in this measure?

MEMBER SAIGAL: This is another validity question. So why don't we vote on the first one, importance to measure and report?

CO-CHAIR FLEISHER: Thanks. Any other comment? I appreciate you sticking with the criteria. Did we lose Christy? Do you agree with that?

MS. SKIPPER: What we're going to do actually, this is a maintenance measure. And due to our new process, there's a decreased emphasis on evidence if there is no new evidence. So you all may elect to use the vote, carry over the vote from the previous endorsement of this measure. Then we will not technically take a vote on evidence.

CO-CHAIR FLEISHER: So does anyone object to carry over the previous endorsed vote?

MEMBER SAIGAL: For what?

CO-CHAIR FLEISHER: For this criteria for evidence. In other words, from the previous endorsement, the new process -- and thanks for reminding me -- is if there's no change from the

previous process, we can just endorse by acclamation, so to speak, the previous vote.

CO-CHAIR GUNNAR: But technically I've heard that the gap has diminished, but it's still substantive. So I think it's easier to vote since this is the first time. Why don't you run us through voting and we'll vote on it?

CO-CHAIR FLEISHER: Okay. And Karen, our expert on evidence, can comment.

MS. JOHNSON: Well, maybe not evidence. But I know the criteria pretty well and it has changed I think since the last time you guys have met and voted. So let me just restate again so that it's clear in what we're doing.

We're with our new maintenance process we're having less of an emphasis on evidence. Okay. So what happened with this measure is it is a maintenance measure. The developer did provide additional evidence, but it seems to be pretty much in the same vein as the other evidence.

So the question for you in terms of evidence, not gap but evidence, is do you feel that there's a need to vote again, or are you happy with just saying that the measure would still pass on evidence? So that will be your first question. Do you want to vote on evidence or not?

Now you haven't talked very much yet about gap. That's the next subcriterion under importance to measure and report. So you will have to definitely talk about and vote on gap. Okay. Does that make sense?

CO-CHAIR FLEISHER: So do we have a yes/no vote that we all do or are you taking just a raise of hands?

MS. JOHNSON: You can be informal on this. I think the question would be, does anybody feel the need to vote this morning on evidence for this measure.

CO-CHAIR FLEISHER: Do you want to just -- okay. We're all okay with that. Larry. Great. Larry, do you have other comments or?

MEMBER MOSS: I just had a question for the developer. Could you please explain the rationale for excluding patients under 18 years of age in the measure?

CO-CHAIR FLEISHER: That's actually I think the next one. Christopher, do you want to go to the next criteria? This always starts like this as we get back into the swing of things.

MEMBER SAIGAL: Go on to number two, is your idea? Okay. So then validity and reliability.

MS. JOHNSON: No, I'm sorry. Let's go back to performance gap and maybe you guys did talk about this and I just missed it.

MEMBER SAIGAL: Yes, we did.

MS. JOHNSON: Okay. So now what you need to do is find out if anybody has any other things you want to talk about about gap. If not, then we'll go to vote on gap. Remember with our new maintenance process we are actually increasing our emphasis in our look at performance gap.

CO-CHAIR FLEISHER: Okay. Questions on gap? No questions. So, Christy, do you want to help us vote?

MS. SKIPPER: Yes. Just a moment. I'm just preparing this slide. Okay. So we'll be voting on performance gap for Measure 0351. Everyone should have received a clicker. When I tell you that polling is open, you will point toward me and press number one, two, three or four, one for high, two moderate, three for low, four insufficient.

We do have two members participating over the phone and they will be chatting in their votes. And Katie and I will be voting on their behalf. So you all, polling is now open for Measure 0351 performance gap: one high, two moderate, three low and four insufficient.

(Voting.)

CO-CHAIR FLEISHER: Okay. Should we just keep pushing and we can't vote more than once.

MS. SKIPPER: We want that number in

the circle in the bottom left-hand corner to hit 22.

(Voting.)

MEMBER SAIGAL: Still waiting on three more votes.

MS. MURPHY: Can you tell which ones you need?

MS. SKIPPER: No. If everyone could just --

CO-CHAIR GUNNAR: Mine is not lighting up. So does that mean it's not --

MS. SKIPPER: I would try again and aim it at me. The remote will record your last vote.

CO-CHAIR GUNNAR: My battery is dead.

MS. SKIPPER: Okay.

CO-CHAIR GUNNAR: Can we get new batteries?

CO-CHAIR FLEISHER: Is everybody seeing a red light?

MS. SKIPPER: Maybe someone stepped out of the room, but we're at 21.

And thank you for your patience on this first round. All right. Now we're at 22. So 22 is at the very --

Yes, so voting has closed. Twenty-seven percent of votes were high. Seventy-three percent of votes moderate. Zero percent low. Zero percent insufficient. 0351 does pass on performance gap.

CO-CHAIR FLEISHER: Before we go on, Lynn, are you prepared to just briefly introduce yourself and any conflict of interest?

MEMBER REEDE: Good morning. Lynn Reede. No conflict of interest.

CO-CHAIR FLEISHER: Thank you.

MEMBER SAIGAL: Should I proceed?

CO-CHAIR FLEISHER: Okay.

MEMBER SAIGAL: So this is scientific acceptability, the next phase. In terms of specifications, it's very clear documentation on how to define the numerator and denominator and the comorbidities required. There are good specifications.

In terms of reliability, the developer used what's called a signal-to-noise ratio which is a measure of the variability between hospital performance and within hospital stability. The measure was acceptable with a ratio of 0.7 or greater for only the largest hospitals, more than 436 discharges. But the developer feels that lower signal-to-noise ratio standards are actually acceptable which would include all the hospitals in the datasets. So there's a bit of a discrepancy there in terms of what's an acceptable signal-to-noise ratio.

In terms of validity testing, it did face validity testing using the RAND Delphi process. It passed on that. They also tried to test how these measures performed in a Medicare analysis looking at five other elements of the hospital structure essentially with better outcomes.

MS. MURPHY: I'm sorry to interrupt. Let's complete reliability and do the vote on that.

MEMBER SAIGAL: Okay.

MS. MURPHY: And then do validities.

MEMBER SAIGAL: Okay. So for reliability.

CO-CHAIR FLEISHER: Amy, any additional comments and then we'll go to Collette.

MEMBER MOYER: The only thing I would add is this is a measure that we actually use and have calculated. AHRQ making the software and the code available takes a lot of the judgment calls that might go into play on calculating other measures. So we find that it's really very straightforward to feel like you're consistently calculating this measure.

CO-CHAIR FLEISHER: Collette.

MEMBER PITZEN: Just a question. Is this the part of reliability voting where we talk about specifications, because I heard some questions about exclusion?

MEMBER SAIGAL: That's coming up though actually. That's validity. That's

subsequent.

CO-CHAIR FLEISHER: Let's get a -- Karen.

MS. JOHNSON: You talked mostly about specifications under reliability. So you want them to be very precise and unambiguous and you want to agree with how they're specified.

They do come into play a little bit with validity as well if you feel like the way something has been specified invalidates the measure. For the most part you will be talking about it under reliability.

CO-CHAIR FLEISHER: So the present on admission questions would be under?

MS. JOHNSON: They start under reliability.

DR. OWENS: This is Pam Owens. And I just want to speak to the question regarding transfer patients. Transfer patients are actually part of the risk model. So we account for that potentially increased risk in each of the risk models for each of the strata.

CO-CHAIR FLEISHER: Rick.

MEMBER DUTTON: I'm sorry. Is that in the risk model for the receiving hospital which was A.J.'s question or for the transferring hospital which is my question?

DR. OWENS: So each hospital would have their own rate and the patients in the risk model is if they are transferred in. That is the criterion in the risk model. Does that make sense?

MEMBER DUTTON: Yes. So that's A.J.'s question. So if the hospital like Penn, for example, received a patient in transfer who is desperately ill with one of these surgical conditions and goes to the OR, the risk-adjustment model accounts for the higher risk of that patient dying. That makes perfect sense.

My question was about the hospital transferring out. So the small community hospital we do a colectomy electively. We put a hole in the bowel. The patient develops sepsis. We fail to identify that in a timely fashion.

When we do identify it, we transfer the patient to Penn. So according to the specifications as I'm reading them, that case now doesn't count for us because it was an acute care transfer out. Is that correct?

DR. OWENS: That is correct.

MEMBER DUTTON: And does that create the potential for gamesmanship?

DR. OWENS: I see what you're saying.

DR. UTTER: I was just going to comment. This is Garth. It does create possibly a small window for gaming. However, there is just simply no way to assess the outcome of interest here in those cases of course.

MEMBER CIMA: But this goes to the point of validity as opposed to reliability because on face validity that makes no sense. And that's the concern. If you're looking at processes of care to identify patients in your hospital this is how we started this conversation was we're looking at processes of care to identify people early and things.

And then if the receiving hospital gets a desperately ill patient, they had nothing to do with the process of care. And if the patient dies, how can they be held responsible for that? And if you're excluding those hospitals as you just said you are, there's no validity here.

DR. UTTER: No, we're simply excluding the cases in which we can't assess whether a death occurred or not because the episode of care really hasn't completed. And as Pam emphasized, the issue of receiving a patient from another hospital, we don't want to exclude those cases because they contribute really to the signal that can be detected. Therefore, we handle them with the risk adjustment.

MEMBER CIMA: But you can't risk-adjust all that away. That's the problem with risk adjustment. There is some risk adjustment that you can't handle. And the validity of this is receiving hospitals are going to be faced with this issue of, how can we be responsible for

this. And now you're talking about the practicality of this.

DR. UTTER: Sure. And we certainly understand that. We've received numerous queries from users over the years on this issue. And I guess one piece of information that may contribute to our understanding of this is that actually these transferred patients end up having lower rates of death than ones that occur inpatient. So it's not perhaps quite the issue that it's perceived to be.

CO-CHAIR FLEISHER: Larry and I think -- okay. A number of questions. Larry, why don't you start?

MEMBER MOSS: I have a new topic. I think Karl had something on this topic. So why don't you go first?

CO-CHAIR FLEISHER: Okay. Karl.

MEMBER BILIMORIA: I think leaving transfer patients in just because that improves your signal doesn't really make a lot of sense. I think that we all would rather see those

excluded. I mean just including a complex population because that improves your event rate doesn't seem to make sense. I think that's what was said.

MEMBER YATES: My clarification was is that transferred in with the diagnosis that's being evaluated as the principal -- not as the principal, but as the present-at-admission diagnosis. My concern being that if the patient dies from something else during the hospitalization, that then becomes the principal diagnosis. Whatever was found to be causing the sepsis, might be gangrenous gallbladder, that becomes the principal diagnosis. But the word sepsis is then a secondary diagnosis just because of the way they coded it.

MEMBER BILIMORIA: No, I was taking exception with what one of the developers was talking about in terms of why they leave transfers in. Yes, I understand what you are saying.

MEMBER YATES: And I'm just saying

it's not the transfers. It's the transfer with that diagnosis that I would be worried about.

DR. OWENS: This is Pam. And I will say as Garth mentioned, we have grappled with this over the years. And part of our concern here with the failure to rescue originally, by Jeff Silber, my understanding is that included all patients.

And the transfer patients are sort of, if you exclude them, then that's a missed opportunity to see what is going on. And absolutely we've had both sides of the argument conceptually to include them or don't include them.

Our approach right now is risk adjustment. We certainly can show you some additional analyses that shows what happens, not today, what happens if we were to exclude them. But conceptually, we were harmonizing originally with Jeff Silber's measure, and that's where this originated a long time ago if that helps in terms of historic information.

CO-CHAIR FLEISHER: Okay, and I have just for disclosure: as I mentioned, I worked with Jeff. So I have an email into Jeff to find out. Karl, did you have anything else?

DR. OWENS: Excellent.

CO-CHAIR FLEISHER: And then Fred.

MEMBER BILIMORIA: No, I think the notion that this doesn't matter, I think most people when they look at their own data will tell you that it does matter. Most of our hits on this are transfers in. And so the idea that this is a small problem I don't think is valid. And I think it would be worth seeing the data on what happens when you exclude these patients.

CO-CHAIR FLEISHER: So I think since we always see back, Christy, is it okay if we ask if they have data-specific --

MEMBER SAIGAL: Can I tell you? There is data. If you look on page 575 of that big document that was sent around this morning, they do list -- and we're jumping around here. We're in another part of the review, right? We're in

validity.

If you look at this page, they show you how many people have been excluded based on various criteria. So 3 percent are excluded based on they were transferred out to acute care facilities. Then there is the number of people who are excluded because their primary diagnosis is a complication; that ranges from 3 percent to actually 13 percent of patients who had a principal diagnosis of septicemia who were excluded. That's the principal diagnosis.

So they do provide some data to address some of these concerns. But I would suggest that we vote on reliability before we get into this part of it because we're really jumping around, and it's going to be a big mess.

CO-CHAIR FLEISHER: Thank you. Fred, do you want to -- you'll pass. Okay. Larry.

MEMBER MOSS: Sorry for my misplaced question before. But if I understand correctly, exclusions belong under reliability. So I'd like to ask my question about the rationale for

excluding patients under 18.

DR. UTTER: Yes, this is Garth. So I think some of this goes back to the origin of this measure and the work of Needleman's group to base it on these five conditions, not all of which would necessarily apply to the pediatric setting, and most prominently DVT/TE and PE. And so I think it's just the construct of the time that this was developed. It was really more applicable to the adult population is the main reason.

MEMBER MOSS: I appreciate that, and I share and accept your rationale, although I would like to see these measures include in the text a specified reason for excluding patients under 18 rather than just the assumption that that's okay.

CO-CHAIR FLEISHER: Okay. Karen, I mean that's pretty standard, the exclusion for 18. But it could be that we include that in gaps, that this might be an important measure to measure in the pediatric population. Can we,

Melinda? I think that's a -- we should look -- one of the gaps could be that this should be extended into the pediatric population. A lot of the time, they don't have the data. Would that be okay?

MEMBER MOSS: Yes, and I think that's reasonable. And I'm not trying to advance personal agenda. I just wanted to raise the issue that if we excluded black people or poor people, we'd be all over it.

And we routinely exclude children and say it's standard. And I'd like that not to be considered standard.

CO-CHAIR FLEISHER: So let's not make that a personal agenda. Let's include that in gaps that whenever you see a measure that should be extended to the pediatric population, we identify that and put it in our document.

MEMBER MOSS: Thank you.

CO-CHAIR FLEISHER: Cliff.

MEMBER KO: So I hate to bring up issues without a solution, but this is just a

blanket statement for measures that use claims data. I know a lot of people around the table have registry-based data. And it's our environment.

But our environment is changing in that most reliability and validity studies that compare claims data to the gold standard of what happened in the medical record showed the lessened reliability and lessened validity of using claims data, especially -- we're a surgery group -- for surgical complications.

Usually when the patient dies, we know it in claims data. But it's the accuracy of the complications. And I know that Fred and STS and in NSQIP and the VQI have shown the problems with claims data. I'm not sure what the solution is, but it is something that we should have on the record.

CO-CHAIR FLEISHER: So another potential gap or area for improvement if it could be created in registry data is what I'm hearing. We could mark that. Fred.

MEMBER GROVER: Yes, we feel strongly in our specialty that fail to rescue really needs to be examined. As I listen through this, this is really a tough problem because we've got patients at these outlying hospitals that probably the best thing for the patient is to refer them out to a center that has more complex facilities to take care of complex issues. So that's the right thing to do.

And yet then you're potentially -- and I see this every day, you're penalizing a tertiary care center that receives these patients unless you can totally risk adjust. As Cliff says, I don't think in this complex patient set you can totally and accurately risk adjust unless you have clinical data. I mean, it's really tough.

I don't know what the solution is. But those are my concerns.

CO-CHAIR FLEISHER: So we will have to vote on whether or not we feel that invalidates the measure. Barry, Karl and then Amy.

MEMBER MARKMAN: In terms of the reliability, how many codes deep do you go in the discharge diagnosis? I mean how do you pick up the shock and the sepsis, and how do you differentiate between the two?

CO-CHAIR FLEISHER: That's a question for the developer.

DR. OWENS: So the measures are developed on data from the Healthcare Cost and Utilization Project and the State Inpatient Databases. Some of those databases have up to 50 diagnoses. So it could be up to 50.

Because in itself, the SAS program, you can set it to whatever number your administrative data has. I believe that 25 is the number for the CMS Medicare data. That's just what happens to be on Medicare.

So it could be any number that you want it to. We develop it on up to 50. The majority of the states have between 25 and 30.

MEMBER MARKMAN: Okay, and these codes are generated by the billing departments of the

respective hospitals? It's ability to get those things.

DR. OWENS: Right. The Healthcare Cost and Utilization Project is based on billing data or discharge abstract data. The coder that enter it are all certified coders. And they take tests, and they look at the coding clinic for additional guidance or guidance that's put out by the federal government. So there's a standard for how to code what's in the medical record.

MEMBER MARKMAN: Right.

MEMBER CIMA: This is Bob. But that being said, there's a number of studies that look at -- well, I don't think the word standard is actually appropriate for coders. There are very significant variations in how coding is applied by institutions. That's been shown multiple times.

Just because one coding institution has their own internal way of reviewing their records, another institution may not. So that introduces variability between institutions and

how they code.

And given the complexity of this, as has been said, when you're looking at sepsis, sepsis is multifactorial. There's a lot of issues. That may not be as good a choice to discriminate between institutions if you're going to use coding as your main thing. And that's been shown multiple, multiple times.

MEMBER MARKMAN: One last question. Now have you performed an audit just to do your own validity test? Have you audited some charts to make sure that diagnosis codes were appropriate?

DR. OWENS: Internal to the project, yes, we have done some validation studies. Some of them have been published, not all of them, where we do individual hospitals or we partner with various consortiums across the country to do mini hospitals. In terms of Medicare, yes, they do their own auditing. So there are various places that audits are done.

CO-CHAIR FLEISHER: Thank you. We

need to keep moving. So I want to get a vote on this. But Karl, Amy and then Cliff. Excuse me. Liz and then Cliff.

MEMBER BILIMORIA: Yes. So I think the variability in coding between hospitals is a definite threat to reliability. That really relates to the validity then of the definitions.

And there have been studies that have compared, as Cliff said, the clinical data to the administrative data and found out only a false negative rate, but a high false positive rate. So it's all over the place. I think the threat of validity here, really we should question the validity of this measure entirely.

CO-CHAIR FLEISHER: Amy.

MEMBER MOYER: So I guess I have a couple of things. One, I don't think we can just generally say claims data are bad, and you can't measure things from them. There's been some measures we find a high correlation between the record and the measure, and some we don't. Frankly, we have yet to incorporate any measure

in our program including registry measures where we haven't had at least one hospital then come back to us and say, oh, we were sending the wrong data in. So there are issues with any data. There are no perfect datasets when it comes to healthcare that I've seen.

The comment I had actually originally wanted to make is, I think, new this year in the measure worksheet that NQF actually chose how you walk through the algorithm. And I personally really appreciated that, being able to see the thought process and see how you applied the algorithm in evaluating the reliability and the validity of the measures. That really was helpful to me to see where we are on the same page and where we aren't. Thank you for doing that.

CO-CHAIR FLEISHER: Great. That was -- Karen. No. Okay, Liz.

MEMBER EREKSON: And I apologize for not knowing everything about this measure. But my question is the discharges to hospice. Does

that count as transfer out of the facility? And where do those patients get counted? And can the measure developers provide any information about how many hospice patients actually get included in the outcome?

DR. UTTER: This is Garth. So if I can recall correctly, I believe the hospice discharges, they're included in the way the indicator is designed. But they really represent a very small proportion of the denominator.

CO-CHAIR FLEISHER: Thank you. Cliff.

MEMBER KO: This is again about the claims, and it's not to attack Garth or AHRQ, but just of claims. So at the college, we did a validation study of claims versus registry with NSQIP and with CMS data. This was over 100,000 patients and like 300 hospitals just where we had data. And of the same patient, we had claims data and we had the registry data. So we could do that study where we look at claims versus registry.

Karl mentioned that the big problem is

the false positive. If we took an SSI, surgical site infection, the most common complication for surgery, the sensitivity was around 30-35 percent. So it detected. It was a low detection rate, which it's okay.

Fine. We look better because we don't find them all. But the false positive rate was around 70 percent. So we're inappropriately dinging people for an SSI when they don't really have an SSI. And when we first found that, I was like, why? Why would that happen? And it's the definitions.

And the biggest thing, the best example is that if I write in the chart after I do a colon operation, on the fifth day, the patient is spiking fevers. I write, "Please get a CT scan to rule out infection." The coder will write that that's an infection because I wrote the infection, even though consider infection, rule out infection, or whatnot.

It's not by the CDC definition of a culture or gram stain or anything like that.

It's those words, and that's how the code is working. That's how ICD is. We're not going to change the International Classification of Disease. But that's what it is.

So that discrepancy between clinical and claims is there, and it's just a matter of definition. What we should probably decide is how do we reconcile that, or do we not reconcile that.

I can tell you that when we brought this work with CMS, and if you see the payment programs, of how they're starting to go away from claims and go to registry and go to electronic measures, that is the direction we're all heading. I don't know if this is the time to stop all these claim measures or have a limited time that we have a shorter window to start to reevaluate them again. But that's the direction that seems to be heading for the environment.

CO-CHAIR GUNNAR: Can I ask a question? And maybe to Chris and Amy as well. Is the perspective that the improvement in

performance associated with this measure, is it due to real quality improvement? Or is it due to an enhanced ability to code or to manage the data if you will? I guess that's the philosophical question to the group regarding this particular measure. A.J.

MEMBER YATES: I would second that question, and I would offer my answer, which is I think in the time frame we're talking about, a significant improvement is as likely a paradigm shift in coding as it is an improvement in surgical care. And I think that the coding, now that hospitals are graded by stars and face millions of dollars in penalties in both HAC, VBP and readmission rates, they're a lot more careful about their coding. And I can give several examples, but won't.

CO-CHAIR FLEISHER: What I would suggest, and we're having a lot of discussions up here is Helen Burstin is the best person right now to have a discussion about, is this time to abandon? I mean this is a much bigger question

than this measure, this committee. This is a question for CMS, Luanne being on the phone.

MS. MUNTHALI: And just one more reminder.

CO-CHAIR FLEISHER: Yes.

MS. MUNTHALI: As you're looking at this measure, remember you're looking at it as it's currently specified.

CO-CHAIR FLEISHER: Right.

MS. MUNTHALI: We understand that there are a lot of aspirational hopes for measurement in general in this measure. But you're evaluating it against NQS criteria as it's currently specified.

CO-CHAIR FLEISHER: I actually had one quick question because Jeff Silber did answer me. And the original failure to rescue -- this is for the developer -- transfers were kept in and adjusted for, it sounds like as a covariate, they just transfer in. But transfer out, they assigned death and failure to rescue to where the original surgery occurred. Do you not do that in

your failure to rescue, which, I guess -- the Needleman failure to rescue as opposed to the Silber?

DR. UTTER: This is Garth. If I understood you correctly, the question is, do we assign death where it occurred? Was that the question?

CO-CHAIR FLEISHER: Actually assign death to the original hospital that did the original surgery in which the complication occurred. It sounds like what you said is you don't do that even though the original failure to rescue measure did do that.

DR. UTTER: No, at least on a broad basis it's not possible to do that without some ability to link hospitalizations. So, no, we don't do that.

CO-CHAIR FLEISHER: So you don't. We need to take a vote. We will come back to this really critical issue. Does anybody have critical comments to the vote on reliability? Karl, did you have one?

MEMBER SAIGAL: I think we should specify what we're voting on now because we had a pretty broad discussion.

CO-CHAIR FLEISHER: We're going to ask -- I say staff, but really the other experts in the room. Do you have a comment?

MEMBER PITZEN: Thanks. I do have a comment. I'll put my measure developer hat on for a second. And I might be traipsing into validity. But I think we have to be careful in disclaiming all ICD-9, ICD-10 diagnosis codes because there are very reliable ways of identifying a population of interest in using those codes.

I think part of this speaks to validity. If there are concerns about the sepsis portion of this measure, what can the developer perhaps tell us about the validity when comparing that administrative code against again the medical record? You're validating.

So the difference in this measure is we're using ICD-9, ICD-10 for the numerator. So

one needs to understand if I'm pulling these particular ICD-9 codes, ICD-10 codes to identify sepsis, what is my reliability in having that true measure?

DR. UTTER: This is Garth. Just one very important issue. I don't know if it was just a matter of misspeaking. But these codes, the diagnosis codes, establish the denominator and not the numerator, which determined by death.

And maybe just to address more broadly the concerns about use of administrative data, it's not quite as clear what biases may be introduced by over and undercoding of these diagnoses since they function just in the denominator specifications. That said, we do have some concerns that over time, hospitals are in effect potentially trying to whittle down their denominator by more selective use of the codes. And that's a valid concern.

Maybe in the long term, the ultimate consideration is, do we go back to the Jeff Silber approach of really including all

hospitalizations with the procedure?

MEMBER PITZEN: Thank you.

CO-CHAIR FLEISHER: Thank you. What I've heard -- Do you have critical comments to reliability?

What I've heard is that we clearly have significant concerns about the use of claims data, which when Helen comes in, she'll talk about the transition and how much longer we continue to accept claims data for measures. And we will clearly put that in the surgical arena, which was at the forefront -- thank you, Fred, of doing these registries. That that was really critical.

So that will be in the report no matter what. That's a general consensus of this Committee, those two questions. Surgery has gone to registries. And how long should we do claims? We'll ask Helen.

Christy, do you want to tell us what we're voting on?

MS. SKIPPER: You're now voting on

reliability for Measure 0351.

CO-CHAIR FLEISHER: Define exactly what we're voting on so that people know what that means or Karen.

MS. JOHNSON: Let me take a shot at it, Christy. You are voting on the precision of the specifications and the testing that you saw under reliability. So all of the discussion that you had, you guys did go into a lot of validity discussions. Please put those to the side. Vote now on, are the specifications understandable to you? Are they precise? And is the testing that you saw in the results of the signal-to-noise testing adequate in your opinion?

CO-CHAIR FLEISHER: Thank you. That's very helpful.

MS. SKIPPER: So voting is now open. One, high; two, moderate; three, low; four, insufficient.

MEMBER KO: This would include the discussion we had about exclusion and the transfers? That's this part?

MS. JOHNSON: Yes. So do you understand what they're doing is the question.

MEMBER SAIGAL: Really. There's a different section under validity for all that. I think it's more --

MS. JOHNSON: Right. Whether you agree with it or not, you will talk about under validity. But the question for you now is, is it precise enough? Do you know what's going on with the transfers out? Do you know what's going on with the transfers in?

(Voting.)

MS. SKIPPER: And we're waiting for just one more vote.

(Voting.)

If everyone could just re-vote and aim your remote in this direction.

(Voting.)

Polling has closed. Results are 4 percent high; 65 percent moderate; 22 percent low; 9 percent insufficient. Measure 0351 passes on reliability.

CO-CHAIR FLEISHER: Okay. Validity is now open.

MEMBER SAIGAL: Okay. So we talked a lot about this already. I'll just say that from the point of view of what was submitted, validity testing was done by looking at teaching status and seeing how the measure performed with and without adjustment. And as teaching status and a few other areas became -- were risk factors for risk performance before adjustment and then became effective after adjustment. So I was testing the use of the risk model here.

But the other issues around validity testing we think about in terms of, does this measure measure what you think it's measuring? There was no data specifically presented about that.

I looked at, in the literature, there was a couple of articles from many years ago that said that this, as Cliff mentioned, was not a very sensitive approach to finding problems compared to chart reviews from single

institutions.

I didn't find the data around the specificity issues Cliff mentioned, but I believe they exist. I guess their application was silent on those issues.

And the rest of this validity, we talked about a little bit in terms of whether the transfer issues are big enough, or whether it's addressed adequately to understand whether it's a significance threat to the validity of the measure.

CO-CHAIR FLEISHER: Other comments? Okay. Fred.

MEMBER GROVER: I think failure to rescue is an important topic. But I don't see how it can happen unless it's within one institution where you can resolve meaningful data in terms of the face validity. The idea being that if you have a complication of an operation you've done, what do you do save that patient from a very, very serious complication?

When patients are being out or

transferred to other centers, it just seems to me like it makes it impossible to validate that that indeed is taking place.

CO-CHAIR FLEISHER: Can you explain? Do you feel it's not valid to not include the transfers out? Is it an appropriate way to look at it?

MEMBER GROVER: I think if we want to learn about is how to improve survival following a complication, there probably needs to be how they're treated within a single institution. If you take somebody who a lot of times isn't transferred in a timely fashion, they're hung onto, and then you get an absolute disaster, what do you learn from that? It's hard to analyze.

CO-CHAIR FLEISHER: Allan.

MEMBER SIPERSTEIN: Kind of a different dimension of complexity. Within a single institution, if you want to track your improvement over time, I see some validity to the measure. If I want to benchmark my institution against yours, that's where I start to have more

questions.

CO-CHAIR FLEISHER: So that, it's measure for intended use, Marcia, is that I'm hearing? Are you questioning whether or not PSI -- are you concerned about how it's being used? Or how the measure is specified? Those are different things that we still have yet to fully address.

MEMBER SIPERSTEIN: I'm just addressing the validity part of it. Whereas, if I'm looking only at my own institution over time to drive process improvement, I feel more comfortable with that. If I'm looking at the validity of the -- I know this crosses a little bit into usability. But if I'm looking at, is the measure valid as a comparison tool, that's a different question.

CO-CHAIR FLEISHER: Thoughts from staff in this area?

MS. JOHNSON: I will just say that measures that are endorsed by NQF are expected to be useful for both quality improvement and

internal efforts as well as different kinds of accountability applications. So you would think about whether you feel like this measure is not valid for those kind of applications.

I would say that -- let me stop there.

MEMBER HANDY: Not to throw the baby out with the bathwater, when we look at their data, there is over 300,000 patients that we're talking about. 3 percent of them were transferred and therefore excluded.

I think that the discussion here is colored by the bias that most of the physicians practice at tertiary care centers, and they receive these patients. So they feel kind of snake-bit by it, so I wonder if we're getting distracted.

CO-CHAIR FLEISHER: Remember we sit here to judge the measure, not how it will affect us, but whether or not the measure is valid. A.J., and then is that Rick?

MEMBER YATES: My comment is just following up on what Allan said, which is that

the validity is in fact tied to the end use because there are -- and again, this is the concept of tiered validity and the fact that you may need only a 10 times microscopic view to see something is either black or white. But to see very fine shades of gray, you might need a 100 times microscope.

And the question is, does it have the degree of focus and clarity in terms of risk adjustment, in terms of the dataset, in terms of the signal-to-noise variables and all of those things? All of those things tie into what the end use. And if you're determining whether a hospital is about where you expect them, about slightly worse than you expect, or just slightly better than you expect, you're talking about gross categories of classification.

When you start to define hospitals by very fine percentage points or tenths of a percentage point in percentile rankings in terms of how they're affected by value-based purchasing or HAC, it does make a difference how fine the

instrument is attuned.

CO-CHAIR FLEISHER: So let me be -- I'd like, Barbara, if you could comment, because you were there. We had a major discussion that got all the way to the Board of whether or not there should be another level of endorsement if it's used in accountability in which pay for performance -- in fact, we didn't say accountability versus pay for performance should be different. Quality improvement potentially should be different. But there is this, developing this idea of further testing and what that means.

But right now we're not -- how CMS implements it in pay for performance is a separate question. Am I getting -- Barbara, for you and Marcia. But Barbara was sitting in the room.

MEMBER LEVY: Yes, I think that's exactly right. As we did a deep dive into this, it became clear that the specifications of the measures and the criteria that we use in the

science of quality measurement shouldn't be different depending on the use of the measure. The science that we apply as we look through things is the same.

However, as we sit around here and think about the practical implications of what we're doing, that's where I think we get into really muddy waters. It's impossible to do a perfect measure. And the question is how far along that spectrum of imperfection we're willing to tolerate.

Clearly, if we're using it internally for our own purposes, it matters not. We use the data. We determine whether the data is good data, bad data.

When it's being used to judge or to pay, that's fundamentally a different problem. And Lee and I were talking about this this morning. The payment world and the quality world are separate. And we're in trouble because we're confounding them in our minds.

We have a job to do with respect to

the science of these measures. And that's different than extrapolating that to what happens in an accountability world. It's really hard for us to do that because we know what the consequences are.

And yet measurement science is telling us something different really than our practical side is telling us. And we're surgeons. So we're practical.

CO-CHAIR FLEISHER: Yes, and I think Barbara said something really important. If we feel how it's employed in value-based purchasing, that's what our societies can do. We're sitting here as individuals saying, is this measure valid?

I think it was Rick, Christopher and then Fred and then Cliff.

MEMBER DUTTON: On the transfers out, as Fred said, the transfers out can be good if you're saving the patient's life by sending them greater resources and can be very appropriate. It can be bad if you're doing it for gamesmanship

to avoid getting dinged by a publicly reported measure.

The perfect measure specification, you would follow that patient. You know whether they lived or died. So it should count. But we should have the outcome counted in both the numerator and the denominator. Obviously, we can do that in some registries. It's harder to do that in administrative data.

So practically maybe that's not a valid answer. And I think that's what the developers said. They're excluded because we can't track them in the data. We don't know whether they lived or died.

What I might suggest -- I'm reassured to hear it's only 3 percent that are excluded on that basis. Perhaps the developers could provide some sensitivity around that. Is there variability across facilities? Are there some facilities that exclude a lot of patients because of that? Does it vary by rural/urban, high tech/ low tech, big/small?

If we could look at that and say, no, these exclusions are happening at random, we would probably be okay with letting this measure go forward as specified. If it's clear that some hospitals really are doing this badly, that would send the opposite message.

CO-CHAIR FLEISHER: Answer from the developer.

DR. OWENS: Certainly, we could do some additional analyses and bring that back regarding transfer status. I will tell you, within the Healthcare Cost and Utilization Project, we actually can follow some patients for some states. And the patients aren't lost. The question is you don't want a patient counted in two different hospitals, right, both the transfer out and the transfer in. But certainly we could come back with some additional analyses.

CO-CHAIR FLEISHER: Recognize that we can vote, but then we will see additional information back. We have post-calls before we make a decision. So it sounds like that is

request from the Committee.

DR. OWENS: This is Sheryl Davies, has another, she is from Stanford and she is the lead contractor for AHRQ. Sheryl, would you like to speak?

MS. DAVIES: Yes, I just wanted to mention a couple of things about the transfers. Again, those analyses are important analyses, and some of them, as Garth has mentioned, we have explored. And some of them as Pam just mentioned, we will be able to explore using the data.

I do want to make sure that it's very clear that the only issue here isn't just the ability to follow-up. There's also, so when we look at transfers out and patients that are transferred, for those patients, they have an attribution issue in that the patients, it could be the care that the hospital previously provided or it could be the care in the receiving hospital that ends up resulting in that complication.

Now recall that we do exclude patients

with a principal diagnosis of the complication. So if a hospital receives the patient for specifically that complication, they would be then excluded from the measure at that receiving hospital. So really, there is a little bit of an attribution for those that aren't excluded in the receiving hospital. Certainly for those that are transferred without a complication, there is an attribution issue.

MEMBER MOYER: Really quick. Is that admitting diagnosis or discharge diagnosis?

DR. OWENS: These are discharge diagnoses.

CO-CHAIR FLEISHER: Christopher, Fred, Cliff.

