NATIONAL QUALITY FORUM

+ + + + +

SURGERY STEERING COMMITTEE

+ + + + +

THURSDAY, MAY 29, 2014

The Steering Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:00 a.m., Lee Fleisher and William Gunnar, Co-Chairs, presiding.

PRESENT:

LEE FLEISHER, MD, University of Pennsylvania, American Society of Anesthesiologists, Co-Chair (in person and via telephone)
WILLIAM GUNNAR, MD, JD, National Surgery Program Office, Veterans Health Administration, Co-Chair
ANTHONY ASHER, MD, FAANS, FACS, Carolina Neurosurgery & Spine Associates
ROBERT CIMA, MD, MA, Mayo Clinic
RICHARD DUTTON, MD, MBA, Anesthesia Quality Institute
ELISABETH EREKSON, MD, MPH, Dartmouth Hitchcock Medical Center
FREDERICK GROVER, MD, University of Colorado School of Medicine
JOHN HANDY, MD, American College of Chest Physicians
MARK JARRETT, MD, MBA, North Shore-LIJ Health System*
CLIFFORD KO, MD, MS, MSHS, FACS, American College of Surgeons, UCLA School of Medicine, American College of Surgeons
BARBARA LEVY, MD, FACOG, FACS, American College of Obstetricians and

Gynecologists
BARRY MARKMAN, Aetna
KELSEY McCARTY, MS, MBA, Massachusetts General Hospital
LAWRENCE MOSS, MD, Nationwide Children=s Hospital
AMY MOYER, The Alliance
KEITH OLSEN, PharmD, FCCP, FCCM, University of Nebraska Medical Center, American Society of Health-System Pharmacists
COLLETTE PITZEN, RN, BSN, CPHQ, MN Community Measurement
LYNN REEDE, DNP, MBA, CRNA, American Association of Nurse Anesthetists
GARY ROTH, DO, FACOS, FCCM, FACS, MHA Keystone Center
CHRISTOPHER SAIGAL, MD, MPH, UCLA
ROBERT SARWIN, MD, MS, Seattle Children=s Hospital, Organization of Children=s Hospital Surgeons-in-Chief
ALLAN SIPERSTEIN, MD, Cleveland Clinic
LARISSA TEMPLE, MD, Memorial Sloan-Kettering Cancer Center
A.J. YATES, MD, University of Pittsburgh Medical Center

NQF STAFF:

HELEN BURSTIN
ANN HAMMERSMITH
VY LUONG
ANDREW LYZENGA
AMARU SANCHEZ
REVA WINKLER

ALSO PRESENT:

MATT BRENGMAN
KEZIAH COOK *
DEBORAH DEITZ
ELIZABETH DRYE
CAROLINE GALLAHER *
LEIN HAN
JEFF JACOBS
JOHN MORTON *
SEAN O=BRIEN
DONNA SLOSBURG *
LISA SUTER
KIM WOOD *

*present via telephone

A-G-E-N-D-A

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

8:00 a.m.

CO-CHAIR FLEISHER: I want to thank everybody for a great first day of the meetings. I think we=re starting to get some clear signals of how this committee is thinking and I think there=s a lot of thought processes that I certainly and I know the NQF staff will take up to CSAC.

What=s important is the thought processes that you have over things like the temperature measures and others of how you think about reserve is actually probably the critical thing. Because people are still wrestling with these issues and recognize again from yesterday that we are the approvers of measures. We do worry about how people use them but it=s actually out of our hands in how people use them.

So, just a recap of what we recommended and what we did not. Importantly we did not

recommend urinary catheter removal, the vaginal suspension at the time of hysterectomy. The perioperative temperature management was withdrawn. The antibiotic timings were put on reserve for three of them. The SCIP measures were all put on reserve.

The clinician antibiotic selection, there was no consensus. The antibiotic timing readministering clinician was deferred, and the discontinuation was put on reserve.

We have a lot of things to get through today. The goal for me is to be done at 3:30. The latest is 3:45 per the agenda so we will get started.

MS. WINKLER: We have a couple of measures that are carrying over from yesterday, but we=re going to do them after we do the first two measures on mortality for CABG. So the first measure we=re going to do is measure 0119, Risk-adjusted Operative Mortality for CABG. That=s the STS measure. You guys here?

Okay, we=re going to flip the order of these measures and do 2558, the Hospital 30-day All-cause Risk-standardized Mortality Rate Following CABG. This is a new measure submitted by CMS.

CO-CHAIR FLEISHER: And I acknowledge that Dr. Grover has chosen to recuse himself given that this is a competing measure to the STS measure.

Who is actually the primary reviewer? Not here yet? Okay, who=s the secondary? There was no secondary. Dr. Jacobs is here? Would you mind starting and hopefully the reviewer for the CMS measure -- if you don=t mind?

So we are going to start with 0119. I am actually the primary and Cliff is the secondary.

DR. JACOBS: Good morning, everybody. And the first measure we=re doing this morning is risk-adjusted operative mortality for

coronary artery bypass grafting. And this is a measure that reports the percentage of patients aged 18 or older undergoing isolated coronary artery bypass grafting who die, including both all deaths occurring during the hospitalization in which the coronary artery bypass grafting was performed even if after 30 days, and second, those deaths occurring after discharge from the hospital but within 30 days of the procedure.

So that=s the standard definition of operative mortality. And this basically reports risk-adjusted operative mortality after isolated coronary artery bypass grafting.

Advantages of this measure include the fact that it=s derived from a clinical data set. So the denominator of patients undergoing isolated CABG is a relatively pure denominator because of the clinical nomenclature utilized to isolate the patients undergoing isolated

CABG. And also the clinical nomenclature allows for a fairly sophisticated risk adjustment.

And I guess with that introduction I=d turn it over to the discussion and questions.

CO-CHAIR FLEISHER: Great. If we could go through the criteria. The evidence. This is probably one of the strongest from an evidence perspective. Multiple papers over the years since 1990 outlining the development, validity and predictive value of this database.

MS. WINKLER: This is an outcome measure so the criteria for evidence for an outcome measure is whether there=s a rationale to support the relationship of this outcome to at least one healthcare structure, process, intervention, or service.

CO-CHAIR FLEISHER: That makes it simple. Any comments? Can we vote?

MR. SANCHEZ: Voting will now begin for subcriterion 1a evidence. One is yes, two

is no. The voting timer starts now.

We have 20 for yes, zero for no.

CO-CHAIR FLEISHER: So, performance gap. This is well defined in the measure. Considerable variation from the 10th percentile of 0.89 to the risk-adjusted rate of 1.67 in the latest data set from June 2012 with quite a bit of variation depending on race and gender.

Any comments, questions? Can we vote?

MR. SANCHEZ: Voting will now begin for subcriterion 1b performance gap. One is high, two is moderate, three is low, four is for insufficient. The timer starts now.

We have 15 for high, 5 for moderate, zero for low and zero for insufficient.

CO-CHAIR FLEISHER: Next criteria is high priority. Addresses a specific national health goal priority. Certainly mortality, important. I don=t think there=s much else to

say. Anybody? No comments, okay. Let=s vote.

MR. SANCHEZ: Voting will now begin for subcriteria 1c high priority. One is high, two is moderate, three is low, four is insufficient. The timer starts now.

We have 19 for high, 1 for moderate, zero for low and zero for insufficient.

CO-CHAIR FLEISHER: Reliability. We heard a lot about how this has been tested over the years from the STS database and how -- and the data sources, the auditing via the STS database.

Do you have any comment?

DR. JACOBS: I think yesterday we covered the penetrance of the STS database which is about 90 to 95 percent of all programs in the country doing coronary artery bypass grafting. And we also discussed in detail the audit process to confirm the completeness and accuracy of the data. I can answer any

questions about that if anybody has any.

CO-CHAIR FLEISHER: I guess in relationship to the next measure do you know what percentage of all cardiac surgeries?

DR. JACOBS: I do. So, when we look at the percentage of all programs, first, we know that the penetrance has increased every single year so that at this point in time depending on what we consider the denominator of programs it=s between 90 to 95 percent of programs in the country.

And as far as cases go the percentage of cases is higher than that because the programs that we=re missing are lower-volume programs. So, therefore I would say the percentage of programs is 90 to 95 percent and the percentage of cases I would probably say is right at 95 percent.

CO-CHAIR FLEISHER: Questions or comments?

MS. WINKLER: I=d like to -- if

everybody else is done. A couple of questions on specifications, particularly as we know we=re going to look at another measure that=s very similar.

In this particular case this is data that=s submitted to the database. What are the specifications around capturing death in operative mortalities? Which deaths do get submitted? Which ones might not? Since you are focusing in on operative mortality who makes the assessment of whether a death is appropriate to submit as operative mortality or might not be.

DR. JACOBS: Right. Well, I think we=ve published several manuscripts that clearly define the criteria of operative mortality right down to the very minute of the 30-day cutoff.

So one is if you die before you go home from the hospital that=s clearly a definitive death. And the other one is 30 days from the

date of surgery with a clock that turns over at midnight the day after surgery. So those definitions are pretty tight and specified and subject to a fairly aggressive audit.

The cause of the mortality is not considered so it=s all causes whatsoever. Because that could start a slippery slope if you say only deaths related to the operation itself. And it=s all-cause mortality.

Which if you take it to the ultimate extreme, if you have a coronary artery bypass graft operation, you go home on day 7 and you=re hit by a car walking across the street on day 15 and die, that=s an operative mortality.

MEMBER DUTTON: It may be that your cognitive dysfunction caused you to be hit by the car of course, so it could be an operative mortality.

DR. JACOBS: And therefore it is.

MEMBER DUTTON: Yes. No, I totally applaud the all-cause mortality. It is

difficult to capture deaths post discharge in some patients. And I wanted -- I know you devote abstracter time to doing that and you work very hard at it. What percent do you think you get?

DR. JACOBS: Right, so we=ve looked at this quite closely and we=ve added fields in the database to capture not only vital status at 30 days but beyond capturing vital status at 30 days the method of documenting that vital status.

There=s a field in the database that says how do you know they were alive at 30 days. Was it because you saw them in the office? Was it because a referring cardiologist saw them in the office? Was it because you called their home? So we go to that level of detail and then we confirm that at the level of audit.

And I would say from our audit process, our adjudication process and from the methodology we use to confirm vital status by

linking with the Social Security Death Master File the accuracy of discharge mortality is well over 99 percent. And I think the accuracy of 30-day mortality probably is between 98 and 99 percent. So both quite good.

Clearly, as you=re alluding to, discharge mortality is more complete than 30-day mortality, but both are extremely good.

CO-CHAIR GUNNAR: So I=d be interesting in Dr. Handy=s thoughts. The perspective from -- at least my perspective is that 30-day all-cause mortality is an accepted definition. It extends across all specialties.

The oddness here, and it=s been historical in the STS, is if I have the patient, he=s in the hospital and he stays for six months but he never leaves and he dies, it=s applied to your mortality rate which does a couple of things.

It takes many more patients out of the

reporting cycle so you have to almost go back. And so there=s some mechanics about that.

And the second is that there=s disparities between hospitals depending on how good they are at offloading patients who may need a lower level of care but don=t go home. So if you=re able to send a patient to a rehab center or to some ventilator center they=re discharged from the primary facility but they really died a related death which is what you=re trying to capture.

So I guess my question has always been has STS looked at going back and sort of harmonizing that with the same data definition that=s used across all surgical fields which is all-cause 30-day mortality.

DR. JACOBS: Right. So the reason we use operative mortality instead of that which is kind of a standard definition for most surgical training programs and for most surgical databases is because it eliminates a

kind of perverse incentive that if you have a clearly very ill patient that=s going to die to keep them alive till day 31.

And by having survival to discharge it eliminates this incentive to say, okay, we=ve got somebody who=s dying and it=s day 28 but we=re going to put a trach in, put a PEG in, keep him alive and get him going till 31 days.

And I think in our training program as cardiac surgeons the definition of operative mortality is a fairly standardized definition that=s been in our literature since the beginning of cardiac surgery.

And we actually have mechanisms in place right now that allow for the fact that if somebody dies in an acute care facility where the surgery did not take place that=s still an operative mortality.

CO-CHAIR FLEISHER: So, attribution is to the original site.

DR. JACOBS: Absolutely. Just like

the readmission measure, same concept.

CO-CHAIR FLEISHER: Okay. And secondly, to your question, you have a date of death?

DR. JACOBS: Yes.

CO-CHAIR FLEISHER: So that if there was ever harmonization with the other measures you could get to 30-day all-cause mortality, not just --

DR. JACOBS: Absolutely. That=s a variable that=s in our database and that=s quite accurate.

CO-CHAIR FLEISHER: So that actually is useful from the perspective of you in a clinical registry.

And do you know how many additional variables you have compared to what=s on the discharge summaries, what=s on the U-BOLTS-92s?

DR. JACOBS: Right. So I think -- I couldn=t give you a number. What I do know is

that the clinical database allows for creation of a cohort of isolated CABG which is much more pure than one would get from trying to identify the cohort of patients undergoing isolated CABG from a billing database.

And that comes from a number of variables. We can look at associated operations with the CABG. We can associate -- we can look at whether or not a ventricular assist device was planned versus unplanned and really create a very pure cohort of isolated CABG patients.

And I think when we talk about harmonization, harmonization across discharge date versus 30 days is a relatively easy thing. The challenge is harmonizing across the definition of the denominator.

And that I think is the big strength of using a clinical registry to identify this cohort. Number one, defining the denominator appropriately, and number two, having precise

variables to do meaningful risk adjustment which are variables that really aren=t in a billing database.

CO-CHAIR FLEISHER: Do you know how many variables are in the model and how many are not available in a billing database? Do you know that?

MR. O=BRIEN: I=ll have to recount but I=d say around 30 variables in the model and in terms of the variables available in billing data, I mean, basically an ICD-9 diagnostic code for thousands of conditions. So I don=t know if you can count it that way.

CO-CHAIR FLEISHER: Fred, any chance you know the answer to that specific question?