MEMBER SAIGAL: The audience or the Committee has some questions about comparability. I should make a note that this measure could only distinguish about 25 percent of the hospitals at being above or below the threshold. So it's not so great. It's moderate to be able to distinguish between hospitals.

Another question I had was for the developer. The gentleman made a comment that they had information on the transfers in, in terms of how they did. I would like to hear if that was the case if they have done specific subanalyses on the impact of those transfer patients and if they've ever looked at a hospital without that data included.

CO-CHAIR FLEISHER: Comments from the developer?

DR. OWENS: Garth, I think this is a question for you.

DR. UTTER: Yes. I don't have that information right at my fingertips at the moment.

CO-CHAIR FLEISHER: Then why don't you get it to the Committee before the next phone call. Cliff.

MEMBER KO: I wanted to follow up on what Barbara said and what we were talking about in terms of how do we think in the context of this wide spectrum of how these measures might be used. On the one hand, like you said, there's

quality improvement internal in one hospital, like okay, whatever. To pay it, where we're going to get penalized up to whatever, this huge 10 percent of what we're bringing in. Because if we vote towards the mean, then we're going to have really atrocious measures when it comes to accountability and public reporting and all that.

But if we're told to vote for the mean, then we should all vote the same way. But then it's just recognizing that that's an issue of when it comes to these measures where they are out there. And these are the things that we get hammered on. Whether it's a professional organization or the NQF or CMS, that we start using these measures that have some level of scientific acceptability. But it's here at the mean rather than an acceptability for what you would want it to be and a payment where you absolutely do not want to inappropriately ding people.

CO-CHAIR FLEISHER: Comments? I mean I would vote personally on whether or not I think

it's a valid measure. I mean it's that simple. I think that within a 5 percent -- the fact that they implement where they take the top and the bottom and that there are people at the edges who make it, heard or not heard. If I think independent of how it's implemented, there is too much -- the risk model makes me uncomfortable about the validity to show it to the public, not whether or not I'd win or lose at the margin, then I would vote on low validity.

MEMBER KO: So is that what you're saying? So we should think about this in terms of like you said showing it to the public, this data?

CO-CHAIR FLEISHER: Right.

MEMBER KO: Is that the mindset we should have? Is this measure good enough to show on a website for 10?

CO-CHAIR FLEISHER: About the validity. I mean that's how I personally vote. Elisa, do you want to?

MEMBER KO: Because that's a different

level than just saying, all right. We're going to use it in our hospital in our department.

CO-CHAIR FLEISHER: Yes, I think whether or not I'd feel comfortable that it's valid. Whether or not I was at the 24th percentile versus the 26th percentile, whether I'd get a penalty or not, I don't think that's how we should vote. That's my gestalt, but --

MS. MUNTHALI: That is true. You should be looking at the scientific merits of the measure. We always say that our criteria should be use-agnostic. But we do recognize that it's very difficult to have these conversations about the scientific merits without thinking about the implications and the application of these measures. With that said, every measure that comes to NQF must be suitable not just for quality improvement but also for accountability purposes.

What you're looking at here is the questions around validity. There is another criterion on use and usability. And that's where

this discussion should take place.

CO-CHAIR GUNNAR: Yes, if I could only jump in. I think this is an issue, to give the 24 and 26 percent comment, which is if I get dinged on this measure, is it valid data? That connects the dots, right?

So we're voting on the validity of the measure itself. Its application is downstream but tightly connected to that. I don't think that's lost, your bias, Lee. But I think if we can put ourselves back in the bubble here and just vote on the validity of the measure independently, then the rest will happen.

CO-CHAIR FLEISHER: I'm trying to say I shouldn't be biased of whether or not I'm at the margin. I should be biased about whether or not my data looks valid and forget the fact that they happen to use these arbitrary cut points at CMS.

CO-CHAIR GUNNAR: Independent of that, you want to know that if I was at a cut point, somebody made an independent cut point, that what

you're being measured on is actually valid information.

CO-CHAIR FLEISHER: Correct.

CO-CHAIR GUNNAR: That's what we're trying to --

CO-CHAIR FLEISHER: That is correct. Barbara, do you have a comment, and then we'll go --

MEMBER LEVY: I do. So I just want to bring us back again. I'm a very practical person. For this particular measure, we have to vote whether we are confident that both the risk adjustment and the exclusion of a death that occurs related to the admitting diagnosis, or the reason the patient came in, are sufficient to make this a valid measure. And that's really where the rubber meets the road.

Is the risk adjustment and the exclusion of people who die -- if it's a saddle embolism and they're sent to a tertiary center, quaternary center, to do management of that very, very dangerous complication and the patient dies,

is that exclusion sufficient for us to have confidence in the validity?

CO-CHAIR FLEISHER: So Christopher, Karl, Amy, Larry, and then we're getting to the end of the one hour we had dedicated. And we're still only on 2b.

MEMBER SAIGAL: Yes, I would just say given the comments we have, you have to judge validity against all uses. I think you have to go to the highest criteria you would use for validity. And I have not seen validity data presented regarding the false positive rates for this measure that would be used for payment. The evidence isn't present.

CO-CHAIR FLEISHER: Remember, vote your conscience and what you believe, and the way Barbara nicely framed it, how valid you think it is, and if there are questions, the developer is clearly hearing some of the concerns here as far as whether they need to come back. There's always an opportunity to readjudicate should that change.

Karl, Amy and Larry.

MEMBER BILIMORIA: For this particular measure, I want to be clear there are concerns about administrative data, and we can have that higher discussion. But this particular measure, I think we're hearing about a number of validity threats. There's the POA transfer status, the risk adjustment. And some of it is amplified because the denominator requires these complications to be included.

And all of the complications have research around them that shows that there may be issues, and not just in one hospital's chart reviews. I mean Cliff said it's in hundreds of hospitals. I think that really is a threat to validity.

I think one of the other things -- and I'm biased on this because our team does the bulk of the research on this -- is we've shown that VTE is not a valid measure, and it's all about how much you look. And hospitals will end up with more cases in the denominator here because

they are doing a good job looking for VTE.

Those are the hospitals that perform better on other objective measures of quality. So you've seen VTE dropped as an outcome measure from a number of programs, most recently, U.S. News, UHC. I think that trend is continuing with a number of other groups. States are dropping it. I think having VTE here is a validity threat.

CO-CHAIR FLEISHER: Thank you. I love this Committee. It's the most academic committee.

MS. DAVIES: This is the measure developer. Can I just mention something about VTE, because I think it's important? This measure just recently went through the Patient Safety Committee, which recommended for endorsement the VTE measure. The denominator for this one does closely align with that PSI 12.

It is important to note that in that submission, and we're happy to provide additional information. I don't want to spend the time to

go through all of it right now. It does specifically address the issue of surveillance.

And there's also, just remember, there's also competing I guess polls on hospitals to reduce their PSI rate in order to perform better on complications measures, whether they be the PSI specific measures or other measures, as well as those that would work to increase their denominators with less severe cases in order to improve their PSI 4 rate. That does provide a little bit of a cross-check I guess on hospitals when you're reporting both of them.

CO-CHAIR FLEISHER: Great. What I would ask is rather than continuing to question individual threats to validity that if there are comments to help us vote, because those individual threats will give you your personal determination of its validity. That's how you should vote.

Amy and then Larry.

MEMBER MOYER: I was actually going to focus us back in, that we're not being asked to

evaluate validity in a vacuum. There are NQF criteria standards around this. They're hopefully summarized in algorithm number 3, which walks you through the questions that go against the NQF criteria and then how they judged the validity according to those set criteria. And it's outlined in the measure worksheets.

We're not being asked to just, I don't know what's validity, and how does this fit? It's really pretty straightforward.

Now if we don't agree with those criteria, then I'm sure there's some method to address those or revise those in some future state. But they are out there, and that is my understanding of how we're supposed to be guiding our voting informed by our expertise.

CO-CHAIR FLEISHER: Thank you for saying that. I'm watching Karen smile. Probably about every two years if not yearly, annually, we actually reevaluate those criteria which is why they're so complex.

Larry, last comment. And then we're

going to call for a vote.

MEMBER MOSS: My question is for NQF staff to help give me guidance for voting. I thought Barbara articulated very clearly the distinction between scientific acceptability and use. And I understand the importance of distinguishing those two.

However, scientific acceptability is not binary. It's not yes or no. There is a spectrum and we all have to determine individually our own threshold for acceptability on that spectrum.

I have difficulty coming up with that point in the spectrum for me without taking into account how that measure might be used. Can you give me guidance on how to vote?

CO-CHAIR FLEISHER: I think we use the -- I think Amy said it really nicely. We go through the criteria for validity, and then you determine. But do not determine -- do not look at the question of the cut points that CMS is utilizing for payment just from the criteria.

So that gets back to -- can we put up the algorithm please?

MS. SKIPPER: And you also have a copy in your packet. But we'll also display it on the screen.

MS. JOHNSON: And while you're pulling that up, I will -- Amy, you did a great job. Some of you guys know this criteria almost as well we do which is great.

Just remember that you guys talked a lot about what we would call here at NQF data element validity, going back and validating those data elements against the records.

That is something that we would love to see here at NQF. It is not something that is required. So we require either testing at the data element level or at the score level. So we would again love to see both, but we don't have to have both.

And I think that will be clear on the algorithm. Do you want me to walk through this, or --

CO-CHAIR FLEISHER: Yes, please.

MS. JOHNSON: Okay. So let's pull up the algorithm. The first question is, are measure specifications consistent with the evidence? If not, then you rate it as low, and you're done.

If they are, then were all potential threats to validity addressed? In other words, did the developers tell you things about exclusions, about risk adjustments, about missing data, and how they handled it.

If the answer to that was yes, then you go down and ask, box three, was empirical validity testing conducted using the measure as specified with an appropriate statistical test? And the answer to this I believe was yes because they did empirical data testing.

So we would go down to box six for that. Then the next question is, it relates to the level of testing that was done. So was testing at the score level done? And in this case again, the answer was yes.

That takes you over to box seven. Was the method described then appropriate? I believe they did a construct validation kind of testing. And tell me if I'm wrong on that, team. I didn't look as closely at the testing. So it was construct validation.

So that takes you over to box eight. And there you look at the results of that testing. And that testing hopefully would tell you if you have a high certainty that you have a valid indicator, a moderate certainty or a low certainty.

I will say that -- and I don't know your name over here, but you're exactly right. It's not a yes or no. It definitely is on a continuum. So you have to vote in a way that you feel is most appropriate. But I believe if there is nobody from the project team or anybody else disagreeing with me, you need to look at box eight. Your choices will be high, moderate or low.

MEMBER SAIGAL: Karen, didn't we also

say that we didn't see validity testing for the very basic, the false positive/false negative data?

MS. JOHNSON: Right.

MEMBER SAIGAL: So I don't think we can go that far. I think we may have to stop much earlier than that.

MS. JOHNSON: Well, the false positive/false negative is what we would call data element testing. That's where you check your claims against the medical record. Right. That if you go down box -- I believe it's box 10. So we don't even get to box 10.

If they have done the score level testing, that is enough to potentially meet our requirements. In other words, we don't have to see that to be able to say that it meets NQF requirements for --

MEMBER SAIGAL: Score level testing wasn't done like that. They didn't do -- They tested the scores, at least in what they presented, they looked at the ability of the

score to discriminate in risk adjustment and looked at an omega-statistic. That wasn't really -- the score testing wasn't done at the level of specificity and sensitivity of chart review.

MS. JOHNSON: Okay. Can we pull --

MEMBER BILIMORIA: This is only talking about what they did. I mean there's other evidence available, and that's not in the measure packet. So how we reconcile that part?

MS. JOHNSON: Right. So you have to look at what they have submitted to see if it conforms to our criteria. I definitely hear you. You would like to see this yes or no data element testing. And hopefully the developers have heard you.

MEMBER SAIGAL: Score testing. Score testing is what I'm talking about.

MS. JOHNSON: Okay. And let's pull open the -- Bear with me a little bit -- validity testing section. And let's just go to the PA section if you will. Let's look at --thank you. Let's see. I apologize for not having done my

homework on this one.

Okay. So what they did was they -- I'm looking at that very first -- sorry. I'm looking at the PA. In 2007, there was some construct validity. There was some face validity, which we still allow.

And then updates to testing. So the most recent thing, the testing has been updated since the last review by using the HCUP data.

Teaching hospitals --

MEMBER SAIGAL: We discussed this already. We have reviewed all of this, and basically they didn't do any kind of false positive/false negative testing on the score.

MS. JOHNSON: Right. It doesn't have to be the score. The false positive/false negative we would consider at the data element level. So what they did is what we would consider score level testing. And we would generally call it construct validity testing.

MEMBER SAIGAL: That's not what everyone else thinks about validity though.

Basically, the concern that people in the room have had is, is this going to pick up things that didn't happen primarily and unseen evidence about that and the score?

MS. JOHNSON: I understand. Right. They would not have it on the score. They would need to be doing that at each individual level test, right, at each individual data level element. You're saying, did they actually have the complication of sepsis? Is that what you're talking about?

MEMBER SAIGAL: The score is determined by that.

MS. JOHNSON: Right.

MEMBER SAIGAL: So score level testing would include understanding whether the score was right or wrong. I just don't see that. Maybe I missed it, but I don't know if co-developer looked at that and saw that. But I didn't see it.

MS. JOHNSON: The typical way that you would test at the score level for validity is to

do things like correlational analysis or construct validation. Basically, the idea is to see if the scores are tracking in the way that you expect. It's not always easy to do because there's not always great variables for other measures to correlate against.

You have score level testing. You do not have data element level testing. So you do not have the thing that tells you that if the claim said sepsis, it was actually sepsis. Or if the claim said sepsis, the claim -- let me flip it. If the medical record said sepsis, does the claim say it?

They did not provide you that information. You can certainly ask for that information. I'm just pointing out that on the algorithm, we accept either type of testing, score level or data element testing.

CO-CHAIR FLEISHER: Are we ready to take -- well, we are ready to take a vote. Yes, A.J. Microphone please. Thank you.

MEMBER YATES: This is maintenance

and I don't want to open up a can of worms. But the SDS question wasn't addressed, disparities, SDS. Is that because it's sort of a separate argument within the hospital, that it doesn't have any application in this case?

CO-CHAIR FLEISHER: There has to be a theoretical construct. For most in-hospital measures, the answer is that -- I think the Disparities Standing Committee, right, Marcia, has pretty much said the in-hospital there is much less of one?

Perfect. There will be lots of questions for you. Some after the break. But the question is, this is an in-hospital measure, and A.J. asked they simply state that there is no theoretical construct to apply SDS. Do you want to give any comment?

DR. BURSTIN: Yes. So what the SDS Committee proposed and said, and what we've moved forward with, is the idea that to move forward with SES or SDS adjustment, you should have both a conceptual basis for it as well as an empiric

basis. That conceptual basis could be based in literature, for example, a clear relationship between an SDS factor and the outcome. Or it could just be logically there is a relationship between them.

At least what we've seen to date is most purely inpatient measures, where all the control happens inside the context of the hospital, unless there is something unusual clinically that would explain a logical relationship to an outcome, has not usually been deemed appropriate to have a conceptual basis for adjustment.

MEMBER YATES: That's why it was only a point of order. I didn't want it to go without mention.

CO-CHAIR FLEISHER: Thank you. And we'll see that later on some of the measures you'll be discussing. So we'll get back to the real fun that you're going to have after the break. Can we call for a vote?

MS. SKIPPER: Yes. We're now voting

on validity for Measure 0351. One, high; two, moderate; three, low; four, insufficient.

(Voting.)

CO-CHAIR FLEISHER: We need one more.

MS. SKIPPER: Voting has closed. Measure 0351, 0 percent votes high; 39 percent votes moderate; 43 percent, low; 17 percent, insufficient. This measure does not pass on validity, and we will stop at this point. The measure does not go forward.

CO-CHAIR FLEISHER: Okay. What I would suggest is we take a break. I will confer with staff whether or not, because the developer will likely come back. I can probably assure you they will come back with some answers. So we may discuss a little bit of other issues that the developer should be aware of. But why don't we take a 10 minute break.

(Whereupon, the above-entitled matter went off the record at 10:45 a.m. and resumed at 10:59 a.m.)

CO-CHAIR GUNNAR: All right. We'll

begin gathering back again to try to get back on track here.

CO-CHAIR FLEISHER: Okay. We all set? I don't know if I'm going to make my comment, starting on time but procedure --

(Simultaneous speaking.)

(Off the record comments.)

CO-CHAIR FLEISHER: Okay. As an anesthesiologist I do have to say that we tried to start on time, but the procedure was longer than booked. And I think this has happened every time with this group is that we should try to make the comments succinct and what data we need from the developer.

We were discussing this. This measure has not passed one of the must pass criteria, but I think what would be very helpful because I'd say there's a 99 percent chance that they will come back with a lot of answers to our questions is if, Christopher if you could and Amy, just go through other questions that you may have for the rest of the criteria.

And unless somebody has any specific comment about something they may want the developer to come back with, then we'll move on to the next, and I'll turn it over to Bill.

And then at some point, Helen can comment about the question that was raised about registry versus electronic. But Christopher, any other critical issues that the developer must address going through the other criteria?

MEMBER SAIGAL: These are on feasibility and usability, comments about that. A lot of it's feasible, obviously. It's been used and it requires AHRQ software to use it, but that's free.

The unintended consequences are relevant to what we discussed earlier. But that's no new data. Just the data we discussed needs to be presented.

MEMBER MOYER: No, as I said, we've run these measures. It's very easy to do. I know it's widely in use. The only unintended consequences I have personally seen is, and I

think this is more on the people reporting the measure, I have sometimes seen patients misinterpret what it is this measure is reporting.

But I think that's more about clearly explaining from a reporting perspective and not on the measure developer.

CO-CHAIR FLEISHER: Any other comments on any other? So I think we are done then. Helen, do you want to say anything now about registry versus claims data?

DR. BURSTIN: I mean, I missed the conversation so I don't want to repeat a lot of what was said, but I think again just you need to evaluate what's in front of you. It's hard to evaluate something to a theoretical comparison that's not before you.

So we would just ask you to look at what's presented, look at the data presented, look at the reliability and validity of the measures in their own right.

Whenever we can get to better data

sources, that's always our goal, but I think we've seen that in the interim at least we are still relying fairly heavily, at least in the safety arena, on claims based measures as others come into being. But again, we just encourage you to look at them on their face value for what they are, not in comparison to something which we don't yet have before us.

MEMBER MCCARTY: Earlier in the discussion we were talking about separating the quality from the accountability, but my understanding is that what happens once we endorse measures and put them out there in the library for anyone to see is that anyone then says oh, this was endorsed by NQF. This is a great measure. Why don't we use it for these purposes?

And I understand that we're supposed to look at them separately, but doesn't that create some risk and accountability on our part knowing that when people see our stamp of approval, they feel like they have the green

light to go forth and use them kind of at will? And so how do we reconcile that?

DR. BURSTIN: We did bring together an expert panel to specifically look at this question of whether measures should be endorsed for their specific applications and uses.

And ultimately, they didn't feel like there was enough there to do that yet. We don't have logical cut points, for example, for when a measure would be useful in one application versus another.

But we do have a whole other process, which is our Measures Application Partnership. So as measures come forward and they're being considered for different federal programs, endorsement is one thing they will consider as a factor, but then they will also look at the characteristics of the program, the characteristics of the measure and see if it's a good fit.

So I don't want you to feel like what you're doing is sort of the last step in the

process. You're really looking at the measure in terms of the measurement properties. Is this overall a good measure?

And I think, keep in mind there are other groups and some folks around the table sit on both groups, will have that discussion as an application comes forward for should we be using this, for example, in value based purchasing for hospitals.

Should this be part of the MIP program? There will be groups like you, clinician workers, hospital workgroups and the overall coordinating committee of the MAP who will then make that recommendation based on what you have said in addition to having the information in front of them of how the actual payment program works because without having that information, it's hard to make that assessment at this table.

CO-CHAIR GUNNAR: I guess one thing back to NQF would be recognizing the NQF measures are used in performance, with specific to

validity, is it -- and back to what was said a second ago, don't we have a personal reflection regarding or evaluation of whether if this is used as performance is it valid and fair, which is -- I mean is the validity of this measure such that if used in performance there is fairness to that?

DR. BURSTIN: You would look at it across the all the criteria, and I think at the end of the day if there's a sense that you would want to ensure that this measure, if used for one of the accountability applications, works. But that would be based on your criteria for reliability, validity and evidence.

CO-CHAIR GUNNAR: Perfect, so to go back to that initial comment, I don't think they're dissociated. I think what we've gotten from guidance here is they actually, in relationship to validity, they actually are connected.

DR. BURSTIN: Yes. There's distinction between a measure that's currently

used for -- we don't always know, for example, when a measure comes forward to us what the actual use will be.

Sometimes it could be measures in use for years and quality improvement only, could be measures like come to us and within a year they're in a pay-for-performance program.

We won't necessarily know that, which is why at these tables you want you to stick to the criteria that are before you and not -- we can't sequester you like a jury and say don't think about all these other things, but this table is really constituted to specifically look at the measures against the criteria.

CO-CHAIR GUNNAR: Thank you. A.J.?

MEMBER YATES: Yes, and just a follow up question generically. Some of the measures used, technical expert panels and committees for face validity or consensus validity statements and the precursor to this process, which is the standing committee but the committee that was brought up ad hoc before would have to vote on

usability.

And in those situations when the face validity was given by a consensus vote or by -- and usability was given a pass by the ad hoc committee and the end use was public reporting only, is it fair to ask if that's a valid -- remains a valid face validity point to bring up six years later if the end use has changed?

DR. BURSTIN: You bring up a really important point. We've had lots of discussions about whether face validity is the appropriate floor. It's certainly not the ceiling for validity, but is it a reasonable floor?

And we've worked a lot with developers. We've worked a lot with end users, and it's very difficult, particularly in the outcomes space to do validity comparisons to another outcome to -- really coming up with other empiric methods of assessing validity of outcomes.

So many of us aren't terribly satisfied, to be perfectly honest, with face

validity. It is, at least in our current process for our current criteria, an acceptable floor. It is not something that would get you high validity, but it will at least get you moderate.

And one of the recommendations that did come out of the committee that we did convene around intended use is we need greater transparency so people can, in fact, see this is a measure that was face validity only but passed based on the strength of other criteria. And we will work make sure that's available.

CO-CHAIR FLEISHER: Thank you. Colleen?

MEMBER PITZEN: I just wanted to add on Helen's comment. We spent a lot of time on the validity criteria today, and I'm just wondering if there's any future thought about requiring data element level validity testing, especially when many outcome measures don't have great statistical score testing that can be done.

CO-CHAIR FLEISHER: Thank you for that question. It's obviously stimulated some

discussion up here. And again, I participated. Helen ran a phenomenal committee that had multiple stakeholders. We discussed this issue, measures for intended use including CMS at the table. So it was quite robust. It's your turn.

CO-CHAIR GUNNAR: So the next measure for discussion is 1550, hospital level risks, standardized complication rate following elective primary total hip anthroplasty and total knee anthroplasty, Centers for Medicare and Medicaid Services.

Developers are on the phone, or they're here. Very good. My apologies. They are coming to the table. Please introduce yourselves, and you have three, four minutes to introduce.

DR. SUTER: Thank you. I'm Lisa Suter. I'm a rheumatologist and associate director at the Center for Outcomes Research and Evaluation, and I really appreciate the opportunity to present our measures today.

I'm here with Karen Dorsey. On the

phone we have Dr. Leanne Hahn from the Centers for Medicare and Medicaid Services as well as Sophia Chen at CMS and Jeph Herrin who's an analyst who's worked with us on these and other measures.

So as you introduced, we're talking today about the 30-day unplanned readmission after elective primary hip and knee replacement procedures as well as the complications measure that is paired with it.

The complications measures, the events of AMI, pneumonia, sepsis that occur within seven days of an elective primary hip or knee replacement. It also captures death, pulmonary embolism, surgical site bleeding within 30 days of an elective procedure, hip and knee replacement procedure, as well as mechanical complications, prosthetic infections within 90 days.

Those 30 and 90-day events are associated with hospitalization, not death but certainly surgical site infection. Mechanical

complications and periprosthetic joint infections are only considered complications if they are associated both with an admission as well as with the diagnostic as well with procedural code for either revision or debridement.

These measures were put in front of NQF in 2012 and were endorsed and we're back for endorsement next. They've been in public reporting since December 2013 for the complication measure and July 2013 for the readmission measure.

The complication measure received medical record validation of its complications outcome, and both measures were overseen through intense involvement from a clinical expert panel as well as a formally engaged, diverse technical expert panel of stakeholders.

There are many topics that we could choose to spend our remaining two minutes to talk about, and we're happy to talk through some of the issues that were brought up earlier today.

But I think based on some advice,

we're going to focus a little bit on the sociodemographic status analyses that we presented as part of the endorsement maintenance application. But we're happy to address any and all concerns from the committee.

So we know that these measures are somewhat different from the other CMS sorted measures that have gone in front of NQF in that they address elective procedures where surgeons have an opportunity to filter access to the procedures, which is somewhat different than the acute medical conditions that are measured with other CMS measures, so we understand that SDS may play a more distinct or unique role in these measures than they do in other measures.

In order to look at SDS in these two measures we followed the same analytic pathway that we did for our other measures. We investigated a number of aspects of the relationship between SDS and these outcomes.

We aim to answer the extent to which providers with more low sociodemographic status

patients perform worse on the measures. The extent to which there's a relationship between SDS and the outcomes for patients, particularly within the measure's multivariable models.

We investigated the influence of risk adjustment for SDS on hospital level scores and whether the influence of SDS was primarily a patient or a hospital level effect. And those analyses are presented in the testing forum.

To just summarize them, we do see that hospitals serving high proportions of dual eligible African American or low AHRQ SES Index score patients, those are the three measures we were able to find national data for, have largely overlapping performance distributions although the median outcome rates are -- there's a slight increase in outcome rates of 0.2 percentage points for hospitals with high proportions of underserved or low SES patients.

When we include any of these variables in a multivariable model for the readmission or complication measures we see model level effects,

odds ratios that range from 1.07 to 1.22.

Those effects are attenuated in the multivariable model from univariate analyses, so there is some effect that's being captured by the clinical variables that are in the multivariable model but they remain statistically significant with up to a moderate level of effect in multivariable models.

When you include SDS, any of these variables, dual eligibility, race, African American race or low AHRQ SES in the risk model the median change in a hospital's outcome rate with adjustment is around two one-hundredths of a percentage point change for both of those measures, regardless of the variable used.

In decomposition analyses, clinical risk variables for these measures are more influenced by patient level components as we would expect, and the SDS variables are more influenced by hospital level factors.

This does not explain the discordance and the disparities, but we do think that it

raises the possibility that there are both patient and hospital level effects playing in this situation, and it's hard to tease those out even with our extensive analyses.

We've also looked at the disparities over time. These measures have been in public reporting since 2013, and we're happy to share these results with you. They are actually publicly available in the annual chart book -- medical chart book publication that CMS posts.

But since 2013, there has been no change in this median difference between hospitals serving high proportions of low SES patients versus those serving high proportions.

It's about a 2.2 percentage point difference, which as I stated originally, is the difference we see in the current measure. So we're not seeing a worsening of disparities with the public reporting of these measures.

We're happy to add any discussion, and we're also happy to do additional analyses. We look for guidance from the committee if there are

specific analyses that you're interested in seeing. I think there was a concern about the variation of some of the different variables that we investigated for the SDS analyses.

The reason we chose dual eligibility, so Medicaid and Medicare status, African American race and also the AHRQ SES Index is that those are available indices and validated indices in terms of the AHRQ SES Index.

They're available on all Medicare beneficiaries. The AHRQ SES Index utilizes a nine-digit zip code, which takes us down to the census tract level, and there are about six to 12 census tracts in any given five digit zip code. So it's a more granular division of geographic location than a five-digit zip code.

I think the other thing to acknowledge is that while we're not recommending SDS risk adjustment, we understand that this will be a robust discussion today.

We also want to flag that in this year's IPPS rule, CMS signaled that they will be

exploring stratification of measures in the future, and so we do think that there has been formal signaling from CMS that stratification is something that they will investigate for its appropriateness for use in these measures. So I'll stop there. Thanks very much.

CO-CHAIR GUNNAR: Thank you. Our discussants are Dr. Yates, and Dr. Cima is on the phone I believe.

MEMBER YATES: My wingmate is practically in the air I would hope by now.

CO-CHAIR GUNNAR: So you're flying solo.

MEMBER YATES: Well --

MEMBER CIMA: Oh, I'm here.

MEMBER YATES: Good to hear from you. I should have taken Delta being in the hot seat, but I'm going to take the lead if you don't mind on this Robert because I was led first. Or do you want to go?

MEMBER CIMA: Oh, sure.

MEMBER YATES: The description has

already been given. I'm following this script. This is a maintenance measure. The critical thing is that the evidence has not changed. Again, this is in regard to our last conversation. The use has advanced, however.

And just to put it in perspective but not to make it a part of the voting, the utilization of this measure is in hospitalcompare.gov.

It's also used in the value-based purchasing with estimated effect in terms of payments in 2019 currently being selected for a look back at that time and in the Health Care Planning and Action Network they have as a consortium decision decided to use this measure and promote it for utilization in the private sector as well as probably for Medicare Advantage when they go to value-based purchasing as well. But that's in the future and that's conjectural.

It should be noted that the distribution of the rate of complications is relatively rare at the median. This is something

that's happened at 4 percent, and as an arthroplastic surgeon I'm happy to say that it does happen rarely and that the distribution is tight, that the bell curve is high and steep between two and six for the most part.

CO-CHAIR GUNNAR: Can I stop for a second?

MEMBER YATES: Yes.

CO-CHAIR GUNNAR: So your comment regarding evidence, since it's a maintenance and our new process is to take a hand vote on whether anybody believes we should have a vote regarding evidence.

MEMBER YATES: I think it's reasonable that the evidence is unchanged and still reasonable to vote for rather than it's being acceptable.

CO-CHAIR GUNNAR: Anyone want to oppose that? Hearing no objection, we will move past evidence to the gap, which is --

(Simultaneous speaking.)

MEMBER YATES: There's a subset of the

evidence section which has to do with SDS disparities, but in conversation with Dr. Fleisher beforehand, we decided that we would put that discussion into the reliability section in terms of risk adjustment.

CO-CHAIR GUNNAR: Very well. So we'll stick to gap. Now we have to vote next on gaps, so if you'll discuss the --

MEMBER YATES: I would say that the gap is reasonably sufficient. We would like to see these complications be zero, and we would like to see every plane that takes off land and have that kind of effect in our arthoplastic.

CO-CHAIR GUNNAR: From a quality improvement point of view and taking sort of the simple approach, do you think this measure has enough influence on the quality improvement of an organization.

MEMBER YATES: Yes.

CO-CHAIR GUNNAR: -- that the current gap in performance is such that it's still not topped out.

MEMBER YATES: Yes, sir.

CO-CHAIR GUNNAR: Very good. Do we need to then vote on gap? We do. Set us up.

MEMBER YATES: And then under evidence I would just add one sub note, and this just to advise the developers and stewards. It's just fascinating to me that the complications don't have some sort of weighting.

I think that there are incidental pneumonias that get better in a few days and were only picked up on a fever work up that shouldn't be weighted as much as perhaps a deprived periprosthetic infection.

And so the lack of some sort of Delphi process by patients looking at the weighting of these complications would be a valuable insight in the future. And the other thing I would point out is that along the lines of weighting it's curious to me that neither neurologic nor vascular injury, including amputation, are captured by the codes.

CO-CHAIR GUNNAR: Dr. Suter, any

comments?

DR. SUTER: First of all, thanks very much for the feedback about weighting of complications. In terms of the vascular injury, those were raised during development. There was a discussion of whether or not to capture neurologic or vascular injuries.

Part of the challenge in developing this measure was ensuring measure complications that could be attributed back to the original surgery. So for example, we don't have DVT or UTI or some other very common due to the ascertainment challenges with those.

And the feeling was we couldn't drill down to those complications appropriately at this point, but certainly with the increase in clinical data and maybe moving toward EHR data, that could be an expansion of the measure in the future.

MEMBER YATES: Yes, and Lisa, that's said in the context of being weighted. If it were weighted, those would be weighted very

heavily by the patient.

CO-CHAIR GUNNAR: Very good. We ready to -- any other discussion? Carry on. Vote for gap is open.

MEMBER SAIGAL: Yes, voting for gap on 1550 is now open. One high, two moderate, three low, four insufficient.

(Voting.)

CO-CHAIR GUNNAR: Looking for one more.

MS. SKIPPER: And Barbee, if you can hear me, we're now voting on performance gap for measure 1550, one high, two moderate, three low, four insufficient.

CO-CHAIR GUNNAR: Be under validity in terms of risk adjustment. Fair enough.

MS. SKIPPER: Voting has closed for 1550 on performance gap. Thirty-five percent votes high, 65 percent moderate, zero percent low, zero percent insufficient. This measure passes on performance gap.

CO-CHAIR GUNNAR: Okay. Off to

reliability. Dr. Yates?

MEMBER YATES: I would just like to stop and make sure Robert doesn't have anything to add.

CO-CHAIR GUNNAR: Good point.

MEMBER CIMA: No, I mean one thing that's notable is that the performance has improved over the two measurement period that they've been using, so this is obviously having some type of impact.

CO-CHAIR GUNNAR: Very well. Ready to move on to reliability.

MEMBER YATES: It should be noted that the calculation of the numerator, denominator it's predicted to the number of expected admissions with a complication multiplied by the national observed complication rate.

That smooths things out a little bit, but bottom line is that the numerator is the number of the listed complications which are time-dependent in terms of their occurrence. The reliability testing, there is some noise and the

measure an ICC score of 45 percent. It would be ideal to have a higher score than that.

The c-statistic for risk adjustment is only 0.65, and that has been a constant over the last three or four years of different measurements of this over time. And that seems to be a fixed statistic.

It has been noted that if certain orthopedic-specific risk factors are added to the risk adjustment that the measure can, and this has been published, can be raised to a c-statistic that is 0.7 or above which would make it more acceptable.

And that's something that the developers are well aware of and are working with them on that from other professional societies. The next paragraph that I had written, and this is not Robert's, but I was worried about the question of tiered validity.

And tiered validity has been taken off of the table. I still would argue that the c-statistic and the relatively low ability to risk

adjust does have something to do with the potential for unintended consequences, which we'll come to later.

But for the most part I would say that the reliability has been proven, and I would advise the panel to vote for reliability.

CO-CHAIR GUNNAR: Dr. Cima?

MEMBER CIMA: Yes, that was my major concern. This was a hospital-based measure. You're looking for patients to try and decide where to go from there, and this is a very -- 0.45 the developers say is moderate, but that's basically flipping a coin.

So I have real problems using hospital-based comparisons. One of our goals is to allow these measures for people to make choices, and they really, really can't with this measure with that data.

Although Dr. Yates has recommended that we pass on the reliability, I really have a hard time saying that we're endorsing a measure where it's like flipping a coin. And that was

what I was hoping in my comments to get people to discuss that. Is that really a reliable measure for a hospital-level measure?

MEMBER YATES: Dr. Cima, I would agree with you that it's disappointing that it's not a better risk adjustment. And I'm just happy that it's risk adjusted at all given --

MEMBER CIMA: But our goal is to get a measure out there to the population where people can make informed decisions. And this is a hospital measure. So it's saying this hospital compared to this hospital as far as hips and knees go, for primary hips and knees, this is the difference.