MEMBER GROVER: I don=t know the answer to that specific question. It=s not just the variables but it=s the definitions.

In other words, in the timing if somebody has an MI and they=re in the hospital, they weren=t admitted with that diagnosis.

They have an MI. You operate on them two days later, that=s very important in terms of the risk adjustment model. So you=ve got to have that level of detail.

CO-CHAIR GUNNAR: The other is my understanding is that the variables actually that force the model change every three years when you reset the model. So that you have a whole bunch you can -- a couple hundred you can look to but that would --

DR. JACOBS: So the fields in the database and their definitions are modified every three years as part of a database upgrade process. The actual variables used within the model have stayed relatively constant.

So, the fields in the database change and that allows us to do a variety of other quality improvement and research initiatives. But when it comes to which of those fields are used for this model we=ve tried to maintain consistency within that subset of fields over

time.

CO-CHAIR FLEISHER: Other questions? Shall we vote?

MR. SANCHEZ: Voting will now begin for subcriteria 2a reliability. One is high, two is moderate, three is low, four is insufficient. The timer starts now.

We have 16 for high, 5 for moderate, zero for low and zero for insufficient.

CO-CHAIR FLEISHER: Okay. Validity. And we=ve heard a lot about testing. Do you want to make any further comments?

DR. JACOBS: I think it=s the same discussion we had yesterday with 10 percent of the sites getting audited every year.

And we know that these particular fields that go into the risk-adjusted mortality measure are amongst the most complete and accurate fields in the database.

CO-CHAIR FLEISHER: Okay, so risk adjustment. Any comments on risk adjustment?

I mean, you=ve had multiple papers on this and I find probably one of the best risk-adjusted models available. And I=m trying to remember, I actually unfortunately did not bring what your C statistic is in your model. MR. O=BRIEN: I think the C statistic is in the low eighties, like 0.80, 0.81 depending on the data set.

CO-CHAIR FLEISHER: Questions on validity or on risk adjustment in this model? We can vote.

MR. SANCHEZ: Voting will now begin for subcriterion 2b validity. One is high, two is moderate, three is low, four is insufficient. The timer starts now.

We have 19 for high, 2 for moderate, zero for low and zero for insufficient.

CO-CHAIR FLEISHER: Feasibility. So I think the feasibility is -- perversely this is -- data collection is difficult in that it=s actually a nurse-driven protocol. It=s

actually a clinical registry.

However, they have between 90 and 95 percent penetrance because basically the hospitals are paying to be part of this data set uniquely which I think many other data registries are having difficulty replicating this because of the resources necessary.

In fact, that=s probably the biggest issue that I=ve heard is whether hospitals can continue to maintain the infrastructure needed to actually complete this data set.

DR. BURSTIN: This is a question for Jeff and Sean. From conversations with Frank Opelka he tells me there=s actually been an effort to define the EHR elements associated with the registry. I think that would be helpful for the feasibility discussion going forward.

DR. JACOBS: Absolutely. STS is involved in ongoing efforts to collaborate with leading vendors of electronic health records to

develop a methodology where an EHR can be linked to the STS database and allow direct importation of whatever fields one could get out of the EHR.

That=s an initiative that=s important to STS to minimize data entry burden but to assure the completeness and accuracy of the data through a direct EHR import.

MEMBER HANDY: So, from the leading vendors what percent of the data could be directly abstracted? Do you have an idea at this point?

DR. JACOBS: I think it would be too early to say because it involves harmonization of definitions and adding certain precisely defined fields into the EHR that the vendors maintain. So I don=t think we could give an exact number yet.

But some things are pretty obvious and easy like name, date of birth, date of surgery, weight and height. Some things are a little

more challenging like the presence or absence of preoperative renal failure, the presence or absence of preoperative pulmonary dysfunction. And that=s just a matter of harmonizing with the EHR. And that=s an ongoing process. I think it=s a little too early to answer that question with an exact number.

MR. SHAHIAN: This is Dave Shahian. Can I make a comment?

CO-CHAIR FLEISHER: Sure. Please, Dave.

MR. SHAHIAN: Yes. We actually have a day-long meeting planned for later this summer with a major EHR vendor. And we=ve been thinking quite a bit about this issue over the past few years.

As Jeff has correctly pointed out there are some things that will be no-brainers.

But the thing that makes a clinical database like the STS registry unique is the granularity and specificity of its data

elements and the standardization across the country.

That distinguishes it from EHRs in which the data collected are generally not standardized and don=t have that degree of granularity.

We have to be able to do what we're doing with the STS database. We have to be able to know that something selected as post operative renal failure in Massachusetts means the same thing as it does in San Diego.

So, you know, it=s probably somewhere a little south of 50 percent I would guess that are actually going to be able to be extracted from an EHR. But even that would substantially reduce data collection burden. So we=re working hard in that area.

CO-CHAIR FLEISHER: Other questions? Comments? Vote.

MR. SANCHEZ: Voting will now begin for criteria 3 feasibility. One is high, two

is moderate, three is low, four is insufficient. The timer starts now.

We have 10 for high, 11 for moderate, zero for low and zero for insufficient.

CO-CHAIR FLEISHER: Usability. It=s used for public reporting. It=s part of, what, Consumer Reports and Quality Benchmarks.

DR. JACOBS: Right. It=s publicly reported on two different websites. One is the Consumer Reports website and the other is the STS website. Consumer Reports provides web access that=s at an arm=s length from STS. And then the STS provides web access that is accompanied by educational literature provided by STS about risk adjustment and about the measures that we publicly report.

So this measure is used for public reporting. It=s also used for a variety of quality improvement initiatives. And therefore I think it has a fair amount of usability in all important domains.

MS. WINKLER: Jeff, I haven=t seen it recently. Do you happen to have a graph or data of the change over time in mortality of the database participants? I saw some earlier and it=s quite impressive. So I was curious what the most recent data looked like.

DR. JACOBS: Yes, I think you=re absolutely right. We=ve demonstrated a number of graphs over the years that have showed the dramatic decrease in risk-adjusted mortality after isolated coronary artery bypass grafting over time from the beginning of the database up until now.

And you=re absolutely right that year after year it continues to decrease. The increments are a little less as the numbers get lower.

But I think that=s one of the biggest arguments in favor of participating in a multi-institutional clinical registry is that participation in and of itself is potentially

associated with quality improvement as demonstrated by this measure.

CO-CHAIR FLEISHER: Do you know -- it=s aggregated by program and participant.

DR. JACOBS: Right. So, it=s reported, stratified by two ways. Stratified by hospital and stratified by participant. The definition of participant is a practice group which rarely is an isolated surgeon, most commonly is a group of surgeons.

And most commonly the relationship between a hospital and a practice group is one to one. But sometimes one hospital will have more than one practice group and sometimes one practice group will go to more than one hospital.

So, by reporting both by practice group and by hospital one can then get at the differences in performance in the rare situations where a hospital has more than one practice group or a practice group goes to more

than one hospital.

CO-CHAIR FLEISHER: And Sean, does it have the statistical power to actually comment by individual surgeon or not? And then Jeff, you can comment whether you=ve chosen to do it by group because of a strategic decision.

MR. O=BRIEN: I think small and variable sample sizes are the challenge for a lot of outcome measures. So that it has the ability to differentiate performance, at the extremes particularly.

I think that it really depends on your choice of criterion, or identifying differential performance. So when providers are classified in performance categories that it involves a tradeoff of kind of types of errors, false positive and false negative.

When you do analyses to actually estimate the true amount of variability that exists, whether or not you can detect it with a given sample size, just asking the question

of how much variability is there, it=s really a substantial amount of variability.

If you estimated the risk-adjusted mortality rates at the top 10 percent of the distribution compared to the bottom percent of the distribution that would be a 3 and a half-fold difference potentially, estimated difference between the outcomes at those extremes.

The STS approach uses a fairly conservative criterion for identifying outliers. And that could be increased by adjusting that criterion. So there=s a power to identify performance to the extremes.

DR. JACOBS: And I=d just like to add a little bit about the concept of reporting via hospital and practice group versus the concept of reporting by individual participant. And this discussion will apply to all of our measures because it=s going to come up repetitively with each measure.

First of all, STS has traditionally and up till now chosen the strategy of reporting by hospital or practice group because we believe that the outcomes after cardiac surgery are reflective of the performance of the entire team rather than the performance of one individual provider.

So, all outcomes are affected not only by the surgeon but by anesthesia, nursing, cardiology, preoperative care, postoperative care. And that=s why oftentimes we=ll see the phrase that cardiac surgery is a team sport and therefore outcomes are reflective of the performance of the entire team.

That being said, there=s clearly a substantial desire to get at outcomes and outcome measures of individual providers. And although currently this measure is reported stratified by hospital and practice group we=re in the process of working fairly aggressively to develop mechanisms to report cardiac

surgical performance stratified by the individual surgeon.

Sean already alluded to some of the issues that we face in doing that which relate to sample size. Any individual surgeon does less cases than an entire hospital or practice group.

And we=re developing some methodologies to overcome those challenges which include not only using composite measures but potentially creating a composite of composites where the performance of an individual surgeon can be examined by their overall performance across a number of operations rather than just isolated CABG.

And that=s something we=re working on fairly aggressively right now to address the desire to get at performance measures for individual surgeons.

I realize that=s a little bit long discussion but it applies to about five or six

of the measures we=re going to discuss today.

MEMBER YATES: Question. Given the fact that you=re being asked about going to a more granular level of individual surgeons, do you have any data or have you ever polled the members of STS as to their acceptance of the STS registry as being (a) fair and (b) accurate?

DR. JACOBS: That=s a great question. And I think, first of all, most surgeons realize that this is inevitable and that our discussion about cardiac surgery being a team sport and the outcomes being reflective of the performance of the entire team, while we believe that=s true is not going to prevent the need and ultimate release of outcomes stratified by individual surgeons. So it=s somewhat inevitable.

And therefore most STS members feel that if this is inevitable it certainly would be better coming from a clinical data source with precise definitions of the denominator and precise methodologies of risk adjustment.

And as far as the question you=re asking about polling members, I think where we are right now is we=re finalizing our methodology to create a technique to report outcomes stratified by individual surgeons for risk-adjusted mortality.

And that=s going to be presented to the membership at our next annual meeting. Dave Shahian is leading that initiative. Sean=s the leading statistical expert on that initiative.

And I think prior to polling the members I think it=s important to share with them exactly the methodology we plan to use to accomplish this feat.

And then after that I think the next step would be to go back and see what kind of buy-in we get from the membership.

MEMBER YATES: If I could just make a comment to the NQF staff. It=s not one of the criteria for usability and use, but the

acceptance by the specialty is an important canary in the mineshaft if you will.

And if there=s a sense of this is fair for us then its utilization across very different potential purposes such as state and various different federal reporting mechanisms is going to be more readily enforced or more readily participated within.

And what you don=t want is a lot of inertia and a lot of -- because there=s a sense that the measure isn=t really capturing quality by the profession itself.

And if you have 90 percent of cardiac surgeons feeling this is not the case with STS, but if 90 percent of cardiac surgeons felt like the data didn=t make any sense to them you=d want to know that as a process for determining its quality.

And that might be something to add to your usability and use criteria. Because I think that would be important moving forward.

DR. BURSTIN: It=s a very fair point. I think again being a multi-stakeholder organization that also has to be balanced by the usability by the end users as well, by consumers and purchasers. So, it is always a balance but I completely see your point.

DR. JACOBS: And I would just add to that. You=re absolutely right, if the users of the database, the cardiac surgeons, don=t buy into the actual use of the database that=s a big problem.

And the fact that over 90 percent of programs, close to 95 percent and certainly 95 percent of operations are reported, I think that alone demonstrates a fair amount of substantial buy-in to what we=re currently reporting.

The fact that the number of programs publicly reporting has doubled over the last three years also demonstrates the buy-in. And I think we=ll have similar buy-in with the

individual participant measures once they=re fully developed.

MR. SHAHIAN: This is Dave Shahian. Can I make one brief comment?

CO-CHAIR FLEISHER: Go ahead, Dave.

MR. SHAHIAN: One of my jobs is unofficial chair of the complaint department database. And I must say it=s been astonishing how few complaints or areas of pushback from our participants.

They=re usually fairly sophisticated nuanced kinds of issues around the statistics and we usually have been able to answer them quite easily.

I think the reason that there=s been such widespread acceptance is twofold. First of all, I think the STS database participants recognize both the value of clinical data and the accuracy with which it=s collected in your institution and then presumably across other programs.

They also recognize that absent the kind of data and accuracy and models that we have they=re subject to a lot of other black-box rating mechanisms out there which they=d like to avoid as we would.

And then the third aspect of this is the education. Before we roll out any measure we go through an extensive educational campaign. The papers are published in the literature. We present them at national meetings. Everything is spelled out in great detail. So I think education is a large part of this.

So we really have very few areas of pushback and hopefully it=ll continue to be that way going forward. Thank you.

CO-CHAIR FLEISHER: Collette=s next and then Anthony.

MEMBER PITZEN: Collette Pitzen, thank you. I just wanted to make a comment about the usability of the measure. And this

comment is not to pick on STS. It's a great, wonderful database. It's also in general for complication rates for surgical procedures. With the national rates of CABG mortality at around 2 percent it's really difficult to discern with this as an individual measure differences between practices.

If you go on STS's website for this particular measure every single practice gets a two-star rating. So I just wanted to share that from that perspective about identifying excellent providers, even though there is differentiation at the decile level it is very hard to discern when you have a complication rate that is so low. Thanks.

DR. JACOBS: Yes, there's a couple of comments I would make related to that. First of all, someone in this room yesterday talked about what kind of complication rates would be accepted in the airline industry. And I think that's great. That was a tremendous analogy.

A 2 percent complication rate in the airline industry would be completely unacceptable. And I think we're trying to get that 2 percent death rate after isolated CABG to be a lot less.

But second of all, regarding your important point about differentiation, the isolated CABG risk-adjusted mortality measure is a component of the overall CABG composite. And the overall CABG composite even on the publicly reported site shows substantial differentiation between one-star, two-star and three-star.