But there's a great amount of inaccuracy in the measurement, so is this a patient-centered view? Is this really giving the patient adequate information to make an informed decision? Or am I just going to send them a quarter so they can flip between two hospitals?

MEMBER YATES: And I agree again with you. You're using the example of the end use,

and if I was allowed within the rules of engagement to use end use as an important criteria, I would agree with you even more. And I would say it's a very debatable point, but I will be quiet now and let people speak.

CO-CHAIR GUNNAR: Can I ask a naive question? These total knees have move to ambulatory environments, so if you're an ambulatory center and then you get admitted with a complication to another hospital, does it track?

MEMBER YATES: I can answer that because currently Medicare does not recognize a total knee replacement as an outpatient procedures. There is a proposed rule that came out about a month ago commenting on their possibly allowing for it to be an outpatient procedure, but currently --

CO-CHAIR GUNNAR: So it won't be --

(Simultaneous speaking.)

CO-CHAIR GUNNAR: Got it.

MEMBER YATES: At this point in time,

it's a moot point until they come out with a final rule that would say that they're going to accept total knees as an outpatient procedure. Not to interrupt --

CO-CHAIR GUNNAR: No, it's perfect. That's the answer I wanted. Thanks.

MEMBER CIMA: Well, I just wanted to go back to one point that Dr. Yates raised. It's about the usability, but if I look at the measurement sheet, the questions for the committee, there's three questions.

So the test could be an adequate sample, yes, but then it goes, do the results demonstrate sufficient reliability so that differences in performance can be identified. And then in the next one, do the results demonstrate meaningful differences in performance that could be identified.

It's unclear if an ICC of 0.5 is adequate to meet those two questions. That's what I was going at, not the end result. I'm just saying what the end result would be, but it

does address in reliability these very two questions.

CO-CHAIR GUNNAR: Dr. Suter? One minute, please.

DR. SUTER: Okay. This is Lisa Suter. So I think I heard two specific concerns, one about the reliability testing, specifically about the test, retest. So the value of ICC of 0.5 is an established moderate value for this kind of use.

I think we can all assume that if we were looking at a blood pressure value and we knew that thee blood pressure was being repeatedly assessed and it varied a certain amount but over a mean of values you got a reliable sample, we would accept that.

If a measure was being based on a blood pressure result and that measure had a hard line where you assign a quality value to that blood pressure, that may be more of a problem.

What we're looking at here are aggregated results at the hospital level. And

when you split those results at the hospital level and do two random samples and test the hospital's measure result using half of its sample and then compare it to the measure result with the other half sample, you get an ICC that's dictated in the moderate range by established criteria.

So it's not a coin flip that you're getting an association. That 0.5 is not an ROC curve. It's not a c-statistic. It's a correlation coefficient that indicates a moderate level of correlation between the hospital performance values when half of the sample is compared to the other half of the sample at the hospital level. Karen?

DR. DORSEY: Let me just add that we do a very conservative test of reliability of the measures where we're actually correlating the score itself, and when these measures are publicly reported, they are publicly reported in categories that include a confidence interval.

So we hold ourselves to a very high

standard when we're calculating reliability at the measure score level. But there is cushion in the public reporting such that if we were only comparing which buckets hospitals fell into, right, we would -- as Lisa was saying, the correlations would be much high. So just want to point that out to the committee, thank you.

CO-CHAIR GUNNAR: So on this side we have -- I think I saw Rick, Collette, Barry and then Karl and Chris. I'm sorry.

MEMBER DUTTON: Very quickly for the developer, why just patients over 65? Why shouldn't this measure be specified for everybody?

DR. SUTER: So that's a great question. Specifically this has been implemented with Medicare beneficiaries, and for Medicare beneficiaries under the age of 65 your qualifications for Medicare are usually disability and dialysis.

And we -- this measure was originally specified testing just above 65. We have

validated it in all-payer data and shown that you do not see substantive differences, nor do you need to make substantive changes to the risk model other than changing your age variable in the risk model to have it applicable in all-payer data under the age of 65.

MEMBER DUTTON: So why not specify it for everybody?

DR. SUTER: We're specifying it for the use that we have systematically tested. We have not followed up with all-payer testing on a systematic basis, on an annual basis.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: Just a question. So in terms about reliability score testing, the staff had questioned the split-sample methodology and I'm a statistician.

We use a beta-binomial statistic to compare our medical groups when we're doing this kind of comparison in Minnesota, so I'm wonder what the staff concerns were with the model and maybe there isn't any concern.

DR. SUTER: Does the staff want to address that?

MS. JOHNSON: I can give it a shot. So they did do the split-sample methodology, and that's something that we have to date accepted as an appropriate methodology. And they have a moderate range correlation.

We put out the 0.7 threshold as just a rule of thumb that's out there but do not consider that as if you have to see something greater than 0.7 because that is not the case. It's just something out there that some people use.

The other piece that they did was look at their model variable frequencies and odd ratios over time, and we don't -- while it's interesting, we don't really accept that as score level reliability.

Frankly, we expect numbers to change over time, or at least scores, maybe not so much model frequencies and that sort of thing. So they did two different things. One of them right

now we're saying is an acceptable method. The other we're saying we don't interpret it as an acceptable method.

DR. SUTER: And to just clarify, so from NQF staff's perspective since we've met the measure result reliability criterion where it's an appropriate measure for moderate reliability by NQF staff.

MS. JOHNSON: You have done score level testing, so according to our algorithm you are eligible for high. But the results aren't necessarily stellar, so you don't have to rate it as high, but you could. So that's where it is.

I will say that we are as staff learning all the time, so we're learning about new methods and that sort of thing. And I've actually been in discussions with another statistician, another developer, who is very interested in this idea of stability over time and is trying to talk to us about maybe adding that in or in some way looking at that.

So be prepared. We might at some

point say yes, we see the utility in that and we'd like to see, that sort of thing.

CO-CHAIR GUNNAR: Yes?

MEMBER YATES: For the developers, a clarification question. You mentioned confidence intervals for public reporting. Do those confidence intervals have fine enough detail for the percentile rankings when it's used for the quality metrics in the CJR and when it's applied to the percentile rankings that are applied to the value-based purchasing?

DR. SUTER: So my understanding of the calculations for the CMMI, CJR bundled peanut program are that they do stratify the percentile rankings. They are using the point estimate. They are not using the performance categories that are on Hospital Compare.

I believe that they adjusted their methodology in response to public comment precisely about that concern, but that there is some wiggle room on the point estimates and therefore using point estimates as a sole

measure.

And so that I think that's why they moved to a stratified scoring system as opposed to an all or nothing black line cut off. You either meet a criterion for payment or you don't, in the original proposed rule.

CO-CHAIR GUNNAR: Barry?

MEMBER MARKMAN: Is your data patient-specific? I mean, do you capture those patients that are operated on by one provider in one hospital but then have the complication within the 90-day period but go to let's say a different provider at a different hospital? And what's your confidence in that?

DR. SUTER: Yes, so the measure captures every event for a patient no matter what hospital it occurs at, and every hospital that has a reported result receives a hospital-specific report from CMS every reporting period, every year. And that report gives each hospital every single patient that was in measure, every patient that was excluded from the measure, so

the fracture patients that didn't get measured, and tells them what happened to them, whether or not they had an outcome event.

For the readmission measure, it tells them what hospital they were readmitted to, so they have all of that information. So even if it's not transparent to them at their own hospital that they're having complications in outlying hospitals and they're not coming back to their own hospital, they do see that information.

MEMBER MARKMAN: Very good.

CO-CHAIR GUNNAR: Chris?

MEMBER SAIGAL: Yes, clarification Dr. Yates. You mentioned that there is a c-statistic that was reported, used for the risk adjustment model I assume, that was insufficient, that basically you couldn't discriminate.

MEMBER YATES: It's not fair to say insufficient. It would be considered low. It depends on your definition of what's a good c-statistic, and that's sometimes in the eyes of the beholder. But the literature that at least

I've reviewed would usually like to see a c-statistic of at least 0.7.

The c-statistic in this case is 0.65, which is low, but given the nature of the database is probably where it is right now. And I can leave that to the developers to answer better than myself.

MEMBER SAIGAL: Okay.

CO-CHAIR GUNNAR: Any other comments?

DR. SUTER: We were thinking maybe we would address that with the larger question about validity and the risk adjustment model. Would that make sense to pause for that?

CO-CHAIR GUNNAR: Amy?

MEMBER MOYER: So one thing I wonder about with the reliability, and I know the feedback had come from the MAP Clinician Workgroup that this is a measure we'd love to see at the individual surgeon level because there are probably some factors that are hospitals and factors that are surgeons.

Having seen admittedly non-risk

adjusted surgeon level data from your measure, it appears there is in some cases some wide variation among surgeons, so I don't know if that could kind of muddy the waters in terms of attributing things to hospitals. But that's something that we feel could make this more useful and would like to see in the future.

CO-CHAIR GUNNAR: Dr. Yates, any other comments? Dr. Cima?

MEMBER CIMA: No.

CO-CHAIR GUNNAR: Very well. Ready to vote on reliability.

MEMBER SAIGAL: Voting is now open for reliability on Measure 1550, one high, two moderate, three low, four insufficient.

(Voting.)

CO-CHAIR GUNNAR: Everyone in the room has voted. We're just waiting. Oh, we're good? All right.

MS. SKIPPER: Measure 1550 on reliability, 13 percent votes high, 83 percent moderate, 4 percent low, zero percent

insufficient. The measure passes on reliability.

CO-CHAIR GUNNAR: Moving on to validity.

MEMBER YATES: Several different analyses are offered for validity. The majority are very satisfying. Two that I would just raise questions about, the original technical expert panel was convened, I believe, with the impression that it was going to be public reporting as a process.

And again, they're not going through the NQF process in terms of going all the way to the end as to usability, but that was the original focus of that convened group was public reporting and not the utilization and value-based purchasing. Is that impression correct?

DR. SUTER: So when we originally developed this measure and asked the technical expert panel to weigh in on the NQF criteria of reliability, validity and usability, we presented it with the question of does this measure assess quality in a meaningful way, not necessarily does

it assess quality appropriate for public reporting versus pay-for-performance.

The use was not indicated in any way, so committee members may have made assumptions in their mind in that validity vote just as we have heard in prior discussions here the challenge of teasing out those issues.

But the question was stated explicitly as do you this TEP based validity for a measure of quality, does it measure what you think it should measure, and do you think it's a meaningful thing to measure and unanimously endorse that?

MEMBER YATES: Then the one question that has already come up in other discussions earlier today was the question of the database or the data source, and the data source in this case is administrative data from coding for CMS.

And I will leave it to the committee to decide on the validity of that data, but the one thing I want to point out is the only validity study that was done was just done with

six hospitals, and it was done and reported in Version 2 of the measure.

In Version 2, they took something. There were 319 reported complications out of the six hospitals, and when they went and looked at the charts there were, I believe, 97 discrepancies amongst 86 patients.

Now that's the -- that's a measure of the specificity of what's found. Now when you add in the fact that very few actually know, charts were found that missed a complication.

That would be the sensitivity being added in, and it's reassuring that hospitals weren't hiding complications in their coding. I think that's great.

But when hospitals are being compared for public reporting and for advancement of quality, they really only care about the numerator, which is what complications are captured and are they accurate.

So I would argue that although the agreement may have been calculated to go to 99

percent, which is your accuracy of your test if you will if this was a lab test, it's the true positives versus the false positives, i.e. the specificity of the test that's important.

And when first looked at, there was a 30 percent discrepancy when some of the outcomes were changed and some of the complications were redefined.

That was brought down to 10 percent difference, 30 true complications, but yet there was still a 10 percent specificity discrepancy. And it's the 10 percent that I'm asking the developers to address.

And in addition to that, having changed the complications and given the fact that out of those six hospitals I didn't see a subgroup analysis but there have been some huge, unique intrinsic variability in terms of how they were coding some of those codes or some of those complications that were dropped, doesn't this measure which affects a whole lot of hospitals deserve another test of the validity of the

database?

So the two questions are, isn't it true that the specificity of the test is the more important question here and that it actually is about a 10 percent discrepancy according to the only valid validity study done.

DR. SUTER: Thank you, A.J. So this is Lisa again. So I think it has been -- which has been stated earlier by Barbara is that it's really hard to make a perfect measure. And we, as developers, are humbled by the challenge in front of us.

So what I will say is this has been a learning process. This is a live measure that we evaluate every year. The validation test with medical record data was a process that we learned from, and we adjusted the measure.

It coming in and out for dry run with hospitals, we learned from that process. We made adjustments to the measure based on how we identified fractures in order to exclude those from the measures.

Those were measure adjustments that were made in response to feedback from hospitals in use prior to public reporting. We made changes to the measure in response to NQF's initial committee review where our planned readmission algorithm which has been vetted by dozens of surgeons and other clinical experts.

We included, I think it was angioplasty procedures as an elective readmission and therefore as a planned readmission it was excluded from the readmission measures and the NQF committee said that's not acceptable.

If I take a patient to an elective joint replacement and they have angioplasty within 30 days after that, that's -- I shouldn't have been taking them to surgery, right. That's a complication of care for this elective surgery.

So this measure has been modified over every opportunity to do so. I think our experience is when it comes down it and the NQF committee that voted in response to all of this information last time felt that this was a valid

measure.

It continues to be in use. It continues to, we think, contribute meaningfully to the quality information available for hospitals caring for these common -- patients undergoing these common procedures.

We can debate whether those numbers justify an invalid measure versus a valid measure. I think there are many, as you said, pieces of information to pull together for this validity.

And we're eager to hear what that information is where you get to learn and evolve the measure appropriately so that it stays a meaningful measure as we go forward.

MEMBER YATES: I guess my point is that it has been reported I think in presentation to this committee and also in the Federal Register and on CMS' web page the different times that it's been put into print that it's got a 99 percent agreement rate, which is absolutely true the way it's been statistically used or

presented.

But I would just point out that it's really the false positives that we're worried about, and there's a 10 percent error rate in terms of that defining the numerator based on administrative data sets and which is, quite frankly, better than what studies have shown for other administrative data sets.

So I'm not debating the validity based on that. I would just argue that it ought to be presented as such that it's not -- there's an error rate. It's not 99 percent agreement.

DR. SUTER: Fair enough.

MEMBER YATES: I think point of contention is to sort have been begging the question.

DR. SUTER: Great. Good input. Thank you. I appreciate it.

CO-CHAIR GUNNAR: Dr. Erekson?

MEMBER EREKSON: So it's actually very reassuring to me to see that the 90-day follow up for patients staying in the fee-for-service

Medicare is only at 0.15 percent.

In doing a lot of claims analysis and trying to track patients through Medicare, it's very surprising to me to see how many patients drift in and out of fee-for-service and into the HMO Medicare plans.

And that's just something for the committee to consider and especially it goes to what you've already said which is this can be used in other paired data sets. But even in this Medicare population we tend to think that the patients are staying in Medicare, and they're drifting in and out all the time.

CO-CHAIR GUNNAR: Any other comments regarding validity?

MEMBER YATES: Well, we're now in the question of c-statistic and also disparity or SDS risk adjustment because that's part of the validity. So we've already talked about the fact that the c-statistic is 0.65, and that's been fairly steady over time.

The discussion about the variability

due to SDS or social risk factors, I think that the group from Yale should be congratulated for doing an extensive work on that.

My -- several comments that I would make though is that it's reassuring to see that at the median in the stratification that there's no difference in outcome for all practical purposes because of race.

And I would agree with that because I don't think complications happen because of the color of someone's skin, and I don't think complications happen because of someone being poor per se.

I would argue that the -- those two though do not define the rest of the population that the hospital takes care, and they go to great lengths to show that it's a hospital effect and not that patient.

And I think that's a reasonable assumption from what they showed. However, without getting into the weeds, the problem is that just because you don't -- the problem is

that you do show that there is an increased risk of complications that is measurable between the lowest percentile or the lowest quintile of hospitals that are populated more -- less heavily with African Americans, people with dual eligibility and hospitals that have a lower EH or AHRQ SES score versus those that have a high proportion of African Americans or the highest quintile proportion of African Americans dual eligibility and the AHRQ cut off of 42.7.

My argument is that the rest of the people that are taken care of in that community or in that hospital don't automatically become rich. And unfortunately in our society an African American status in an urban environment is a surrogate for poverty in a lot of cases.

Certainly dual eligibility is a surrogate for poverty, but just because there are other people in that community that aren't dual eligible, they don't automatically become more rich.

And finally, the cut off of 42.7

doesn't describe the entire population. So I would argue that the hospital effect may in fact be a community effect, and my question is, did you test for the spectrum of the AHRQ SES scores to see whether or not it played an effect in terms of -- as a single variable, that being poverty.

Did the poorer hospitals or the hospitals that take care of the larger proportion of the poor, even those are just over 42.7, are those communities not affecting the hospital score? And shouldn't that be something that's risk adjusted?

CO-CHAIR FLEISHER: Just one point. The CSAC has made it very clear that to even address the issue of race with SDS risk adjustment is inappropriate.

It may be disparities is something we want to look at it, so I think I would ask that we actually not go into that issue because it's not something that we should appropriately be addressing with the issue.

MEMBER YATES: Right, and just to clarify, I made that point that I don't think it's an issue.

CO-CHAIR FLEISHER: Right, so --

MEMBER YATES: And I think that actually the data shows that. I think it has to do with it being a possible surrogate for --

CO-CHAIR FLEISHER: Right, so that's what we should be discussing.

MEMBER YATES: Right, and so I'm making that point right now.

CO-CHAIR FLEISHER: Great. Thank you.

DR. DORSEY: Can I just clarify that we did not use race in our analyses as a proxy for SES? Those are two completely separate issues. We use race really as a comparator variable because we believe that many of the mechanisms by which race is associated with outcomes are parallel, not all but some, with mechanisms by which SES is associated with outcomes.

And it helped us to understand and

interpret our analyses to have those types of variables in the model, but we were not using race as a proxy for SES which is the explicit direction of the disparities committee and the SDS trial guidelines.

CO-CHAIR GUNNAR: Any other comments? Barry?

MEMBER MARKMAN: Was the -- well I noticed in one of your bullet points that's removed other post-operative infection which is a pretty common code that's with this complication. Was that -- and now it's just wound infection or just specific periprosthetic joint infection?

Can you comment on that because if you go through the codes, that's much more specific? And was that a result of your audit, or was that --

DR. SUTER: So that decision originally to define the complications in that specific way was based on feedback from orthopedic surgeons and the clinical experts to try and -- that was reinforced by empiric

analyses to try and avoid including complications that were -- that would contribute to the false positive concern -- false positivity concern that A.J. brought up.

And so really we just look at periprosthetic infections or joint infection codes. In addition, you have to have a readmission for that. So it can't just be in the outpatient setting, and you have to have a concomitant surgical either debridement or revision code.

So there is -- it's a very high bar to achieve that, which is one of the reasons this complication rate is very low. I mean I think if you included UTIs in this complication rate, it would be a very different outcome rate.

But UTIs are messy, so those decisions were made with the influence of the clinical experts that helped us develop this measure. I'm happy to address the c-statistic, but do we want to collect other comments first or --

MEMBER CIMA: Can I just ask a

question of the developer about the 90 days for the hip infection? Just as a point of harmonization, CDC is now requiring reporting for one year for prosthetics.

Is there a reason for being different? Is it harder to collector or is there too much noise, or does it not add much? Just because again requirements for institutions for reporting and looking were using different definitions, and I just was wondering why the CDC wants to use one year for this, but you are choosing 90 days?

MEMBER YATES: Robert, this is Yates. I think you may have it backwards because my understanding, and I live in these waters, the CDC actually went from one year down to 90 days as their cut off for the periprosthetic being accounted for by the surgery.

That's the current ruling that I'm aware of. It used to be one year, and now it's 90 days.

MEMBER CIMA: I thought it was the opposite.

MEMBER YATES: Yes, no I mean I look at our surgical infection rates every month, and I know that it's currently -- and we go strictly by CDC standards and CDC standards are 90 days.

MEMBER CIMA: Oh, okay.

MEMBER YATES: So they're actually harmonized with that at this point.

CO-CHAIR GUNNAR: Lee?

CO-CHAIR FLEISHER: So the dual eligible status has an odds ratio of 1.21, and then you make the comment that you decided not to include it in the measure. That's actually I believe a decision of NQF. So can you tell us how you decided not to include the measure since it was, I believe, statistically significant? Correct?

DR. SUTER: So every single variable that we looked at for the SDS analysis, dual eligibility, the AHRQ SES and race, which we were not looking at as a proxy but as an additional variable, are all statistically significant in the multivariable model.

They all have reduced odds ratios from univariate analyses, so they are attenuated by combining them in a robust clinical model. We did not make the decision to recommend exclusion or inclusion of these variables based solely on their odds ratios in the multivariable analysis.

We think it's a much broader discussion. When you include all of those variables in, you don't change the c-statistic from 0.65. I think acknowledging that there are models out there that have achieved higher c-statistics, those models have incorporated patient reported outcomes, such as functional status and position reported concomitant lower extremity disability or pain.

We think those are really meaningful risk variables. We don't have access to them on a national level. They are not adequately coded in claims data, so we can't include them in our model.

We're working with the societies, as he mentioned, to encourage the collection of that

information so that maybe someday we could include them. Maybe we'll get to EHR data well before that that may be more meaningful.

In terms of the decision whether or not to include SDS, I think there are lots of things that influence that. One is it's not -- it does not make a difference. It does not change the c-statistic.

It does not change the median point estimate for a hospital's complication or readmission rate. The median change is two one-hundredths of a percentage point. So for some hospitals, it may change.

The thing to recognize about our measures is they are measures of relative performance compared to the national average. So when one -- when you add SDS into the model, one group of hospitals' rates will go down and another group of hospitals' relative rates will increase.

And so we don't think that risk adjustment for these variables necessarily

addresses the issue that physicians legitimately have concerns about, which is are there things that are not captured that influence but even if you included in the model, we're not seeing a marked change.

And I'm not sure. It doesn't address that, and it certainly takes the ability to see the disparities away, and we do know that there are disparities. There's an article in JBJS that came out last week or the week before that indicated that minority patients have less access to these procedures, and they have higher complication rates.

So I think it is a reality in this situation, and we just think that there may be other options available to us, like stratification that may be --

(Simultaneous speaking.)

DR. DORSEY: Can I just add something? I just want to add specifically to your point that the guidance that we've received and other measure developers received about the SDS trial

was that in making a decision about whether to include SES risk factors or not we had to include -- consider two things.

One was the conceptual relationship, and the other was the empirical relationship. So the empirical relationship alone was not sufficient to recommend that we include the variable.

And the conceptual relationship was that potentially socially disadvantaged patients had a higher disease burden. If that were the case, then our current model would adjust away the independent association between SDS risk factors and the outcome.

We saw some attenuation but not a complete disappearance of the effect. The second was that socially disadvantaged patients get worse care or get disparate care, sometimes don't get adequate care, right. That will be measured at the patient level.

And the third is that there's some source unrelated to hospital quality of care,

things like competing economic priorities which hinder adherence to post-discharge instructions, et cetera.

What we did to look at that was the decomposition analysis where we tried to see how much could be measured at the patient level and how much could be measured at the hospital level.

When we looked at that we saw that the hospital effect tends to dominate with SDS variables, which is in contrast to the clinical data elements where the patient effect tends to dominate.

Because that is true, because we see a dominant hospital effect what we conclude from that is by including the variables we risk adjust away a component of hospital quality uniquely the SDS variables as compared to clinical variables.

So because of that issue with a conceptual relationship, the empirical association alone was not enough. And I agree that we also see that it just makes no difference which adds further weight to the decision not to

recommend.

CO-CHAIR FLEISHER: I actually ask are any of the other clinical variables on the same level. I mean we know about severity of clinical variables, and we've discussed that this morning that you've included in your model because building a model is something that is well used, but I would argue that it is driven by the data. It's not fully driven purely by the data, and we make assumptions.

The other question is in your Bayesian model because of the way you create your Bayesian hierarchical model, do you have -- do you know how that may influence how your potential risk adjustment might change because of the shrinkage, because of where the SDS factors may or may not play the most influence?

DR. DORSEY: So again, we were just talking about the effect size, the odds ratio. We do include risk factors with higher and lower odds ratios in the current risk model.

But again, I'll circle back to the

fact that it's not just an empirical association that we are charged to consider, but it's also the conceptual relationship, which is the point I just made earlier, right.

So the SDS variables are distinct as we could see in the decomposition analysis from clinical variables. And so we're applying a different conceptual relationship than we are with comorbidities there.

With respect to the hierarchical model, as is true with all the covariates in our model or the model if we consider it overall as a whole, we acknowledge the limitations that you're bringing up with a hierarchical model.

This is something that's been considered and brought up by many statistical experts and stakeholders and an active area of reevaluation for us where we're constantly looking at new techniques and tested some of the fixes that have been recommended by statistical experts.

We have not yet found something that

is better than our current model, which I think many experts have acknowledged that there's not evidence that something is better than our current model.

But this, I would say that in terms of the empirical relationship and the shrinkage issue that you talk about, it's equivalent to the other risk factors that we include. Correct.

CO-CHAIR FLEISHER: If you analyze that within the shrinkage model itself, SDS or the AHRQ.

DR. DORSEY: Jeph, do you want to jump in?

LT. MOORE: Yes, this is Jeph Herrin. I'm not sure I follow the question. These results that we're presenting are from a shrinkage model, from a Bayesian model, so if you're asking whether these results are the same in this model. I guess that answers your question.

If you're asking whether SES is somehow influenced by shrinkage in a way

different from other clinical risk factors, I'm not quite sure how to answer that question.

I think that they -- the numbers of low SES patients are not necessarily concentrated in larger or smaller hospitals. So I wouldn't expect to see any kind of differential impact to clinical factors, but we haven't dug into that.

MEMBER YATES: Can I add something?

CO-CHAIR GUNNAR: Please.

MEMBER YATES: We -- you can't -- just leaving it at the dual eligibility and the AHRQ SES score, it's a dichotomy whether or not you're dual eligible or not. You can't look at the rest of the population the hospital cares for and say well, they were really close to being dual eligible so ignore that.

So you're really left with the zip code analysis and the AHRQ SES score, and you chose to look at the population of 42.7 and below. What about the rest of the population?

And if you were to look at it as a variable from the entire scale of AHRQ across the

entire population of the hospital and use that as a variable, wouldn't that not maybe perhaps explain the hospital effect that you're seeing. In other words, some really great hospitals live in some really poor neighborhoods.

DR. SUTER: And so we have looked at this. We have looked at the AHRQ SES index as a continuous variable. We've looked at it --

MEMBER YATES: You have or have not? I'm sorry.

DR. SUTER: We have for the readmission measures. I don't believe we've done this analysis for the complication measure, so we'd be happy to run those analyses for this committee.

We did not see any difference in the results when we ran it as a continuous variable, so wherever you were on your AHRQ Index, that was your -- influence of the AHRQ SES Index did not change the results for the readmission measure.

MEMBER YATES: As a hospital?

DR. SUTER: At the hospital level.

MEMBER YATES: And that's presented in 1551? I don't believe that's in --

DR. SUTER: No, it's presented with -- we presented that to the readmissions committee with -- several readmission measures went forward to the readmission committee.

There was a more formal presentation of SDS in front of that committee. And those results were shared with that committee. But we can certainly rerun those analyses for these measures.

MEMBER YATES: Well, let me step outside of the statistical world and just ask you as a statistician and you as a statistician. There's this debate over the hospital being the cause of the -- or that the patients with SDS compromise tend to go to less good hospitals.

Yet, you're still talking about some of the world's greatest hospitals being in some very poor neighborhoods. I mean as point of contention, I'm not there anymore, so I will use it as an example.

But ten years at Hopkins, I'm well aware of the neighborhood around Hopkins being very poor and being -- I mean the gunfire and everything else at night, it's a poor neighborhood.

Yet no one would say that Hopkins is giving horrible care to those patients, or at least I would hope you wouldn't. So if in your heart of hearts do you feel like we're really saying it's the hospital's fault for some of these great centers that are taking care of some of the poorest patients in the country?

Are we missing something, or is there some way of just still discerning the two that would still maybe make you want to put SES in there just to make sure it's followed on a parallel basis so that we don't miss anything?

DR. SUTER: So I don't actually know what Hopkin's results are, but we could investigate. I think the other point of information that's helpful to think about is that there is a distribution curve.

So when you look at hospitals by any measure of socio-demographic, be the AHRQ SES Index, dual eligibility, you can look at unemployment rates. You can look at a whole host of other SES variables that may or may not be available for all Medicare beneficiaries.

But when you look at those distribution curves, you see distribution. So you see people who are lumped -- hospitals that are lumped in the middle and then you see hospitals at either end of the curve.

And what's fascinating is that there are hospitals, and maybe Hopkins is one of them, where they are caring for patients in a highly underserved area with lots of underserved patients. And yet they are achieving remarkably good outcomes.

And we know those hospital exist. We also know the other hospital that's dealing with mostly non-minority, non-vulnerable patients that are doing very poorly, so the hospitals at either end of the spectrum exist no matter what group of

patients they're caring for, which at least suggests to us that it's possible to do well on these measures, regardless of your patient population.

I think the question is do you want -- how do you want to incentivize hospitals. Do you want to incentivize them to take care of patients and to acknowledge okay if you have dual eligibility or AHRQ SES or whatever the measure of SDS is, if you have that group of patients you're going to get an extra bye for that proportion of patients because we're going to risk adjust for those patients?

Or do you want to want to say no, we're going to publish a transparent non-adjusted measure results and then maybe acknowledge that payment for those hospitals may -- I guess we're talking about the measure right now.

We're not talking about payment policy, but there are lots of downstream ways to acknowledge that you do not want to limit resources to hospitals serving for under-

resourced populations based on their quality.

You want to acknowledge the societal implications of those measure results, but I'm not sure that you want to hide those measure results.

DR. DORSEY: And let me just add that the results of our analyses indicate that putting one of these indicators into the risk model would actually not resolve a problem of making -- based on how under-resourced hospitals look.

MEMBER YATES: And it would be nice to have -- adjust the -- what you did with the AHRQ scores for readmission. It would be nice to do it for the 1550.

CO-CHAIR GUNNAR: We're going to need to move on, but Rick and Barbara?

MEMBER DUTTON: Sorry. I just wanted to agree with Lisa's comment earlier that --

CO-CHAIR GUNNAR: It was nicely stated. Wasn't it?

MEMBER DUTTON: We don't -- I would not recommend risk adjusting for SDS because

that's one of the things we want to find when we report this measure.

CO-CHAIR GUNNAR: Barbara?

MEMBER LEVY: I just want to point out that finding it doesn't imply that the payment should be different or that people should be penalized. And I think we need to be advocates to very strongly state that finding those disparities should not and perhaps as a committee we can say that, translate into a reduction in payment for these facilities.

And looking at the Hopkins example, some of the facilities have resources. They have other resources that they can apply to this population and some catchment hospitals have nothing.

The only thing they have is the payment for the services that they deliver. And I think that we need to make a very clear statement from this committee that says we want to see what those disparities are, but we do not think that it is inherently a measure of the

quality of the institution that is remediable.

For example, if we look at staffing, if you don't have the money to mind the nursing ratio that is what you have to do to take care of this population, if you're in an area where you don't get enough payment from CMS to drive that nursing ratio, you can't correct for these things.

So I think again we need to separate out the payment piece of this from the scientific assessment of the measure so that we do what's right. We discover disparities when they're there and then we can address them in a multitude of different ways.

MEMBER YATES: And I would just add that it would be very -- I don't want to give the developers another 100 hours of work, but it would be great to stratify the top performers and the low performers and then look at whether or not some top performers are weighted down by the challenges that they face, which with limited budgets and cut budgets do make a difference.

And I think that the -- I would argue though that over time because the rubber is hitting the road now for those reasons, I think that over time this has to be continuously monitored because what you've shown in your data is that there are -- in the upper quintile proportional hospitals, that there is an effect of some sort.

And that effect could change, and it's an effect that's demonstrated I think more dramatically than what's been presented in the quality papers and the quality reports. The bell curves --

MEMBER YATES: So I think that you have to -- you've picked up something here that there is something in the noise, and then if you can get to follow it over time because it's going to be a big distinction in terms of possibility of unintended consequences --

MEMBER YATES: -- more of one population versus another that does --

CO-CHAIR GUNNAR: Can we move to vote

on validity?

MEMBER SAIGAL: Voting is open for validity, one high, two moderate, three low, four insufficient.

(Voting.)

MEMBER YATES: If I could make a public comment, Dr. Fleisher.

CO-CHAIR FLEISHER: Yes.

MEMBER YATES: Or Dr. Gunnar.

CO-CHAIR GUNNAR: Yes.

MEMBER YATES: I just want to reassure the committee that almost everything that we've been talking about on this measure will be very similar to the next one.

CO-CHAIR GUNNAR: My assignment was to get done with 50 and 51 by 12 o'clock.

MS. SKIPPER: Voting has closed on validity, 13 percent votes high, 83 percent moderate, 4 percent low, zero percent insufficient. The measure passes on validity.

CO-CHAIR GUNNAR: We have to open up at 12:30. We'll stick to the schedule, 12:30 for

comment. All right. We can move on now to feasibility.

MEMBER YATES: Feasibility, it's been demonstrated to be feasible because it's been happening. And that's -- I know that's a tautological argument, but it's actually very true. So having seen it in action for going on four, five years, I think it's reasonable to say it's feasible.

CO-CHAIR GUNNAR: Rather than take a -- well, I guess -- do we have to -- it's a maintenance measure.

MEMBER YATES: Yes. We have to with maintenance.

CO-CHAIR GUNNAR: We have to vote, so carry on.

MS. SKIPPER: Voting is open for feasibility on Measure 1550, one high, two moderate, three low, four insufficient.

(Voting.)

CO-CHAIR GUNNAR: You've got two missing, so that's all right. Close it.

MS. SKIPPER: Voting on feasibility has closed, 90 percent votes high, 10 percent moderate, zero percent low, zero percent insufficient. And the measure passes on feasibility.

CO-CHAIR GUNNAR: Next is usability and use.

MEMBER YATES: I can promise you it's usable, and it has been used in public reporting now through hospitalcompare.gov. It's also being used, as I said, as one of the -- it actually modifies as a quality metric for the CJR bundle.

Twenty-three percent of the hospitals have their price point set in part by some of the percentile rankings from this measure as well as others. And it's going to be used in -- it's already in the IPPS for 2019 for usability as part of the value-based purchasing.

So it's got usable endpoints that have been -- that we're not going to debate the accuracy of or the like. But under usability there is the question of unintended consequences,

and I would argue that given the fact that this is an elective surgery and being very elective, it is the sine qua non of the surgical procedures that might face unintended consequences of patients with slightly higher or marginal risk finding their access to care blocked by a sense of -- a lack of risk adjustment and by a sense of lack of accuracy of the database.

And I say that in conjunction with its use because its use is in a competitive zero sum environment. And in game theory, everybody's going to be assuming that the other person is lowering their risk and their risk population.

And it is now a commonly enough used word that it's in literature, so I'll use the word cherry pick as being something that this is at high risk for inducing as time goes by.

And with a 2 to 6 percent range, and a median that's a very, very -- has a very sharp bell curve, the differences between one hospital and other or heaven help if it goes to the level of provider.

The differences are very small that make a difference in terms of hospital penalties and other public disclaimers. So I continue to be worried about the unintended consequence of risk shedding. I will leave it at that. That's the usability portion.

CO-CHAIR GUNNAR: Oh, we have Amy, then Collette.

MEMBER MOYER: So a couple of comments on this. So our population is entirely commercial, so while it's useful to have this measure there, the fact that it's Medicare limited use and we're seeing just a massive ramp up in ambulatory surgery hopping up and saying, well, we're do joint replacements.

Well, that's great, but what's your quality. And so the ability to really feel like we're making as robust a possible comparison and holding them to the same standards is very important to us, especially as the joint replacement age has kind of dropped, we're seeing more patients have their first replacement on us

and then Medicare is left to kind of deal with the fall out and revisions if things didn't get right first time.

And the other thing I would add is that we hear from a lot of patients okay, it's great that you can tell me something about the hospital, but I'm going to pick a surgeon. And so not knowing that additional level of granularity for them is a challenge.

MEMBER PITZEN: Just a question. In the data that's provided back to the hospitals, are they able to see the detail of which complication occurred? Thank you.

DR. SUTER: Yes. Sorry for the recording and those on the phone, the answer is yes to Collette's question.

CO-CHAIR GUNNAR: Barbara, do you have a comment? No. Are we ready to vote on usability and use? Does Robert have anything to say? There you go. Proceed.

MS. SKIPPER: Voting is open for usability and use, Measure 1550, one high, two

moderate, three --

MEMBER CIMA: I was muted. I have no problems with this. I don't think these are all the right discussions, but I think it's going to be used as it is and the risk of cherry picking is there. But it's always going to be there.

(Voting.)

MS. SKIPPER: Just waiting on two more votes.

(Voting.)

MS. SKIPPER: Voting has closed on usability and use, 39 percent votes high, 57 percent moderate, 4 percent low, zero percent insufficient. The measure passes on usability and use. We will now need to move on and take a vote on overall suitability for endorsement.

CO-CHAIR GUNNAR: Any other comments before we take a vote on the measure? Proceed.

MS. SKIPPER: Okay. One yes, two no for overall suitability for endorsement of Measure 1550.

(Voting.)

MS. SKIPPER: One hundred percent of votes are yes for overall suitability for endorsement for Measure 1550.

CO-CHAIR GUNNAR: So I'm going to take a -- yes, Helen?

DR. BURSTIN: Just a quick comment. This isn't about directly this measure, but since so many of you have raised this issue of cherry picking, this would be a great group to maybe have a small workgroup off cycle just to talk through are there companion measures.

How do you actually kind of keep track of that issue because I do think -- it's something you guys are working on already?

DR. SUTER: No, but actually as a developer thinking that we'll need to bring these measures back, understanding exactly what analyses you guys would like to see from us would be really helpful in terms of guiding the surveillance on this.

DR. BURSTIN: So we will pull a workgroup together for those of you who are

interested just to have a conversation. Again, it's going to keep coming up.

And I think if we have a handle both on what the analyses are or some people have even called for -- sometimes for some of these kinds of measures, particularly if they involve costs, the concerns if you actually have companion measures that look at concerns around stenting in addition to cherry picking.

So I think it might just be a good discussion. This seems like a good group to start that with. Thanks.

MEMBER YATES: Could you spearhead that --

DR. BURSTIN: Yes, absolutely.

MEMBER YATES: -- because this is a temporal thing, and we have to capture it as it occurs.

DR. BURSTIN: Yes, I agree. And we are also trying to very much as part of our new strategic plan think about how we increasingly get feedback in real time as things happen on the

ground. But to the earlier point, we need to know what we're looking for and how you would even measure it. So, thank you.

CO-CHAIR GUNNAR: I'm going to take the chair's prerogative and push through the next measure we have. The measure developer's here, and I think we can move through it fairly quickly.

So this is the companion measure, 1551. It's the hospital-level 30-day, all-cause risk standardized readmission rate following total knee and hip replacement, CMS.

DR. SUTER: This is Lisa Suter. We have no additional comments and happy to answer any questions.

CO-CHAIR GUNNAR: Discussants are Dr. Erekson and Dr. Yates.

MEMBER EREKSON: So I would say from my reading of this measure, the only tangible difference in this measure is the outcome which is a 30-day readmission rate where they used a very standardized readmission look for all-cause

readmissions.

There are very strict exclusion criteria for planned readmissions for maintenance chemotherapy and some of those things, but otherwise the measure is almost identical to the last.

MEMBER YATES: I have no more comments on evidence, and I think the evidence can be accepted as presented. And I would also argue that the discrepancy and future validity arguments versus the SDS question have been discussed ad nauseam already and that we can accept those as being discussed. So I think you can vote on the evidence.

CO-CHAIR GUNNAR: So we don't need to vote on the evidence. We accept the evidence. It's a maintenance measure.

MEMBER YATES: I accept that distinction.

CO-CHAIR GUNNAR: Anyone want to vote for the evidence? Seeing none, we can move on to the cap.

MEMBER EREKSON: So what is presented in the gap is that when this measure was first reported the hospital performance score was 4.9 in the first year. It dropped -- actually, maybe my numbers are slightly different.

It dropped slightly and right after the measure was first instituted, and it's remained at a four point -- well, this is 5.3, 4.9, 4.4 for an overall mean of 4.9. So it is continuing to drop.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: Nothing to add.

CO-CHAIR GUNNAR: We'll go to vote. Yes? Christy?

MS. SKIPPER: We're now voting on gap for Measure 1551, one high, two moderate, three low, four insufficient.

(Voting.)

MS. SKIPPER: Just waiting on one more vote from the phone.

CO-CHAIR GUNNAR: All right. You good?

MS. SKIPPER: Voting has closed for gap, Measure 1551, 38 percent votes high, 62 percent votes moderate, zero percent low, zero percent insufficient. The measure passes on gap.

CO-CHAIR GUNNAR: Moving on to reliability. Any further discussion based on -- beyond what was --

MEMBER YATES: The c-statistic is amazingly the same.

CO-CHAIR GUNNAR: Yes.

MEMBER YATES: And the ICC jumps to 0.49, so it's the same level of moderate in the correlation and you have the same moderately low lucky to model.

CO-CHAIR GUNNAR: Dr. Dutton?

MEMBER DUTTON: Just a quick question for the developers. Transfers out are excluded again. How big an effect does that have? Does it transfer out include sending a patient to rehab?

DR. SUTER: It does not. A transfer out is a patient who is transferred from --

transfers in are excluded. Transfer out they're -- you're -- the outcome of readmission is assigned to the hospital that discharges the patient. So it's shifting the attribution but not direct.

CO-CHAIR GUNNAR: Any further discussion? Carry on. Christy?

MS. SKIPPER: Voting is open for reliability, one high, two moderate, three low, four insufficient.

(Voting.)

MS. SKIPPER: Twenty-three percent votes high, 77 percent votes moderate for reliability of Measure 1551. This measure passes on reliability.

CO-CHAIR GUNNAR: So Lee?

CO-CHAIR FLEISHER: Validity.

CO-CHAIR GUNNAR: Yes, validity, calling it. Dr. Fleisher?

CO-CHAIR FLEISHER: Yes, I just want a point of clarification. Helen can correct me if I'm wrong. Lisa or Karen, you made a comment

that we would only see the risk adjust rate. It is a requirement that SDS is incorporated, that both the risk adjusted with and without SDS.

So your statement is inaccurate just for the committee's purpose that we will always be able to see whether or not, at least under the trial period whether or not --

DR. BURSTIN: Right, so when this al went through a couple years ago, the idea would be that if a measure was endorsed with adjustment for SDS factors it was required that there would also be stressifications for the stratified results.

So you would always, in fact, see what was underlying and driving those differences while allowing potentially for a level playing field. Again, we'll see how that plays out.

What's not clear is exactly how that gets played out, for example, on Hospital Compare although that may be our recommendation. So again, I think there's still a lot of play in this space. I think we're all learning

collectively over the last year.

MEMBER YATES: I've got to ask a question. We didn't vote on whether SDS should be required or not be required. We just voted on the validity of the measure as is, so I don't understand your last comment because you said that if it's endorsed with the --

DR. BURSTIN: If you had made the -- let's play this out just for the sake of argument.

MEMBER YATES: Right.

DR. BURSTIN: If you had said, back to Yale, we disagree with you. We believe that given your odds ratio of 1.21 for dual eligibles we want that to actually be in the model and that happened to be what got voted on, then if the measure was endorsed with adjustment in the model, we would require, we being NQF, that there would also be specifications to see whatever variable was actually adjusted for for STS stratified to get at this issue of masking.

You've basically agreed with the

developers to leave the model as is, so there's nothing with which to stratify.

MEMBER YATES: Actually, we were told to vote on it as for validity and we didn't get asked a question as to whether or not it should be modified.

DR. BURSTIN: Yes.

MEMBER YATES: In other words, you're saying that if it had flunked validity because it didn't --

DR. BURSTIN: Correct.

MEMBER YATES: -- included SDS. And see this was part of the technical questions I was asking on the phone --

DR. BURSTIN: Okay.

MEMBER YATES: -- a month ago was how do we handle that. Is it supposed to be a parallel process or a process that's embedded into the question?

DR. BURSTIN: At this point, we have embedded SDS adjustment into validity, at least for the two year trial period, which we're

halfway through. And so you could have rejected that measure theoretically on the basis of the fact that you do not think the information presented suggested the measure should remain unadjusted for SDS.

The conversation you had around this table didn't suggest that was the direction you were going.

MEMBER YATES: I would argue that as a point of clarification it wasn't really made clear --

DR. BURSTIN: Okay.

MEMBER YATES: -- that we were doing it that way.

DR. BURSTIN: We will do a better job of that. Thank you.

CO-CHAIR GUNNAR: Would you handle a new versus a maintenance measure differently in that space?

MS. MUNTHALI: No, we would handle them the same way, so yes.

CO-CHAIR GUNNAR: Any other comments

regarding validity?

MEMBER YATES: On validity? I would just point out the same question about the original TEP and face validity.

The second comment is that I'm really not sure how -- since this is a dichotomous result, either readmitted or not readmitted, I'm not sure that it can borrow the validity of 1550 in its original validation study and then again the Federal Register and in multiple things it's given a 99 percent agreement rate based on a study that was really looking at readmissions, I mean excuse me, complications.

So I think that's a little bit of a slide on that. The -- to look at the validity of this versus the charts, you'd have to look at all of the risk factors and you'd have to go back through the charts of 12 months.

So I don't think that study can be done. So I'm giving it a pass on validity since it's a dichotomous endpoint. And then since we now have this other question then, I would still

say it should considered by the committee whether or not the inclusion or non-exclusion of SDS is an important factor in terms of voting for the validity of the measure because I don't think it was phrased that way in the last process.

DR. SUTER: This is Lisa. So in response to A.J.'s input, so a couple things. The validity of the planned readmission algorithm was performed through a chart, a medical record review.

It's distinct from the complications validation. That information is not included in every single readmission measures application to NQF.

We're happy to share that with the committee, but it is -- the planned readmission algorithm has been vetted by a number of different committees, mostly the admission and readmission committee that's investigated those other readmission measures.

This measure is paired with its complications partner and is in the surgery mouth

of admissions/readmissions committee. In regards to the SDS, we see similar results in the readmission measure as with the complication measure.

MEMBER YATES: And correct me if I'm wrong, but when you say it's been validated for the risk factors, that validation is based on the previous validation for the HCC model. Correct?

DR. SUTER: The key measures have not had their risk adjustment models validated in a clinical record review. They have had their outcome assessments validated in medical record review.

MEMBER YATES: Okay.

CO-CHAIR GUNNAR: Any other comments, concerns regarding validity? Seeing none.

MS. SKIPPER: We're now voting on validity for Measure 1551, one high, two moderate, three low, four insufficient.

(Voting.)

MS. SKIPPER: Voting has closed, 9 percent votes high, 78 percent votes moderate, 9

percent low, 4 percent insufficient. This measure passes on validity.

CO-CHAIR GUNNAR: Moving to feasibility.

MEMBER YATES: Yes, it's feasible.

CO-CHAIR GUNNAR: Any other discussion? Christy on feasibility?

MS. SKIPPER: Voting is open for feasibility, one high, two moderate, three low, four insufficient.

(Voting.)

MS. SKIPPER: We have 20 votes and we need three more, so if you can aim your clicker in this direction. Eighty-seven percent votes high, 13 percent moderate, zero percent low, zero percent insufficient. The measure passes on feasibility criteria.

CO-CHAIR GUNNAR: Use and usability.

MEMBER YATES: It's used as part of hospitalcompare.gov, and it's also used in the Readmission Reduction Program from CMS. It has -- if it's payment quality implication having

started this year, it's part of the basket of conditions and procedures that make up that basket of procedures and conditions at risk for that.

That's a weighted basket, so for some hospitals it becomes a very large part of the weight. And the penalty is 3 percent across all CMS reimbursement as a potential worst case scenario.

And in terms of gaming and unintended consequences, I would argue the same things that I argued before. And I would also argue that the research that's been done before, not seeing an increase in observations days preceded probably the attention level of hospital administrators in that the OBS base have possibly gone up at least in my perception and that more recent data than 2013 would be more indicative of that as an unintended consequence.

CO-CHAIR GUNNAR: Any other comments?

MS. SKIPPER: Voting is open for usability and use, one high, two moderate, three

low, four insufficient.

(Voting.)

MS. SKIPPER: Fifty-seven percent votes high, 39 percent moderate, 4 percent low, zero percent insufficient. This measure passes on usability, and we can move on to -- the committee can move on to a vote for overall recommendation for endorsement.

CO-CHAIR GUNNAR: It's open.

MS. SKIPPER: Your options are one yes, two no.

(Voting.)

MS. SKIPPER: Thank you, 95 percent votes yes, 5 percent votes no. This measure is recommended for overall suitability for endorsement.

CO-CHAIR GUNNAR: We'll open up the line now to public comment.

OPERATOR: If you have a comment at this time, please press star 1. There are no public comments at this time.

CO-CHAIR GUNNAR: So we will break for

-- what do we need, ten, fifteen minutes for bio break and gather lunch, bring it back and we'll carry forward. Thanks for your efforts this morning.

(Whereupon, the above-entitled matter went off the record at 12:48 p.m. and resumed at 1:05 p.m.)

CO-CHAIR FLEISHER: We actually promised the developer, for those in the room, we promised that at 1:05 we would be back on. I don't like to have my surgical colleagues be concerned about starting on time.

So we're going to start with 0713, VP Shunt Malfunction Rate in Children. Is the developer on the line?

DR. GOUMNEROVA: Yes, we're both here. My name is Lili Goumnerova, and I have Sarah Henderson with me, as well.

CO-CHAIR FLEISHER: Okay, great. We can hear you, but it would be great if you could speak a little louder.

DR. GOUMNEROVA: Okay, can you hear me

now?

CO-CHAIR FLEISHER: Yes. If you could just -- three to four minutes about the measure, itself, from your perspective.

DR. GOUMNEROVA: So, the reason we decided to come up with this measure, it is a relevant measure, since hydrocephalus is treated primarily with ventriculoperitoneal shunts, and they have a very high failure rate.

And this is probably one of the most common neurosurgical procedures that is performed across pediatric hospitals. It is a good way to evaluate how we, as hospitals, do amongst each other, and also to get a good sense of what the malfunction rate of these very commonly inserted devices is amongst hospitals, not necessarily just looking at individual experiences in academic hospitals, but across all hospitals.

The measure was submitted a number of years ago, and we have been using it here for at least five years. And it is very reliable in our institution, and it provides us good data on how

we're doing. We have made changes in our practice based on the measure, so we find it useful.

Additionally, I have submitted this measure to the pediatric neurosurgical group. There is the pediatric section of the AANS, which is the American Association of Neurological Surgeons, and Congress of Neurological Surgeons. There has been interest from other institutions in the hospital to use it. So it is being also used in other institutions.

I'm not quite sure what more you would like to hear from me. Anything specific or not?

CO-CHAIR FLEISHER: No, that would great. Our discussants are Larry, correct, and Cliff? And this is of concern or not of concern, the potential conflict --

MEMBER MOSS: As you suggested, I'll just state it briefly. Does that make sense?

CO-CHAIR FLEISHER: Yeah, and Cliff can --

MEMBER MOSS: As the developers

mentioned, this is a measure that was recently adopted for use by the Pediatric NSQIP program. My secondary reviewer, Dr. Ko, has recused himself completely because of his leadership of that program. I want to make the Committee aware that I'm a founding member of the steering committee of Pediatric NSQIP. I was not involved in the selection of the use of this measure and don't believe it prohibits me from providing an objective review, but I did discuss with Dr. Fleisher in advance, and I'll recuse myself from voting, even though I'll discuss the measure.

CO-CHAIR FLEISHER: Before we start, is everybody comfortable with Larry -- and I honestly think, Cliff, if you have any comments, to comment, they will just recuse themselves from voting and giving us a formal recommendation. But discussing would be fantastic.

DR. GOUMNEROVA: And Larry who? I'm sorry; I didn't catch the last name.

MEMBER MOSS: Moss.

DR. GOUMNEROVA: Larry Moss, okay. I

used to be on the executive committee of Pedi NSQIP until a year or so ago. So I'm familiar with it. Okay.

MEMBER MOSS: Okay, I'll move forward. I'll just begin with a couple of brief content comments because I know most of the audience are not pediatric providers. Hydrocephalus is accumulation of cerebrospinal fluid around the brain in excess, which can cause permanent neurologic injury.

It is treated with a diversion procedure where there's a mechanical device that takes CSF out of the cerebral ventricles and puts it into the peritoneal cavity for absorption. As Dr. Goumnerova stated, these fail commonly, and thus the initial rationale for the measure.

I want to begin with a discussion of evidence. There is some new evidence provided by the developers for this outcome measure, which is submitted for maintenance. It's submitted under 1b.5, but I think it's more relevant to 1a, as we're discussing here. And that's a paper by

Rossi et al., who looked at a single institution retrospective review trying to identify risk factors for shunt malfunction or failure.

Of note, this paper concluded strongly that -- in fact, I'll just read from the conclusions. That would be the best way to do this.

In this study, none of the risk factors that were examined were statistically significant in determining shunt failure within 90 or 180 days. Given the negative findings and the fact that all other risk factors for shunt failure that have been proposed in the literature are far beyond the control of the surgeon, i.e., non-modifiable, the use of an institution's or individual's global shunt revision rate remains questionable and needs further evaluation before being accepted as a quality measure.

So I think it calls into question the initial evidence that was presented in 2011. I'll add to that, before I'll stop for the developer's comments, that I would ask the group

to consider that the intent of the measure is to identify shunt malfunction rate, but what is actually measured is redo operation for a failed shunt. There are not uniformly accepted, evidence-based criteria for the indications for a redo, so the redo rate is potentially heavily influenced by local practice, or the local approach and definition of shunt failure, as much as it might be by the actual outcome.

So I'll pause there for the developer.

DR. GOUMNEROVA: Yeah, so I am familiar with the paper by Rossi. It is very interesting. When we looked at risk adjustments and risk factors -- I should say risk factors -- we do agree that there are some historically thought to be risk factors which did not prove to be valid. So I'm not going to disagree with their comment. However, I think that, nevertheless, when we looked at our data, there do appear to be trends.

And I think that's one of the things that Pedi NSQIP shunt measure is going to be

looking at that in some granularity, also, is going to try and address this issue of risk stratifications, risk factors. I'm not going to be able to answer that completely, but I'm familiar with that, and we're working on addressing that.

I do want to comment: we're not looking at re-operation patients here. The way we actually look at these patients, patients who have had an immediate prior operation, are not the ones that are looked at in this dataset. So I don't understand why you think that is the case.

MEMBER MOSS: So, correct me if I'm wrong, but my understanding from the coding definitions provided is that a patient becomes part of the denominator if a shunt is a de novo procedure, and then becomes part of the numerator if a second operation is done within the specified period. Is that correct?

DR. GOUMNEROVA: Where do you have that? Where was that in the --

MEMBER MOSS: I've got that from the development worksheet. I'll try to find the specific place, but I'm not sure I can do that off the top of my head.

DR. GOUMNEROVA: Right. We specifically tried to actually get around that. Because you are correct, you do not want to be looking at re-operation rate for the same patient over and over again. So, in actual fact, that is not the way it is. And I specifically look at these data periodically, at a quarterly basis. And I know that is not what is being captured. That is not the way the measure was defined --

(Simultaneous speaking.)

MEMBER MOSS: I'm sorry, go ahead.

DR. GOUMNEROVA: I'm just trying to find where that is in the --

MEMBER MOSS: Yeah, I've got the numerator statement here, so I'll just read it so we're all on the same page. The numerator statement is the number of initial ventriculoperitoneal shunt placement procedures

performed on children between the ages of zero and 18 years of age that malfunction and result in shunt revision within 30 days of the initial placement.

DR. GOUMNEROVA: Correct, yes. So what we say is initial placement. So a patient who, for example, has had an operation within -- say, for example, a patient comes in, has a shunt placed, and then it's revised within 30 days, that is exactly what we're capturing. However, if that patient then goes back to the operating room in another 15 days, we're not counting that. We're avoiding the patient who has the multiple repetitive procedures.

MEMBER MOSS: I want to make sure that I understand you clearly. As I understand the measure, what becomes a positive, or what enters you into the numerator, is a shunt revision, correct?

DR. GOUMNEROVA: Correct, yes.

MEMBER MOSS: Okay, thank you.

CO-CHAIR FLEISHER: So, right now,

we're on evidence, because this is a new measure, but you're discussing a little bit of validity.

MEMBER MOSS: Yeah, I understand this goes a little bit into validity, but I think it does speak to some changes in the core evidence. My understanding is that evidence for an outcome measure needs to relate to whether there are processes of care that can influence the outcome, and that's what I'm getting at with this discussion.

CO-CHAIR FLEISHER: Okay, Barry?

MEMBER MARKMAN: I think it's a maintenance measure, isn't it? It's a maintenance measure. But then the question I was just going to ask the developer is --

CO-CHAIR FLEISHER: Your mic's not on, is it?

MEMBER MARKMAN: I think so.

DR. GOUMNEROVA: Hello? Yeah, I can hear it.

MEMBER MARKMAN: Okay, so why did you pick 30 days, when the retrospective cohort

study, the large study, shows almost a five times complication rate after 90 days?

DR. GOUMNEROVA: That is a very good point, and we have talked about extending that to 90 days. It is true, when we first started out with this measure, we wanted to look at 30 days, only, but we are in the process of extending it to 90 days. We have not done all of the work on that, but exactly to address those concerns.

MEMBER MARKMAN: Okay, good. And what actually is the malfunction? Is it a device malfunction? Is it a clogging of the shunt? What actually is the malfunction?

DR. GOUMNEROVA: There are a number of things that can go under the category of malfunction. By and large, it is a clogging of a device.

The device is essentially a tube, a piece of plastic tube with holes on one end, through which the fluid enters. There is then a component called a valve, which is then attached to another piece of tubing that goes into the

distal cavity, whether it's the peritoneum or atrium or some other space. Anywhere along that valve/tubing area, there can be a clogging. In the majority of cases, it is a clogging of the tubing that is in the brain, in the ventricles..

MEMBER MARKMAN: It's not the actual pump? It's not the actual pump, it's the tubing that's clogged, right?

DR. GOUMNEROVA: It's the tubing, that is correct. The valve -- so, 80 percent, about 70 to 80 percent of the malfunction occurs because clogging of the tube in the brain. The rest are related to malfunction of the valve that regulates the flow.

Now, when people talk about malfunction, they also frequently will code it as malfunction, where in actual fact the tubing was placed in the wrong location. With this current measure, we do not have a way of separating the clogging of the tube which happens normally, or because there was an error where the tube was not placed in the correct location.

CO-CHAIR FLEISHER: Thank you. Are there other questions about evidence? Because he major question you had asked, from my understanding, is for a process, or this outcome, could it be influenced?

MEMBER MOSS: Correct. So, my understanding is that for evidence to be revisited in a maintenance measure, it needs to be because there is new evidence. And the submitted paper that I referenced was submitted by the developers as new evidence, And I think that and some other factors at least call into question whether this measure is a valid indicator of processes of care at an institution.

DR. GOUMNEROVA: I still think it is a valid measure, despite the paper. I think if we -- there needs to be a rethinking of how we evaluate hydrocephalus treatment. And that's a big question. I understand your concern. However, this is not necessarily accepted amongst the neurosurgical community as a way of looking at treatment of hydrocephalus. But I'm willing

to -- open to discussion on that.

CO-CHAIR FLEISHER: Well, I think what I'm hearing is -- and I actually think, based upon those comments, we should actually vote if people would agree on this one. Does anybody want to vote, or does everybody agree that the evidence is sufficient from the -- okay, we've gotten calls for votes. Christy?

MS. SKIPPER: Voting is now open for Measure 0713, whether the rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service. 1 yes, 2 no.

(Voting.)

MS. SKIPPER: We have 20 votes. And we're just waiting on one more, so if you can aim in my direction. Oh, two that aren't voting. We'll move on.

We do have 20 votes, 55 percent votes yes, 45 percent votes no. This measure is consensus not reached, but we will continue the evaluation of the other criteria.

CO-CHAIR FLEISHER: With consensus not reached?

MS. SKIPPER: No? So this measure would be moved to our post-comment call where you all would then make a decision.

MS. MUNTHALI: Yes, you will vote on the rest of the major criterion, except the overall vote. You'll vote on the overall during the post-comment call.

CO-CHAIR FLEISHER: Okay, everybody got -- good. Next?

MEMBER MOSS: Okay, we're moving on to performance gap now. The evidence provided is a study comparing the rate of shunt malfunction at the developer's institution, Boston Children's Hospital, with what they define as a benchmark, which is derived from ten hospitals in the PHIS database, which is an administrative database of the Children's Hospital Association. That's provided, but my question for the developers are a few.

First of all, since this measure's

been endorsed since 2011 and there's been five years of duration, why isn't there more information available comparing institutions against each other?

Second question is why have the developers not taken the PHIS database and used the administrative data to compare institutions against each other?

DR. GOUMNEROVA: I think that there are other institutions that have used it. I guess it is our failure in not wanting to compare ourselves to other institutions. It has been, essentially, trying not to want to compare each other. That's basically the reason.

MEMBER MOSS: My other comment here is regarding sociodemographic disparities. The developers state that these data are not required to be collected at the institutions utilizing this measure. But I wanted to point out that a significant portion of this patient population is premature infants, which have been shown in many other domains of care to have marked differences

that are related to socioeconomic status. So I just question the developers on that decision and if they could take us through the thought process a little bit.

DR. GOUMNEROVA: I think that you are correct, and we're aware of those. However, when we looked at some of the data, it did not appear to be a valid -- in our data, it did not appear to be valid, looking at that and outcomes. I think that, although in other data that is supported, it did not appear to be relevant in our VP shunt malfunction data. We're happy to look at that again, but it's not been a significant factor.

CO-CHAIR FLEISHER: Thank you. Comments? I have a question for Karen or staff. Your internal rating, can you give the rationale between your preliminary assessment, so we can actually understand how the staff think about this?

MS. JOHNSON: I think it mainly has to do with the fact that it was one hospital, 46

eligible cases, and only two repairs or revisions. So that's not to say that there's not a problem with quality, it's just we weren't quite sure that there was enough information for you to make that rating.

So you can decide if you need more information. You can decide if you think that is good enough to indicate a gap in care. And we will point out again our new guidance from the CSAC, dated July, so hot off the presses.

You are kind of given permission, if you will, to think about some of these outcome measures, particularly mortality measures or patient safety measures. And I don't know if you consider this to be a patient safety measure or not. I think we had that conversation earlier. You know, you might not have to feel that there is as huge a gap as you might for a process measure.

CO-CHAIR FLEISHER: Are we ready to vote? Other comments? And what I've heard is this has been endorsed for six years, and there's

data from one center. Am I repeating that correctly?

DR. GOUMNEROVA: That's correct, yes.

MS. SKIPPER: We're now voting on the performance gap for Measure 0713, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Voting has closed for performance gap. Five percent votes high, 15 percent moderate, 40 percent low, 40 percent insufficient. This measure does not pass on performance gap.

CO-CHAIR FLEISHER: So we do not go further, correct, at this point?

MS. MURPHY: So, what you can do, if you feel that the measure is important and the information that is provided thus far provides you that information, is that you can go ahead and look at the other criteria. Karen's shaking her head no.

MS. JOHNSON: Maybe not that emphatically. I think since it completely

failed, you're not in that gray zone. We generally would not go forward. It's your call. If you want to, we could. I think the other question that you may want to answer is, is there something that perhaps the developer could come back with on post-comment that would potentially make you change your mind about this?

CO-CHAIR FLEISHER: I think that's the key thing. I will actually defer to the committee first. Barry, did you --

MEMBER MARKMAN: Yeah. I mean, I would like to commend the developer for at least attempting to address a significant surgical problem. A 30 percent complication rate within 180 days? I mean, that's unacceptable. So I commend the developer for trying it. Maybe she can restructure it or something else, but that's my feeling. That's my personal feeling.

CO-CHAIR FLEISHER: Okay, Rick and Larry and Fred? What I'd like is specific instructions to potentially get past the performance gap as the first comment.

MEMBER DUTTON: Having been in this position myself, I know that the developer will appreciate whatever feedback we can give about how to make it better.

DR. GOUMNEROVA: Absolutely.

MEMBER DUTTON: I guess it's good, generally, that a measure that's passed before is now not up to our standards. That suggests evolution of our process.

I would say this needs to come back with more data attached to it, more performance data from more than one hospital, larger sets of data, whether through NSQIP or other mechanisms, to show, to tell us more about what that performance gap is.

DR. GOUMNEROVA: Okay.

CO-CHAIR FLEISHER: I think that's very helpful. And I don't know how quickly you can get that. When's our post?

MS. SKIPPER: Our post call is August 25th, so about a week from now. That's our post-meeting call, is August 25th.

MEMBER MOSS: I wanted to just echo Barry's comment that this is a critical area of need with respect to patient care. And I also commend the developers for the effort. And a far more favorable outcome than seeing this go away would be for NQF staff to work with the developers and ultimately come back with something that could be endorsed.

I wanted to make one specific suggestion to the developers. There is way more compelling evidence about the relationship between VP shunt infection and processes of care than there is about malfunction. That may be something to consider that would meet the NQF bar and keep VP shunt care on the radar.

CO-CHAIR FLEISHER: Fred.

DR. GOUMNEROVA: I agree. We can -- yes.

MEMBER GROVER: I think, generally, the measures that we have approved, at least historically, tend to be more than one. And to test reliability and usability, it needs to be

beyond one center, to see how easy and reliable the data is as it's collected across centers. And I guess my advice would be to find a number of other children's hospitals that would be willing to be involved with this and maybe put that together and come forward.

And I would ask the NQF staff, though, would they need -- I would think they would need to have data showing performance over a period of a few years, across several hospitals, with consistent data and reliability to validate the process.

I think it's an important project. I think we all agree on that.

CO-CHAIR FLEISHER: I think that's -- hopefully the developer's hearing that. Do you have data? Cliff, do you know, data in the Pediatric NSQIP for more than Boston Children's?

(Off the record comment.)

DR. GOUMNEROVA: I'm sorry, I can't hear.

CO-CHAIR FLEISHER: Dr. Ko is saying

that you're talking about putting it in. It sounds like the Committee would like to see, if not data across years, at least data across more than one hospital. Is that the consensus of the Committee? What I also think you would probably hear is the Committee agrees this is an important measure, just it's insufficient is actually what a lot of people, 40 percent, said.

Why don't we go Amy, then Liz, then Barbara. And you'll also have the expertise of having Cliff being at NSQIP.

MEMBER MOYER: This is Amy. I'm sensitive to the fact that sometimes our limitations on our measures are placed there by the data sources. But barring that, I'd like to see some sort of rational time period for the measure. Because looking at the failure rates that are in some of the studies cited, it feels like longer than 30 days might be an appropriate period to capture more of the rates, if feasible.

DR. GOUMNEROVA: Okay, I appreciate those comments.

CO-CHAIR FLEISHER: Liz?

MEMBER EREKSON: Yeah, thank you for bringing this measure. And I would just -- when you go back and reset, when you're looking at a measure where a device has malfunctioned, the point is very well taken that the device can be misplaced, and so that is a reason for a failure, but there are so many times where our medical devices fall down on us and our data about changes in medical devices fall down us. So while you're resetting this measure, I would really advocate that you collect data on the device itself, and the manufacturer.

DR. GOUMNEROVA: Yeah, that is fairly easy, since it's a relatively low incidence, so we can get those data.

MS. SKIPPER: This is Christy Skipper. I'd just like to correct myself for the record. Our post-comment call is November 3rd. I don't know if that will give you enough time, but I just wanted to correct and notify that our post-comment call, where a developer could

present additional information to the Committee, is November 3rd. Thank you.

CO-CHAIR FLEISHER: Melinda, can you help us if -- what I'm hearing around the room, or feeling around the room, is that there's a desire to see this back. If they don't meet the post-comment call, what happens in this context?

MS. MURPHY: If they don't -- I guess I'd like to back up one step from there and ask whether or not, at the time the Committee has its post-meeting discussion, if there's anything that the group would want to hear back from the developer, in terms of, having heard what they've heard today, that they have some plan or some concept of how they would move forward to address the concerns you've raised.

And then if it goes for a post-comment call, post-member comment, so it would go out to the membership in a draft report saying where it sits at that time of the call of this Committee. Given where it is today, it wouldn't be a positive set of comments, in terms of moving

forward without some plan.

You would expect that without the support of the Committee, there wouldn't be a high likelihood of a positive vote, a positive comment, I should say, around moving the measure forward. That time window gives more time to get information, but it will appear in a draft report, given where it is right now, that it's not supported, with a different set of words, but essentially, it would say that.

CO-CHAIR FLEISHER: Other comments? Barbara?

MEMBER LEVY: I wanted to go back to the evidence, because I really think that we've got a couple of issues here. One is, as a quality measure, is there anything that the provider can do differently, or the institution can do differently, to impact this outcome? That's Question No. 1, and a major for this. It's not that the condition isn't really important.

And secondly, to reiterate what Liz

said, having a registry of these things to look at device malfunction is different than having a quality measure that looks at performance and gaps in performance.

Whereas I think it's very important to track this and to try to figure out what the issues are, a device registry might be a better way to get at some of the malfunction. I, personally, given the newest evidence, as presented to us, would have a problem with this measure, even with additional data from additional institutions, rather than just one institution, because, getting back to the fundamental issue, is this really a measure of quality, or is this a measure of how seriously ill these kids are, or some other -- that the devices are just inadequate to resolve the problem for them?

CO-CHAIR FLEISHER: Collette?

MEMBER PITZEN: I just wanted to add on to Barbara's comments. I was thinking about that as I was voting for evidence, trying to

think of one process that would perhaps change that outcome. And came up with good monitoring of that child over time might prevent a malfunction. No?

CO-CHAIR FLEISHER: Larissa, did you -- no?

MEMBER TEMPLE: It's been a long time since I've put VP shunts in, but I do remember that they do malfunction. To my recollection, it's usually not the device. And I think what we're hearing, though, is that when you come back with that measure, you really need to demonstrate that it's not a device issue, or that you have mechanisms to track that. Because I don't want the developers to get discouraged by the fact that they need to focus on the device, as opposed to other pieces. But I think you're hearing we need to hear the evidence that it's not device, or that it's a very small component of it.