And then if we get to the non-publicly reported programs, all of the programs which is the remaining half of them there's even more differentiation. So, about 12.5 percent are one-star, 75 percent are two stars and 12.5 percent are three stars.

And then when one goes beyond the star rating to the actual percentage there's even

more variation because one can stratify not just into one-star, two-star and three-star, but the actual numbers and those numbers are publicly reported to accompany the stars.

MEMBER PITZEN: Right. I'm sorry, I forgot to mention I think there's extreme value in inclusion in the composite.

CO-CHAIR FLEISHER: Thank you.

CO-CHAIR GUNNAR: Can I jump in just real quick? Just real quick. Because I think Collette's point is a good one and one for NQF to think about going forward.

My understanding is ACC, American College of Cardiology and STS have for some time tried to collaborate on integrating their information.

Isn't the real question about quality since it's a cardiac service team sport really isn't it about when you present to the hospital with a cardiac -- a coronary syndrome of some sort and how you then exit the hospital. And

that now incorporates so many more technologies than it ever did in the past.

And it really relates to the integration of a cardiac cardiology team. So I wanted to kind of have the comment about that.

DR. JACOBS: Yes, and I can comment on that a bit. The STS and the ACC are currently involved in a number of collaborative initiatives that link data from the STS database with data from the NCDR, the database of the ACC, and also with Medicare data for longitudinal outcomes. And that's been funded through a federally funded grant called the ASERT grant. It's led to a number of publications some of which have been published in the New England Journal.

And what that allows is for the assessment of performance across the entire spectrum of cardiac care, the whole team, not just surgery but surgery, cardiology and the whole team in a programmatic examination of

outcomes.

I think that the work done with those linkages up till now have been to look at comparative effectiveness research and overall quality of healthcare delivered by the cardiac team.

And it's I think a reasonable possibility that that initial work is ultimately going to lead to some measures that could be proposed that would not just be cardiology or cardiac surgery quality measures, but cardiac team quality measures. It's not a big leap from where we are now to get to that point.

CO-CHAIR FLEISHER: I just have a quick question for staff. The star rating. Depending on different committees is that part of the measure or is that not? Just for clarification for me.

MS. WINKLER: No, the star rating is actually not. The measure includes the point

estimate as the measure result.

How it's portrayed to any particular audience could be quite variable depending on the user. This is STS's chosen method but it is by no means something that automatically goes with the measure.

This is something that's been discussed throughout NQF the last time we went through this. So just to be clear.

DR. JACOBS: And we at the STS look quite closely on what's the best way to portray this information to patients, patients' families and other consumers of healthcare.

And on our website and on the Consumer Reports website one of the goals is to be able to portray this information in a fashion that it can be understood by somebody with a sixth grade education.

And then to allow a deeper dive into the same data with more detail for those that have the sophistication and the desire to learn

more.

So therefore, the star rating allows one level of examination but by right-clicking on those stars, or double-clicking on those stars, all of the detail numbers that go into choosing those stars is also provided.

CO-CHAIR FLEISHER: So, Anthony, do you still have a question?

MEMBER ASHER: Dave, are you still on the line?

MR. SHAHIAN: Yes, I am.

MEMBER ASHER: Dave, when you were talking -- I'm just looking at this from the standpoint of the individual users. And of course the goal here is to develop more relevant measures of performance. And I would imagine also to ultimately relieve individuals of burdens that aren't relevant.

And so in this present period do you believe it's the case that the measures that you've developed, particularly in the context

of the registry, has this overall reduced the burden on thoracic surgeons with respect to relevant measures programs? Or is that more of a future goal?

MR. SHAHIAN: No, I think it has. I mean there are -- if you're talking about the other sort of proprietary commercial rating organizations that I have alluded to, they're out there and programs are continuing to receive ratings. But our members have highly validated and we hope NQF-endorsed measures that stand as the gold standard. And that's' the advantage we have.

I guess I would have to say that there are fewer of the less desirable ratings applied to cardiac surgery now simply because folks recognize the validity and the strength of our measures.

And it's one of the arguments that when other societies come to us and ask about this whole measure development process we always

point out to them that you need to do it, you need to do it right. You need to do it with the best possible science. And that's your best defense against some of the less desirable, less accurate rating systems. So I think we're on our way to achieving that goal.

And we're expanding our portfolio of measures to include more and more procedures and more and more aspects of care like appropriateness which hopefully you'll be seeing in the future.

CO-CHAIR FLEISHER: Thanks. I think we need to vote on this measure. So, usability.

MR. SANCHEZ: Voting will now begin on criteria 4, usability and use. One is for high, two is for moderate, three is for low, four is for insufficient information. Timer starts now.

CO-CHAIR FLEISHER: And I think this is a great discussion. As we go to the other

measures if it's similar hopefully we can just say as discussed first thing this morning. It will be a way to continue.

Okay, we're now onto -- as soon as this is done.

MR. SANCHEZ: We have 12 for high, 11 for moderate, zero for low, zero for insufficient information.

CO-CHAIR FLEISHER: Are we ready to go onto vote? Any objection? Okay, go ahead and vote.

MR. SANCHEZ: Voting will now begin for overall suitability for endorsement. One is for yes, two is for no. The timer starts now.

CO-CHAIR FLEISHER: Okay. And if we can switch and get CMS and I guess the Yale group for the next measure. All yours.

MR. SANCHEZ: We have 23 for yes and zero for no.

CO-CHAIR GUNNAR: So this is 2558,

correct? Hospital 30-day all-cause risk-standardized mortality rate following coronary artery bypass graft surgery. CMS. Who is the discussant?

My apologies. CMS.

MS. SUTER: Actually, my name's Lisa Suter. I'm a physician with the Yale Center for Outcomes Research and Development. We developed this measure under contract to CMS.

MS. DRYE: Hi, Elizabeth Drye. Also director of quality measurement there and a physician who worked with Lisa on this measure.

CO-CHAIR GUNNAR: So, you have three to five minutes to sort of provide an overview.

MS. SUTER: Great, thank you. Thank you very much for giving us this opportunity to speak in front of you.

As was recently discussed this is a high-impact, high-priority procedure. It has been conditionally MEDPAC supported as a measure.

You may or may not be aware that this was developed in combination with a readmission measure based on administrative claims. Those two measures were developed in very close harmonization with a complementary readmission measure based on registry data developed by STS that involved weekly workgroup meetings with our group and the STS measure developers which led to very closely harmonized readmission measures.

And the mortality measure that we developed benefitted from this close harmonization of the readmission measures. And is intended to complement the claims-based readmission measure.

Although we were unable to validate this measure against the STS database due to contractual issues the cohort definition of isolated CABG was validated for the readmission measure against the registry-based measure with a 97 percent accuracy, much higher than

previous published reports of administrative claims achieving an accurate definition of isolated CABG. And we're sure that this is due to the close collaboration between STS and Yale.

In addition, we were able to validate the risk adjustment model for this measure against a well known state registry database in New York with the assistance of Dr. Edward Hannon and received a high correlation between the two risk models.

I think that's all I need to say and I'm happy to answer questions from the committee.

CO-CHAIR GUNNAR: Ms. McCarty?

MEMBER MCCARTY: So, this is an outcome measure that reflects the process of care coordination, particularly for peri- and postoperative care and care at the time of transitions, for example, discharges to home or skilled nursing facilities.

It is an all-cause mortality with a risk-standardized mortality rate for patients 18 years and older discharged from the hospital following a qualifying CABG procedure where mortality is defined as death from any cause within 30 days of the procedure after having an admission for CABG.

We've heard a lot about this already this morning so I think I can probably stop there.

MS. WINKLER: Why don't we go ahead and start going through the criteria. So the first one is evidence. This is an outcome measure. This is sort of the same measure you just talked about. It would be interesting if you voted differently for this time than the last one.

CO-CHAIR GUNNAR: So a challenge has been posed by NQF.

MEMBER GROVER: So I'm abstaining on this one too.

MEMBER CIMA: Can we just ask a question of sort of the gorilla in the room or elephant in the room? Why two?

MS. WINKLER: We'll get there.

MEMBER CIMA: Okay. I didn't know where that was going to fall in the criteria.

MS. WINKLER: We'll get there.

CO-CHAIR GUNNAR: Amaru, we're ready.

MR. SANCHEZ: Voting will now begin for criteria 1a evidence. One is for yes, two is for no. Timer starts now.

We have 23 for yes, zero for no.

CO-CHAIR GUNNAR: Proving we are a trainable group. And honest in our responses. So, performance gap.

MEMBER MCCARTY: Yes, so performance gap. So, in the data analysis they did they showed an average of 3.3 percent mortality with a fairly large range I think for this type of measure where it ranged from 1.5 percent for certain facilities up to 9.3 percent. Which is

again I think a very wide amount of variation for this type of measure. So I do think that there is a high opportunity, high performance opportunity with this measure.

CO-CHAIR GUNNAR: Any other discussion? We're ready to vote.

MR. SANCHEZ: Voting will now begin for subcriterion 1b performance gap. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: Waiting for one more? Or we're good. Olson, Dr. Olson stepped out. So we're good at 22.

MR. SANCHEZ: Sixteen for high, six for moderate, zero for low, zero for insufficient.

CO-CHAIR GUNNAR: I think that tracks with our previous response.

MEMBER MCCARTY: In terms of priority we know that CABG surgeries are high-cost procedures and they account for the majority of

cardiac procedures performed nationally. And then again from what we've seen with the data it's also associated with considerable morbidity, mortality and healthcare spending. So my assessment was that this is a high priority.

CO-CHAIR GUNNAR: Any further discussion? All right, ready to vote.

MR. SANCHEZ: Voting will now begin for subcriterion 1c high priority. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR FLEISHER: In case people were wondering I just asked -- there is an obligation to make those statements for the transcript purposes.

CO-CHAIR GUNNAR: We have 22 so let's --

MR. SANCHEZ: We have 21 for high, 1 for moderate, zero for low, zero for insufficient.

CO-CHAIR GUNNAR: And now for the discussion.

MEMBER MCCARTY: So, for reliability the developers did use the exact same risk stratification methodology as was used in STS because they wanted to keep these two measures in harmonization if I understood that correctly from the description.

In order to review the data they looked at a combined data set with 2008 to 2010 data and randomly split it into approximately two equal subsets of patients and calculated the risk-stratified mortality rate of each sample and got agreement between those two sets. So they did significant reliability testing and showed good results from that.

And again, the methodology is the same one that I believe we heard described earlier today.

CO-CHAIR GUNNAR: Any other -- yes, here we go. Dr. Yates.

MEMBER YATES: The data for STS is through the direct registry. How are you collecting the data for CMS? Is the data coming from that registry data directly, or is it coming from harmonized collection of data through administrative data set?

MS. SUTER: This is Lisa Suter with Yale CORE. The data for this measure, entirely administrative claims data. There's no burden on hospitals. These are already submitted using billing codes.

The risk adjustment strategy looks back 12 months prior to the hospitalization for the CABG procedure, includes all appropriate diagnostic and procedural codes in the risk adjustment. And this risk adjustment model was validated against a registry sample using the New York registry with Ed Hannon's assistance.

MS. WINKLER: When you look at the data you presented for the gap in variation

compared to that from STS they are quite different. So, how are the specifications of this measure differ from the STS measure? I mean, what can account for those differences?

MS. SUTER: So, I think there are -- this is Lisa Suter from Yale CORE. I think there are several possible discordances between the results you saw from STS and our results, the first of which is these were results from 2009 through 2011 for our data. We can't speak to the STS data but certainly there may be trends over time that we're not capturing.

The second of which is as you're aware of the last discussion the outcome is slightly different with ours being a truncated measure at 30 days all-cause even if you are hospitalized beyond the 30-day mark.

And the last of which is that you're measuring a slightly different collection of hospitals. This measure captures all

hospitals submitting Medicare claims for CABG procedures. And it is also looking solely at among Medicare beneficiaries. So, patients who are 65 years and older. Whereas the STS data includes younger patients.

We have validated our measure specifications in an all-payer data set using California all-payer data and found similar results. But the results that were presented in the application are for Medicare beneficiaries 65 and older.

CO-CHAIR GUNNAR: Dr. Fleisher. Go ahead.

CO-CHAIR FLEISHER: So, I'm going to ask the questions I asked of the STS group. Which is how many additional elements or less elements are there in your model and how do you handle transfers within your model.

MS. SUTER: This is Lisa Suter from Yale CORE. The claims-based measure is harmonized with regards to transfers with the

STS measure. The mortality is attributed to the initial hospital performing the initial CABG procedure.

So, any subsequent transfers are not considered separate hospitalizations. The outcome is attributed to the initial hospital performing the CABG procedure.

In terms of -- your second question was the differences in terms of risk adjustment factors. So, on page 37 of the technical report we list the risk variables included in our model. I'm not actually sure numerically which ones are -- how they differ from STS.

I will say that there's a significant overlap between the mortality risk adjustment model and the readmission risk adjustment model which was validated against the registry data in a matched cohort of patients and yielded very similar C statistics for readmission.

Again, it was validated against a New York State registry and revealed very similar

C statistics in that database with the clinical registry data yielding a C statistic of 0.75 and the claims-based measure, 0.74.

With our whole cohort our C statistic is 0.84. That includes the national data. That's the C statistic for our model in national data.

In terms of the differences between a clinical risk model, as you can imagine claims data captures very different information. A COPD diagnosis does not necessarily present the same information as an FEV1 might. So, they do capture different information.

But when they are aggregated up at the hospital level in a matched cohort of patients they produced correlation rates for the readmission measure of in excess of 0.90, close to 0.95 in some cases depending on the statistic used.

We saw similar correlations in the New York State registry data when we looked at a

comparison of the risk model performance in the clinical data versus the registry data.

CO-CHAIR FLEISHER: So, have you actually compared the rankings between yours and the STS data set? Do hospitals change in rank?

MS. SUTER: So, we have not done a comparison of our mortality measure compared to the STS results.

We did do a recategorization analysis in the New York data set. As you can imagine this is a challenge. There were only 36 hospitals that had matching and of those 36 only 35 had at least 25 cases in the data that we had available.

Among those 35 hospitals, 33 were identically categorized in terms of performance between the 2 measures, and 2 were categorized differently.

The two that were categorized differently were identified as low-performing

outliers by the registry model and were categorized as no different than average, New York average by the claims-based measure.

We tried to investigate that a little bit. The predictions of the performance estimates are highly overlapping but the cutoffs just meant that for the claims-based model they fell close to, you know, just below the cutoff for being considered worse than average.

So I think in comparison we know from our readmission measure that the claims-based measure does seem to err on the side of conservatism in terms of identifying outliers which may be an appropriate approach given its application potentially.

CO-CHAIR GUNNAR: How do you manage missing data?

MS. SUTER: So, there are patients who are not enrolled in Medicare for the sufficient period pre-operatively to allow us risk

adjustment are eliminated from measurement. And that's probably the largest exclusion to our measures.

In terms of unreliable data, a small proportion of patients are eliminated if they have what we consider data that might be in error. So, for example, age over 115 or things like that. It's a small proportion of patients, it's probably 0.1 percent of cohort, but they are excluded with a careful algorithm that identifies outlying data.

And otherwise there are no -- I mean, these are administrative billing data. There are no missing data as long as you're enrolled and we have captured your claims.

CO-CHAIR GUNNAR: And for perspective, Dr. Grover, STS average age for CABG, is it 68? Sixty-seven? The median.

MEMBER GROVER: It's in the sixties I believe, but I can't remember.

CO-CHAIR GUNNAR: But it's at or

around Medicare age at 65.

MEMBER GROVER: Well, yes, but there are younger patients. It's just they're a lot, you know, that's a disease that presents later.

CO-CHAIR GUNNAR: But I think for the perspective of the committee about half. It's roughly about half of --

MEMBER GROVER: Yes.

CO-CHAIR GUNNAR: -- the patients would be of Medicare age that undergo coronary artery bypass grafting. Okay. We can agree on that I think? All right.

Collette?

MEMBER PITZEN: Hi, Collette Pitzen. I just had a small question that came up during our workgroup call. Thanks, Dr. Suter, for all the explanations.

If I'm understanding correctly patients that have a repeat CABG procedure during the measurement period are not included. Or their initial CABG is the one that's included

but a redo CABG, then potentially mortality after that would not be included. And I was just wondering about the overall impact of that particular exclusion. Thank you.

MS. SUTER: Thank you. This is Lisa Suter from Yale CORE.

In terms of the decision to identify the first CABG in the measurement period and exclude subsequent CABGs this was a lengthy discussion with the measure developers that included our workgroup with cardiothoracic surgeons including Dr. Shahian who I believe is on the phone.

And the thinking and what has always really impressed me about cardiothoracic surgeons involved in this measure development is the tremendous amount of responsibility for patient outcomes in this group of physicians.

And the feeling was any subsequent events in the year following a CABG, particularly a redo CABG were likely a

significant failure of the initial CABG. And the feeling was from a clinical standpoint to capture that initial CABG as the primary event and obtain an appropriate measurement and performance estimate for that hospital performing that CABG.

Certainly other approaches are justifiable but for this particular measure the recommendation from our clinical experts was to appropriately pick that.

A distant history of prior CABG is captured as long as it's captured in the codes, the RV codes for history of CABG procedures.

MEMBER KO: I wanted to ask a question about data source. So clearly the prior measure was registry. This is claims. And there's pros and cons of both. If we could marry the best of both that would be terrific.

When we try to do that at the college, merge clinical and claims, one of the problems we have is the delay in getting the

administrative. We can't seem to reconcile that and give back timely data.

How do you handle that? What do you see as the time frames if you're going to report however, semiannually or annually? How delayed is that data? And do you see -- and does it even matter if it's delayed? If it's a year late, you know, how large are the differences from year to year? Because I suspect it's not that large.

MS. SUTER: So this is Lisa Suter from Yale CORE. It's a great question and I'll present and answer and also allow CMS to comment as well if there are additional comments.

First of all, the exact reporting mechanism for this measure hasn't been determined. Other similar measures have been reported across a three-year time frame. So a hospital has a three-year collection period of data.

There is usually about a 18-month

delay between the end of the measurement period and the report of the results. And the reason for that extensive delay is that hospitals have a legal opportunity to debate their claims and their payments. And so that gives hospitals an opportunity to make sure that their claims are accurate and to ensure that the administrative claims that are being used for performance estimation are final.

And therefore, it's a tradeoff. And this is something that measure developers are challenged with all the time which is the tradeoff between accuracy of our estimation and the rapid cycle with which we can provide feedback to hospitals for performance improvement.

In terms of generating reliable estimates at the hospital level we have for the purposes of the outcome measures that we have developed thus far and that are in public reporting because of the high-profile nature of

these measures we have favored longer periods of data collection in order to provide stable estimates using final action claims in order to produce the most acceptable estimates to stakeholders possible.

We recognize that that's a tradeoff. And this is certainly one of the areas in which having complementary measures that measure a similar outcome and a similar cohort such as the two measures we're discussing today provide measurement opportunity in different spaces that are both very important to healthcare.

CO-CHAIR GUNNAR: Any other discussion? Ready to vote.

MR. SANCHEZ: Voting will now begin for 2a reliability. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 12 for high, 10 for moderate, 1 for low, zero for insufficient.

CO-CHAIR GUNNAR: Validity.

MEMBER MCCARTY: So in terms of validity this measure was tested for validity at the data element level.

In addition, the developer cites that the validity testing has been completed similarly for six other NQF-endorsed varoius 30-day mortality measures that are currently used now for public reporting.

And then we heard that fom various testing that they did they actually have used data from five different data registries in terms of validating the data at various levels. So I believe that this measure has been highly validated.

CO-CHAIR GUNNAR: Any further discussion? We're ready to vote.

MR. SANCHEZ: Voting will now begin for 2b validity. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: Just vote again

real fast everybody. There we go.

MR. SANCHEZ: We have 14 for high, 9 for moderate, zero for low, zero for insufficient.

MEMBER MCCARTY: Okay. In terms of feasibility all data elements required for this measure are captured in electronic claims. And as stated earlier, this type of data collection and analysis is already being collected at facilities for other types of 30-day mortality metrics that are currently publicly reported measures. So, feasibility for this I believe is high.

CO-CHAIR GUNNAR: Any other discussion? Ready to vote.

MR. SANCHEZ: Voting will now begin for feasibility, criteria 3. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: Everybody one more time. We're missing one. There we go.

MR. SANCHEZ: We have 21 for high, 2 for moderate, zero for low, zero for insufficient.

MEMBER MCCARTY: Okay. In terms of usability and use I have a little bit of difficulty thinking about the usability because there are competing measures that are evaluated. But I know we get to that discussion later today in terms of comparing them.

It would be used for public reporting, or that is CMS's plan though I guess they're still considering it. But they will be looking to make that happen.

And this is a new measure so at least in terms of the way it's defined from CMS so there's no -- it's not a continued one. But it would be used for public reporting.

MS. WINKLER: Dr. Cima, you had said ask when. Now might be part of your question would be reasonable to ask, the added value, or

usefulness of two, or something like that.

DR. BURSTIN: Or actually just frame it the usefulness of this particular measure.

MS. WINKLER: In the context of others.

MEMBER CIMA: Well, I mean the question becomes -- it's not really a data burden issue because it's administrative data. But we already know that 90 to 95 percent of all cardiac programs are in STS. Hospitals are doing that. It's more granular data.

It does have the advantage of not having a 30-day cutoff. And for those of us who have spent time in cardiac ICUs a number of patients that do expire there greater than 30 days. And so, you're going to miss some.

Having two measures out there with sort of the same title that are going to be used in the public reporting space using different methodologies doesn't really do -- it may do a service, but it may also do a disservice because

you may pull away with different ideas.

And I think as we're pairing away things just having a measure because we have another measure that does the same thing, that's going to capture a measure that has more granularity, has more detail, that you can get a subset of those greater than 65, I'm not sure why -- I'm just trying to understand why. Why two.

It just seems like that's one of the things we should try and do is either harmonize them to one or basically say one is better than the other and we're going to go with the one that's better. And just decide which one. I don't have a dog in the fight but I'm just saying just go with one.

CO-CHAIR GUNNAR: Dr. Siperstein.

MEMBER SIPERSTEIN: I mean, obviously there's an issue comparing this to the STS measure. But the issue I have with usability even as an isolated measure to assess the

quality of the program is that this just looks at the Medicare subset. And that's the major factor just as an isolated measure that I have issues with usability.

CO-CHAIR GUNNAR: Tagging onto that question historically if you underwent a trach after a coronary artery bypass procedure you actually fell into a different ICD-9 code and you actually didn't get tracked under CABG.

Under a DRG code, yes. So, apologize for that, yes. Different DRG code. So you wouldn't get tracked for that reason.

MS. SUTER: So this is Lisa Suter from Yale CORE. We do not use DRGs in the definition of our measure cohort so you're identified by your procedure code which does not disappear even if your trached following surgery.

MEMBER SIPERSTEIN: Perfect. Thank you.

CO-CHAIR FLEISHER: So to follow up on Allan's question or was it Bob's. So I look at

two things.

One is is it fair to look at a program. Why would you think that only measuring the Medicare patients, are they uniquely different? You actually have a smaller sample size for a given population. So that why is there advantage to only saying that to Medicare beneficiaries. So CMS may want to answer that.

Two is there's a very small cohort of small programs that don't participate in the STS database. Given the size or the volume of patients, especially with the hierarchical model that you use can you make any statements about quality in a hospital that's so small that they wouldn't participate in the STS database.

Because my understanding of your hierarchical model would actually place them as no different than average which gets to the third statement. In the STS model you actually have outliers and in your model do you know what percent really fall as outliers?

We heard 12.5 percent low, 12.5 percent high statistically from the STS group. Where do you fall out as far as high and low?

MS. DRYE: Hi, this is Elizabeth Drye from Yale CORE.

The focus on Medicare patients is in part -- I'll just speak as a measure developer and CMS, I think you might want to come in because it's partly pragmatic and partly programmatic focused.

The pragmatic piece is as you know, that's the one national data set that we have where we have every patient covered for a defined age group, 65 and older. So we are able to pull in those straggler hospitals and programs that don't participate with STS. And those are of particular concern.

In terms -- that's a great question. Is there enough volume at these hospitals. I don't think, and I would turn back to both my colleague Lisa Suter and our STS colleagues.

I don't think we were able to really fully characterize the uncaptured STS hospitals because of data agreements. We really would love to be able to answer that question, exactly how many patients are at the hospitals that don't participate with STS and how concentrated are they within those hospitals.

But I think we're particularly concerned about the quality of care. This is a Medicare beneficiary-centered approach. And that's an important group that we're all going to be in at some point.

CO-CHAIR FLEISHER: Well, I would actually ask the robustness of the model. So, from year to year given sample size of Medicare patients who actually have a larger, maybe anywhere from a 100 percent larger sample size if it's really only 50 percent Medicare. So the robustness from year to year to say anything about Medicare, is it really -- is a larger

sample size in any given year going to give you more a reflection of the program, or do you actually think that the ability to just track the Medicare cohort really reflects the patient's care in that hospital?

MS. DRYE: We're trying to get at the quality of care for those patients specifically. And I think we will have enough patients. You know, all of our outcomes measures are focused on the subset of patients in a hospital that are Medicare patients.

We are -- we don't have a reason to think that care doesn't track with the overall care in the hospital. So, I mean, it's a research question that we've tried to look at in different contexts.

But I think we will have enough to have outliers in our population. I think there were not a lot of -- it's very hard if you -- whether you see outliers or not depends as you know on how conservatively you classify outliers.

So if you use, for example, a 95 percent confidence interval and say we're only going to put somebody in a better or worse category if they're very -- we have a really high level of confidence then in these kinds of outcomes measures you're not going to see a high number of outliers.

But I think we would have the ability to identify very poor performers, even if they have relatively few cases.

CO-CHAIR FLEISHER: I guess another way to ask my question and I'll be brief is if they're an outlier in your data set and not an outlier in STS the next year, has anybody looked, would they remain an outlier in your data set? Or was it just a statistical fluke of that year?

MS. SUTER: So, we haven't had the opportunity to look at that. I will also just say I tried to see whether we've had an opportunity to run the actual performance

categorizations on the whole U.S. sample.

And we don't have those bootstrap results that actually slot hospitals into categories. It was only done for the small subset of hospitals during the validation process. So I can't actually tell you using the methodology for existing publicly reported CMS mortality measures how many outliers there are in the CABG measure.

But Elizabeth's right that you could create more outliers depending on what cutoff you use to define an interval estimate around the national average in order to define what's statistically significantly different from national average.

MS. DRYE: I would just -- I'm trying to answer your question. I apologize because I feel like I'm not answering your questions as directly as I want to. And it's not because of the way you're asking them, I think I'm just -- they're really, really tough questions.

Would that same hospital -- let's say that, you know, we're trying to get at a latent variable of quality in these hospitals.

And let's say it's a bad hospital. You know, if it shows up one year as a bad hospital in our measure or an STS measure would it necessarily be an outlier in the next year? I don't think we can say for sure it will.

What we think about with these outcome measures, we try to produce them so we get the most reliable score we can. But if you don't want to be a bad hospital you better be good. If you want to be a really good outlier you better be great. I mean, the really, really great and the really, really bad will be consistent, but there's some movement year over year. And I think that's just the nature of using patient outcomes to capture a latent variable of quality.

CO-CHAIR FLEISHER: Dr. Yates.

MEMBER YATES: Just a follow-up on the

duality question that came up. And this may be something that CMS answers more readily. But it came to my attention the other day that I saw NSQIP data being advertised on HospitalCompare.gov as being now one of the things being reported there.