CO-CHAIR FLEISHER: Larry?

MEMBER MOSS: Just in reference to Barbara's point, well stated, that there are not

existing evidence that processes of care influence the rate of shunt malfunction. That's not the same as saying the processes of care doesn't influence it. It's just that there isn't evidence currently that links those two, which goes to my suggestion that if we focused on something such as infection and put attention on shunt care, there are often parallel improvements outside of the specific factor being measured, and overall shunt failures may go down. We may be able to have an overall positive influence on the problem by focusing on a different outcome.

CO-CHAIR FLEISHER: Fred, final comment?

MEMBER GROVER: Just briefly, going back, I really do think you need to enlist other hospitals to take on your database. I agree, you need to at least list the device, so you can, if you see a discrepancy in results, you can at least identify that. But it's like anything else. If you have a high -- hospitals are high outliers, whatever you want to call it, are

having more difficulty than others, then that leads you to examine the processes of care at the high end and the low end and to see what differences there are. But I think you have to start there, but you really have to make sure this database works across more than one center.

DR. GOUMNEROVA: I appreciate that, yes, thank you.

CO-CHAIR FLEISHER: Thank you. I would suggest you get back to Melinda and the rest of the staff to determine how to move forward.

DR. GOUMNEROVA: Absolutely.

CO-CHAIR FLEISHER: Staff will contact you on the best approach to moving forward.

DR. GOUMNEROVA: Okay, thank you.

CO-CHAIR FLEISHER: Cardiology colleagues, we're up to 3024: Carotid Endarterectomy, Evaluation of Vital Status and NIH Stroke Scale at Follow-up. Collette and Liz are the two discussants.

If we could get from the measure

developer, if they could introduce themselves and give us a three-minute overview of the measure, that would be fantastic.

DR. GRAY: Dr. Bruce Gray, from Greenville, South Carolina.

MS. CONNOLLY: I'm Traci Connolly. I work with measure development at the American College for Cardiology, within the registries of the NCDR.

CO-CHAIR FLEISHER: Do you want to give us --

DR. GRAY: Sure, try to give a brief background in regards to carotid revascularization. Two standard treatment alternatives for patients with significant atherosclerotic disease include carotid artery stenting and carotid endarterectomy, two different type of procedures.

Carotid endarterectomy is an open procedure, requiring an incision, performed by surgeons, in which the main goal is stroke prevention and correction of an underlying

atherosclerotic lesion that was felt to be causative or perhaps a risk for subsequent stroke.

Traditionally, the procedure is performed with a median stay of one to two hospital days. And it's important that we not only follow patients out beyond their hospital course, but out to 30 days, because approximately one third of neurovascular events after the index hospitalization, symptoms can occur in about a third of patients who do develop stroke after the procedure, after discharge.

And so the proposal is to use a certified stroke trainee, it does not have to be a physician. The NIH Stroke Scale certification is easily attainable, can be performed within about ten minutes of time, and with great inter-rater reliability. This NIH Stroke Scale would be performed pre-procedure, during the hospitalization, and at 30 days follow-up.

This is important because for hospitals to be able to show potential outcomes

and safety of the procedure, measurement needs to be continued out. We want to focus on stroke and death as simple measures.

The current, recent randomized trials use 30-day end point as a reference. And this process measure is in direct parallel to the previously approved process measure for carotid artery stenting. So this is a direct parallel type of measure proposal.

CO-CHAIR FLEISHER: Thank you. Collette, you want to --

MEMBER PITZEN: Sure, thank you. I commend the developers for bringing this measure to us again. It was presented in 2010, as I understand. I do have some concerns about the overall measure construct as it's currently specified and tested. I'm going to share those, and then we'll go into evidence.

Good outcome measures based on either patient-reported outcome tools or clinician-reported assessment tools can be built and implemented into clinical practice. Along

with these measures, it's a good idea to have an accompanying paired process measure to understand the rate at which the assessments are being performed. The process of simply administering a tool really can't stand by itself and should be considered with an outcome measure.

The follow up of vital status of assessment of alive or deceased, I'll also note that the time frame for follow up could be stated more clearly. The registry algorithm credits 21 to 60 days; the recommendation is a 30-day follow up. So in terms of assessing patients, it's just important to clearly state what your window of time is that you're allowing for numerator credit.

And then I have, also, some concerns about the fields that were listed in the specification in terms of the actual calculation of the measure. I think that there could be some enhancements or improvements there.

So, going into evidence. The evidence, as stated, is insufficient for a

process measure. As a reminder, process measures need strong evidence that they should be done for each and every patient going forward. The first citation relates to an upgraded guideline, with recommendation to monitor neurological outcomes. And the second relates to non-invasive imaging, which is not a part of this measure.

If it was an outcome measure, then some of the steps in the process flow submitted make sense, but it is simply a process of administering a tool and assessing for mortality, which are not linked strongly to the desired outcome. The desired outcome is improved quality of care, so how do we know when we've hit that outcome? I'll open that for discussion of evidence.

CO-CHAIR FLEISHER: Liz, and then Fred.

MEMBER EREKSON: I would just echo, slightly, what Collette's saying, that this measure, the way that I'm reading it and the way that it's written, is that what we are measuring

is if the stroke scale was performed at 30 days and we have that, not what the stroke scale was. So it is a process outcome.

And so when I'm reading through all of the information and evidence presented, multiple societies agree that it's a good idea to do this exam, but I did not find evidence that said that patients who had this exam done did better at a year, two years, five years than patients who did not have this exam done.

CO-CHAIR FLEISHER: Fred, did you have a comment? And then Cliff.

MEMBER GROVER: Yeah, I'm just kind of curious how this came about. What surgical involvement do you have, as in people that are actually doing the procedure?

DR. GRAY: Restate that, please?

MEMBER GROVER: This is out of the American College of Cardiology. I thought it might be out of the Society of Vascular Surgeons. You have cardiologists doing endarterectomies?

DR. GRAY: No, sir, surgical, but over

180 hospitals are involved with the registry. The registry is inclusive, with over 12½ thousand stenting procedures involved in the registry, as well. So I think the value to overall healthcare is the complete collection of datasets from hospitals, not just divided according to turf and specialty.

MEMBER GROVER: I just was, I suppose, wondering if you incorporated them in the process. I knew about your following the endovascular ones and their college.

DR. GRAY: Sure. Certainly, the multispecialty CREST trial utilizes similar data points, and acceptable end points, if you will. Having an adjudicated neurologic assessment is terribly important. It just hasn't been done. In order to get to outcomes, we've got to measure something. And to say, well, we don't have any consistent use of this stroke scale yet, how are we going to get it if we don't start measuring something? And as you know, there's tremendous contention in regards to who's performing what

and let's make it clear.

They're surgeon-performed endarterectomies, but that doesn't mean that the collection of that information, done in a systematic, comprehensive, quality way, wouldn't provide a meaningful tool that potentially could be used for outcome assessment in the future. And I think that's what we're driving for in the process part, is we need to measure it, and not just say, well, it's not been done yet, or somebody else needs to do it.

MEMBER GROVER: Okay, I get that. Yes, thank you.

CO-CHAIR FLEISHER: Cliff?

MEMBER KO: I have a couple questions. Number 1 is that this for carotid endarterectomy; does this also include stenting, or is that going to be a separate measure?

DR. GRAY: That's already been an approved measure, No. 2396.

MEMBER KO: So it's basically this same measure, except this is for carotid

endarterectomy?

DR. GRAY: Correct.

MEMBER KO: Isn't there a post-carotid endarterectomy stroke measure, an outcome measure? Because we keep on hearing from CMS that they would rather have outcomes than process, and so this is a process measure to evaluate it, but it's getting to the stroke issue of post-operative stroke. Is that a measure?

DR. GRAY: It's getting there.

MEMBER KO: But is there not measure already, or is that on the list?

DR. GRAY: No. There's no comparable measure.

(Off the record comment.)

MEMBER KO: So there is a measure already for post-op? So this is basically, if we fulfill that measure that's maintained, then this is kind of a duplicate, because it's the process to the outcome?

CO-CHAIR FLEISHER: Karen, do you want to make some comment? No? You're staying quiet?

Okay. Sal?

MEMBER SCALI: If I may, I'm a vascular surgeon. I think you're probably looking at trying to standardize the way that the strokes are adjudicated, because there's always a lot of variability in the literature about who's measuring what, just like wound infection, etc. Is that sort of the spirit of this, is you're trying to institute some type of more strict criteria of who says or who's qualified to make the diagnosis of stroke?

Because one of the issues, whenever you look in the literature, is just like wound infection, just like any other complication after procedures, who's measuring it, and how's it being measured? And what is the definition of a stroke? Is it neurologist says it? Is it radiologic? Or is it just some clinician, whether it's a physician extender, resident, whatever, who sees the patient at those time points and says, bing, they hit some NIH Stroke Scale assessment, so we're going to call it a

stroke?

DR. GRAY: Extremely germane comment. And that's why the use of a certified NIH Stroke Scale person is required, so that you have a standard. And that standard is universally accepted, I would believe, in the clinical arena, certainly is the standard for randomized control trials. And I think the extrapolation into real world makes sense as a routine.

CO-CHAIR FLEISHER: I think that's probably validity, when we get back to -- but other comments on evidence? A.J.?

MEMBER YATES: Is there any literature at all stating that there's a relationship between this outcome with a better patient result?

DR. GRAY: Certainly, the NIH Stroke Scale relates directly to long- and short-term outcomes.

MEMBER YATES: But the process occurring, is there anything to say that changes the outcome?

DR. GRAY: No, sir, that's not been measured to-date.

CO-CHAIR FLEISHER: Okay, we will vote, but Karen's going to talk to us first. Now she's ready.

MS. JOHNSON: Let me explain a little bit about our process for these kind of measures. If it's sounding like you may be thinking that this measure doesn't have the evidence required, and that an exception may be warranted, we do have an option for what we call an exception to evidence that you can invoke.

If you are thinking that's how you want to go, we need you to vote insufficient on this vote, and then we would have a separate vote that you would say up or down, yes or no, we want to invoke the exception. If you are not interested in invoking the exception, then, of course, you would not say insufficient there. It would be more low. Actually, that's not exactly right. You could still say insufficient, but we need more than 60 percent of the voting members

to land on insufficient before we could even talk about going to exception. So, I hope I didn't confuse you. Do you need me to repeat it again?

CO-CHAIR FLEISHER: Collette?

MEMBER PITZEN: This is Collette. I don't want to be too negative, but there has been a space of six years between when this measure first came, and now it's coming to us again, pretty much with the same feedback. So I'm not sure that the measure itself is strong enough to warrant insufficient with an evidence exception, unless there's a strong feeling that this would get the NIH stroke tool out there. I have some concerns around that to a hospital-based measure. Their follow-up rate, I forget the rate list, but the actual administering of the stroke tool is less than 2 percent.

CO-CHAIR FLEISHER: Okay, Christy?

MS. SKIPPER: We're now voting on evidence for Measure 3024: 1 high, 2 moderate, 3 low, 4 insufficient.

Did someone leave the room? If we

could just all aim once more.

(Voting.)

MS. SKIPPER: Okay, with 22 votes, we have zero percent high, zero percent moderate, 59 percent low, 41 percent insufficient. The measure does not pass on evidence, and we will stop discussion of it here.

CO-CHAIR FLEISHER: So, John?

MEMBER HANDY: I just want a point of clarification. In the evolution of measure development, process is where we started. Historically, I think many of the processes had very good evidence behind them, internal mammary artery and coronary artery bypass, but is there a precedent for this sort of process measure being approved? Because I'm sympathetic to the developer's statement that we've got to start measuring something. And this is a meaningful outcome, if you can demonstrate stroke.

CO-CHAIR FLEISHER: The staff can comment, but we have said if there's no outcome in this space, process is certainly sufficient.

Cliff?

MEMBER KO: If the process is to evaluate for stroke, and there is a stroke outcome measure, but it's done, as someone was saying, all kinds of different ways, is this potential harmonization, that if it's going to be a stroke, it is done this way, if that is the issue? Or is that not the issue?

If we're measuring stroke in the outcome stroke measure, and everyone does it -- someone was like, I'm a colon surgeon, that looks like a stroke, it's a stroke -- if that's wrong, and we have to all do it using this process method, is that worthy of a harmonization of the process leading to that outcome measure -- within the outcome measure? Just throwing it out there.

CO-CHAIR FLEISHER: It's up to the developer to decide.

MEMBER KO: When the college was asked to harmonize something with the CDC, it came from the NQF.

CO-CHAIR FLEISHER: Right, but that

was -- I assume that's an approved measure.

MEMBER KO: No.

CO-CHAIR FLEISHER: No?

MEMBER KO: While it was going through this.

DR. BURSTIN: Cliff's referring to efforts we did early on to harmonize and actually come together on the CDC and the college's measures around surgical site infection. My understanding, at least from the neurology side -- and we've had a lot of these discussions about potential inclusion of the NIH Stroke Scale in risk adjustment, for example, for stroke patients. The NIH Stroke Scale, I think, is clearly the one that had been most identified as the leading candidate. The question that has come up, I believe, when we've talked about this in other committees, as well, is the idea that it requires a certified person who knows how to administer the process, administer the test.

But you need to speak to what happened in -- did this come up in cardiology, as well?

Do you want to speak to the discussion there, just for consistency?

MS. MUNTHALI: The cardiology committee landed in much of the same places as this committee. We missed much of the discussion, but it sounds like, from the recap that Karen gave me, they struggled with many of these same issues.

CO-CHAIR FLEISHER: So we don't move forward. Any other recommendations from the committee or staff? Barbara?

MEMBER LEVY: I just want to point out that this process is something that's validated and used in clinical trials. That doesn't necessarily mean that it should be extrapolated and used clinically for all patients, at all times. It's a costly -- Fred was just doing this, but this would be an added cost that may or may not be clinically relevant.

To Cliff's point, if we have a measure already in existence that looks at clinical stroke, we may tweak that measure to make it a

little bit better, but to say that we need to use a research tool that's excellent and perfect for research, and use that clinically, I think that's a stretch that probably we're not ready to go that far.

CO-CHAIR FLEISHER: Thank you.

DR. GRAY: Could I just comment to that?

CO-CHAIR FLEISHER: Please.

DR. GRAY: I'm a little bit concerned that I didn't clarify the ease and ability to perform an NIH Stroke Scale. It's not just the purview of research people, but it needs to reach the bedside on a regular, consistent basis. And it just, quite frankly, is not. For me to say that I performed a procedure and the patient had no stroke, but to have a third-party independent observer say the same thing using objective, simple criteria that takes than less ten minutes, does not involve an MRI scan, does not involve the cost of an ultrasound, even, and could be part of the routine follow-up in the office, of

which most practitioners will see their patient within that time frame in the office, we're not talking about much additional cost.

Now, in regards to the acquisition of that skill, it's readily available online, and it would promote providers to obtain that skill level. And so my plea isn't to say that this is adding a lot of extra burden. It's going to put some objectivity to the bedside and extend it out beyond the hospital window.

CO-CHAIR FLEISHER: Okay, so just for staff's purposes, I want to just ask, the vote -- because the majority felt it was low evidence, there is no exception to the evidence rule. If the majority had voted insufficient evidence, we could revote on whether there's an exception. Does anyone want -- people understand that? So essentially, we've said it's not going forward. Does anyone want a revote?

MEMBER BILIMORIA: Can I ask to clarify?

CO-CHAIR FLEISHER: Okay, Karl.

MEMBER BILIMORIA: I think process measures are great. We don't have an outcome, that's fine. It's also great in the setting of a well-documented processes or a consensus panel that stated that a particular process is good for an improving an outcome, even if we don't have evidence to back it up.

There are certain things in the wound infection bundle that don't have a tremendous amount of evidence behind them, but they're good processes to adhere to. I think one of the things that I was looking for was to hear whether there is, for this specific tool to be applied in every circumstance, whether there's broad ranging society approval for that, and maybe that, for me, might change the way I would think about this.

DR. GRAY: Once again, we're trying to take a developing tool and make it more -- to create the value by recognizing that following patients objectively with some type of measure out beyond their hospital window is of value.

MEMBER BILIMORIA: But have multiple societies said this is --

DR. GRAY: Yes.

MEMBER BILIMORIA: -- the thing to do?

DR. GRAY: Without a doubt.

MEMBER BILIMORIA: Can we go over -- this specific instrument should be used all the time to evaluate stroke patients -

(Simultaneous speaking.)

DR. GRAY: It's a part of every randomized clinical trial assessing stroke. Now, you say is it used consistently in the real world? I would answer that no. I think that's no. Can it be? To the question in regards to cost, I think it was a very important question, but I thought the leading into the usability and the cost makes sense.

CO-CHAIR FLEISHER: Okay, Liz, and then --

MEMBER EREKSON: When I look at the entire measure, not just the evidence, we have 12,000 patients who participate in the registry,

and the 30-day follow up is 58 percent. Of the 12,000 patients who participated in the registry -- which is a wonderful undertaking to get that many patients in a registry and be collecting data on those patients -- only 2 percent of them had the stroke scale. What that tells me, while I'm reading the measure, is that the people who are invested in the registry, which are people who are already invested in the quality improvement efforts, are not invested yet in the stroke scale. So why is the actual process so low in the patients that were presented to us? Right now, it's at 2 percent. Can we get it higher without this being a quality measure?

DR. GRAY: I would say that the chances are less likely without some motivation. Now, why are people in the registry to begin with? Because it was mandated by CMS for you to be involved in stenting trials, and to be reimbursed on Medicare population, you had to be involved in the trial. So it wasn't without that carrot. Is it the chicken or the egg? We're

back to the same story.

If we don't start measuring something, we're not going to get any outcomes, and that's my plea. We've got to start measuring something that's objective without significant additional cost. You can't just run an MRI on everybody at 30 days, and you're certainly not going to use duplex ultrasounds as a surrogate endpoint, and you're not going to take the surgeon's or the interventionist's word, per se, without some independent objective confirmation with a very simple scale. So that's my plea.

CO-CHAIR FLEISHER: So --

DR. BURSTIN: I was going to say that in some ways, this is -- I was just looking at Fred -- this is somewhat analogous to the very early days when STS came forward with measures that said, did you participate in a registry with benchmarking? Part of the logic presented back to us by STS at that time was it's what will help drive participation. I think this is a similar argument. The tendency is we think about

measures being fairly distal or proximal to the outcome.

This one's pretty distal, which is not our preference, but I do think the logical chain they're presenting is that if this is the dominant tool -- and I will say, from lots of discussions with the NIH Brain Attack Coalition around the stroke measures in the past, the lack of risk adjustment for the NIH Stroke Scale, there's a great deal of interest in trying to move towards this as the standardized tool. But you need to decide whether you think it's too distal for the outcome, or is it sufficiently close? That's the reason why I think Karen raised the specter of would you potentially be interested in reviewing the full measure and using the evidence exception, that you believe that the benefits to patients would significantly exceed the risks, even if you don't, in fact, have evidence right now that says use of this dramatically effects outcomes?

CO-CHAIR FLEISHER: Let me just call

one question first again. Do people want to revote based upon everything we hear? Is anybody asking for -- Barry would like a revote. Okay. Let's revote, and then we can continue the discussion.

MEMBER YATES: Can I ask a question?

CO-CHAIR FLEISHER: The question is --

MEMBER YATES: No, I had a question --

CO-CHAIR FLEISHER: Oh, you have a question. Go ahead, sorry.

MEMBER YATES: This is the act of doing it, but is there, as part of the measure, a repository where the end result goes? Because that would be more consistent with the STS argument because they had a registry that collected it. Does this require it being captured by the registry?

DR. GRAY: It doesn't require it to be captured by the registry. It is available to be captured by the registry.

MEMBER YATES: See, I would make an argument that it's different than the STS because

they would have required their registry to capture the things that they were asking to be captured. It just needs that extra step.

DR. GRAY: Maybe I didn't understand your question. Without it, yes, it's captured three times. In CS, it's captured three times in pre-procedure, within the hospital stay, and at 30 days.

MEMBER YATES: So the act of doing it actually gets it put into the register, so it could be utilized?

DR. GRAY: Yes.

MEMBER YATES: That's a big distinction.

DR. GRAY: Yes.

MEMBER YATES: That's why I'm asking the question. I think people are focused on the process happening, but there's an end result in which place there's a repository. Okay.

CO-CHAIR FLEISHER: So the vote -- and correct me if I'm wrong -- if you believe that there's low or insufficient evidence, but there

might be a rationale for an exception, then vote insufficient. If you believe there's low evidence, but there is not a rationale for an exception, then vote low. Correct, Karen?

MS. JOHNSON: Or -- and this is my fault for not being clear enough earlier -- if you feel like you do not want to go with an exception, but you still feel that the evidence is insufficient, that's fine to vote insufficient here. If more than 60 percent say insufficient, we will go to the vote yes or no for exception, and you can decide then if you want to do the exception. In other words, you don't have to vote low to make sure that it doesn't go through on an exception.

CO-CHAIR FLEISHER: Thank you.

MS. SKIPPER: We're now voting on evidence for Measure 3024, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: There was an error. We're going to revote for Measure 3024, 1 high, 2

moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Just a second. I'm having a technical difficulty. Let's try this again. Voting is open for Measure 3024, high moderate 1, 2 moderate, 3 low, 4 insufficient -- 1 high, 2 moderate, 3 low, 4 insufficient.

(Off the record comments.)

CO-CHAIR FLEISHER: Yes. Okay, all those who believe high, raise your hand.

(Show of hands.)

CO-CHAIR FLEISHER: Moderate?

(Show of hands.)

CO-CHAIR FLEISHER: Low?

(Show of hands.)

CO-CHAIR FLEISHER: Nobody voted until now. One, two, three, four, five, six, seven, eight, nine, ten, eleven. Is that correct?

(No response.)

CO-CHAIR FLEISHER: All those who vote insufficient? One, two, three, four, five, six, seven, eight.

MS. SKIPPER: Dr. Cima voted as well.

CO-CHAIR FLEISHER: What did he vote?

MS. SKIPPER: Low.

CO-CHAIR FLEISHER: Low. So it's 12 to eight. So, 12 for low.

MS. SKIPPER: That's 20 votes.

CO-CHAIR FLEISHER: I think we've done this twice and I've gotten the same result.

MS. SKIPPER: I'm just going to, for the record, state it. So sixty percent votes were for low, 40 percent for insufficient, so the measure does not pass on evidence, and we will stop the discussion.

DR. GRAY: Thanks for having us.

CO-CHAIR GUNNAR: So moving on to the next measure, 1519, statin therapy at discharge after lower extremity bypass, Society of Vascular Surgery, who are coming to the table.

DR. JOHNSON: Hello, I'm Brad -- go ahead.

CO-CHAIR GUNNAR: Welcome, and you have your three minutes.

DR. JOHNSON: I'm Brad Johnson. I'm a vascular surgeon. I'm chairman of the Quality Performance and Measures Committee for the Society of Vascular Surgery. We're here to talk about five different maintenance issues. I'm going to talk about the first two. 1519, I'm going to limit my discussion on this since the other four have many more questions to be addressed.

So for the second time, I'll limit my discussion on 1519 to the fact that statins have already been shown to decrease the perioperative mortality and morbidity of our lower extremity bypass operations. We do have a compliance problem with range from maybe 1 percent up to -- down to 76 and back 79 percent in 2015, so there is room for improvement with this measure. I'll limit my comments to that.

CO-CHAIR GUNNAR: Discussants are Drs. Olsen and Saigal. Dr. Olsen?

MEMBER OLSEN: As you can see, this is up for a process measure up for approval again,

originally in 2012. The data's really relatively constant over the last four years, with about 75 to 80 percent reporting, ranging from 69 to 84 percent.

CO-CHAIR GUNNAR: As this is a maintenance measure, any other need to revote on the evidence? Seeing none, we will carry forward to gap.

MEMBER OLSEN: The gap original goal was a 90 percent. There's still a range, as indicated, of around 69 percent to 84 percent in compliance with the measure.

CO-CHAIR GUNNAR: Chris, anything else you want to say?

MEMBER SAIGAL: I agree. There's still a gap.

CO-CHAIR GUNNAR: Very good. We can vote. Christy?

MS. SKIPPER: Voting is open for Measure 1519 on performance gap, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: On performance gap, 76 percent votes were high, 24 percent moderate, 0 percent low, 0 percent insufficient. The measure passes on performance gap.

CO-CHAIR GUNNAR: Moving to reliability. Dr. Olsen?

MEMBER OLSEN: Let me find my place here real quick. Reliability has been tested in 100 patients from five institutions or over a two-year period, with 100 percent compliance rate, although it's probably low, considering there's 23,000 patients that have been evaluated over that -- over a five-year period.

MEMBER SAIGAL: I have one comment and question for the developer. They include in here the kappa statistic for the numerator, which I think is more of a validity measure, but it's 80 percent agreement with the -- whether statin was actually prescribed or not. Is that acceptable to the registered participants if they're being used to -- say be measured in terms of value-based purchasing, is that acceptable, in

terms of the validity of the measure?

DR. JOHNSON: Yes, you're correct. It is more validity. I'd like for it to be better than that. That's not acceptable. So hopefully, we can improve upon that aspect.

CO-CHAIR GUNNAR: So we'll address audits and that sort of thing in the next section. John?

MEMBER HANDY: That's exactly my question. Actually, this kind of goes to some of the other SVS measures being proposed. I'm a little confused as to what you're using. You've got a New England database. You've got a Society of Vascular Surgery database. You don't really distinguish between them. You kind of use the New England one as the primary data source, and I don't really know what the penetrance of these databases is and the adherence to them.

DR. JOHNSON: Yes, that's a good question, and hopefully I can answer for the next couple measures because there was a lot more questions. Yazan, go ahead and introduce

yourself.

DR. DUWAYRI: Yazan Duwayri. I'm a vascular surgeon at Emory.

DR. JOHNSON: This was a maintenance -- of course, it's a maintenance measure. All of the development was done based upon most of the data from the New England registry, which eventually evolved into the Vascular Quality Initiative database. We have updated data based upon the Vascular Quality Initiative with validity results and so forth that we didn't send in to you. We can get that data to you. Maybe it's our fault that we were thinking maintenance, and we didn't do the amount of work needed for validity and reliability. But yes, the New England database is now the Vascular Quality Initiative database, which covers the country, which has 400 and some centers in it and, across the board, includes 30 percent academic practices, 30 percent private practices, has cardiologists in it, has interventional radiologists in it. So it's become a very

powerful database. If you need more validity and reliability data from that, we're glad to provide that. Any other comment?

DR. DUWAYRI: Yes, I think that covers most -- VSGNE, which is the Vascular Study Group of New England, is the parent study group of VQI. It basically -- it reflects the same thing. It's currently a regional group of VQI among many other regional groups covering the entire country.

MEMBER HANDY: But it's a self-reported registry, you sign on as a participant, and you submit data?

DR. JOHNSON: Yes, it is self-reporting yet. For instance, in 2015, they went back in and audited. We do audit and validation studies in each hospital and each provider that participates in it. We have quality measures and auditing measures into the system.

DR. DUWAYRI: Yes, the claims data are -- you have to enter 100 percent of your

procedures, it's not selective, and the claims data are tested against data entered in VQI. That's a way of maintaining the integrity of the data.

CO-CHAIR GUNNAR: But in regards to reliability, it's a question -- this is a particular question in the registry that they submit, are you -- is your statin therapy -- is it there, present at discharge, yes or no, and you report that in the registry?

DR. DUWAYRI: Correct. It's a registry entered by data managers and abstractors, yes, correct.

CO-CHAIR GUNNAR: Fred?

MEMBER GROVER: What's your penetration? What percentage of your vascular surgeons in the U.S. participate in this database?

DR. DUWAYRI: I don't have an actual percentage, but the number is growing. We are growing annually, and now we've reached close to 500 medical centers. I think the growth rate is

very fast. Sal may want to --

MEMBER SCALI: I'm Sal. I'm a vascular surgeon from the UF. There are 3,100 vascular surgeons that are board certified in the United States. Currently, there are over 2,000 providers in the VSG or VQI, 48 percent of them are vascular surgeons. If you do the rough math, it's about a third of the vascular surgeons in the United States currently participate in the VQI.

It doesn't mean that they all participate in those modules. Because when you participate, you may just participate in venous disease and not report your carotid outcomes. To your question, how many vascular surgeons in the VQI represent what's going on nationally with carotid-based care, I don't have the specific numbers, but about a third of the vascular surgeons in the U.S. participate in some manner in the registry.

DR. DUWAYRI: Just to answer that question. The data currently has 914 providers

participating in the infrainguinal bypass module, which we are covering in this measure, so 914 out of the total number of vascular surgeons, so around 30 percent -- 25 to 30 percent.

CO-CHAIR GUNNAR: Any other comments? Chris, anything you want to add to reliability?

MEMBER SAIGAL: On this topic, no. My only question was about that validity question, but I guess it's acceptable to the members that are using it.

CO-CHAIR GUNNAR: Any other discussion regarding reliability? Hearing none, voting.

MS. SKIPPER: I would like to note that testing was completed -- reliability testing was completed with patient-level data elements, and according to our algorithm, the highest rating this measure could get on this criteria is moderate. Now, you'll be voting 1 for moderate, 2 for low, 3 insufficient.

(Voting.)

MS. SKIPPER: Voting has closed for reliability on Measure 1519, 82 percent votes

moderate, 18 percent low, 0 percent insufficient, and the measure passes the reliability.

CO-CHAIR GUNNAR: So we'll move on to validity. I'll turn to the developers anticipating the audit questions. Maybe you can fill us in on how you validate the data that's in the registry.

DR. DUWAYRI: So again, we did not do new validity testing. The testing that was performed was performed at the time of the initial submission of the measure, and it was performed by a random sampling of around 100 procedures representative of the procedures enrolled in VQI and abstracting the data separately by separate data abstractors and comparing that to the data already existing in VQI. I don't have the -- the results are present in our submission.

CO-CHAIR GUNNAR: Dr. Olsen?

MEMBER OLSEN: I don't have any additional comments to that.

CO-CHAIR GUNNAR: Chris?

MEMBER SAIGAL: Don't remember what I said.

CO-CHAIR GUNNAR: Any other discussion regarding validity? I think we're open to voting. Oh, I'm sorry.

MEMBER TEMPLE: When we have self-report data that goes into registries and there's an audit process and that's how the registry presented the data before, do you think as this committee, sort of it behooves us every three years that another check of validity's done? In that if a measure becomes reportable, you could see how, potentially, people start changing how they do self-report.

I wonder if, when we ask for maintenance, whether we need to say we need to see the same kind of -- we need to do an audit. You need an audit for the three years between the endorsements? It's a question. I don't know the answer.

CO-CHAIR GUNNAR: Certainly, that audit process is in other registry -- ingrained

in their SOPs. Helen, do you want to comment?

DR. BURSTIN: Karen and I were just having an offline conversation about the need to pull together our stats consultants to kind of have a relook, broadly, at what's required for validity. I think we will add that to the list because I think that's a fair question is if a measure's in use, how does it change, and how does audit fit into that? That's not something we have as a requirement right now.

MEMBER TEMPLE: The other question I had is how much missing data is there with that statin reporting?

DR. DUWAYRI: I don't have that data available right now.

DR. JOHNSON: M2S, or it's called streamlining, is now running our database now. I didn't send this to the committee, but they, in 2015, put out a PQRS Validation Execution Report. They're going in and looking at our data. That one they submitted on behalf of 65 vascular physicians participating in the 2014 PQRS

program. They looked at 12 measures and went through and did an audit on that, looking at ten numbers in NPI combinations and stuff. So there is ongoing auditing and validation going on within our registry.

CO-CHAIR GUNNAR: Is there any feeling that we should delay a vote until we have this information?

MEMBER SAIGAL: There's 2 percent missing data, they report, so it wasn't a very large amount.

CO-CHAIR GUNNAR: We have the answer. Any further -- Collette, yes, and Rick.

MEMBER PITZEN: I have a question. In the reliability section, the kappa statistic for statin at discharge was .8. I don't know if that means agreement, but in the validity, in your random sample of 100, I would expect that you would have a fairly high capture of statin at discharge, so I'm just trying to understand the difference between the two.

CO-CHAIR GUNNAR: Can you answer?

DR. DUWAYRI: Again, this testing was performed a while ago, at the time of initial submission. I'm sorry?

MEMBER PITZEN: The validity? Okay.

DR. DUWAYRI: Yes, the testing we submitted was the previous testing submitted at the time of initial submission.

MEMBER SAIGAL: I think the validity testing, to your question, that they report is really between the discharge summary and the discharge orders, not between the registry and the discharge orders. I think measuring the validity is something else they couldn't quite follow. The key part of validity here was in reliability, the kappa statistics. Looking at what they actually captured, it was 80 percent. That is just a question for vascular surgeons, I think, if that's okay with them, which I think I would have to answer that it is.

MEMBER DUTTON: I'm just going to note quickly looking at the kappa statistics, the reviewers agreed highly on what procedure was

done, 80 percent on whether a statin was given, 100 percent agreement on the patient's age, but only 90 percent agreement on whether the patient died or not. I found that curious..

CO-CHAIR GUNNAR: I believe that's a point of fact and needs to be evaluated in your decision regarding validity.

CO-CHAIR FLEISHER: I urge people to read Tom Lee's editorial in the New England Journal last week about the death of one of his patients.

CO-CHAIR GUNNAR: Any other discussion? Anyone feel that we should not move to vote on validity because we need additional information? Seeing none, I will carry forward.

MS. SKIPPER: All right. We're now voting on validity for Measure 1519, the highest possible rating is moderate, so you will be voting 1 for moderate, 2 for low, 3 insufficient.

(Voting.)

MS. SKIPPER: Voting has closed for validity on Measure 1519, 68 percent votes

moderate, 23 percent low, 9 percent insufficient. This measure moves forward on validity.

CO-CHAIR GUNNAR: Dr. Cima, do you have a question? I see your -- don't hesitate to speak up. He's not locked out, is he, from commenting over the line? You have an open line, so feel free. Chime in whenever he'd like. Next, we move on to feasibility. It's a maintenance measure, but evaluated independently every time it comes up. Dr. Olsen?

MEMBER OLSEN: I was wondering we should use the reporting, it's high, so usability from that standpoint is very easy. If you're not, you're not involved. It's less than 2 percent missing data, as we've already indicated.

CO-CHAIR GUNNAR: Chris, anything else?

MEMBER SAIGAL: We discussed this before. This is a registry measure, so it's only feasible if you're part of the registry, and usable. To the degree to which that's a barrier, that's a barrier, but I think it's feasible and

usable if you're in the registry.

CO-CHAIR GUNNAR: Do you report this data anywhere else, other than to participants?

DR. DUWAYRI: No.

CO-CHAIR GUNNAR: All right, I think we're ready to vote.

MS. SKIPPER: We're voting on feasibility for Measure 1519, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Voting is closed for feasibility, 9 percent votes high, 86 percent moderate, 5 percent low, 0 percent insufficient. This measure passes feasibility.

CO-CHAIR GUNNAR: Moving on to usability. Dr. Olsen?

MEMBER OLSEN: Yes, usability is reported through the PQRS web payment website. There's no other, really, data available. Developer doesn't report whether it has any potential harms in this area. Really pretty straightforward.

CO-CHAIR GUNNAR: Chris, anything else?