And is there any possibility for the greatest transparency for consumer stakeholders that you would have dual reporting, side-by-side reporting of STS reporting and also the CMS Yale CORE reporting processes so that there would be a way for consumers to compare, say, the Medicare population versus population at large, or see that the two correlate?

MS. HAN: This is Lein Han from CMS. I would like to address the question of what you call duality or competing measures.

From CMS's perspective we don't see this issue as either/or. I think I would like to frame it as more like short-term and

long-term. And see it as a progression.

CMS very much is supportive of measures using different data sources, clinical data, EHR data. And our goal is to develop measures, use measures in the EHR environment using electronic health records in the future.

But you can see that -- I can see the vote feasibility is pretty high for these measures from the panel. So that's the current concern right now, short-term, is the feasibility for us.

We take into account the burden to hospitals, burden to taxpayers, and a lot of legal factors too, to implement a measure. So, that's how we see why we proposed using the claims-based measures as a short-term goal here.

CMS has developed not only claims-based measures. We have hybrid measures in terms of clinical and the EHR data

with the claims data. And we are working toward develop -- or we call it reengineer, the de novo measures using EHRs.

So, I think for us it's a progression issue. It's not either and or. It's not dual thing.

And the other thing is that CMS is not the only organization develop or use measures. And we'd like to see these measures -- these are both very good measures. Like other organizations may have the capability to use other measures. I think that's one of the reasons we support these types of measures.

And to answer the question about voluntary reporting right now on the Hospital Compare using the ASC measures. And it's just a voluntary reporting. So we don't require hospitals to submit data. And then it's really hospital has an agreement with a registry whether they are participators. CMS, like I said, it's a feasibility issue. We like to see

this data but we don't really require hospitals to submit the data to CMS. Thank you.

CO-CHAIR GUNNAR: Dr. Levy?

MEMBER LEVY: This is just a really practical question but this has all been specified with ICD-9. Have you tested it with ICD-10? Just from a practical standpoint and our being able to use this measure over the next several years. I can foresee some administrative issues and reliability and validity issues with this measure.

MS. SUTER: This is Lisa Suter with Yale. The measure has been crosswalked to ICD-10. We don't actually have ICD-10 data in which to test it at this point.

The crosswalked ICD-10 as you can imagine is a challenge for this kind of a measure and involved both using a standardized metric or software tool that crosswalks and then had extensive physician review in order to ensure that the crosswalk was clinically

coherent. And when data becomes available I'm sure there will be a plan for actually testing it in ICD-10 data.

MEMBER LEVY: My real concern there is ICD-10 PCS as I understand it is not as granular as the ICD-9 for procedure codes. And that even though you did the crosswalks, how hospitals will do the crosswalks and report the data may be problematic.

MS. SUTER: It's certainly a concern that will need to be addressed when ICD-10 is implemented and there's actual data to assess the measures in ICD-10 dat.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: Collette Pitzen. I just wanted to comment on a subject that was a little bit ago. But it has to do with the data being specified and used in the Medicare population.

From my developer's hat if you have really good specifications there's no reason

why you couldn't take these specs and apply them to an all-payer state database, or have that applicability. So I don't think that should be a negative in our consideration.

MS. SUTER: Sorry, this is Lisa Suter from Yale. Just to follow up, this has been validated in the California State all-payer data set. So, it does perform similarly in an all-payer data set.

CO-CHAIR GUNNAR: Dr. Markman.

MEMBER MARKMAN: Using claims data versus a registry you can bullet your claims data down to the individual physician if he is an outlier.

MS. SUTER: This is Lisa Suter from Yale. This measure was developed at the level of the hospital and was tested solely at the level of the hospital. The measure development contract was specified as such.

We have not tested it at the physician level. As had been earlier discussed there are

concerns about sample size in terms of being able to estimate at the surgeon level.

MEMBER MARKMAN: You see, but in the big picture when you compare registries and with the comment before on how it's difficult to get the individual surgeon to comply using claims data you can, I mean you can -- the advantage of using claims data is that you can really narrow it down to outliers in the physician population. Because you need to.

And I think that that's one of the big differences between a registry versus using for CMS.

MS. DRYE: Sorry, I just wanted to add onto that point. And I think it's true for both STS. I know it's true for the outcomes measures CMS has reported for hospitals.

Hospitals do get patient-level data which identify the specific surgeon and the risk factors and outcome for every patient confidentiality. And our hope is that if there

are outlier surgeons we know that are causing an unacceptably high mortality rate that there are discussions that go on in that hospital.

So it's not transparent to the public but it is to the provider group and the hospital.

CO-CHAIR GUNNAR: Ms. Moyer?

MEMBER MOYER: We've talked about the reduced sample size with this, but what we haven't talked about, specifically salient to usability and use, is there's no opt-out of the public reporting.

When, you know, and I applaud STS for the voluntary reporting and all of the hospitals that share their data, but it's a voluntary reporting publicly.

And as someone who purchases in a state where previous voluntary efforts, we have had a geographic area of hospitals send us a letter and say you don't need to know how good we are. We're just, we're not going to participate.

There's an advantage to this measure that takes that off the table.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: My one comment was is that having looked at the release of the individual physician payment data that was put out in 2012 and looking at the local distribution of attribution of different surgeries to different surgeons it wasn't clear that the billing data which may not be the same as the data attributed to individual surgeons that you're using, but it wasn't clear that it was accurate.

We had one partner that had -- one member of the group that had been very, very busy but he was accounted for absolutely no surgeries. Yet somebody else was attributed accurately. I personally was attributed for just my office work.

So, it may be because of common billing or common tax IDs being used. And I think

that's changing with individual tax IDs being more required. But there is a question as to how you capture the individual surgeon through billing data because there has been in the past common tax IDs.

DR. BURSTIN: Just two cautions before you vote on usability.

First, the discussion of the individual physician is really important and very important as we think about the future. It's not applicable to this measure. This is just a facility-level measure. It shouldn't have some of the issues that were just raised.

And secondly, this really is about the usefulness of this measure. As tempting as it is to do the comparison that is a secondary discussion after the decision around endorsement -- recommendation for endorsement of both measures.

So I don't want you to weigh in the comparison and contrast. You're really

looking at this measure and its usefulness, usability and all the other criteria individually.

You'll then have a chance to talk about competing. But I just want to keep it clean.

CO-CHAIR GUNNAR: But the argument to that is if I have -- we know we have two measures. One is 18 months behind the other. Even if they were equal, even if we said they were equal, from an NQF perspective and for the assistance of this committee if you had two exact measures but one was 18 months behind, data was 18 months behind from a usability point of view that has an impact.

DR. BURSTIN: Right and so what -- those are perfectly appropriate for you to rate the usability of this measure. But I wouldn't do it as a comparison to the other one. You'll have an opportunity to have a discussion of the two measures side-by-side.

But certainly fair game to talk about

anything related to this measure in your vote on usability for this measure. But not in contrast I think to the other one, just to be fair.

CO-CHAIR GUNNAR: Dr. Handy?

MEMBER HANDY: Well, could you expound on that a little bit, Helen? Because we really haven't had the harmonization discussion yet and it's not even part of the script actually. So, I'd like to hear a little bit more of the mechanics. Because it does seem unfair that we're sort of simultaneously harmonizing. It seems unfair to this measure.

MS. WINKLER: Let me just add this. This is the reason we're going through them independently. Each one stands on its own, it gets rated as an individual measure.

Once we know they both meet the criteria then we can have that. If one of them doesn't we don't need to have that conversation. So that's why we're doing them

first.

So that's exactly what we're going to do. So you need to get this one evaluated before we can go onto the next step of comparing them.

CO-CHAIR GUNNAR: Let's vote. Dr. Siperstein.

MEMBER SIPERSTEIN: I'd just like the developers to address the concern on the usability of this particular isolated measure.

If it's to assess the quality of the facility and you're only addressing a subset of the patient population doesn't that potentially negatively impact usability?

MS. SUTER: So this is Lisa Suter from Yale CORE.

Similar to the impact that existing Medicare claims-based mortality and readmission measures have had on readmission and mortality rates nationally and we are seeing particularly for AMI as well as for

pneumonia and other conditions that public reporting of these measures among the Medicare population is influencing national rates.

It is likely that hospitals are not individually separating their quality improvement efforts out for the Medicare population, but they are responding to performance estimates across their hospital populations.

And it is likely that surgical care provided to Medicare populations particularly for CABG which is much more common in that age group are certainly a signal of quality of the hospital.

CO-CHAIR GUNNAR: For myself I appreciate the fact that there's an attempt to risk-adjust what could be just publicly reported unadjusted mortality rates. So, plain statement.

DR. BURSTIN: Also a couple of weeks ago the readmission committee reviewed the

claims-based CABG readmission measure as the same population. So there's also I think an issue of the harmonization to the related readmission measure I suspect as well.

MS. SUTER: And in terms of usability of this particular measure the intention is that it is appropriate to measure two domains of care across the same cohort.

And given that there is a claims-based readmission measure we think that it's important when your hospitals are focusing on things like reducing readmissions that unintended consequences such as short-term mortality are simultaneously measured to ensure that we are not influencing the wrong kinds of quality improvement efforts in our work to reduce readmissions.

CO-CHAIR GUNNAR: Any further discussion? Dr. Han.

MS. HAN: Hi, this is Lein Han from CMS.

I want to say that we implement the measure we also monitor the impact. And from what we saw from the analysis they have provided us and by the -- by others too we saw the reduction or decrease in the readmission and also improved the quality of care for the mortality.

Because we have a published paper on how AMI mortality, distribution of the AMI mortality I think narrowed and also shifted, the bell curve shifted to a lower mean.

So, I think in the national level because of the way CMS implemented national level. And we do see the impact and stimulated the quality improvement. I think this is very encouraging. Thank you.

CO-CHAIR GUNNAR: Any other comments? Let's vote. Usability and use.

MR. SANCHEZ: Voting will now begin for criteria 4 usability and use. One is for high, two is for moderate, three is for low,

four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: So we need one more, is that right?

MR. SANCHEZ: Yes, I think we need one more.

CO-CHAIR GUNNAR: Can everybody do it one more time?

MR. SANCHEZ: We have 8 for high, 12 for moderate, 3 for low, zero for insufficient information.

CO-CHAIR GUNNAR: So a vote for does the measure meet NQF criteria for endorsement. Any discussion before we vote? Go ahead.

MR. SANCHEZ: Voting will now begin for overall suitability for endorsement. One is for yes, two is for no. Timer starts now.

We have 22 for yes, 1 for no.

CO-CHAIR FLEISHER: Great. We're going to move onto measure 0264. As we continue we're going to have to balance a richness of the discussion with allowing all of

these to be discussed since there are people in the room who have come to participate. Prophylactic intravenous antibiotic timing. Do we have -- is it Donna on the phone?

MS. SLOSBURG: Yes, Donna is on the phone.

CO-CHAIR FLEISHER: Great. And who's the reviewer? Fred, you can actually participate. It's nice to have you back.

Donna, do you want to start with giving us a brief overview of the measure?

MS. SLOSBURG: Sure. I'm Donna Slosburg. I apologize I couldn't be there in person. I'm the executive director of the Ambulatory Surgery Center Quality Collaboration.

This measure was 0264. It's ambulatory surgery center admissions with an order for a prophylactic IV antibiotic for prevention of surgical site infection who received a prophylactic antibiotic on time.

It was endorsed in 2007 and went through maintenance in 2012. I was listening in yesterday during all the discussion about the antibiotic measures, but unlike those other measures this measure has only been in use in the CMS Ambulatory Surgery Center Quality Reporting Program since October of 2012.

CMS did present the ambulatory surgery data regarding this measure on April 23 of this year. And timely use for calendar year 2013 was at about 96 percent.

Again, this is a little different than the other measures. We are only in our second full year of reporting. And I'd be happy to answer any questions.

CO-CHAIR FLEISHER: Okay, Fred, do you want to take us through evidence?

MEMBER GROVER: I'd just like to perhaps ask the developer here a little bit on the level of analysis before we do that if I might.

In terms of -- it's analyzed at the individual ambulatory surgery center level for 349 centers, but the remainder of the total of the 671 centers are centers that report to a corporate system level.

Not having the information on those it would seem to me doesn't allow you to truly see the trends over time between sites. Could you maybe clarify that? Is that an issue? Or explain that to us?

MS. SLOSBURG: I have someone else on the phone as well. Kim, I don't know if Kim wants to take this question.

MS. WOOD: You are correct that when we initially developed this measure we had to rely on volunteers and we were able to recruit corporate volunteers and individual volunteers.

When you look at this measure, however, currently as it's being used in the ASC Quality Reporting Program that CMS has

developed these results are available to the centers at the individual center level as well as how they stand in comparison to the overall rate.

CO-CHAIR FLEISHER: So Fred, is the evidence different from the other antibiotic measures for this one?

MEMBER GROVER: Things get better. Yes, the evidence, I mean obviously this is a process measure. It's a timeliness the same being within one hour. They review a prospective study. Six were large observational, two were small and there were three randomized trials.

And their overall assessment was that there was good evidence. And I guess I would rate the overall evidence based on my assessment as moderate.

CO-CHAIR FLEISHER: Great. Can we -- any comments? Yes, Barbara.

MEMBER LEVY: I just want to comment

that those papers are all in inpatient settings. And I'm not sure that we're clear on antibiotic prophylaxis in the outpatient setting. So, I agree with you, moderate at best. But in fact those data don't apply to this patient population in this setting.

MEMBER GROVER: Yes, and maybe again you might want to respond to that. But my take on that would be that's what they would say is one of the unique parts of this they want to see if those other measures apply to this.

MEMBER LEVY: Then they need to look at outcomes.

MEMBER GROVER: Yes.

CO-CHAIR FLEISHER: So, are you questioning whether the link to outcome is sufficient enough and therefore whether it even deserves a moderate in the outpatient setting because it's a different population.

MEMBER LEVY: I am, I just don't know that we have the data to provide the link.

CO-CHAIR FLEISHER: Okay.

MEMBER CIMA: The question is is that this is an order for antibiotics but it doesn't question -- the real question is is it an appropriate order for antibiotics. So, anybody can write an order.

The measure says if you wrote the order we're going to give it within an hour. And it has a whole long list of antibiotics that probably shouldn't even be used in an outpatient setting.

So the real question is is it appropriateness. And there's no data to support in an outpatient setting like this in these ACSs whether or not in many of these cases whether you even need an antibiotic. So I agree, I'm not sure there's even evidence base to support this.

CO-CHAIR FLEISHER: So, can we get -- Donna, can you or your colleague comment on -- I assume this is a single measure of timing and

not timing and choice. And whether or not there is evidence in the outpatient setting.

MS. WOOD: This is Kim and I'd be happy to address that. The first question raised was whether there is evidence of impact on surgical site infection outcomes in the outpatient setting. The answer is no.

There is no literature that we're aware of that looks at this in the outpatient setting.

The second question was about whether or not we were looking at the appropriateness of the selection of the antibiotic and no, we're not.

However, the way that we've designed this measure is such that we have attempted to make it, let me think of the right word, compatible perhaps with the physician measure.

And that's why you see the comparable list of antibiotics that ties in with the measure that the AMA PCPI has developed

regarding the selection of the antibiotic.

CO-CHAIR FLEISHER: That's not in the specifications currently, correct?

MS. WOOD: Yes.

CO-CHAIR FLEISHER: Okay. Rick?

MEMBER DUTTON: One of the things that might help clarify this is understanding what specific procedures this applies to. Since likely a lot of the literature does apply to the cases that are now being done in outpatient centers, things like laparoscopic cholecystectomies, a lot of GYN surgery, arthroscopies that were all inpatient back in the day when this literature was written about the value of prophylactic antibiotics.

So the problem here is understanding which cases this measure is intended to apply to. Because I think there is probably good evidence and it probably is appropriate for some of the cases being done in the outpatient center but not all.

CO-CHAIR FLEISHER: But can you comment on the denominator inclusions?

MS. WOOD: So, the denominator is designed to strictly assess whether the facility administers antibiotics that are ordered for prophylaxis on time.

And again, since there is a measure that looks at the appropriateness of the physician's orders that can be used to assess the physician-level use these two can go hand in hand. But --

CO-CHAIR FLEISHER: My question is --

MS. WOOD: -- we're just looking -- sorry?

CO-CHAIR FLEISHER: Do you have any exclusions or is it all procedures? Is the simple question.

MS. WOOD: No, it's all procedures for which --

CO-CHAIR FLEISHER: Thank you. Kelsey?

MEMBER MCCARTY: Yes, just to follow up on Dr. Cima's comment. I'm wondering why it's limited to just IV antibiotics and not administration of all antibiotics. Because you're also getting a sample bias there.

CO-CHAIR FLEISHER: Any answer from the developer? I don't know if you --

MS. WOOD: It is written that way in order to be as harmonized as possible with the related physician measure.

CO-CHAIR FLEISHER: Collette?

MEMBER PITZEN: Just a measure design comment or observation. If it's based only on those patients that have an order for that there is some selection bias.

Perhaps if it was procedure-based then you'd be catching patients perhaps that should be receiving that antibiotic that are not.

CO-CHAIR FLEISHER: Thank you. I think, Fred, any comments before we vote?

MEMBER GROVER: Those are all good

comments. And it's obvious that this is just to see about the timing of the antibiotics when the order's been written. It doesn't address whether the order should have been written or not written. So I think we're probably ready to vote on the evidence.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for high, two is for moderate, three is for low, four is for insufficient evidence with exception, five insufficient evidence. Timer starts now.

We have 1 for high, 7 for moderate, 12 for low, 1 for insufficient evidence with exception, 2 for insufficient evidence.

CO-CHAIR FLEISHER: Okay, so this measure fails on evidence and therefore fails. Is that correct? We are moving onto the next measure. Thank you.

We need STS back. Who's doing 0126?

CO-CHAIR GUNNAR: Okay. So this is 0126 Selection of Antibiotic Prophylaxis for

Cardiac Surgery Patients, Society of Thoracic Surgery. Dr. Jacobs?

DR. JACOBS: Well, good morning, again. Jeff Jacobs. This is again Selection of Antibiotic Prophylaxis for Cardiac Surgery Patients. It's a measure that reports the percentage of patients 18 years or older undergoing cardiac surgery who had an order for or received preoperative prophylactic antibiotics recommended for the operation.

The evidence base comes from a substantial body of literature about the value of the appropriate use of prophylactic antibiotics in prevention of infection and mediastinitis.

And I think that this is a previously endorsed measure that's coming up for renewal. And I think probably the evidence base has been discussed here extensively previously so I think I could just answer any questions.

MEMBER DUTTON: I was the primary

reviewer of the measure. The numerator is the antibiotic order is written. STS also assesses whether it was actually given. The denominator is all cardiac surgery patients over 18. I am not sure why the exclusion to adults only other than to note that there's little evidence in pediatric populations. The evidence is -- Jeff?

DR. JACOBS: Well, there's two reasons. The literature that supports this particular measure was developed based on patients over the age of 18.

There is a body of literature related to the use of prophylactic antibiotics for pediatric cardiac surgery, cardiac surgery in those less than 18. But it's not as substantial and the evidence base is not as solid.

And on top of that there's substantial variability in practice across the country that has been published and documented in a number

of published surveys about pediatric antibiotics. So consequently this measure is focusing on adults where there's a more solid evidence base.

In addition to that the pediatric population undergoing cardiac surgery is in a different component of the STS database. So this is an STS adult cardiac surgery database measure. And the congenital database has the patients that are less than 18.

CO-CHAIR GUNNAR: But just to be clear, adult congenital gets included in the adult population?

DR. JACOBS: That's a real question that we could spend the next hour and a half talking about. But the brief answer is that an adult congenital patient can end up in either database depending on what the predominant component of the operation is. And I'll give a couple of examples.

If a patient with a previous repair of

tetralogy of Fallot comes in and gets a coronary artery bypass graft that's going to end up in the adult cardiac database.

If a patient with a functionally interventricular heart that's had an atrial pulmonary connection comes into the hospital for a Fontan revision to a more hemodynamically favorable Fontan circuit at the age of 35 that would end up in the congenital database. So, it really depends on what the most predominant component to the operation is.

CO-CHAIR GUNNAR: But with regard to this measure would the latter be included or excluded from?

DR. JACOBS: This measure applies to all patients entered into the adult cardiac database.

CO-CHAIR GUNNAR: Only the adult. There you go.

MEMBER DUTTON: As far as the evidence goes I think the evidence is very strong

connecting this process to prevention of particularly mediastinitis.

CO-CHAIR GUNNAR: Any other discussion? We'll go ahead and vote.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for high, two is for moderate, three is for low, four is for insufficient evidence with exception, five is for insufficient evidence. Timer starts now.

MEMBER GROVER: Just for the record I'll be abstaining from all the STS measure votes.

MR. SANCHEZ: We have 22 for high, 1 for moderate, zero for low, zero for insufficient evidence with exception and zero for insufficient evidence.

MEMBER DUTTON: Regarding the opportunity for improvement the current performance on this measure is median of 99.2 percent with 2.8 percent of providers identified as low performers.

This would meet our criteria from yesterday for being topped out. My opinion is it is an important measure. It should continue to be collected.

The cost to collect this in the registry setting is essentially zero as this one is as we were discussing earlier hardwired into the system and routinely collected.

And I think it is reasonable to continue to collect and report this particular measure, particularly when mortality from cardiac surgery is also at a very low level. That was 98 percent earlier and that measure was accepted by this group.

DR. JACOBS: And I would just add to that, first of all, I agree with everything that you said. And the other piece to consider is that the mortality associated with a postoperative infection after cardiac surgery as well as the morbidity of a postoperative infection after cardiac surgery is

substantial, perhaps worse than almost any other postoperative infection.

Postoperative mediastinitis is a big deal. People die from it. Those that survive are often in the hospital for months and months requiring multiple operations.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: Since the inception of this particular measure there's been greater utilization of MRSA screening in some surgical subspecialties.

And my question to you, does the STS database or registry allow for the recording of the actual antibiotic used as opposed to being checked off as appropriate. And does it also allow for, as an exclusion criteria for the use of a first or second generation cephalosporin, the fact that the patient may be MRSA-positive?

And I ask this in the sense that a performance gap could be discerned over time in terms of mediastinitis rates based on whether

or not, for instance, clindamycin or vancomycin are used versus whether or not people are screened for MRSA.

So, given the fact that those are potential end uses which are not part of this measure which is looking at whether or not appropriate antibiotics have been given is the granularity within the system so that that could be an outcome that would justify continuing to make this an important measure?

DR. JACOBS: So, currently no, but that's certainly an excellent idea for an area of future investigation within the database. I think that's a great idea.

Currently we track whether or not the appropriate antibiotics are given and there's a whole list of criteria on what's appropriate and what's not based on the evidence base. But I think what you're suggesting is a great area of future investigation.

CO-CHAIR GUNNAR: So, the question,

Dr. Jacobs, if this process measure was placed on reserve is it from STS's perspective that mediastinitis or the rate of mediastinitis would really be the end -- it's the end goal, correct? That's the quality goal.

So the question is if this fails or is placed on reserve based on the performance gap which is insignificant at this point my question is isn't -- you still stand on the NQF-endorsed measure of outcome of mediastinitis, correct?

DR. JACOBS: Well, absolutely, we have our NQF-endorsed measure of mediastinitis which is also a component of the NQF-endorsed composite score. So that's all important.

The only thing I would add is that clearly the major endpoint here is prevention of postoperative mediastinitis.

But there's also some subtle endpoints regarding use of appropriate antibiotics and prevention of overuse of inappropriate

antibiotics that could lead to development of multiple resistant strains of organisms within hospitals.

So, by choosing the appropriate antibiotic there's more endpoints than just preventing mediastinitis. There's also preventing using the wrong antibiotics which could then lead to a variety of other problems within the hospital.

MR. SHAHIAN: This is Dave Shahian. The only other point is that the endpoint of major interest which is internal infection of mediastinitis occurs with an average incidence of about 0.3 percent.

So it's a devastating but exceptionally rare outcome and it is very hard to distinguish levels of performance based on something that occurs that infrequently. So I think it's a reason for considering a process measure in this case.

MEMBER DUTTON: Yes, I would echo

that. And one of the reasons I think we should reconsider our stance on some of the topped-out process measures is that the processes are measurable and improvable above 98 percent whereas the actual outcome which is, as you just heard, 3 per 10,000 is not measurable.

MEMBER CIMA: It's the same case though for all other patients having surgery. Or did we decide yesterday something different?

DR. JACOBS: I think the only differentiator here is the potential severity of the postoperative infection for cardiac surgery is probably an order of magnitude worse than most postoperative infections. So that makes it a little bit bigger of a deal.

And you know, regarding the concept of topped out, I would just add that mediastinitis is exceedingly rare, less than 1 percent, maybe 0.3 percent.

But this measure allows evaluation of quality of care to the remaining 99.7 percent

who don't get mediastinitis. Their quality of care is evaluated by assuring that they get the right antibiotics.

And it really does effect a change in the behavior of the operating room team. During the timeout, one has a timeout to make sure the right antibiotic was given at the right time. And that, a lot of it is because of this measure. And I'd sure hate to see that go away.

MS. WINKLER: I just want to make one comment. You're talking about a lot of things that are outside this current criteria which is the gap that may be appropriate in some of the other criteria. So I'd really ask you to focus in. Right now we're talking about performance gap.

And I would also caution you that the criteria should be applied equitably and equally along all the measures. So, you know, the question about gap is the same question for each measure. There may be some of these

issues you're raising apply in other criteria that would be appropriate.

CO-CHAIR GUNNAR: Dr. Temple?

MEMBER TEMPLE: I have a gap question. The current measure includes received and ordered. And what I'm wondering is in your data set are you seeing a gap -- if we separate the received and ordered are you seeing a gap in the number of patients receiving the antibiotic? Because really that's the real gap that may be a potential place for improvement.

DR. JACOBS: I think that's a good question. We track both. But I don't know that we've looked at the gap between received and ordered.

MEMBER TEMPLE: Because I think that if we're looking -- if there's any gap that would be the gap that would be important for us to consider in looking and evaluating this.

CO-CHAIR GUNNAR: Ms. McCarty?

MEMBER MCCARTY: I'm wondering if you know of the estimated 0.3 percent of patients who do end up with mediastinitis, what percentage of those failed this measure of getting the right antibiotic on time.

DR. JACOBS: That's another good question. I don't know that we've looked at that.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: I don't know if STS can answer this, but just as a point of clarification. And this is an impact as opposed to a gap question but it ties into the conversation.

It's not so much the current rate of mediastinitis as reported. It would be the rate of mediastinitis that you would have without antibiotics.

The corollary is in total hip replacement it's a 5 percent incidence of infection without antibiotics and 1 percent

with. So the real question is what would be the consequence for the population of patients as was already asked that didn't get it and what's their risk.

And in addition to mediastinitis there's a certain amount of morbidity and a cost that's not inconsequential from graft harvest sites which have a higher infection rate and are certainly probably affected by antibiotic administration.

DR. JACOBS: Sure. There's graft harvest site infections. There's infections of prostheses like prosthetic valve infections and endocarditis. The big one is the mediastinitis but there are other postoperative infections, I would agree with that.

CO-CHAIR GUNNAR: All right. Collette?

MEMBER PITZEN: I'm sorry, I'm going to sound like a broken record but I just want

to beg the question. We're asked to evaluate the rate in the data that we're seeing for this measure as it's specified, and at 99.2 percent where is there to go from there?

DR. JACOBS: I'd answer that with the same answer that other panel members provided yesterday. 99.2 percent seems pretty high but it would be totally unacceptable in the airline industry and I think it should be totally unacceptable in the cardiac surgery operating room.

CO-CHAIR GUNNAR: Noted. All right. We'll go onto vote. Oh, I'm sorry. Ms. Moyer.