MEMBER SAIGAL: Yes, I agree.

CO-CHAIR GUNNAR: All right. Any other discussion? Yes, Larry?

MEMBER MOSS: I understand the measure is being used. I have a question about the impact it's having. When the measure was initially submitted, the rate of statin use was in the 40 percent range, went up very quickly to 79 percent, and then over the last five years has essentially been unchanged.

That still means that 20 percent of patients aren't getting this clearly valuable therapy. Any comments on how we could do better? And a corollary question is this is in registry participants, I've got to assume, across the globe, outside of the registry, the rate's even worse than this.

DR. JOHNSON: Yes, especially since reimbursement is going to be changing. We think that we can get this measure up higher than we

currently are at 80 percent. The good thing about VQI is we are doing clinical improvement activities across the registry, and we do it in regions.

For instance, we realize that our surgical wound infection was higher in people that use the old-fashion betadine prep for prepping, so we sent that message back to our participants and told everybody they needed to change that.

Subsequently, our surgical wound infection went down. The same thing, I think, will occur here, now that people will be a little more incentivized to follow recommendations of the group, as a whole, that we know statins decrease mortality and morbidity with our lower extremity operations. We're at 80 percent.

This is a measure we need to address and get up into the 90s, 90 percent. I think that we certainly have room for improvement, and I think our society in VQI has the ability to communicate back to not only the participants in

this, but taking data from this database, when we go to our society meetings, we're able to communicate that same message back to all the members of our society, all the vascular surgeons.

MEMBER SCALI: I'll just add to that. Part of the issue with the number not changing all that much is that the VQI has seen a rapid proliferation of centers coming online in the last three to four years, the most rapid growth. If you look at the more mature centers, over time, they're actual statin utilization rates approach 90 percent.

It's just that with newer centers, that's not -- you're sort of bringing them into the fold, telling them about the initiatives that are already -- and then they have to sort of put that into their process of care. There's been a bit of a lag that you'll see on things like wound infection, statin utilization, etc.

MEMBER SAIGAL: One comment about that. If you consider that the reliability is 80

percent, or the validity's 80 percent of the measure, and you're peaking at 80 percent, in terms of use, I wonder how much of that is measurement error? Maybe you're getting better performance than you know.

DR. DUWAYRI: That's one of the things that has to be improved. I think it gets improved when we keep this measure going.

MEMBER KO: Can you clarify one of your statements? Because the information we got from your packet was that you're reporting this publicly through PQRS. Is that --

DR. DUWAYRI: Yes, I wanted to correct myself to say that. I thought there -- other than this public reporting, if you were asking about other entities that we report to.

MEMBER KO: Sorry, can you say that again? That was my question. This is a PQRS measure?

DR. DUWAYRI: Yes, all five measures we are submitting today are PQRS measures.

MEMBER KO: Okay.

CO-CHAIR GUNNAR: Yes, A.J.

MEMBER YATES: Since it applies to all the measures, is that PQRS through QCDR?

(Simultaneous speaking.)

DR. JOHNSON: Through the measures group, yes.

MEMBER YATES: The distinction being is that an eMeasure or something that's reported by the self, it's being reported twice, then. If it's being reported to the registry and reported by the surgeon, himself, then if doesn't make sense, but if it's QCDR, then it's going through your registry, and then to CMS.

DR. JOHNSON: Correct, it's going through our registry, then through that, yes.

MEMBER KO: So this is QCDR, not qualified registry?

DR. JOHNSON: Right.

CO-CHAIR GUNNAR: Usability for this maintenance measure.

MS. SKIPPER: We're now voting on usability for Measure 1519, 1 high, 2 moderate, 3

low, 4 insufficient.

(Voting.)

MS. SKIPPER: Thirty-two percent votes high, 68 percent moderate, 0 percent low, 0 percent insufficient, Measure 1519 passes on usability and use. We will now vote on overall recommendation for endorsement.

CO-CHAIR GUNNAR: Any further discussion before that? Okay.

MS. SKIPPER: 1 yes, 2 no, for overall suitability for endorsement.

(Voting.)

MS. SKIPPER: One hundred percent votes recommend this measure for overall suitability for endorsement.

CO-CHAIR GUNNAR: Very well. Next measure -- so that's 1519. Next measure is rate of open repair of a AAA, where patients are discharged alive, Society of Vascular Surgeons.

DR. JOHNSON: You all asked that question. The reason we changed that -- because we did have mortality -- was the Quality

Performance Measures Committee, in a discussion, felt like we needed to project a more positive outcome versus having death from open procedures, so that's the reason that is worded -- that measure. Tim Kronowitz, you all thank him for that.

CO-CHAIR FLEISHER: There is data in the business literature that that's actually a good thing to do. That actually stimulates people in a positive way when you create value statements.

PARTICIPANT: CMS generally encourages positively-phrased messages just because there's less confusion about the specifications.

DR. JOHNSON: Yes, Tim Kronowitz is one of our more experienced members, who participates a lot, came back to us and told us that was the case, and we need to project a little more positivity within our group. Vascular surgeons can be a little cynical, at times.

Regarding the rest of the measure, a

lot of questions have been answered on the first measure concerning our risk adjustment and our validity. Concerning risk adjustment with AAA repairs, we don't do risk adjustments because it gets all about certain selection. That's the reason we don't. Again mentioned was concern concerning reliability and validity.

As mentioned earlier, our kappa scores are all 1, indicating a strong agreement for identification of the correct procedure AA perform, the diameter of our aneurysms, our elective repair, and our hospital mortality. Concerning performance gap, many mortality still has a large range across our centers. It ranges anywhere from 0 to 14.3 percent. The other thing we found is also -- and we don't understand why we do have a higher rate of mortality in Hispanics. We have, also, being Hispanic as an independent risk factor for increased, not only mortality, but also higher hospital charges, which we can't explain. We're hoping to continue this measure as a maintenance and hope to

determine why these things are occurring.

CO-CHAIR GUNNAR: Discussants are Drs. Grover and Levy. Fred?

MEMBER GROVER: First of all, I'd like to congratulate the Society of Vascular Surgeons for really stepping up to the table and coming forward with a lot of measures and respectable penetration in your vascular colleagues. I think this is still the best way to get doctors to buy in and their teams to buy in to improving care when they develop their own measures.

This is relatively straightforward, about as straightforward as something can be. It's basically looking at a gray area, in terms of the size of the aortic aneurysms in males and females and the ones that are enlarged like for females from five to five and a half centimeters for males to five and a half to six. Whereas it' not as indicated as much move ahead just on the size of the aneurysm. These are asymptomatic patients. Obviously, the caveat is that if you're going to operate on one of these

asymptomatic, smaller aneurysms, you really need to have excellent results. That's what they're measuring.

It's not risk adjusted, and I might make some suggestions later about a broader way to look at your aneurysms, where it is risk adjusted. But for this particular measure, I don't believe that's necessary. You're monitoring whether or not your surgeons are making the correct assessment of these asymptomatic patients in achieving a low mortality.

By and large, that's being done on the average. But as you point out, there are variations geographically and from center to center. I don't know whether you want to just move straight on. I think the evidence is supported for what they do. The performance gap, as I mentioned, and as mentioned by the measure developers, is a geographic one. It's low, but it should be. The specifications or, I think, the reliability in regard to specifications, this

is clinical data. It's verified against some administrative data, as well, in the New England group. You audited, I believe, as well. The kappa score is one or approaches one, so it's high. All of these things, I felt, were positive and passable. In terms of --

CO-CHAIR GUNNAR: If we get back to just evidence, itself, because that's what we'll vote on. First, your recommendation is --

MEMBER GROVER: Yes --

CO-CHAIR GUNNAR: -- there is --

MEMBER GROVER: -- it passes.

CO-CHAIR GUNNAR: -- sufficient evidence to support this measurement? Dr. Ko -- or actually, Dr. Levy, and then Dr. Ko.

MEMBER LEVY: I agree.

CO-CHAIR GUNNAR: Okay, Dr. Ko?

MEMBER KO: Is there another AAA measure that's out there for mortality that --

(Off the record comment.)

MEMBER KO: No, not from the SVS. I thought there was another mortality AAA measure,

CMS or AHRQ.

(Off the record comments.)

MEMBER KO: Because the problem with that other measure was that it incorporated both elective and urgent -- emergent AAAs. This is probably a better one, but then just knowing that there's something --

CO-CHAIR GUNNAR: We looked at it. There were no other competing measures. We'll look for that while we continue.

(Off the record comments.)

MEMBER KO: It was something that Pat Romano put out many, many years ago. There was a lot of heat to it because it put all these things together.

DR. JOHNSON: I'll have to look. I'm not sure.

MEMBER KO: It was statistically done well, but clinically, it didn't make sense.

(Off the record comments.)

(Simultaneous speaking.)

DR. DUWAYRI: I can give you the

number here. In five years, for open, we had 4,266 in the registry. In the vascular, we had 13,487, so three times, four times as much, so 75 percent, 25 percent.

CO-CHAIR GUNNAR: All right. Larry, go ahead?

MEMBER MOSS: I've got a question about the exclusions. Is this the appropriate section to raise that?

CO-CHAIR GUNNAR: No. We'll stick to evidence. Is there sufficient evidence to support this measure? Can we move to vote on that?

(Off the record comments.)

CO-CHAIR GUNNAR: Oh, yes, it's a maintenance, that's right. Thank you. Does anyone want to vote on evidence? No. Can we go to the gap?

MEMBER GROVER: We can go to the gap. As he mentioned and I mentioned, the average mortality is very low, but there is a range of difference from hospital to hospital in

geographic areas, so there's still a gap there. There's room for improvement.

MEMBER LEVY: I also think it's critically important, as we're shifting more and more to EVAR, that the ones that are being done open, we definitely need to track outcomes in asymptomatic patients that are being done open.

CO-CHAIR GUNNAR: Helen, is there a competing measure?

DR. BURSTIN: There is a competing measure. Cliff's right. The AHRQ measure is still endorsed, which is the claims-based measure of AAA repair mortality rate. It's IQI 11, number of deaths amongst cases, and it's got four stratum, open repair of a AAA, open repair of an unruptured AAA, endovascular repair of a ruptured AAA, and endovascular repair of an unruptured AAA.

MS. MURPHY: They withheld submitting that one at this time because they're looking at doing some work with Leapfrog to combine the Leapfrog measure with the AHRQ measures.

CO-CHAIR GUNNAR: So remind me the NQF process for --

DR. BURSTIN: You still evaluate this one on its own, on its own merits. When that other measure -- we will have an opportunity to do a competing measures discussion afterwards, but this measure should be reviewed on its own. I think the other distinction is I think the Leapfrog measure was claims-based, as well, so I think both of those are working towards a single claims-based measure. This measure's obviously registry based.

CO-CHAIR GUNNAR: So the fact that the two screens went blank in the front of the room, is that a predictor of the way the afternoon's going to roll out? Those screens all went blank. We're going to vote on gap. Gap is next.

MS. SKIPPER: Voting is now open for performance gap on Measure 1523, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: We're waiting for one

more vote in the room, unless someone stepped out, which I didn't see. Measure 1523 passes on performance gap, 32 percent high, 68 percent moderate, 0 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: Very well, we move on to reliability. Dr. Grover?

MEMBER GROVER: Again, I got a little ahead of myself earlier. This is clinical data, so it's chart reviews, which in my humble opinion leads to likely a higher level of accuracy. The discharged alive is a very clear outcome and easily defined, so from that standpoint, the end point is very simple to collect and collect across centers.

I personally believe -- and let me just digress here. I'll only take one minute -- that the direction here really should go -- would be to risk stratify. You've said in your proposal, several times, you haven't had -- you've got the population to do that, or there aren't risk models out there. My suggestion would be you do your own, based on the number of

patients you have.

Now, you can have a model that would cover a broad spectrum of aortic aneurysms, where you consider size, comorbidities, the whole nine yards, the location and symptomatology, but that's just aside. For this particular study, I think what you're trying to do is to monitor the mortality in a very low-risk group and be sure it is indeed low and stays low, and you don't need risk adjustment for that particular thing. I gave the reliability a moderate score.

DR. JOHNSON: Can I make one comment?

CO-CHAIR GUNNAR: Yes.

DR. JOHNSON: Your point was very to the point because nowadays, when you're training -- I'm in a vascular track residency training program. As we increase the number of endovascular aneurysm repair, we are concerned that our trainees won't have as much experience in open aneurysm repair and what is going to happen to the mortality in an elective procedure, whose only benefit is based upon the natural

progression of that disease to rupture and lead to the death of the patient. That's the reason we think this is a very important maintenance measure to continue.

MEMBER YATES: Yes, question. All surgical specialties have seen a reduction in length of stay with ERS processes and interventions. The discharge of the length of stay has dropped by a day, say, over half a decade, average. It's taking away one less day for an opportunity to not be alive, to be positive. Why stop -- since this is a registry, and I'm assuming that the registry would pick up someone that dies five days after they leave the hospital, why does it stop at the door, as opposed to some sort of time course, like the first 30 days, that would be more in harmony with other measures? This is sort of begging the question, if you get them out day half you really reduce -- or increase your leaving alive.

DR. DUWAYRI: The registry, itself, currently requires that we have long-term

follow-up data that exceeds nine months post-operatively. However, still in compliance with this long-term follow-up data entry, there's not as much as we want. We are in the process of improving that. For accuracy of this measure, we chose the in-hospital death, but yes, we understand the importance of long-term follow up, particularly in those prophylactic measures. Hopefully, we will have, sometime in the future, long-term data.

MEMBER YATES: As part of the data collection for this measure, do you collect which day they're discharged, so that you can monitor whether or not there may be the possibility that people leaving on day two versus day four makes a difference, in terms of the numbers, as opposed to being an improvement?

DR. DUWAYRI: You're asking if we're collecting length of stay as an --

(Simultaneous speaking.)

MEMBER YATES: -- as part of this measure, just as one of the covariables?

DR. DUWAYRI: No, we are not collecting it as part of this measure, but we are collecting it in the registry.

MEMBER YATES: Okay, I'm just pointing out it's a temporal variable that's important right now.

CO-CHAIR GUNNAR: Any other comments regarding reliability?

MEMBER KO: I just have a quick question. If this is a raw rate of the survivors over all the open AAAs, this can be done with a lot of different datasets. You can do this with claims. Can people participate in this measure without belonging to VQI?

DR. DUWAYRI: Claims does not have diameter. We chose aneurysms of moderate size here.

MEMBER KO: Okay, got it.

CO-CHAIR GUNNAR: Dr. Temple?

MEMBER TEMPLE: This is just a point of clarification. Most of the mortality measures that we've been looking at are 30-day

mortalities, am I right? How many of them are a discharge?

CO-CHAIR GUNNAR: The three sessions I've had, I can't independently think of another one.

MEMBER TEMPLE: Again, I guess it's just a question for the group. Do we want to harmonize that everything's 30 day versus discharge? My impression with AAAs is that there is a real mortality between discharge and 30 days, as well. I wonder if you could comment?

DR. JOHNSON: Usually with open AAAs, if they leave the hospital, they usually are going to be alive in 30 days. The ones -- this is an operation where if you don't recover very well right after the operation, you're going to be in the hospital for a month or two with some of these open aneurysms repair. I just don't think there's that high of a -- people that leave the hospital -- and I don't have the data, but people that leave the hospital, the percentage that survive 30 days is pretty high, or should be

high. Do you want to comment?

DR. DUWAYRI: As we know from other registries, 30-day follow up after discharge is very -- we think it's easy. You can do it with a phone call or any other way, but it is not. We have a lot of missing data in all registries at 30 days. I think this will -- as we look at that for surgical patients, we have, first, to make sure that we are able to obtain accurate data at 30 days.

CO-CHAIR GUNNAR: Amy?

MEMBER MOYER: In looking at the links you've provided to your specifications, it's actually not very clear to me what you're measuring, but it seems even beyond the 30 days, you could get around this mortality measure by discharging someplace other than home, based on the description of it. If I send someone to a SNF, or send someone to some other unit, that would potentially escape the measure?

DR. JOHNSON: Yes, in this measure, you could get around that. Hopefully, in the

registry, in our follow up, we would pick that up, but in regard to this measure, yes, if you send them to a nursing home facility and they died, we would have no idea.

CO-CHAIR GUNNAR: Fred?

MEMBER GROVER: I just reiterate the 30-day importance, realizing that it's a little more difficult. We just ran an analysis of our open aortic valve replacements for if you just do it within hospital versus 30 days, if they're discharged prior to 30 days.

We did it for TAVR, as well, the transcatheter aortic valves. Collecting what we call the operative mortality, which includes the 30 days, there's about a 1½ percent to 2 percent increase in mortality if you get the 30 days.

CO-CHAIR GUNNAR: I guess my concern is when you have a dwindling number of cases being electively sent for open repair, if you have an event, it then lingers forever. It also has a huge impact on your overall rate.

You never collect enough cases to

actually overwhelm the fact that you may have had one unanticipated outcome that was just unfortunate. So I guess the question for this particular measure actually comes down to the numbers and the impact of having small numbers, and an event that then overwhelms -- there's this huge noise in the system, then, if you will, per institution.

MEMBER KO: Sorry, so this is at a facility and individual surgeon level?

DR. JOHNSON: Correct, we were able to take the data back to the individual surgeon, and we can take it back to the facility.

MEMBER KO: So if this is a PQRS measure, it's exactly what Dr. Gunnar said. It's going to be you have one case every three years. I think that issue, the technical statistic piece, should be almost separated between facility and individual provider.

CO-CHAIR GUNNAR: I guess the question goes back to when it gets -- this needs to be considered independently, but when it gets

potentially harmonized with the AHRQ measure, which measures far more than just a single elective, open AAA, there may be some value in trying to get these -- it really would -- from my perspective -- and from a quality improvement perspective, don't you really just want to know what the overall rate of discharging patients who have been treated for AAAs is, or elective AAAs? What's the need to necessarily separate open and EVAR, given the fact that now the numbers in open are so low? Go ahead, Fred.

MEMBER GROVER: I could give you some and probably a lot of my vascular colleagues here in the room, I would think probably a lot of the opens, maybe not in this size group, of aneurysm, but they may relate to higher risk patients. I think at this point in time, we need to collect both. I think if you're looking at individual surgeons, you have to lump several years together to come up with a reasonable number of cases. You can't do it on an annual basis, unless maybe you're at the Cleveland Clinic or something.

MEMBER HANDY: I was just going to say that with a longer post-operative follow up, you uncover camouflaged patients that you didn't identify. As a matter of fact, most specialties are moving -- or many specialties, I should say, not most -- are moving to 90-day follow up because there's a body of literature looking at that increased time frame that has plausible morbidity and mortality associated with the operation, as opposed to shortening it.

CO-CHAIR GUNNAR: In your registry, what's the number of AAAs performed by any one, single provider in a year?

DR. DUWAYRI: We don't have the data, but there is -- it's probably not per provider, per medical center is a better way of looking at it, and there is a huge variability. There are medical centers who perform none. So it's hard to provide -- I don't think an average will represent that.

CO-CHAIR GUNNAR: The back of the envelope calculation here is that 25 percent of

the cases in this dataset now are open, and 75 percent are EVAR. If the average facility does 30 cases a year, a single mortality is going to make them an outlier immediately. Barry?

MEMBER MARKMAN: The next measure I'm presenting is pretty similar to -- in fact, it's the same language, actually, throughout the whole thing. Dr. Grover, you said 1 percent or 2 percent? Because the mortality rates we're talking about here are .3 or .4 percent.

(Off the record comment.)

MEMBER MARKMAN: I know. He's talking about 1 percent carrying out for 30 days, but --

MEMBER GROVER: I was using the aortic valve --

(Simultaneous speaking.)

MEMBER MARKMAN: Yes, that's an important point because --

(Simultaneous speaking.)

MEMBER MARKMAN: -- if you do extrapolate out, you would want to look at that data. Do you do any re-admission in your

registry, re-admission rates?

DR. DUWAYRI: Not for now, but this is coming in our registry. Currently, no, there is no re-admission data in our --

MEMBER MARKMAN: Yes, because you can capture a lot of this stuff, this 30-day stuff, with re-admission data. Are there any other measures, other than mortality? Do you look at other datasets, other than mortality?

DR. DUWAYRI: In open aneurysms?

MEMBER MARKMAN: Yes.

DR. DUWAYRI: We also have length of stay is one measure that we -- but it's not submitted here.

MEMBER MARKMAN: Right, no, I'm just saying, you said you have a registry, you have some data points --

DR. DUWAYRI: Yes, length of stay usually reflects other -- the number of complications and the severity of complications that occur after open aneurysm repair. This is one other way of tracking those. Of course,

everything is tracked, including pneumonias, MIs, and other things. The main feedback that we give to centers and providers is death and length of stay.

MEMBER MARKMAN: Going back to what Dr. Gunnar said is that when you begin to harmonize, I don't know what the variables are in the other measures, but you may want to be forward looking and try and see if you can capture --

(Simultaneous speaking.)

MEMBER MARKMAN: Yes.

CO-CHAIR FLEISHER: Getting back to reliability, you mentioned that if they're discharged to SNFs -- I'm not really concerned, but LTACs is where you would send some patient who could linger. One of the nice things -- and we spent a lot of time, and I think it was with the STS group, because if I'm not mistaken, your measure is always 30 days or in hospital, so LTACs could be an exclusion, too? Maybe that's validity, and I have a lot of questions about

validity.

CO-CHAIR GUNNAR: Yes, I think that's validity.

CO-CHAIR FLEISHER: Okay, then I'll --

(Simultaneous speaking.)

CO-CHAIR GUNNAR: We've crossed the chasm between reliability and validity a few times here. Why don't we -- is everybody ready to vote on reliability? This is a maintenance measure currently being collected. Reliability.

MS. SKIPPER: Yes, and again, the highest rating here is moderate, so you'll be voting 1 moderate, 2 low, 3 insufficient.

(Voting.)

MS. SKIPPER: Measure 1523 passes on reliability, 74 percent moderate, 26 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: So now we'll proceed to validity, which I think has been, really, the focus of the discussion for the last five minutes. Lee, you had comments you wanted to --

CO-CHAIR FLEISHER: I gave one. The

other thing is the risk assessment. You actually have data to say that at the extremes of age is just one example where there's wide disparity, but you don't risk adjust. The validity of the measure to measure what you're talking about without risk adjustment has me concerned because with that change, could there be -- if you risk adjust it, then that would at least stratify. That would be more comfortable. But you even are able to show one variable that does have an extreme, you're saying, influence. Then the other validity is I just wanted to know, do you have a plan for -- do you have any data on the clinician level assessment?

Because as you heard, the statistical validity of -- maybe the facility, I think you have, maybe, some data to say it's valid at the clinician level. I didn't see any data to show validity, so that would just be -- three questions, sorry.

DR. DUWAYRI: I'll answer on the age. Again, we're starting just moderate to small

aneurysms. Regardless of what the age is, the risk of rupture in those aneurysms is the same. So if we are, again, looking at patient selection, if you're doing an open aneurysm and you're an 80 year old or 90 year old, then your rate of death is higher than 10 percent for an aneurysm that has a risk of rupture less than 5 percent. I think that is what this measure is looking at. Although you can risk adjust for age, yes, I think still, what we are trying to look at is if the surgeons are performing repairs in patients who are high risk because of their age or other risk factor, other comorbidities and, therefore, have a higher, although acceptable for age, mortality rate. In our opinion, this probably is not justified in a small aneurysm. That's why we are limiting this measure to a small aneurysm. That's the answer.

MEMBER GROVER: I hear what you're saying. I do think, though, that you might get some risk aversion because you aren't taking the age into account. Whereas, if you had a model

that showed -- indeed reflected that even in these small aneurysms, the risk goes up somewhat by various comorbidities, age is one of them, that would be useful. I'm not saying that negates the value of this, but in the future, I think that would be a way I would move.

CO-CHAIR GUNNAR: Alternatively, if I just verify, if I somehow identify a symptom, then it takes you out of this evaluation?

DR. JOHNSON: Yes, because once an aneurysm becomes symptomatic, the risk of rupture substantially goes up, and along with it, the mortality.

CO-CHAIR GUNNAR: Thank you.

MEMBER MARKMAN: I'm just going to hit another point because I assume you submitted the same five articles as new evidence. One of the articles that you submitted was -- had a total of 128,000 patients, 14-year period.

We're talking about risk assessment, but when we submitted that article, the comment or the conclusion on that article goes beyond

patient risk, but it said that hospitals that complete fewer than five OARs and eight EVRs annually have a significantly greater mortality compared to their counterparts, which I think is part of what you're doing in this measure.

You're saying if you don't do enough, your mortality increases. That's why it varies from 0 to 14 percent. What you're doing is you're kind of isolating out the hospitals that don't do enough or the providers that don't do enough. Explain a bit more about the validity of that statement. Because what you're doing is really kind of -- and this is public. This is a public reporting. When it comes to validity, I'm just, to your measure, can you explain a little bit more about it?

DR. DUWAYRI: It's a real point. High volume center, high full volume surgeons do better. That's one of the things that surgeons, before embarking on those procedures, should consider. It will first affect their outcome because it's going to show faster than it is

going to show in a higher volume center, but it's also possibly likely better for the patient to consider a referral to undergo treatment in a higher volume center.

MEMBER MARKMAN: Does your data show that, too? Because I'm looking through your data. Is it statistically significant in your data, in your registry?

DR. DUWAYRI: As far as I recall, I'm not sure that analysis has been done, but it's doable.

MEMBER MARKMAN: Okay. Because that goes to validity because it's all over this measure about low volume. I didn't see it addressed specifically in your measure.

DR. JOHNSON: I agree with you the importance of that. If you look across the nation, more and more open aneurysms are difficult aneurysms to repair and, therefore, more and more open aneurysms are getting referred to tertiary centers. I'm in a tertiary center, so I get more open aneurysms sent in from people

who don't do a lot of open aneurysms anymore. As we progress on, you're going to see that -- we hope that will happen.

To continue to report this data, we have the ability to identify those centers that don't do a lot of aneurysms, and we have an ability to report back to them that you're not doing a lot, and your results aren't that good, and maybe you need to transfer, or you need to refer these patients to other centers. That's the direction that I hope this is going to take is VQI has grown from 160.

We're up to 400 or 500 centers, and it's going to get bigger. Our goal is to let people know what the results are, because we can provide them to the providers and to the facility, and then hopefully encourage people to refer patients to the appropriate center to do the operation.

MEMBER MARKMAN: I don't want to backtrack, but that's basically how you want to close this performance gap because you want to

focus on doing that. Okay, I got it.

CO-CHAIR GUNNAR: I'm going to go to Rick, and then Larry.

MEMBER DUTTON: Just quickly, I wanted to speak in support of the SVS concept of an un-risk-adjusted measure. What we're trying to measure is surgical and facility judgment. Should we do this case? These are elective and discretionary cases. I think that's really getting at the key question. So measuring it in unadjusted fashion does that. If it's a high-risk case, you shouldn't do it, and you wouldn't be in here.

CO-CHAIR GUNNAR: Actually, Larry first, and then Sal.

MEMBER SCALI: I just wanted to sort of follow up and say, again, I think that philosophically, if you think the quality measure is there for patient selection, then I think for something that has a low annualized rupture risk to begin with, then I think you should probably vote to say yes, it should be non-risk adjusted.

If you think the spirit of the measure is to say once you've decided the operation and you're now doing this operation, did your operation do what was intended, which was make sure the patient's alive and they got home and all those other things and didn't die, then it has to be risk adjusted. But it's about patient selection, and it's because the annualized rupture risk of a small aneurysm is so low that the risk/benefit window, therapeutic window for surgery is also very low.

So you're trying to drive or change surgeon behavior to not select the octogenarian with renal insufficiency with a 5.3-centimeter aneurysm, and to do the EVAR, who has a life expectancy that's not going to be long enough to derive benefit of the prophylactic repair. That's sort of the spirit of the measure, I believe.

CO-CHAIR GUNNAR: Larry?

MEMBER MOSS: Now, I'm going to ask a question about exclusion. I'm slow, but I'm

teachable. Related to the point that Barry made, I'm interested in your decision to suppress all data on surgeons who do less than ten cases for an elective procedure with a demonstrated volume outcome relationship, where patients have a lot of time to choose their surgeon.

DR. DUWAYRI: Why we're excluding low-volume surgeons?

MEMBER MOSS: Less than ten, why did you make that decision?

DR. DUWAYRI: Just basically to avoid the problem that we were just talking about, in terms of how a low number will affect the mortality rate.

MEMBER MOSS: Yes, I understand that, but if the underlying concern is that --

DR. DUWAYRI: It's to pick up the --

MEMBER MOSS: -- experienced people should do these cases, are we missing that?

DR. DUWAYRI: It is possible.

CO-CHAIR GUNNAR: Larry, it speaks directly to validity, thank you very much.

Collette?

MEMBER PITZEN: Thanks, this is Collette. I guess I would just recommend, for future, I would like to see a measure that has a larger time frame than discharge.

CO-CHAIR GUNNAR: Cliff?

MEMBER KO: I don't want to belabor the point of risk adjustment for an outcome measure, but I'd like to ask any of the NQF staff, are there a lot of outcome measures that are unadjusted, they're raw rates? Is this something that I'm missing?

DR. BURSTIN: It's not usual. Actually, our criteria specifically say outcome measures should be risk adjusted, unless there is justification why you're not. I think part of what we've heard -- and correct me if I'm wrong -- is that essentially, you're saying the patient selection, by selecting this lower-risk group, inherently, they're doing a bit of poor man's risk adjustment.

We've seen that, for example, with

cataract surgeries. We've seen measures come forward looking at complications after cataract where they limit it only to the lowest risk patients. I think it's almost a way of stratifying as a way of getting at risk adjustment. But it's not common, and it's certainly not -- I think it's an issue that gets raised pretty commonly --

(Simultaneous speaking.)

MEMBER KO: To the measure developers, is that what volume is -- you have a high--volume surgeon, they pick their patients better?

DR. JOHNSON: No, I don't think high-volume surgeons pick their patients better.

MEMBER KO: So there's something beyond selection?

DR. JOHNSON: Right.

MEMBER KO: Which means that the risk adjustment is probably important.

CO-CHAIR GUNNAR: Lee?

CO-CHAIR FLEISHER: As I'm thinking through this -- because there's both selection --

and the crazy thing about selection, when I hear stories, is actually the low-volume surgeons get the worst patients because the high-volume surgeons have the volume to continue to do it. I'm worried about unintended consequences. I don't know where that fits within the criteria -- in usability? Okay, so we'll go back to that because -- thank you. We haven't addressed that. The holy grail in this area, besides patient selection, is actually appropriateness. I'm wondering how your measure gets to appropriateness, as opposed to patient selection by the surgeon of what they'll do. Will there be -- we can get to this in usability. Will patients who should be operated on not get operated on because of the way you've constructed the measure? All you're saying is the patients they chose, not the patients they refused.

MEMBER LEVY: So in an elective case, in a patient who's unlikely to rupture, you're not denying care to someone where there's an algorithm that says these people should have

surgery. I think that's the whole point of the risk adjustment, or lack of risk adjustment for this. These are highly elective procedures, where you really need to be able to do shared decision making with a patient and say, in my hands, these elective procedures for small aneurysms that are unlikely to rupture gives you a better outcome than not doing anything.

I think that's really the issue. I don't think there's an unintended consequence here of denying surgery to people -- the more likely scenario that you're talking about, Lee, is the patient who comes in with symptoms, or the patient who's ruptured in a rural setting. You've got a low-volume surgeon who has to deal with an emergency right then. That's a totally different scenario.

MEMBER MARKMAN: Just one brief comment. It may not just be the surgeon because you're looking at the facility. It may be the fact that the facility that's not doing enough of these cases has the capacity, the nursing, the

ICU and everything else in a community hospital that can't -- it's not appropriate for the facility. They're focusing on the provider. Some facilities have an elective procedure but still a major procedure. That's just a question.

CO-CHAIR GUNNAR: But they exclude low-volume centers.

MEMBER MARKMAN: But the conclusion -- but you just said that's kind of what they're aiming at. They're not studying it, but it's still mentioned throughout the entire measure.

MEMBER CIMA: Isn't that the problem, then, with validity here? You'd have to be doing 40 or 50 aneurysms a year to get ten open. That's a lot of aneurysms for a single site, other than a big academic center.

CO-CHAIR GUNNAR: Do you have any -- back to the developers, do you have any information regarding what percentage of your centers actually meet the threshold for submission or evaluation of this measure?

DR. DUWAYRI: No, not now, but this

can be obtained.

CO-CHAIR GUNNAR: Does anybody feel that information would help us in making a decision, that we should delay the decision on validity or vote on it until we actually have that information?

MEMBER TEMPLE: I have one more question, I think, that fits -- may require coming back to -- again, I'm confused about the registry, in terms of how death is recorded. Is it self-report, and is every physician -- it's a procedure specific for provider and for institution.

Do we know that each provider within an institution reports, or is it just 60 percent of providers that work at that institution? I'm having a little bit of a problem understanding how the mortality rate is actually reported in the registry. That may be more information that we need coming back to before making a final decision.

DR. DUWAYRI: Your question is about

if a provider -- as I understand your question, I'm going to answer it. Basically, yes, if a provider is participating in VQI, he or she will either participate individually or under the hospital umbrella. But regardless, if I start participating in this open aneurysm module, I will have to continue to participate, and I will not choose to enroll this patient who died or avoid to enroll him.

MEMBER TEMPLE: So do I put in the patient died, or does Cobra put in that a patient died?

DR. DUWAYRI: You do, or your data manager. It's basically removed from the medical record.

MEMBER TEMPLE: So, then, if I, as a surgeon, am going to do the open AAA module, is every surgeon in that institution also committed to that module?

DR. DUWAYRI: Most of the participants are participating under institutional umbrellas and, therefore, yes. But that does not -- I

cannot say that 100 percent of the surgeons in all of the institutions are participating. So it is possible that one institution will have some surgeons that are not participating.

MEMBER TEMPLE: Thank you.

DR. JOHNSON: But in general, yes. For instance, I'm with the University of South Florida. Tampa General is the main hospital. They pay for the registry fee, and all the surgeons in Tampa General are underneath that institutional thing, so we all get reported through that. Then when --

MEMBER TEMPLE: You all use the same module?

DR. JOHNSON: All use the same module, yes. We all use the same module, and when it gets audited, we all get audited. When VQI sends in somebody to look at our patients to make sure that each physician has recorded each of his cases, and then recorded and make sure it is reliable and valid data.

MEMBER TEMPLE: I guess I'm cynical,

and I'm kind of wanting to see more -- a little bit more information about how the coding works for the groups and for the providers and the auditing. I think that is affecting my sort of sense of validity. Yet, I really, really recognize how hard these registries are to get off the ground, and I think it's the right thing, to be using registry data, but I think it would be helpful to get a little more information.

MEMBER SCALI: I can clarify. All of the mortality events in the VQI are linked to the SSDI. Every six months, those mortality events are updated. It's like 99 percent accurate for mortality. That's why there's --

(Simultaneous speaking.).

MEMBER TEMPLE: At discharge or a 30-day --

(Simultaneous speaking.)

MEMBER SCALI: -- so that's the issue. There's a six-month lag. So at discharge, they capture, you would have to read that data abstractor for the surgeon who didn't enter the

mortality for that event on that admission.