MEMBER MOYER: I just wanted to go beyond just the 99.2 percent. We're at 100 percent performance on this measure by the 30th percentile. So, it's not just the overall measure performance, it's looking at the variation within those percentiles.

You have to go to the bottom decile to get below 98 percent. And I know that's not

acceptable for the airplane industry although I'd get on a plane -- we're all getting on planes tonight but we're not all going to have CABGs tonight. It's a different risk between the two.

DR. JACOBS: Sure, but patients are going to that bottom decile to have their CABG. And there are patients that actually have their CABG there.

MEMBER MOYER: That's true.

DR. JACOBS: And I think that emphasizes the importance of this.

MEMBER MOYER: And I think it's important from a QI perspective. I absolutely think it should still be available to those hospitals.

But I mean, from a gap perspective I'm just not really seeing, you know, if we want to be consistent with what we did yesterday I'm not seeing it.

CO-CHAIR GUNNAR: That is the

instruction. So, shall we vote?

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

We have 1 for high, 5 for moderate, 16 for low, 1 for insufficient.

CO-CHAIR GUNNAR: So, at this point we then ask is the committee by a raise of hands willing to consider this as a reserve measure? Yes. All yeses in the air. How many is that?

(A show of hands)

MEMBER MCCARTY: In the same way we evaluate measures to determine if we should have more than one of the competing measures, if we move measures into reserve status do we do the same thing for those?

MS. WINKLER: Yes.

MEMBER MCCARTY: Okay.

CO-CHAIR GUNNAR: So, we have 20 yeses and so we'll carry on.

MEMBER DUTTON: As far as the reliability of the measure goes this is a measure renewal. The mechanism for collecting the data has not changed in decades. So it is -- it remains reliable and --

MR. SANCHEZ: Sorry, Dr. Dutton, high priority first.

MEMBER DUTTON: Oh, sorry, priority. This is important.

(Laughter)

CO-CHAIR GUNNAR: Any further discussion? I think that discussion actually we've had so we can carry on.

MR. SANCHEZ: Voting will now begin for 1c high priority. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: Dr. Ko has stepped out so there will only be -- and Dr. Grover has recused himself. I think that's correct at 21.

MR. SANCHEZ: We have 17 for high, 3

for moderate, zero for low, 1 for insufficient.

MEMBER DUTTON: Sorry, I now recommend that this measure is reliable.

CO-CHAIR GUNNAR: And any further -- we've had that discussion numerous times. Shall we vote?

MR. SANCHEZ: Voting will now begin for 2a reliability. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 17 for high, 4 for moderate, zero for low, zero for insufficient.

CO-CHAIR GUNNAR: We've discussed the validity of the STS data at great length and I believe it is highly valid.

CO-CHAIR GUNNAR: Go ahead and vote.

MR. SANCHEZ: Voting will now begin for 2b validity. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 21 for high, 1 for moderate,

zero for low and zero for insufficient.

MEMBER DUTTON: The feasibility of STS data collection has also been discussed. This requires a substantial investment by the institution in doing it but in this high-profile area this has been justified.

And I think indicated by the fact that that 90 to 95 percent of all institutions do participate in this. I think it's been adjudged to be usable for the institution.

Given that mechanism this data is very feasible to collect and as I say pretty much hardwired. This will eventually transmit directly from the electronic records.

CO-CHAIR GUNNAR: Any further discussion? We'll go ahead and vote.

MR. SANCHEZ: Voting will now begin for criteria 3 feasibility. One for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 14 for high, 9 for moderate,

zero for low, zero for insufficient.

MEMBER DUTTON: Regarding usability I think the decade-long history of this measure with substantial improvement in performance over that time indicates that it is usable to drive improvements in performance at the local facility and practice level.

I have one additional thought for STS which is it might perhaps be appropriate at some point to create a composite of process measures as well as the composite of outcomes. In other words, what would be a bundle of good performance as a process measure that may be more discriminating than the individual measures themselves.

DR. JACOBS: I think that's an excellent idea.

CO-CHAIR GUNNAR: Any other discussion? Hearing none let's vote on usability and use.

MR. SANCHEZ: Voting will now begin

for criteria 4 usability and use. One is for high, two is for moderate, three is for low, four is for insufficient information. Timer starts now.

We have 14 for high, 5 for moderate, 3 for low, 1 for insufficient information.

CO-CHAIR GUNNAR: And so the overall suitability for reserve status. Any further discussion? Let's go ahead and vote on that.

MR. SANCHEZ: Voting will now begin for potential for reserve status. One is for yes, two is for no. Timer starts now.

We have 22 for yes, 1 for no.

CO-CHAIR GUNNAR: So the recommendation of the committee is that this measure be placed in reserve status.

CO-CHAIR FLEISHER: Okay, we're going to keep you here. We're going to keep you here. We have five minutes left before the break to get through 0128. Who's the discussant? Okay. So, similar measure. Jeff, it's all

yours.

DR. JACOBS: This is duration of antibiotic prophylaxis for cardiac surgery patients. It's a partner measure to the one we previously discussed, percent of patients over the age of 18 undergoing cardiac surgery where prophylactic antibiotics were ordered to be discontinued or were discontinued within 48 hours after cardiac surgery end time.

CO-CHAIR FLEISHER: Evidence.

MEMBER REEDE: Thank you. So the evidence for this measure is similar to all that we have looked at from the STS database. I believe it would be rated moderate to high. If we want to discuss more I can.

CO-CHAIR FLEISHER: Comments? Questions? Okay, let's rate the evidence.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for high, two is for moderate, three is for low, four is for insufficient evidence. The timer starts now.

CO-CHAIR GUNNAR: Can you try to vote one more time?

MR. SANCHEZ: We have 17 for high, 5 for moderate, zero for low, zero for insufficient evidence.

MEMBER REEDE: As is indicative of all the successful measures this too is probably topped out with a minimal gap in performance.

CO-CHAIR FLEISHER: Any comment, Jeff?

DR. JACOBS: No, same comment as before. The performance gap is minimal but the stakes are high because postoperative mediastinitis is a bad infection.

CO-CHAIR FLEISHER: Okay, vote.

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

We have 1 for high, 4 for moderate, 17 for low, 1 for insufficient.

CO-CHAIR FLEISHER: Okay, as is our tradition now how many would like to put this on reserve status as opposed to -- okay. I'm going to short-circuit the discussion and ask is there anything different that anyone would like to discuss regarding any of the other criteria for this measure? Any?

MEMBER MOYER: I just have one quick question. For the bottom decile of hospitals if I'm reading this correctly, there's two and they scored zero percent? That's not a data error? They're literally not hitting this measure ever?

MR. O'BRIEN: I don't know the details of those particular sites but I think it's possible that perhaps is a data error.

DR. JACOBS: It could be a data error. Could be low-volume sites that aren't doing a lot of CABGs and just haven't complied. Without going back and looking at the individual records it's impossible to know.

CO-CHAIR FLEISHER: Can you please take us through each vote?

MR. SANCHEZ: Voting will now begin for 1c high priority. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 17 for high, 6 for moderate, zero for low, zero for insufficient.

MR. SANCHEZ: Voting will now begin for 2a reliability. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 17 for high, 5 for moderate, 1 for low, zero for insufficient.

MR. SANCHEZ: Voting will now begin for 2b validity. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 17 for high, 5 for moderate, zero for low, zero for insufficient.

MR. SANCHEZ: Voting will now begin

for criteria 3 feasibility. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: We still need a couple of more votes.

MR. SANCHEZ: We have 14 for high, 9 for moderate, zero for low, zero for insufficient.

Voting will now begin for criteria 4 usability and use. One is for high, two is for moderate, three is for low, four is for insufficient information. Timer starts now.

We have 12 for high, 8 moderate, 3 low, zero insufficient information.

Voting will now begin for potential for reserve status. One is for yes, two is for no. Timer starts now.

We have 22 for yes, 1 for no.

CO-CHAIR FLEISHER: On that note we are only four minutes behind with one measure behind. So we keep it at 10:45 to restart so

that we can continue to stay on time. Thank you.

(Whereupon, the foregoing matter went off the record at 10:32 a.m. and went back on the record at 10:46 a.m.)

CO-CHAIR GUNNAR: So the expertise measure is 0131 Risk-adjusted Stroke/Cerebrovascular Accident, STS. Dr. Jacobs.

DR. JACOBS: Hi, good morning again. Jeff Jacobs again. This is measure 0131 Risk-adjusted Stroke/Cerebrovascular Accident. It reports the percentage of patients over the age of 18 undergoing isolated coronary artery bypass grafting who have a postoperative stroke, i.e., any confirmed neurologic deficit of abrupt onset caused by a disturbance of blood flow to the brain that did not resolve within 24 hours.

This measure is being considered as an individual measure now but it's also part of our

composite measure for coronary artery bypass grafting as well.

It has a low percentage of occurrence but the consequences are devastating, very similar to postoperative mediastinitis.

The measure has been publicly reported in the composite. It's also utilized for a variety of quality improvement initiatives and in fact STS currently has a stroke workgroup within the STS task force of quality improvement that is trying to identify variables that are present in hospitals with very low stroke rates so that those variables can be shared across across the breadth of STS.

And I think that summarizes the use of the measure as well as its scientific basis. And I think I can answer any questions.

CO-CHAIR GUNNAR: Discussant?

MEMBER ASHER: So as mentioned this was -- this is an outcomes measure that was originally endorsed in 2007, most recently in

2011.

With respect to evidence several studies were referenced by STS. Specifically studies that document reductions in the rates of stroke post CABG with implementation of various preoperative strategies.

They also demonstrated that the implementation of these modalities is highly variable among groups performing CABG and also that the stroke rates are significantly variable.

The conclusion was many opportunities exist to decrease stroke rates by increasing implementation of these evidence-based strategies. And I think that there is sufficient information here to suggest a relationship between the measured outcome and a number of different healthcare processes.

CO-CHAIR GUNNAR: Any discussion? Vote on evidence.

MR. SANCHEZ: Voting will now begin

for 1a evidence. One is for yes, two is for no. Timer starts now.

CO-CHAIR GUNNAR: Let the record show Dr. Grover has recused himself.

MR. SANCHEZ: We have 23 yes, zero no.

MEMBER ASHER: With respect to performance gap it's already been mentioned that the stroke rate is low perhaps due to the implementation of this measure. But that does represent a somewhat limited opportunity for improvement.

In addition, and as we'll discuss in more detail in a few minutes in the empirical validity testing section almost all participants fall into the mid-performance category. And I mean like 99 percent participants. So in that regard this doesn't provide an opportunity for a lot of distinction between providers.

So we've had this conversation in the context of a number of these different measures

but this one in particular seems to have a lot of providers falling within one particular category which is that moderate range.

So I would rank this as low to moderate on the basis of the room that we have for further performance improvement.

CO-CHAIR GUNNAR: So, I need some clarification here because on a performance metric the act of doing something in relationship to, you know, the outcome.

We can easily see that if you're giving antibiotics 99 percent of the time that's one thing. That's different from evaluating the outcome itself. So the outcome here is low. The measurement of that outcome you would look at as does -- so the question is as part of this registry it is mandatory that you assess whether or not this outcome occurred. I don't think the two equate.

MS. WINKLER: I think it's an excellent question. I'd ask the same question

myself is, you know, when you're looking at important outcomes does it matter that everybody's doing well, or is that a particularly good thing?

And I think particularly in these low-incidence adverse outcomes we see a lot of this in patient safety. Don't you, Andrew? Andrew oversees our patient safety portfolio. And so this is not an unusual situation to have low incidence, and in fact that's highly desirable.

So I think you're right to ask the question that when it comes to desirable outcome measures can you -- the whole concept of topped out doesn't really I think have the same meaning as it does with a process measure.

So, I don't know if Helen wanted to weigh in on that either. The question about high levels of performance with outcome measures.

DR. BURSTIN: In general outcome

measures that are considered safety events, even low-volume, are still very appropriate to track just because they're important adverse events. So we don't worry as much about low events there.

MEMBER ASHER: Well, then is there an issue here about the way performance is being evaluated? Over 99 percent of people are falling into a particular performance category.

So, if we're looking at -- I mean, what is the gap we're looking for? If we're looking at a gap among providers there just isn't much of a gap among providers even in achieving this particular outcome. So, I'm just looking at a way of assessing this particular measure.

CO-CHAIR GUNNAR: Let me take Dr. Asher to the logical endpoint which is tomorrow there is no more postoperative stroke. Does the measure then become irrelevant?

DR. JACOBS: And I would say even if

there was no more postoperative stroke it's still a very relevant thing to keep track of because a stroke is a devastating event.

MEMBER ASHER: I think that it would only -- we've made some arguments about some things that are either intermediate outcomes or performance measures and talked about how various healthcare processes have just become part of what we do.

And so if everybody is doing those things then the question would be are those hardwired into the processes of taking care of patients with cardiothoracic surgical issues. And I don't know the answer to that.

But all I know is that most people are now -- we have low rates. Most people are falling into one category.

So I'm just trying to find a way for us to logically address this issue. Is there a performance gap.

MS. WINKLER: Just from sort of a

perspective on the criteria I think you've raised an important question around dealing with very important adverse outcomes that are very low incidence, thank you.

And so the issue around the gap I think becomes less of a criteria under these circumstances. But we haven't provided that guidance for you.

Andrew, I'm just wondering in terms of your patient safety issues around this where have they gone?

MR. LYZENGA: I mean, typically we haven't had much pushback on endorsement of low-incidence adverse outcomes.

I don't want to push us into this area but there has been discussion often about the application of those measures subsequently. Which is again maybe not in the purview of this committee. Some people feel they're not as suitable for, say, accountability purposes. But again, I don't want to get us into that

discussion necessarily. That's my impression of the previous discussions on this kind of issue.

CO-CHAIR GUNNAR: So if it's all right we'll start on that side of the room and then we'll work to this side of the room. So, Dr. Saigal.

MEMBER SAIGAL: I was just going to say that I think that a performance gap in terms of it being an outcome measure is still important to report because I think that consumers do look for these things. And even if it's a very low rate I think it's meaningful to people that are looking for surgeons and for hospitals.