That being said, mortality, in general in terms of capture within the registry is literally 99 percent, but there is a six-month lag when the SSDI gets updated to that. That's why some of the capped statistics that people have talked about are reviewed. So it is possible that you can get 30-day mortality events from the registry, if that's specifically what the committee recommended for you guys. That's the clarification on mortality.

CO-CHAIR GUNNAR: So validity speaks to the participants for not participating?

MEMBER SCALI: For open AAA, and specifically, obviously, the vast majority of open AAA repair at least in the registry is done by the vascular providers who are doing that. There are about 10 percent -- I'm the chair of the EVAR committee, but I also do a lot of work with the open aneurysm repair committee for the VQI, so I know that at any one institution that participates in these modules for data capture,

it is theoretically possible, if you're a single surgeon in a five-surgeon group and you wanted to participate and you open AAA repair module, you could be a single person that does that.

Then the VQI doesn't force -- because it has to be voluntary. It's less than 10 percent of all of the people putting data in the registry for open and endovascular AAA repair that have that specific scenario. The overwhelming majority of participants who are in the module, it's at a center level, so they capture all of the cases at the level.

MEMBER TEMPLE: So should we be looking at center rather than provider level?

CO-CHAIR GUNNAR: Again, I think we're voting on validity recommendations back to the developer on how -- this is a changing landscape because of the relationship between -- an asymptomatic patient's relationship between referrals to open versus EVAR. We're going to get to that in the next measure. The validity of this particular measure, as far as the data

that's in the registry, the information that's in the registry, I think that's been addressed. Does everybody agree? Okay, which means we can then carry on and vote? Okay.

MEMBER KO: When we vote for validity, we're voting for both the -- to what Larissa was asking, both facility and individual?

CO-CHAIR GUNNAR: That's how it's specified.

MEMBER KO: Exactly, unadjusted facility/individual --

(Simultaneous speaking.)

CO-CHAIR GUNNAR: Correct.

MEMBER KO: Okay.

CO-CHAIR GUNNAR: And does it actually measure what it's measuring?

MS. SKIPPER: Voting is open for validity on Measure 1523, 1 is the highest rating, moderate, 2 low, 3 insufficient.

(Voting.)

MS. SKIPPER: Measure 1523 passes on validity, 61 percent votes moderate, 30 percent

low, 9 percent insufficient.

CO-CHAIR GUNNAR: We'll move on to --

CO-CHAIR FLEISHER: If I may add --

CO-CHAIR GUNNAR: Yes.

CO-CHAIR FLEISHER: Just a point of letter. I would strongly urge the measure developers to consider the clinician issue because you may want to consider whether or not -- I think there were many of us who felt uncomfortable with that aspect of the validity.

MEMBER YATES: You mean in terms of the volume?

CO-CHAIR FLEISHER: In terms of validity and whether this can be measured, I ask them to consider, during the post-comment period, getting us data on that validity because it just squeaked by. I agree with Larissa, and I've heard it from others.

MEMBER YATES: I just want to make sure I understand; it's the volume question being suppressed that you're worried about?

CO-CHAIR FLEISHER: Whether it was

tested, and whether there was sufficient volume to say anything at the clinician level.

MEMBER YATES: Thank you. Just making sure I understood.

CO-CHAIR FLEISHER: That's what I'm asking. I don't know if others are asking that.

CO-CHAIR GUNNAR: Looking at it sort of globally, I would recommend that it be brought together with this other -- this is a measure that actually is claims based, right? You don't have to necessarily -- it's not currently risk adjusted. It's really just dead or alive at discharge. It's CPT code. I had an open aneurysm. The only thing, the caveat to this is asymptomatic and some moderate size issues that clarify the cohort.

I don't know how -- I think it's worth a discussion later downstream whether or not these two measures are so disparately separate, or they can be brought together. That's the way I would look at it. It's an important thing to measure. I don't think anybody disagrees in the

room. The question is does it need so many of these excluders that it actually gets us down to a group of patients that now, I'm not even sure exists that often, quite frankly.

(Off the record comment.)

CO-CHAIR GUNNAR: All right, next is feasibility. It's currently measured. It's been so long that -- who's our commenters again? It's Fred and Barbara, right? Yes. Any issues on --

MEMBER LEVY: No --

CO-CHAIR GUNNAR: -- feasibility from either one of you?

MEMBER LEVY: -- other than you have to participate and register.

CO-CHAIR GUNNAR: Shall we vote?

MS. SKIPPER: Voting is open for feasibility, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: We're just waiting on one more vote for feasibility. We'll move forward with 22. Results are 45 percent high, 55

percent moderate, 0 percent low, 0 percent insufficient. The measure passes on feasibility.

CO-CHAIR GUNNAR: All right, we'll move on to usability. Comments, Dr. Grover? Dr. Levy?

MEMBER GROVER: It's very easy, it's very straightforward, it's very simple.

CO-CHAIR GUNNAR: Right. Any other discussion? Yes, Dr. Moss?

MEMBER MOSS: I wanted to raise a comment about unintended consequences. I'm glad Barbara brought up patient choice. I can't think of a better, more exemplary example of a situation where patient choice is highly relevant.

I could imagine patients that are moderate risk that don't want to walk around every day with the risk of sudden death, and they might choose to take that risk up front and have the operation. The question is does this take patient choice off the table and make that value judgment for them?

MEMBER LEVY: I think not. I think what it does is it gives patients valid, we hope, data to say in the hands of a certain institution, certain surgeon, given their circumstances, that the risk of mortality from the procedure is greater than or less than the risk of walking around with the aneurysm. Definitely, you're right, we live in a world where our patients think more intervention is better. To the extent that we can provide data for patients to show what the down side to intervention is or can be or you could project that it would be for those patients, I think that's really important.

MEMBER MOSS: I guess I'm also asking are we disincentivizing centers and surgeons from taking on moderate risk patients who want an operation and make that choice for themselves?

MEMBER LEVY: Shouldn't we be, if the risk of death for them is higher than the risk of rupture and mortality? The question is shouldn't we be? Shouldn't we be collecting enough data to

be able to answer that question?

DR. JOHNSON: The discussion that occurs at a patient level -- in fact, we train our vascular surgeons this way -- is what is the natural progression of the disease if we do nothing at all versus intervening? In that discussion, I need as much data as I can tell them about what happens when I intervene, as far as mortality and morbidity.

Because I know what happens with the natural progression of rupture without intervening. I don't think we're going to de-incentivize -- surgeons like to operate. I don't think we're going to de-incentivize them from that. Hopefully, we're just going to de-incentivize them from operating on people that the outcome is going to be worse if we do nothing at all.

MEMBER CIMA: The one thing is given the volume threshold requirement, a lot of surgeons and institutions are not going to be captured by this, so we're not providing patients

with any reasonable information -- the vast majority of patients.

DR. JOHNSON: I would say that the vascular surgeon community is so small -- there's only 5,000 of us -- that the majority of open aneurysms nowadays, since the communication levels -- already open aneurysms are getting moved into tertiary centers. I think most of the information is getting out to patients and other surgeons to say that if you're not a high volume -- if you're not doing this operation ten times, fifteen times a year, you don't need to be doing this operation.

CO-CHAIR GUNNAR: Just remind me again, this is reported every six months in a rolling -- it's a rolling 12 month, with 6-month submissions, right?

DR. JOHNSON: Right, correct.

CO-CHAIR GUNNAR: So the unintended consequence is if I have -- if I was going to do ten or eleven open AAA repairs, and I had one death in number four, the one thing I'm not going

to do is -- seven more. I might do six or five, but I'm not going to do -- I'm going to refer two of those out. Okay, Amy?

MEMBER MOYER: I had a slightly different scenario for the usability to a patient and where can we really effective decision making in a patient. I'm wondering if by splitting out open repair and endovascular repair, if we're not making it more challenging to give that patient usable information? Because I'm guessing they're not like, I think I need this kind of a specific repair. I'm going to look and I'm going to make my surgeon choice.

They're probably going to have a conversation with the surgeon first, and once that conversation, unless it goes very poorly, will happen, it's hard to redirect care at that point. It really does kind of come down to the surgeon then saying, you need this type of intervention. You should go to this area. I don't think that --

(Simultaneous speaking.)

CO-CHAIR GUNNAR: I think there's an intersection, which is what we're getting at, is the intersection between informed consent on a case-by-case basis. Your obligation or your participation and reporting your outcomes in these cases. That's separate and distinct, but not really germane, I think, to this discussion, although it is from an ethical and quality improvement point of view, if you will.

I don't see -- I guess I don't see how we evaluate a measure based on, as we've come to a number of places here, the unintended consequences or how it might be manipulated or this idea about gaming. That's for a separate discussion that NQF is committed to have. I think we have to evaluate them for what they are on face value. Yes?

MEMBER MCCARTY: Just another comment about the way that the volume can play out at the individual level. It's not clear to me if this particular measure gets reported by facility, and then by surgeon. Some of the metrics I'm

familiar with are that way. I've encountered the scenario where a surgeon who wants to participate and do a procedure at an off-site facility runs into this issue of that satellite facility doesn't have the volume, so then they get dinged if something happens at that satellite facility, or there's more restrictions about there has to be two surgeons, and then it depletes resources because you have to have two people.

I don't know if there's a way for the developer, when looking at getting that data, to look at the surgeon level, if it can be across facilities, and not just by each individual facility where someone might practice.

CO-CHAIR GUNNAR: Again, this goes back -- informed consent and what you're bringing up is actually an ongoing professional practice evaluation and the constant need to update your credentialing and privileging of an individual. Personally, I see that the data exists, but the processes are separate and distinct, if you will. But I appreciate the perspective. Lee?

CO-CHAIR FLEISHER: Perhaps the one thing I can suggest is one of the advantages of the steering committee is continuity, as well as being able to ask for certain things when this comes back either for maintenance in a year or maintenance in three years. If I can ask that we collect some of these concerns that people have, assuming this measure passes, that we would like to see. There's actually, now -- Helen and the group created this wonderful database, so that it's a living document about some of the concerns we have, so the next time it comes up, that will exist.

This is something relatively new, the full extent. Does that sound reasonable to people, that we would get some of these concerns of what we'd like to see if this measure comes back, either in one year or three years?

MEMBER YATES: I think that's a great idea. I think it should be captured -- you could technically call it the captured concern database and non-technically call it the worry list or

anxiety closet, but somehow or another, having a collection of those, so that the -- there's better ways of making an acronym for it, but I think that you have to come up with something that collects those.

(Off the record comment.)

MS. MUNTHALI: You mean the annual update process or --

MEMBER YATES: No, the data --

(Simultaneous speaking.)

MS. MUNTHALI: We don't have a clever name. I would suggest that you give a specific time frame to the developer. It would be for us to say during your annual update period, in a year, if you can come back with the changes, we'll make sure we capture all of that.

CO-CHAIR FLEISHER: Great --

(Simultaneous speaking.)

MEMBER YATES: Yes, but Elisa, don't you intend that to be something that's attached to the Phase 3, then Phase 4 --

CHAIR SHORT: Yes --

(Simultaneous speaking.)

MEMBER YATES: It should be in that dataset, as opposed to being outside of the dataset.

CHAIR SHORT: Correct, that is what they created. Yes, that has been a major project, if I remember correctly.

MEMBER LEVY: Just to follow up on that, as I look through these measures and look at things that are coming back for re-endorsement, it would be helpful to have an intro to the measure, when we first click on it, that gives us the worry list, because it frames for us how to look at that measure as we're looking forward, rather than embedding it in the discussion of the different elements that we're looking at. I think having a frame as we're beginning to look at that measure would be really helpful.

(Simultaneous speaking.)

MEMBER YATES: Then along those lines, when I was asked to prepare a preliminary data

sheet, no one asked me to list a set of concerns separately that would be good to add to that. It would be good that this script allows for a list of concerns to be created, so that there's an easily retrievable list for the staff to put together into that list.

Maybe the group, as a whole, could knock some of them off and say, don't worry about this; don't worry about that, from the conversation, but leave these. That would add something to the script and the preparation by the prep ahead of time. I think that'll go to what you're talking about.

CO-CHAIR FLEISHER: That's great and maybe the lead discussants can even take a role in working with staff to solicit input to create those data elements.

MEMBER YATES: Given the fact that we didn't do it this time -- because I have lots of things I'd like to put in my list

(Simultaneous speaking.)

MEMBER YATES: Can we be given the

opportunity, or have it sent to us, to submit those questions or worries, so that they can be collected from this meeting, because we didn't do that prospectively?

CO-CHAIR FLEISHER: We can do whatever we want. Sounds great.

MS. MURPHY: Can I just say one thing real quickly? We can do that. What I think will be really important for the group to do is be certain that whatever you submit as concerns are really important concerns to you that you positively want information back on in some period of time. Otherwise, you will drown yourself in this kind of information.

CO-CHAIR FLEISHER: Perhaps if we do that, then we will collect it, send it back to everybody, and check that it has the consensus of the committee, sort of the editorial work. Go ahead, sorry.

CO-CHAIR GUNNAR: Experience drives wisdom. Anything else, Dr. Yates?

MEMBER YATES: No. I mean, when in

doubt, don't. Benjamin Franklin.

CO-CHAIR GUNNAR: Yogi Bear? Usability and use, can we vote?

MS. SKIPPER: Voting is open for usability and use, Measure 1523, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Measure 1523 passes on usability and use, 22 percent high, 65 percent moderate, 13 percent low, 0 percent insufficient. We will now vote on overall suitability for endorsement.

(Voting.)

CO-CHAIR GUNNAR: We'll wait for this, but I would not, I would just keep pushing. We have now moved into the woefully behind schedule phase.

MS. SKIPPER: Just for the record, Measure 1523 passes on overall suitability for endorsement, 78 percent yes, 22 percent no.

CO-CHAIR FLEISHER: Just a point because we have STS waiting, which is why we're

pushing through. Tomorrow, we'll start at 8:00. I don't know when the food will get here, but the STS, our colleagues, need to be out of here by 10:00, so we're going to start very promptly.

CO-CHAIR GUNNAR: Very well. We'll move on now to in-hospital mortality following elective EVAR of AAAs, Society of Vascular Surgeons.

DR. DUWAYRI: This measure parallels the one that we were just talking about. It discusses EVAR, endovascular repair of abdominal aortic aneurysm. It is not worded as positively as the open aneurysm repair measure. We're looking at mortality rate after EVAR. EVAR, over the last 15 to 20 years, has become an alternative to open aneurysm repair.

Multiple randomized control trials have shown its effectiveness in preventing death from aneurysm rupture. Perioperative mortality has been found to be lower than open aneurysm repair and, therefore, it has become the most common method of repairing aneurysms. This

measure looks, again, at repairing aneurysms that are at relatively low risk for rupture. Again, the key concept in this measure is that patients who are at low risk of aneurysm rupture should only be offered elective EVAR if their procedural mortality rate is low. Therefore, again, here, no risk adjustment is performed. I will leave the rest for your discussion.

CO-CHAIR GUNNAR: Our discussants are Drs. Dutton and Scali.

MEMBER MARKMAN: I think you're jumping forward.

CO-CHAIR GUNNAR: Did I -- oh, I did. My apologies. Thank you for the correction, Dr. Markman --

(Simultaneous speaking.)

MEMBER MARKMAN: In fact, in this 1,000-page document, I don't think 1534 was in there -- 1,000 pages, it's unclear, we're trying to find it. You can bring it up, but it's exactly right. It mirrors the preceding measure. It's outcome maintenance, data collected by PQRS,

publicly reported. The evidence is the same. They gathered five new studies which were similar to the ones in the other, so I think we could bypass the evidence.

CO-CHAIR GUNNAR: As this is a maintenance measure, we can -- any other discussion regarding evidence? Any further need to vote on evidence? Dr. Yates, you're good?

MEMBER YATES: Yes.

CO-CHAIR GUNNAR: You have your -- we will carry on to gap?

MEMBER MARKMAN: Right, I asked my questions on the gap before, in the preceding, very similar --

CO-CHAIR GUNNAR: Other than the denominator being significantly larger?

MEMBER MARKMAN: Yes.

CO-CHAIR GUNNAR: Any other discussion on gap? All right, shall we vote? Is Christy around?

MS. SKIPPER: She just ran out.

CO-CHAIR GUNNAR: We can do it by hand

for this one if you want, unless you --

MS. SKIPPER: We might want to do that. Let's do it by hand.

CO-CHAIR GUNNAR: You want to do it like this? We're ready for gap.

MS. SKIPPER: Voting is now open for performance gap on Measure 1534: 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: This measure passes on performance gap, 38 percent high, 62 percent moderate, 0 percent low, 0 percent insufficient.

MEMBER MARKMAN: In terms of reliability, you answered our questions about the percent of the members that are in it. I don't think we have to go over that again, and we felt like it was okay in the last measure, so.

CO-CHAIR GUNNAR: Any other questions about or comments about reliability? Hearing none, I think we can vote.

MS. SKIPPER: We're voting on reliability for Measure 1534. Again, the highest

rating is moderate at 1, 2 low, 3 insufficient.

(Voting.)

MS. SKIPPER: Measure 1534 passes on reliability, 82 percent moderate, 18 percent low, 0 percent insufficient.

MEMBER MARKMAN: In terms of feasibility, we discussed that, too.

CO-CHAIR GUNNAR: Reliability -- validity.

MEMBER MARKMAN: Validity, yes.

(Simultaneous speaking.)

CO-CHAIR GUNNAR: Validity is next.

MEMBER MARKMAN: Too many -ilities there. We've discussed the validity question, which was the major question that we had raised before. We've addressed a lot of these issues.

CO-CHAIR GUNNAR: Any other comments? Go ahead and vote.

MS. SKIPPER: Voting is open for Measure 1534, validity, 1 moderate, 2 low, 3 insufficient.

(Voting.)

MS. SKIPPER: I've got 21 votes, looking for one more. Moving forward with 21 votes. This measure passes on validity, 76 percent moderate, 24 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: Usability?

MEMBER MARKMAN: No, it's feasibility.

CO-CHAIR GUNNAR: Feasibility.

MEMBER MARKMAN: Feasibility. I was just commenting that you're only missing 1 percent of your -- that's in the high category. That's pretty good. That wasn't mentioned before, that 1 percent of the data, so I think there's a high feasibility to it.

CO-CHAIR GUNNAR: All right, any other discussion? Move forward.

MS. SKIPPER: Voting is open for feasibility: 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Moving forward with 20 votes on feasibility, 40 percent high, 55 percent

moderate, 5 percent low, 0 percent insufficient. The measure passes feasibility.

MEMBER MARKMAN: Last time you came before the Committee, you were thinking about putting these two together. You decided to go separate ways, but now we're a couple years later. Do you want to put it together in the same measure, or consider that at some point when you come back? That's my only comment. I read the notes from last time comparing -- yes, EVAR and open in the same measure.

DR. DUWAYRI: I think they are different. They're offered for different reasons. Therefore, we prefer to keep them separate.

MEMBER MARKMAN: Okay. You said that last time, too, but I'm just bringing it up because you did make the comment last time, or the committee made the comment, I mean, you're doing this so why not compare it in one measure. I don't know if there's any comments on that.

CO-CHAIR GUNNAR: Any other

discussion? Go ahead and vote on usability.

MS. SKIPPER: Usability and use, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: This measure passes the usability and use criterion, 38 percent high, 52 percent moderate, 10 percent low. We'll now vote on overall recommendation for endorsement with 1 yes, 2 no.

(Voting.)

MS. SKIPPER: Measure 1534 is recommended for endorsement, 91 percent yes, 9 percent no.

CO-CHAIR GUNNAR: All right, moving on to 1540, post-operative stroke or death in asymptomatic patients undergoing carotid endarterectomy, SVS.

DR. DUWAYRI: The definition of this measure is looking at the mortality and stroke rate after elective carotid endarterectomy performed for stroke prevention in asymptomatic patients. Carotid endarterectomy has been proven

by randomized trials, in the past, to have benefit in stroke prevention in the presence of carotid stenosis.

The key concept, again, here is to offer this surgery to patients who would benefit from this stroke risk reduction and who are not high risk to -- who are not at high risk of death or a stroke in the perioperative period. The usual perioperative mortality and stroke rate in the literature, from randomized control trials, is somewhere between 1 to 2 percent. However, many surgeons and centers who perform this procedure do not meet the high standards that are followed in randomized control trials.

There is a variation in the literature, when looking at the real-world data in the perioperative stroke and death rates, that goes from 1.4 percent in the hospitals that participated in these trials, up to 2.5 percent in low-volume hospitals. We also have VQI data from around 1250 providers, including 27,000 carotid endarterectomies. The median

perioperative stroke and death rate is .4 percent, but it ranges from 0 percent to 100 percent, again, reflecting low-volume centers. It's also found that the rates are higher in females and older patients. Therefore, there is a gap -- there is variability in the literature, and we feel that this measure is a good quality measure in patients who would undergo an elective prophylactic procedure.

CO-CHAIR GUNNAR: Discussants now are Dr. Dutton and Dr. Scali.

MEMBER DUTTON: I think this measure has high face validity. It's obviously what the patient wants to know if they are going to undergo this procedure. My only real question on the evidence, is it the highest level? Compared to maximal medical management -- Dr. Dutton, you have an asymptomatic carotid stenosis. Take statins, take aspirin, versus have a carotid endarterectomy or a stent. When is the endarterectomy or the stent indicated, and what is the most current evidence about that?

DR. DUWAYRI: Well, the evidence that we have from high-quality randomized control trials is relatively old, from the ACAS study that told us 60 percent is a good threshold for treatment. Most surgeons these days do not follow this and go higher, up to 80 percent, again, based on the knowledge that best medical therapy is much better than it was 20 years ago.

There is an ongoing randomized control trial now, CREST-2, which randomizes to carotid endarterectomy versus current maximal medical therapy or carotid stenting versus current medical therapy. I think the data is still to come. We do not have current randomized control trial data to tell us the answer. But most surgeons, in general, follow the 70 to 80 percent rule, but the data that we have available now is old and would support surgeons performing this for more than 60 percent.

The recommendations from the American Heart Association is to not offer this procedure or carotid revascularizations unless you can

predict that your perioperative stroke or death will be less than 3 percent.

CO-CHAIR GUNNAR: As this is a maintenance measure, the question to the two of you is that the evidence really hasn't changed since the last time, and nothing new has come forward. Is your recommendation that we vote on that, or that we proceed?

MEMBER DUTTON: I don't think we need to vote. Dr. Scali?

MEMBER SCALI: Yes, there's an evolution in the data in the post-statin era, but no Level 1A evidence, like Yazan alludes to. There's certainly some literature, and there's naysayers that say that the role of surgery, or stent, for that matter, for asymptomatic patients is no longer, because the annualized stroke rate for a patient on a high-intensity statin, and some of the natural history data trials which are out there, published in Stroke 2010/2012, can have annualized stroke risks of 1 to 1.4 percent. In the old data, we were quoting patients 2

percent annualized stroke risk.

We've reduced that risk to 1 percent, so it's a 1 percent annualized stroke risk reduction benefit for surgery over best medical therapy in the pre-statin era. That was what we hung our hats on, and that's what we still do today because we don't have the Level 1A evidence. Because there may be certain subgroups, those who have carotid disease progression, those who have contralateral occlusion, et cetera, they're younger patients, et cetera, what may still benefit from prophylactic endarterectomy. So most societal guidelines, certainly the SVS, still holds that ACAS sort of governs our decisionmaking, but it does have to be surrounded -- you have to be very circumspect about who you offer this operation to.

I don't think we need a vote on the literature. I just would say that there are a lot of caveats when this discussion rolls out with the subsequent sections about how to measure

and validity and so forth because where a lot of the decisions are being made are based on trials from the 1990s.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: Unlike AAAs, which may be picked up by ultrasound or MRI of the spine incidentally, CT scans, finding a carotid that's in trouble means that somebody's looking. Somebody has to do an elective ultrasound. Likewise, the lower risk from the statins has to be aggressively treated. There's a lot of things going on. I only see a question about disparities that has to do with gender and age. Is there any evidence, or was there any effort to look at disparity in terms of treatment of what is probably the sine qua non of elective vascular surgery, that being the endarterectomy of the asymptomatic carotid? Is there any evidence for disparity?

(Simultaneous speaking.)

MEMBER YATES: -- that's part of the evidence that has -- the maintenance component

has to go through this time?

DR. DUWAYRI: You mean by physician, medical center, region, those kind of things?

MEMBER YATES: No, by STS versus doctors.

DR. JOHNSON: But are you saying in regard to risk of stroke with no treatment at all, or risk of stroke after treatment?

MEMBER YATES: Access to the care, and also the outcomes, perhaps because they come later, or their symptoms haven't been as carefully observed? We're supposed to, in this maintenance period, look at that question, and that question I don't see being addressed in this.

MEMBER SCALI: Great points. I would just say that there are currently no screening guidelines for carotid disease, for asymptomatic carotid disease, so we don't really know the antecedent history, in terms of -- we just know the patients with PAD, by and large, and/or CAD, also have a high preponderance of carrying

carotid occlusive disease. But there's no welcome to Medicare physical, get your ultrasound to screen for aneurysms, so you wouldn't really know about certain disadvantaged groups that may or may not get found to have that disease.

We do know the prevalence of risk factors that predict likelihood of having carotid occlusive disease, but we really don't have any screening guidelines to say these patients may have been vulnerable or didn't get screened appropriately, or there was delays in care or access to care that led to a higher subsequent stroke risk. I'm unaware of any literature or guidelines that would help us clarify those points in the maintenance phase.

MEMBER YATES: I only say it because these things are found either because of somebody with a high degree of medical expertise that wants to be screened, or they are the end result of a high-utilization patient that, because of other procedures, gets screened. As such, it's sort of like a -- it's a composite measure,

really, for disparity.

MEMBER SCALI: Yes, the more vocal majority, as a director of a busy vascular lab, this is one of the most common things that we see, in terms of carotid ultrasounds are pre-op evaluations for cardiac surgeon, pre-op evaluation for some other operation, a bruit, those types of things, or there's symptoms that aren't consistent with lateralizing hemispheric stroke, and there's some dizziness workup, et cetera, syncope events. That's the vast majority of the indications where people come to ultrasound for screening.

CO-CHAIR GUNNAR: Liz?

MEMBER EREKSON: I just want to make one point about healthcare utilization and disparity. There's a lot of work that's been done with healthcare utilization and the geographic variation that happens with it. If you take disparity out, you still have huge variations in the utilization that are not explained by disparity. I think we just have to

be careful with that.

If you look at it, if you take out disparity, the utilizations are more likely defined by the hospital region you're in than it is by your disparity.

MEMBER YATES: It's just a point of reference, though, that we're -- it's a point of continuity here that we're supposed to ask that question in this process, at this meeting. This is the one that's more elective than the procedures because if it's coming out of varied screening processes.

CO-CHAIR GUNNAR: To get back to sequence --

MEMBER YATES: That's part of the sequence in this.

CO-CHAIR GUNNAR: Eventually, when we get to validity, right?

MEMBER YATES: No, it's supposed to be disparities are in the gap, in the first part, in the evidence.

CO-CHAIR GUNNAR: Just for

confirmation, we've all agreed that we don't need to vote on the evidence?

MEMBER YATES: That's fine.

CO-CHAIR GUNNAR: Anyone believe that we do? Okay, now we go to the gap.

MEMBER YATES: And I've made my point.

CO-CHAIR GUNNAR: Any other comments?

MEMBER DUTTON: The gap in measured performance presented is small.

I think we heard about a 1 percent difference between good centers and bad centers. But again, referring to the discussion this morning, this is an outcome measure with strong importance to patients, so I think as a public accountability measure, I think this is very appropriate.

MEMBER SCALI: That, and I also think that there's enough center and region variation that we sometimes pick up in-patient sample papers, Medicare papers, or even our own analysis through the VQI, there's still enough center and region variation that's of import, even though

the surgeon-specific outcomes are -- they're driven incredibly low. The perioperative stroke risk contemporary outcomes, asymptomatic carotid disease, in those patients, it's a 1 percent number, even lower in many circumstances nowadays.

CO-CHAIR GUNNAR: Any other discussion on gap? All right, I think we're ready to vote.

MS. SKIPPER: We're now voting on gap for Measure 1540, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Measure 1540 passes on performance gap, 5 percent high, 82 percent moderate, 14 percent low, and 0 percent insufficient.

CO-CHAIR GUNNAR: Moving on to reliability.

MEMBER DUTTON: Again, this measure's been discussed before. I had no particular questions around it. Dr. Scali?

MEMBER SCALI: I agree. I think your

points were well taken in your critique. I would echo the sentiment that it's technically a construct of two different outcomes that seem very reasonable because obviously individually significant and serious and important for the patient, provider, and institution, so yes.

CO-CHAIR GUNNAR: Any other discussion? Hearing none, proceed with the vote. Oh, I'm sorry, Liz.

MEMBER EREKSON: I'm sorry. I just have one question for the developer, and this just echoes the previous measure that we reviewed from a different society. How do you define stroke, and do we need a validated way of measuring stroke in this population? Obviously, this is a very different measure, and you're looking at 120-day outcomes, but I think it might help the developer of the other measure.

DR. DUWAYRI: Stroke is defined as any post-operative neurologic deficit, whether it's from a hemorrhagic or an ischemic event, whether it's ipsilateral or contralateral to the

treatment site, any neurologic deficit that lasts more than 24 hours in those previously completely asymptomatic patients.

MEMBER EREKSON: Does the surgeon decide that? Does a neurologist decide that? Does the coder decide it?

DR. DUWAYRI: It's a coder slash surgeon. We enter Rankin's score, but this outcome is based on, really, what the medical record states, which is basically what the surgeon says.

CO-CHAIR GUNNAR: Any other questions? Voting on reliability.

MS. SKIPPER: You'll be voting reliability for 1540, 1 moderate, 2 low, 3 insufficient.

(Voting.)

MS. SKIPPER: Measure 1540 passes the reliability criteria, 86 percent moderate, 14 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: Okay, next is validity.

MEMBER DUTTON: Again, I had no issues.

CO-CHAIR GUNNAR: Any other discussion? Hearing none -- oh, Collette?

MEMBER PITZEN: This is Collette. I have the same concerns with the mortality measures that we looked at previously, in terms of having a longer window instant to discharge.

MEMBER SCALI: I can answer that. This has actually been studied pretty extensively, at least in carotid patient population. It sort of speaks to the prior discussion about the 30-day versus in-hospital outcomes now. It may be very relevant to harmonize with what's been reported in the literature. Maybe we can ask the developers to speak more about using the in-hospital versus 30 day, but we've actually studied this in several different methods, whether it be through Medicare, in-patient sample, as well as through the VQI.

Patients who suffered in-hospital

stroke or death is about one third of the events, as it was discussed earlier, that occur out of the hospital, but within 30 days. The identical predictors have been shown and reported in several papers. The patients who get those events in hospital have the identical predictors for the ones who get it out of hospital.

While it is true it dominates the overall denominator of the events, as long as you're comparing across centers and surgeons within the same time points, whether it's the in-hospital event or the 30-day event, you should get the same relative differences as shown. About one third of the events that occur occur out of the hospital, but it's an identical set of predictors that shows patients who get those events in hospital are the same as the type of patients who get them out of hospital.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: I'd just like to reply from a patient-centric standpoint of the measure that could have the potential to have fairly low

rates of mortality, which is a good thing, I think it's important to appropriately handle.

CO-CHAIR GUNNAR: Appropriately what?

MEMBER PITZEN: To include an extended window. You can say theoretically and epidemically everything is all the same, but when you have something that's reported at a very low rate. If, perhaps, the concern for mortality or whatever within a window is more important to them making that decision, I just think that there's some value in considering that.

CO-CHAIR GUNNAR: Dr. Handy?

MEMBER HANDY: That was my point, exactly, earlier. Every specialty that looks at this, the longer you look, the more you find.

MEMBER DUTTON: The problem with -- I'll speak to the opposite point of Collette, and I don't mean to be facetious, obviously. Extend the window long enough, it's 100 percent, and it's easy to count. One reason to make an argument for keeping it shorter, if it's the same -- as Dr. Scali said, if it's getting the same

result, in other words, the same reading of hospitals and doctors, even at a lower level of mortality. The burden of collection is lower earlier.

It's easier to get more accurate data earlier. You heard the discussion earlier about follow-up at 30 days, and that starts to decline, and you can have, then, strokes or neurologic events for other reasons. You can be hit by a car and be dead for a different reason. You get greater specificity in the lower data collection burden earlier. That's one reason why in-hospital measurement might be appropriate in some cases.

MEMBER SCALI: I think the issue of the stroke adjudication outside the hospital, if those events were to occur, are very difficult for providers within that 30-day window. If you say that you did your carotid endarterectomy, they left and pos-operative Day 1, two weeks later they had some bizarre event at Hospital X that's not a VQI hospital, and then they see you

at 30 days and they say, I had a stroke, but they look okay to you, then how do you code that? That was, I think, probably one of the difficulties, I would imagine, perhaps the developers could probably speak to those challenges. Because mortality, yes, but I think the stroke adjudication out of the hospital makes this very challenging.

DR. DUWAYRI: There is no question that 30 days is better than the one day or two days after a surgery, but it is -- I think it will be inaccurate data. It will not be uniform across centers. Some centers will have the opportunity to follow their own patients, and other centers, due to their locations, will have to rely on follow-up outside.

It will increase the cost of participation in such registries and, therefore, will decrease our access to such data. So yes, it's much better to have 30 days, but I think it is -- it is not as practical as to have the short-term period.

CO-CHAIR GUNNAR: Larissa.

MEMBER TEMPLE: I completely appreciate the concerns with the accuracy and the costs of the data, the registry, but I think that in time, the patients are going to demand it. I think that when they see that all of the other outcome measures are 30-day, and you're presenting at discharge -- carotid endarterectomies go home very soon.

As a patient, you may wonder what are you hiding, and why aren't you giving 30-day? Clearly, you're not ready to capture 30-day in an objective way for all patients, but I do think that probably the next time this measure comes up, this will be a longer discussion item. I'm curious; we had a patient advocate on this committee in the past. Are we going to continue to do that? It would be interesting to hear that perspective, as well.

MS. MURPHY: The answer to the advocate is that, yes, we still have someone who, as it turned out, was an excellent

representative. She had a personal circumstance that was quite serious that prevented her from being able to be here.

CO-CHAIR GUNNAR: Focusing on validity. Fred, sorry.

MEMBER GROVER: It seems to me like if you're doing procedures or surgeries or whatever, we darn well ought to know how our patients are doing in 30 days. I mean, give me a break, but I think there's a certain ethical obligation. If I were the public, I'd probably demand it.

CO-CHAIR FLEISHER: One of my questions to the committee is at CSAC, there is patient representatives in a more diverse group. Do we want to put into our report that there's some -- that this is one of the concerns, that it's clear to the CSAC that this needs to be taken into consideration when they look across measures? Does that make sense as a solution? Am I overstepping?

MS. MUNTHALI: No, you're not overstepping, but I would like to have

clarification from Melinda. Will that patient advocate be able to weigh in, perhaps, before the CSAC?

MS. MURPHY: I'm not sure.

MS. MUNTHALI: That might be an option, as well, in addition to --

(Simultaneous speaking.).

CO-CHAIR FLEISHER: That would be -- okay, in the report, what I hear from many members of the Committee is, should this pass, we still want to reflect concerns of the Committee, so when we go out to voting, our membership groups, particularly trying to get the patient groups to comment, that that might be put in place.

DR. DUWAYRI: If I may add, again, this is a registry measure, and participation in the registry requires that you have -- although we are not including this in the measure, itself, but participation in the registry requires that you have long-term follow-up data, again, exceeding nine months, not only 30 days.

If you do not have that follow-up, more than 80 percent of the time, that is more than 80 percent of the patients need to be seen or documented, whether they're dead or alive, at more than nine months. If you do not follow that, you will be put on probation in the registry itself.

The registry takes care of this concern. I know the measure does not, really, because to keep it as accurate as possible, to keep it in this short period of time, but to address the concern of patient safety, the registry requires that we have long-term follow up.

CO-CHAIR FLEISHER: Should that be in some part of the specifications of the ability to be part of the registry? Is that in the document?

DR. DUWAYRI: I don't think we included it, but we can include this.

(Simultaneous speaking.)

DR. JOHNSON: This document here?

DR. DUWAYRI: This document, yes.

DR. JOHNSON: We could include it.

(Simultaneous speaking.)

MS. MUNTHALI: We will open up the measure submission form. So by the post-comment call, which I think is -- I don't know the date -- November 3rd, if you can update the submission form by then.

CO-CHAIR GUNNAR: Sorry, Karl?

MEMBER BILIMORIA: I just wanted to echo that a little further. We've done a few things. We've looked at what proportion of cases have a post-discharge event within the first 30 days, and it's a huge proportion. You know, it's something like complex surgeries, 40 percent of people, their only complication happens in the outpatient setting. A quarter of deaths happen in that setting. Not only that, it does actually change hospital rankings on quality. I think it takes away a lot of the noise about whether you discharge people early or late or whatnot. So you have the data, it sounds like. I think there

are ways to standardize at 30 days, and you will just improve the measure.

CO-CHAIR GUNNAR: Any other comments? I think we're ready to vote.

MS. SKIPPER: We're now voting on validity for Measure 1540, 1 moderate, 2 low, 3 insufficient.

(Voting.)

CO-CHAIR GUNNAR: Are we re-voting? Oh, there we go. That was reliability.

CO-CHAIR FLEISHER: I just asked the developer if they would add -- if they have beyond-three-month data, to add it to the measure submission form, in addition to the fact of you have to have 80 percent compliance. We will see that on our post-call.

MS. SKIPPER: I need everyone to please re-vote on the validity for Measure 1540, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Measure 1540 passes on validity, 9 percent high, 57 percent moderate, 26

percent low, 9 percent insufficient.

CO-CHAIR GUNNAR: So up to use/usability -- or feasibility, yes, feasibility. Any comments before we vote? All right, carry on.

MS. SKIPPER: Voting is open for feasibility, Measure 1540, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: With 22 votes, this measure passes on feasibility, 27 percent high, 68 percent moderate, 5 percent low.

CO-CHAIR GUNNAR: Now, use and usability. Any comments?

MEMBER SCALI: I think the use and usability, there's some questions for the Committee that were posed, specifically about the unintended consequences. I think it surrounds similar arguments about the moderate-sized asymptomatic AAA. Again, we're back to the same sort of construct, which is asymptomatic carotid patients have an annualized stroke risk of less

than 2 percent, so the role of operating on patients who would be deemed unfit for open surgery versus high risk for subsequent surgical complications, et cetera. Again, I think the spirit of this was -- and the developers can tell me if I'm wrong. I think the spirit was to look again at patient selection because of the same type of construct that you guys designed for your AAA indicators.

I think the unintended consequence, in terms from a patient-centric view, could be that you would be denying more patients surgery, but that's actually what's occurring in the United States currently. We're actually seeing that, and that's been occurring over the last five or ten years. If you look at claims data, there's sort of been plateaus, in terms of carotid revascularization -- open, that is.

Granted, the carotid stenting has muddied the waters. CMS has put a governor on how often are those procedures are being done because of the requirements for enrollment in

registries to get reimbursed. We have seen plateaus in the rates of carotid revascularization in the field that the risk profiles of patients, at least in claims data and when they do walk-overs between VQI and Medicare -- because we have a current initiative to look at matching between the VQI patients and the Medicare identifier for those patients. So you can get to re-admission and long-term outcomes because a lot of patients don't come back to their VQI hospital. We are looking to do that.

It's not an easy thing to do for a patient safety organization with de-identified data, in terms of how you do those coding algorithms. But we have figured out that, at least if you look at the risk profiles over the last five years for carotid revascularization, by and large, they tend to be younger patients. They tend to have less of the CHF, a lot less renal insufficiency, et cetera.

It all comes from societal guidelines in the literature that came out after ACAS

showing that patients with values greater than two, patients greater than age 75, female gender, et cetera, all things that were in the trial that were either a) under-represented, or 2) were excluded from the original ACAS data, didn't derive any benefit from prophylactic endarterectomy. While there is a potential for the unintended consequence of so-called denying patient surgery, I think I would put it in Barbara's construct. It's about telling the patients that the cure is potentially worse than the disease by doing nothing. I would just only raise those points about the usability.

CO-CHAIR GUNNAR: Any other comments about usability and use? Amy?

MEMBER MOYER: I apologize because this isn't specific to this. I'm looking around on your website. Do you publicly report this anywhere like the STS does, or do you have plans to do that?

DR. DUWAYRI: PQRS, that's it. Let me see how we submitted this, but the registry

provides us the ability to do both, but I think this is both here.

DR. BURSTIN: Will it go on Physician Compare then?

DR. DUWAYRI: Yes.

MEMBER SCALI: And Hospital, is that what he said? Is that what you asked?

MEMBER SCALI: Are the benchmarked facility data in public?

DR. DUWAYRI: Currently, other than PQRS, no.

MEMBER SCALI: Okay, so just in that.

CO-CHAIR GUNNAR: Karl, do you have something else? Fred, anything else? All right, ready to vote on use and usability? Christy?

MS. SKIPPER: Voting on usability and use for Measure 1540, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. SKIPPER: Measure 1540 passes on usability and use, 13 percent high, 65 percent moderate, 22 percent low. We'll now be voting on

overall recommendation for endorsement, 1 yes, 2 no.

(Voting.)

MS. SKIPPER: This measure is recommended for endorsement, 83 percent yes, 17 percent no.

CO-CHAIR GUNNAR: Moving on to the last of the series from SVS is 1543, post-operative stroke or death in asymptomatic patients undergoing carotid artery stenting. Developers want to make any comments?

DR. DUWAYRI: Very similar to the carotid endarterectomy measure, looks at perioperative stroke and death. Carotid stenting is an alternative to carotid endarterectomy as a method of carotid revascularization for stroke risk for stroke prevention. Again, this measure looks at patients who are asymptomatic and, therefore, this is a prophylactic intervention.

The literature suggests that the stroke rate for carotid stenting is slightly higher than carotid endarterectomy, usually at

around 3 percent. Our data analysis from the VQI registry, from 3,342 carotid stent procedures, reveal that the perioperative stroke and mortality rate ranges between 0 to 1.7 percent.

The risk of these events is higher in patients older than 70 and in females. Again, this is a measure looking at the prophylactic intervention and, therefore, the key concept is to offer this procedure only in patients who will have a relatively low risk of events around the time of intervention. Carotid stenting is currently only reimbursed by Medicare in symptomatic patients. However, asymptomatic carotid stenting is being performed. Usually, the indications are for anatomic indications, either for high cervical lesion with previous operation or previous radiation.

CO-CHAIR GUNNAR: Discussants are Drs. Scali and Dutton.

MEMBER SCALI: The developers present, in terms of summary of the evidence, there's actually been a tremendous amount of activity in

this area looking at the role of carotid stents, specifically, in the asymptomatic patient population. On the heels of CREST-1, there were a series of additional trials and large registry analyses that were done. There was a lot of inter-group analyses, and the developers have offered more updated literature to support that this is something that needs to be measured.

One question I had with the developers that I always sort of have difficulty with, and it's in the spirit of measuring this, as you guys alluded to, it says that there's no published guidelines for carotid artery stenting for asymptomatic patients in three out of the four societies, with the exception of one society: the American College of Cardiology. So how do we reconcile that if we're going to make a quality measure about measuring an outcome for a procedure that's not even recommended by three out of four of the major societies that describe care and outline guidelines for asymptomatic carotid disease?

Because what has been shown, unequivocally, is that if you remove the MI composite endpoint in the literature that's there, if you just look at stroke and death, stenting has a two-fold higher risk of stroke and death, compared to surgery. That is Level 1 evidence.

Then you sort of have to reconcile to make a guideline or a comment about a procedure that's been shown to have a higher risk than the operation, which we just vetted for -- said it had a very narrow therapeutic window to begin with. If you can add some comment about that because the literature and the evolution of this literature has been very brisk in the last few years, and it's really muddied the waters about what we do with these interventions.

DR. DUWAYRI: Yes, I can't agree any more. I think it's another reason we need to keep this track. We don't know the value of carotid stenting in asymptomatic patients over -- again, comparing it to not only carotid

endarterectomy, but also to maximal medical therapy. This is, again, a procedure that carries a higher risk of perioperative events in asymptomatic patients than endarterectomy. So again, the answer to this is yes, that's why we need to continue to measure it.

MEMBER DUTTON: Let me ask a slightly different question. Unlike carotid endarterectomy, which we can imagine are being done by a fairly homogenous group of vascular surgeons, are there lots of specialties in the waters here, and how do you intend to get data across the landscape?

DR. DUWAYRI: I don't have the numbers now, but VQI is open for participation, and we have participants who are neuro interventional radiologists and interventional cardiologists. I do not know the exact number of how many of these are participating in this carotid stenting module and in other endovascular intervention modules.

DR. JOHNSON: Yet, looking into the future, our quality committee is already

interacting with interventional radiology, and hopefully in cardiology, in combining our -- putting everybody into one database. The other thing we've discussed with interventional radiology is alternative payment models, is looking with them and doing one with them. You're right. Every specialty or society that participates in this intervention ought to be within one database, and that's what we hope to achieve, but getting there is another point.

CO-CHAIR GUNNAR: Dr. Ko?

MEMBER KO: I just want to pick up on Sal's comment. Am I understanding you correctly that we shouldn't be doing these? If that's the case, the performance measure should be an appropriateness measure that we shouldn't be doing these? Because the reason we have this measure should not be to collect data. It's not participation in a registry to collect data so that we have more knowledge. It should be a performance measure, and so --

(Simultaneous speaking.)

DR. DUWAYRI: There is literature to suggest that the surgeon -- that interventionalist experience and volume significantly influences the perioperative outcomes in these procedures. I think there is value in treatment of asymptomatic patients, and there are anatomic indications for carotid stenting in patients who are not candidates otherwise for carotid endarterectomy. The indication for carotid stenting is not usually the same indication as carotid endarterectomy, and it carries a higher perioperative risk, but if you can perform this procedure with a lower risk than what the natural history of the patient's carotid stenosis will carry, then I think it is okay to proceed with it.

MEMBER SCALI: I think, to Dr. Ko's point, that's what I was struggling with philosophically. I think you have to sort of evaluate it on each of the merits to make sure the NQF criteria for how the measure's been developed is appropriate.

But I think stepping back, if you endorse a quality measure for something that three out of four societies say you shouldn't really be doing to begin with, what is that saying to providers who want to do this or don't want to do this? What's the patient -- the guy, gee, three out of four societies tell us we shouldn't be getting this for this indication, yet NQF says that we should measure it, meaning you should do the procedure, I guess. That's where I struggle a bit philosophically, but at least on point to talking about the data, I think the developers provided enough data to meet the evidence-based algorithm that's needed for vetting this as a measure. But I think the bigger question, philosophically, is one that I struggle with.

CO-CHAIR GUNNAR: To get back to our charge here, what you've stated and claimed is that the evidence supporting this measure has changed since this was vetted the initial time? So we actually, in this case, should re-vote on

the evidence.

If it is insufficient or low to support this measure, then it will no longer carry, unless we then decide to make an exception.

DR. DUWAYRI: I will point out that this procedure --

CO-CHAIR GUNNAR: Just one second. Am I correct?

MS. MURPHY: Yes, except there wouldn't be an exception to that because it's not process.

CO-CHAIR GUNNAR: There we go. We've now gotten down the algorithm. It takes a village. We learn together.

MEMBER SCALI: I would say that for the evidence to support it, with the caveats mentioned with the newer evidence that is given for this, it's not the Level 1A evidence that, historically, things have been based on. The CREST trial looks at -- the primary end point was a composite end point when it compared carotid

endarterectomy and carotid stent.

It was stroke, death, MI, which unfairly -- or fairly, depending on how you look at it -- dings the surgical arm. That being said, when you remove the MI component and you look strictly at stroke/death, based on the only real Level 1A evidence that we have currently, it would tell you that there's a two-fold higher risk for doing prophylactic stenting versus surgical therapy for asymptomatic carotid disease. Now, subsequent sub-analyses in other registry data have clearly shown, as Yazan has mentioned, that in really experienced hands -- like you've done 25, up to 50 carotid stents, in so-called high-volume centers, high-volume providers, looking at plaque morphology, calcium, arch, there's so many anatomic variants that also impact the outcome of this procedure, which have been published and are here represented in the literature, that would say that yes, in those specific providers' hands, you can get an equivalent, or even equivalent stroke/risk

outcome for the procedure compared to endarterectomy.

So the literature that's there does sort of represent the evolution of the data as we understand it, but there's still a -- I guess we'll get to the gaps and things like that. There's still some real issues about which patients would be eligible, which providers should be included and so forth.

The only Level 1A evidence that we have -- currently, CREST-2 is underway. CREST-2 will look at medical, surgical and stenting, but it also has mixed groups. On CREST-1, only about -- it was half or just over -- there's less than half were actually asymptomatic disease.

So the natural history -- yes, so that's the other thing is again, we're making all these assumptions about stenting, but when -- everybody came out after CREST-1, and if you were pro stent, you said, look, stents are as good as surgery. As long as you do a lot of stents, you can get as good outcomes as surgeons.

The devil was in the details to do the sub-analyses. Wait a minute, we pull out the MI from the composite end point and you look at stroke death specifically, the stroke end point, two-fold higher risk across the board, no matter how you looked at it.

This has been looked at in subsequent registries, when you look at stenting for prophylactic endarterectomy. That being said, there's been subsequent analyses and subsequent large registries that have also been produced in the last several years, and the authors present that data, again, showing that you can have equivalent outcomes for the measure with all of those caveats, being you have to have supremely good skillset, meaning you've done 25 to 50 of these things, and you have to be supremely good about patient selection, not just comorbidities, but anatomic specifications of the lesion.

Lesion morphology has a lot to do with outcome, as much as age, gender, et cetera. So therein lies the conundrum about the literature

and where the role of current carotid stenting is for asymptomatic -- that's why CMS doesn't pay for it. You can't do it right now, unless you are in a registry. You don't get reimbursed for this.

CO-CHAIR GUNNAR: Any comments from the developers?

DR. DUWAYRI: The procedure, whether it's endorsed by the Society for Vascular Surgery or not, is still being performed out there. It's endorsed by American College of Cardiology, which is a large society. I think for these reasons, we need to continue to monitor them. I don't think we are approving this procedure or rejecting it. We are just suggesting that we need to monitor its outcomes.

MEMBER YATES: For the sake of clarity, since you have your registry, has the incidence of its utilization gone down over the last year or two?

DR. DUWAYRI: I actually can find that.

MEMBER YATES: While he's looking that up, Dr. Moss?

DR. DUWAYRI: The answer is no. In 2010, there were 100 cases in the -- no, I'm sorry. There were 198 cases in the registry, of carotid stenting, and it has gone down -- gone up every year.

In 2015 -- yes, but this is by adding centers. In 2015, there are close to 2,000 carotid stents entered. Accruing more, but that's one possibility, again.

MEMBER YATES: Your rate per surgeon or rate per center hasn't gone up, or would you say -- you may not have that off the tip of your tongue, but it sounds like it's equivocal right now. Whether the literature says to do it or not, it's happening.

DR. DUWAYRI: Yes, but we did not grow by ten times. The rate of growth on the procedure volume is ten times.

MEMBER DUTTON: You could argue that if the procedure is -- if the indication for the

procedure is controversial, then measuring the outcome is even more important.

CO-CHAIR GUNNAR: Yes. Larry?

MEMBER MOSS: Sal, I just want to ask your opinion. This is not a scientific question, but you articulate your conflict well. In your opinion, are patients with carotid disease better off with this measure in existence, endorsed, or not?

MEMBER SCALI: In my humble opinion, I think the answer is yes because the activity is going to happen. And hopefully we'll have the answer soon enough with Level 1(a) evidence from CREST 2, but I think the answer is yes because we do want to know what those rates are nonetheless, despite all of the misgivings in the literature.

MS. MURPHY: So may I ask a question of Dr. Scali? Are you saying the measure should be endorsed by NQF versus information collected in the registry for analysis?

MEMBER SCALI: I thought the question was more do we think that there should be a

measure of stroke and death after carotid stenting for asymptomatic disease, and I think despite all of the concerns about the literature and where that really sits in the current treatment armamentarium for carotid revasc, I think the answer is yes, we -- the activity is occurring, there are people who do participate in registries and do do a modest amount of volume, and so you do want to sort of have an understanding of what their outcomes are and make sure the people are held accountable, particularly when the risk/benefit window is so narrow.

So, I'm not going to make the statement that NQF should endorse this without going through the formal process, but I would say that there should be a stroke/death endpoint that we're following for carotid stenting. Whether or not we do carotid stenting for the indication is a different issue. I think that is what I was struggling at, that Dr. Ko was mentioning, is do we endorse something that several societies say

that we probably should not do?

CO-CHAIR GUNNAR: And CMS does not pay for?

MEMBER SCALI: Unless you're in a registry or a trial, yes.

DR. DUWAYRI: Not this registry, but --

MEMBER SCALI: Not this registry, it's a carotid-specific stenting registry.

CO-CHAIR FLEISHER: We're getting back to evidence from an NQF endorsement perspective, separate than should this be in a registry, and that --

DR. BURSTIN: Evidence for the measure focus, there is a rationale that, again, this is an outcome. Is there a rationale that there are logically processes et cetera that would influence the outcome? Sounds like there's probably lots you can do to prevent stroke and death among patients getting an endarterectomy. I think the bigger question is, and I think this is what Melinda was trying to get at, does it

send a strange message to have an endorsed measure for a procedure for which the indications are in play, just kind of play it like that?

But the flip side of that is, if in fact it is still being done, is it important to measure the outcomes, even if perhaps it is at times being done inappropriately? And it probably sounds like the better measure we'd love to see you have come back with is a measure of appropriateness as this settles out, because certainly it sounds like we're seeing a lot of it being done based on your database, and it's not clear what guidance they're using to make that selection.

CO-CHAIR GUNNAR: So again, to get back to the evidence question, which I think needs to be re-voted on, is "Is there sufficient evidence, albeit that it has changed, to support a measure of this outcome in relationship to quality and quality improvement?" Dr. Yates?

MEMBER YATES: I would argue absolutely. The new evidence argues all the more

for making sure the measure goes on, especially since there is a whole different spectrum of specialties that seem to feel that the procedure is the correct one to do. And the evidence argues more strongly for continuing.

CO-CHAIR GUNNAR: Collette, and then Fred.

MEMBER PITZEN: I am just struggling a little bit, not being a vascular surgeon, but when I try to take it in the different thing, the criteria for evidence for an outcome measure says that there has to be at least one process or processes in place that support and lead you to getting to that outcome. The denominator of this measure is defining something that is controversial, so the first step in that process, or second step, patients evaluate a decision to proceed towards stent. So I am having a hard time wrapping my head around that.

Let me take another example: with cholesterol management for a vascular measure in our diabetes measure, the guidelines changes

drastically, and even though some wanted to stay with an LDL target, the evidence no longer supported treating to an LDL target of less than 100, so we had to change it. It wasn't supported.

So again, I don't know where things are at in terms of controversy and studies yet to come, but I just have concerns.

CO-CHAIR GUNNAR: So based on that, I think we should vote on evidence because I believe that there's -- there's a -- a change of opinion amongst many of you regarding -- from the moment we started this discussion. Dr. Grover?

MEMBER GROVER: Well, I think we're interested in quality and patient safety, and you have raised some really serious red flags here in this discussion, and I think that to me makes it even more imperative you continue collecting this data just to see what it shows. And perhaps -- I mean, here you've got a database, and you've got a chance to measure that, and you document higher -- worse outcomes with that measure, that needs

to weigh into the guidelines in the future.

CO-CHAIR GUNNAR: So we're going to proceed with the vote on evidence. Christy?

MS. SKIPPER: We're now voting on evidence for Measure 1543. 1 yes, 2 no, that the evidence supports the relationship of the health outcome in at least one action.

(Pause.)

MS. SKIPPER: Measure 1543, we've landed in a gray zone on evidence: 55 percent yes, 45 percent no. So our options here are to -- well, we can move forward to vote on the remaining criteria, but we will not take an overall vote on suitability for endorsement.

CO-CHAIR GUNNAR: We have touched all the different variations today, I think.

MS. MUNTHALI: So you will continue voting, and we hope that the public and member comments that come in will help to inform this criterion where you were not able to reach consensus.

CO-CHAIR GUNNAR: Okay. So we --

we've moved to gap.

DR. DUWAYRI: I had a comment on evidence, just to --

CO-CHAIR GUNNAR: Yes.

DR. DUWAYRI: -- clarify.

CO-CHAIR GUNNAR: Go ahead.

DR. DUWAYRI: So again, there is a recommendation from the American College of -- American Heart Association to revascularize carotid if your perioperative risk is less than 3 percent. There are randomized -- there is one major randomized trial that compared carotid endarterectomy to carotid stenting that has been interpreted in two different ways.

One is that yes, they have equal outcomes and another way is that carotid endarterectomy is better. Now, depending on how you interpret this data, you will recommend carotid stenting or carotid endarterectomy. So I think the data in the literature now does not say that -- I mean, in my opinion, it says that carotid endarterectomy is better than carotid

stenting, but there is a large medical society in this country that recommends carotid stenting as a major revascularization for carotid disease, so I don't think there is any condemnation currently in the literature for carotid stenting that is agreed upon by the majority that we shouldn't do carotid stenting. I just wanted to clarify this point.

MEMBER HANDY: Do we have a procedural way forward, because the spirit of monitoring it, because we don't know if the outcome is pertinent. So how do we get beyond this lack of evidence to say that we do want to follow it, we just don't have the evidence to say we recommend stenting?

MS. MURPHY: What we're going to do is continue the process through the remaining criteria. We will include in the draft report the information about the occurrence. We will hope, expect, that we will get public comment that can then help inform what should be done going forward.

DR. BURSTIN: Clarifying that, I'm really glad you raised that, and I don't want that to be misinterpreted as part of the report that goes out. I would also make it very clear, this vote on evidence has nothing to do with whether or not stenting is appropriate. It is the evidence focus for this particular outcome measure, should you look at these two outcomes for patients who are stented.

But I think it's an important distinction, because my fear is, and I think a lot of the fear in the room, is that people will look upon this measure as it goes forward as NQF coming forward and saying we endorse that this is an appropriate procedure, and that's not what the measure is. The measure is "Is it appropriate to continue to look at the outcomes of this procedure, regardless of where the literature is sorting out?"

MEMBER HANDY: So it's really an opportunity for the pertinent societies to weigh in in a very informed fashion to sway the

discussion.

CO-CHAIR GUNNAR: Barry?

MEMBER MARKMAN: Yes, I mean, I have been here for three years. I have never encountered this situation where we're trying to measure outcomes in something that is sometimes detrimental. I mean, have we had this -- Okay. Okay.

(Laughter.)

CO-CHAIR GUNNAR: I think Collette provided a good example, quite frankly. The evidence changes. Practice changes based on evidence, right?

All right. So gap.

MEMBER DUTTON: Well, there is clearly a gap in understanding about whether this procedure is indicated or not, so yes, it is controversial. That is the very reason we should measure it.

Otherwise, I think a discussion would have the same points as the last measure: one-day stay versus 30-day mortality, that question will

come up again. And again, in the measure data presented, variability is very low, but again, this is a significant outcome measure that the public will want to know the answer to.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: Yes, I mean, do the patients even stay in the hospital for this? Is this an outpatient procedure?

DR. DUWAYRI: Most of the time, no, the patients stay as long as they do for carotid endarterectomy, which is one day. And largely, it is for perioperative blood pressure management, so yes, usually, it is an inpatient procedure. I don't know if CMS approves it as an outpatient procedure.

CO-CHAIR GUNNAR: Any other comments? Dr. Grover?

MEMBER GROVER: I guess I think of things somewhat differently, and I feel like we're a little bit restricted, and maybe it is because I am undisciplined. But I think this is a controversial issue, and that demands

collecting more data than some that are not. I mean, this is patient safety, and it is good to have more than one group collecting that data, I think.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: I just have a technical point, being the one that I was talking about, the 30-day window the entire time. All the data that I can see in the application is actually within 30 days of the carotid stenting procedure, so I am not exactly seeing data of at discharge, like how the measure is. So just technically, they should align.

CO-CHAIR GUNNAR: Barry, any other comments?

(No audible response.)

CO-CHAIR GUNNAR: You have none? Okay. Want to vote on gap, Christy?

MS. SKIPPER: We're now voting on gap for Measure 1543. 1 high, 2 moderate, 3 low, 4 insufficient.

(Pause.)

MS. SKIPPER: Measure 1543 passes on performance gap: 32 percent high, 55 percent moderate, 14 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: Move on to reliability?

MEMBER SCALI: With regard to the reliability testing, I think the original development was done with a trial review analysis through the VSGNE patients, so that is the New England cohort of the VQI, and they did chart review analysis and showed excellent reliability testing, as indicated by their kappa statistic, which was a strong agreement for identifying the correct procedure as well as in-hospital mortality and stroke. That was seemingly done sort of above board.

MEMBER DUTTON: I'd like the developers to answer that question. I put in the discussion here I was confused in reading the specifications. Patients are excluded from the measure if they have criteria symptoms within the one-year proceeding stenting or symptomatic

carotid stenosis less than 120 days prior to procedure or other carotid stenosis 120 days or greater. I couldn't figure out what that meant. Are we asymptomatic for a year, or asymptomatic for 120 days, or what?

DR. DUWAYRI: For some reason, I don't really know the answer for sure for that, but they separated the one year into two exclusion criteria, and one less than 120 days. But the -- this means that any symptoms within one year is an exclusion. I don't know why they put it in two exclusion points. I will have to check back with that.

August 22, 2016I think the code is different for 120 days. For less than 120 days, the code is 9006F, and for more than 120 days, the code is 9007F, so I think this is the reason why we separated it, but I am not sure about the technicality of this.

MEMBER SCALI: So are you considered by the measure, are you asymptomatic if you've been without stroke-like symptoms for 120 days,

or it has to be a full one year? Because that is different than all of the trial data, right?.

CO-CHAIR GUNNAR: Any other discussion? Fred, you've got your -- you've got your card up.

(No audible response.)

CO-CHAIR GUNNAR: Okay. All right. I think we're ready to vote.

MS. SKIPPER: We're now voting on reliability for Measure 1543. 1 moderate, 2 low, 3 insufficient.

(Pause.)

MS. SKIPPER: Measure 1543 passes on reliability, 64 percent moderate, 36 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: Validity?

(Pause.)

MEMBER SCALI: So again, I think this is one of the things in terms of the risk adjustment question now, for carotid endarterectomy and small AAA, we don't tend to risk adjust, whereas for, you know, the patient

selection piece, again, it is critical here for providers offering CAS, so you know, the literature clearly shows that if you risk adjust for provider, hospital, anatomical variance, and patient covariance, then you can get those outcomes. But if the spirit of the measure is specifically to measure the appropriate patient selection of who you offer, then I do agree that probably no risk adjustment would be indicated. So I just wanted to bring up that point because of all of the other concerns already raised about the evidence evolution.

CO-CHAIR GUNNAR: Rick, any comments?

MEMBER DUTTON: No, I agree with that. I think this is all about patient selection, so no -- no risk adjustment is appropriate.

CO-CHAIR GUNNAR: Any other comments?

(No audible response.)

CO-CHAIR GUNNAR: Ready to vote.

MS. SKIPPER: We're now voting on validity for Measure 1543. 1 moderate, 2 low, 3 insufficient.

(Pause.)

MS. SKIPPER: Measure 1543 passes on validity -- or, I am sorry, it does not. It looks like we have landed in a gray zone for the validity of Measure 1543, and again, we will continue discussion of the remaining criteria.

CO-CHAIR GUNNAR: So next is feasibility, feasibility next?

MS. SKIPPER: Yes.

CO-CHAIR GUNNAR: Any other discussion?

MEMBER DUTTON: I think we have shown that it is feasible to collect in a registry, but this is not going to be easily transported to an eMeasure or to abstraction from claims data because the diagnosis of stroke is so specific and requires sort of hands-on testing, not abstract.

CO-CHAIR GUNNAR: Any other comments?

(No audible response.)

CO-CHAIR GUNNAR: Okay. Vote?

MS. SKIPPER: Voting on feasibility

for Measure 1543. 1 high, 2 low, 3 moderate, 4 insufficient.

(Pause.)

MS. SKIPPER: Measure 1543 passes on feasibility, 5 percent high, 71 percent moderate, 24 percent low, 0 percent insufficient.

CO-CHAIR GUNNAR: And now for the interesting one, use and usability, or usability and use.

MEMBER MCCARTY: This is more of a general comment for NQF and CSAC, but I want to go back again to the conversation earlier about quality versus accountability, and I wonder if in some ways we could tie into use and usability, instead of high, moderate, low, different purposes.

So tracking the outcome for self-improvement locally so you know what your outcomes are and you can improve if you need to, or potentially a level up from that is publicly reporting, but nothing is tied to it. And a level up from that is now we're publicly

reporting, and there is pay-for-performance, because in our conversations, that is becoming a theme, but I think there is some nervousness in endorsing. We only have one way to endorse a measure, it is endorsed or it is not.

And then from there, I know Helen described other ways that other groups then decide what to do with that, but I feel like I would be more comfortable if we had a way of being more explicit about these conversations and the intent of some of the reasons why or how we're saying these metrics should be used that are more explicit when the metrics are posted other than kind of buried in the meeting minutes and in the moderate rating or low rating of usability. So just a suggestion for NQF.

CO-CHAIR GUNNAR: Anyone from NQF want to comment?

MS. MUNTHALI: No, we have noted your recommendation, and what we would suggest is that the committee makes a statement, strong statement, about your preference for how these

measures should be used. We cannot dictate how these measures are picked up and used, but I think it would go a long way for the committee, the Surgery Committee, to say that in the report.

CO-CHAIR GUNNAR: Any other -- oh, Dr. Yates?

MEMBER YATES: Yes, I think we said that last time, and that went up to the central committee, and they came back and said, no, we're not going to do tiered usability, tiered validity, so we can say it again, but what I heard this morning from the people that are on the central committee was that they are not going -- that just look at the quality, whether it's a quality measure or not, but don't worry about the end use in terms of how you vote on things. So I think we did say that once, just as a point of clarity there.

MEMBER LEVY: So there is a difference between saying that as a general rule for how NQF endorses measures versus this committee making a recommendation about a particular measure so that

when the MAP looks at the measure, they will have our comments that we do not think it should be used for accountability, for example. So our comments on individual measures about our review of the literature may be useful for the MAP. That is different than saying that NQF in general will have different tiers of endorsement.

MEMBER YATES: I would argue that we're now halfway through the measures between the two days, and saying that now doesn't allow for us to -- as a committee, to say that for some of the measures we have already gone over that there was some questions, so I am not sure that those comments are in a logical enough place for the MAP to pick them out.

MEMBER LEVY: Well, we will still have an opportunity I think when we do the final call so that we could do that with the other measures.

MEMBER YATES: So just so that the staff know, make sure that's on the agenda.

(Simultaneous speaking.)

CO-CHAIR GUNNAR: Yes, I just want to

reiterate, maybe I am missing something, but there are NQF measurements measure process and outcomes that are important to the patient in informed consent. They are important to the evaluation of the physicians' ongoing professional practice. They are important to quality improvement. They are separate processes, yet the information is important to those separate evaluations. In my opinion, we can't confuse them. We can't mix them together.

MEMBER YATES: My comment would be is that when they are used by CMS, they are used with a nomenclature of NQF-endorsed, and it's in their Blue Book in terms of what performance measures to use that it has to go through the NQF process, so the two are intertwined and embedded.

And I understand that there are some measures that are too big to fail, but we can at least put some pressure on at least making them better and making them more focused for some of the more fine details that they are actually delineating us by in terms of performance by

hospital, and eventually, by physician.

CO-CHAIR GUNNAR: And I should say that the bucket -- the second, third, fourth bucket, then, is -- or third bucket, of use of NQF, is pay-for-performance, so I think that we have to figure out, are those unintended consequences of NQF, or are they natural uses of them? If they're natural anticipated uses of NQF measurements, then I think we have to be cognizant of that.

In this case, this is where it actually goes back to what Dr. Scali says, would an OPP want to know the percentage of -- you know, are you actually doing procedures that are not recommending by guidelines by three societies? Is that your management of asymptomatic patients?

MEMBER YATES: In this particular case, and CMS doesn't pay for this --

CO-CHAIR GUNNAR: And that is the --

MEMBER YATES: They're not going to get penalized for it.

(Laughter.)

CO-CHAIR GUNNAR: But I think it has been a great discussion for me. I think clarity going forward is on how, as a standing committee, we evaluate this, we have been told, go back, evaluate it as a single stand-alone measure for the use of quality improvement. We can't help collectively saying, yes, but it is used for -- . And that may not be an appropriate use.

Okay. All right. So I think -- thanks for the side rant. We have to vote on usability and use, or are we

-- yes.

MS. SKIPPER: Yes, and this is the last criteria that we have to vote on. And we will be done with the SVS set.

So we're now voting on usability and use for Measure 1543. 1 high, 2 moderate, 3 low, 4 insufficient.

(Pause.)

MS. SKIPPER: Okay. With 20 votes, 10 percent high, 45 percent moderate, 45 percent

low. We have also landed in a gray zone for usability and use. And since this measure did not reach consensus on validity, we will not take an overall vote for recommendation for endorsement.

CO-CHAIR GUNNAR: All right.

MS. SKIPPER: And -- and we've run over the agenda about half an hour, but we did have time for -- if we can ask the operator to open the line for member and public comment, and if there is anyone in the room who would like to make a comment on any of the measures discussed to this point.

THE OPERATOR: If you would like to make a public comment, please press star 1.

(Pause.)

THE OPERATOR: There are no comments at this time.

MS. SKIPPER: Are there any comments in the room?

(No audible response.)

MS. SKIPPER: Okay. Hearing none, I

just want to note that we did run over a few minutes. We have the STS measures that will be discussed in the morning. We will be beginning at 8 o'clock, so about a half hour earlier, to finish the review of the remaining measures.

And I also want to note that there is a reservation for dinner at P.J. Clarke's at 6:30, and it is located at 1600 K Street. You can come see me if you need walking directions, but there is a reservation if you all would like to have dinner at P.J. Clarke's. And is there anything else I missed?

(No audible response.)

MS. SKIPPER: Yes, you can leave your remotes and your placards and get ready for the morning, 8 o'clock. And thank you all for a good first day.

(Whereupon, the above-entitled matter went off the record at 5:31 p.m.)

H

N

O

W

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Surgery Phase 3
Standing Committee

Before: NQF

Date: 08-16-16

Place: Washington, DC

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

Neal R. Gross

Court Reporter