MEMBER PITZEN: Collette Pitzen. As the naysayer on high-performing measures I do have different feelings about outcome measures. I do think that they do have their place. They may not be well suited for accountability in trying to differentiate

performance between practices but I think that there's value there. I just wanted to share that. Thank you.

MEMBER MOSS: The significance of this event justifies continued measurement and reporting.

But in the patient safety world there's some increased effectiveness for reporting very low-rate events as days between last event versus proportional percentage rate. There are some statistical advantages to that. It tends to make it more meaningful to the stakeholders.

MEMBER DUTTON: I'm kind of concerned by the double standard here, why an outcome measure is not topped out but a process measure would be.

If you think about it the process is actually a much more controllable thing. As a consumer I recognize, and in fact it's part of the informed consent, I could have a stroke as

part of my CABG.

But if I thought that my doctor wasn't going to give me the antibiotics which is something that's completely under his control I would be much more upset about that as a consumer of healthcare.

CO-CHAIR GUNNAR: Dr. Ko.

MEMBER CIMA: Just to go to the point though. Outcome measures are important but 99-plus percent of people have the same rate. So it's not really a good tool for distinguishing. It is part of composite measure which is probably a more realistic one to provide.

And it's not like this measurement is going to go away. It's going to be in the STS database. And the purpose of the STS database, it's you have individual practices look at their performance and say, boy, we're a high outlier in stroke maybe.

So the purpose of collecting this at

a national level and reporting it as an individual measure, I feel if it's 99 percent it's non-distinguishable. It is somewhat topped out in that sense.

And it's not like it's going to stop being measured.

DR. JACOBS: I would say that beyond the star rating again there's a percentage of stroke that's also reported which provides additional information. So it's not just the information one gets from one-star, two-star, or three-star, but the more granular data with the actual numbers that support that.

I would also say that at the present point in time every component of our composite ratings are NQF-endorsed measures. So our composites are all made up of all NQF-endorsed measures.

And if -- we've not had a situation where any of the domains of the composites are non-NQF endorsed measures up till now. In the

CABG composite there's 11 domains -- there's 11 measures that divide into 4 domains and those are all NQF-endorsed measures.

MEMBER KO: My question goes to exactly that. And maybe this is -- maybe we can get some guidance from the NQF.

It seems like everyone is going to move towards composites. So, do we need each of this component of the composite to be NQF-endorsed?

Because then we're suddenly, I mean each little part by itself might not mean as much as the whole together. So how do we do this?

DR. BURSTIN: Yes, that's actually a helpful question. So we updated our composite measure evaluation guidance about a year ago and in fact changed, and this is a change from the prior time. We looked at the STS composite that we no longer require the individual elements within a measure to be endorsed, or

actually really more so want to see how they -- we're actually asking for more of the evaluation at the composite level, how those come together.

And in fact, I think some of the issues around low gap here will come up in reliability and validity potentially is where we often see that some of these very low-volume events have a difficult time of being reliable on their own, although can be good when you combine them with other safety events.

I will say though that we did do a report three or four years ago specifically on the issue of low-volume safety events. And recognized -- and the point that was just given by Dr. Moss was exactly right, that there are different ways to display safety events that are still meaningful even if they are low volume. So changing the denominator. Changing the days since last central line infection.

And I'm actually curious, and Andrew may know this, but for example, in our evaluation of the AHRQ patient safety indicators just a couple of weeks ago I think the denominator was actually different because it's more, for example, I think it may even be per 1,000 discharges.

So there are different ways to display that data that still may be very informative to patients and purchasers. And I'd be curious to hear Amy's perspective in particular of even if it's low-volume is it enough to say it's 3 per 1,000 versus 6 per 1,000 as they're making some of those decisions.

So very good point. I'm happy to share around that paper, this work that was done several years ago specifically on the low-volume safety events because I think it's becoming more and more relevant right now.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: To the question as to

whether or not the outcomes at a low rate are equivalent to process measures, 99 percent of patients can expect to get a total hip or a total knee replacement without infection. But that 1 percent is something that we strive to be 0.5 percent. We strive to be 0.3 percent.

And I would argue that at this part of the composite that is STS, that being stroke, is an outcome that the stakeholder or patient might, depending on their value system, see as a fate as bad or as worse than death depending on how badly the stroke turns out. And they're different for the rest of their life.

And so if you have 10,000 coronary artery bypasses and you have a 1 percent rate, and you have 100 people that have a stroke, the goal would be to make it 20, or make it 10, or make it zero.

I would argue that outcomes of this sort are never really a never event, but the goal should be to hit never. And as such

carving down from 99 percent to 99.9 percent would be an important goal. And I think it's worth keeping this and keeping this as an important endpoint.

I'm sure that cardiac surgeons when they have someone with a stroke it's similar to looking for the black box after a plane crash. They go back with a fine-toothed comb trying to figure out where it went wrong. And I think it's important to have this.

DR. JACOBS: I would agree with that. And I would just add that the logic that we used earlier this morning to endorse risk-adjusted operative mortality really applies here. It's the same exact arguments as far as the percentage of occurrence of the event.

So if one were to be consistent with what we did this morning the same logic would apply here. It's subject to the same criticisms.

CO-CHAIR GUNNAR: Dr. Fleisher?

CO-CHAIR FLEISHER: Yes, I would agree with Dr. Yates and just say that you could also think of this as a patient-reported outcome. It's actually the patient is reporting a functional status change which leads -- but that's the difference between what some of the other outcomes say. It's not -- I'll leave it there.

DR. BURSTIN: Just one more comment. I was just checking in with Karen Pace who's our lead methodologist reminding me that again, performance gap says considerable variation or less than optimal performance.

So I guess the question for STS is if you think there's, you know, is there variation across providers. And that's an adequate reason as well for the performance gap. So, perhaps a question for you guys.

MR. O'BRIEN: I'll just weigh in from the statistician's perspective.

In terms of the variation across

providers I don't have the numbers in front of me. But what's being reported in the table there is the distribution of the estimates of performance over providers. Those are generated from that hierarchical model which has the property of the estimates in the presence of a moderate sample size. To the extent there's uncertainty about the estimates, the estimates are shrunken back towards the average. So it's substantially underestimating the true amount of spread between participants. If each participant had vast numbers of cases you'd actually see a much wider distribution.

And for a lot of the outcome measures we see comparing from the top of the distribution to the worst. When you estimate that true signal distribution it's fairly substantial.

I've heard that NQF staff mention in connection with other measures that I don't

think you're being asked to endorse a particular star rating threshold. And so although you're seeing very few one and three stars, I don't know if that's how they were labeled in NQF measure submission. That actually is a very, very conservative threshold that's being used.

In the NQF submission materials I think they were labeled as confidence intervals. They're actually -- the interval estimates that come out of the hierarchical models aren't confidence intervals in the sense that traditional conventional confidence intervals are. They're more like Bayesian-type intervals. And they're much more conservative.

And so we've looked at other methods of classifying provider performance by using less rigid, strict certainty criteria. So to the extent that you have wiggle room on that you can basically create more outliers.

So, if these were conventional confidence intervals, you know, basically if you had a 95 percent confidence interval and you used another version of this you can have 5 percent outliers just kind of -- even if there was no variation you'd expect to have just false positives rated as outliers.

So this is really controlling, minimizing the probability of falsely classifying as above or below average performance. But there's certainly by -- on conference calls and discussions we have considered other approaches and there's plenty of opportunity to have many, many more outliers than are shown in the submission material.

DR. JACOBS: I would just briefly add to that that a take-home message from what Sean just said is that there still is variation among providers with the postoperative outcome of stroke.

There's enough of a variation that STS

has invested substantial time and energy in creating a stroke workgroup that is looking at the high-performing providers to try to figure out what they're doing better than everybody else to figure out what can then be done to minimize stroke even more.

Admittedly it's a low-incidence complication but there's still enough variation that we've been able to identify high-performing providers and try to learn from them. And the effort is being made to cut this from 1 percent to 0.5 percent to less because it's so devastating.

CO-CHAIR GUNNAR: We'll work on this side and then we'll come back over here. Ms. Moyer.

MEMBER MOYER: We in general have a preference when we do our public reporting for measures that show variations. In this case we would be preferentially looking for the composite.

That said, there are instances where if we -- in the absence of that measure we'll report something like this that shows no variation just to let patients know it's something we've looked at. It's not we don't know anything, we just can't differentiate.

I absolutely think this should be part of that bigger composite. But I mean, from a patient perspective if they're looking at the individual parts of it, oh gosh, do I go to the person who's got the better stroke rate or the better deep sternal wound infection rate.

I mean, that's not something patients can really appropriately weigh. I think they'd be looking more for that overall harm, that overall morbidity.

And I think this is a very important aspect of that. I just, I struggle with it on its own as a useful tool for patients or accountability applications.

MEMBER SAIGAL: I just wanted to

comment to Sean that I don't think -- I mean, the interest in finding performance differences is a take-home that's great to get. But I don't think it's at the expense of sacrificing specificity of the measure. So I think it would be very counterproductive to find variation where there could be question if it really exists as a latent variable.

DR. JACOBS: I would just say that this is not a situation where there's no variation and that we'd be having a lot of errors. When you use a model to estimate the amount of true variation it's very clear that there's substantial variation. And the difficulty is in the ability to estimate that rate precisely. So there's going to be some tradeoff between false positives and false negatives.

But it's not basically where you're accepting just for the sake of having outliers. That you catch a lot of true differences that

we're currently missing by using a very strict criterion.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: Collette Pitzen. Just to comment for future. You've done a ton of statistics in the section where we're looking for reliability testing on the performance score. That information wasn't provided. And I'm sure that you probably have information about that. So for future submissions that might be helpful in helping us determine if there is differentiation between provider groups.

CO-CHAIR GUNNAR: Dr. Asher?

MEMBER ASHER: This has been a helpful conversation for me. I would just ask, going back to the process.

So it seems to me where we've been going with this is that, first of all, we can argue about the amount of variation. And I understand that basically depending on the

statistical models there may be more or less variation that's represented in some of these things, particularly with the star ratings.

It seems like we also -- we need to re-calibrate our ideas of what substantial variation are, particularly in these low-incidence but high potential morbidity areas. And that might -- I don't know that we need formal guidance on that, but it may be almost like an asterisk you'd have next to like 1b suggests that we as a group need to be looking at these areas slightly differently.

I'm just bringing that up because again the numbers are low but maybe we have to think about it in a different way.

CO-CHAIR GUNNAR: Any other discussion? So, if I capture a bit of this just before we vote. We have already voted on an outcome measure which is mortality which shows a low rate of distinguishing data between centers.

And we've had a complete discussion as to the fact that if this was placed in reserve status or eliminated it still would be voted as -- it's already part of a composite measure. But that's not what we're voting on at this point.

So, knowing what we know with regard to performance gap I think we vote.

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 7 for high, 10 for moderate, 4 for low, 1 for insufficient.

CO-CHAIR GUNNAR: Dr. Asher?

MEMBER ASHER: I'd argue this is a high-priority area.

DR. JACOBS: I agree.

CO-CHAIR GUNNAR: Any further discussion? We can vote.

MR. SANCHEZ: Voting will now begin

for 1c high priority. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

We have 19 for high, 3 for moderate, zero for low, zero for insufficient.

MEMBER ASHER: So with respect to reliability the numerator statement is the number of isolated CABG procedures in which postoperative stroke is marked as yes. Denominator is all patients undergoing isolated CABG. There are no exclusions. We all know what the data source is. I have no particular issues with specifications, definitions, or coding.

With respect to reliability testing, this wasn't in my mind well separated out in the measure information. However, I believe what was meant to represent this particular testing, and please correct me if I'm wrong, but was this audit a process involving re-abstraction of data for 20 cases, comparison of 72 individual

data elements with those submitted to the data warehouse.

And there was substantial agreement between those data sets which basically demonstrates that the data contained in the database is both comprehensive and accurate. So is that what was supposed to be represented for the reliability testing?

DR. JACOBS: Yes, I think absolutely that relates to the audit. And when we do the audit that provides the numeric details of the data re-abstraction process to confirm that not only is the data complete but it matches a re-abstraction so it's accurate.

CO-CHAIR GUNNAR: Yes, I would say we've already voted on reliability and validity with regard to the STS registry. So let's carry -- unless there's further discussion let's carry on.

MS. WINKLER: The only thing to think about is yes, you've talked about the testing

of the database. Remember that reliability involves the specifications. So if there are any questions around that.

And validity also includes any threats to validity, handling of exclusions, the risk model. So, those could vary from measure to measure. So if there are any things that are different please bring them up.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: I'm just curious about the numerator and the qualification around an event less than 24 hours. Could you just talk to a little bit about the reliability and the checking of that particular data element? Thank you.

DR. JACOBS: Sure. So, there's a universal definition of stroke. And there's a paper that's called "The Universal Definition of Stroke." And that was a harmonized definition across multiple medical societies. And STS participated in the creation of that

harmonized definition of stroke. And that was the definition that is used in the STS database to track a stroke.

And part of that definition is a temporal cutoff where you have a transient neurologic event that is associated with recovery versus a permanent neurologic event.

And obviously it's a continuous variable that we're dichotomizing. We're dichotomizing that with a cutoff of 24 hours which is a standard universal definition of stroke. And that's part of the audit process that during audit one can confirm if the patient had transient left arm weakness or transient blindness and it went away in four hours that's not going to be a stroke. But if a patient had two or more consecutive days of those findings that would be a stroke. And those are all audited fields.

MEMBER PITZEN: And so the validation around those fields has been acceptable?

DR. JACOBS: Excellent.

MEMBER PITZEN: Great. Perfect. That's all I needed to know.

CO-CHAIR GUNNAR: Any other discussion? Let's vote on reliability.

MR. SANCHEZ: Voting will now begin for 2a reliability. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: Please re-vote.

MR. SANCHEZ: We have 14 for high, 8 for moderate, zero for low, zero for insufficient.

MEMBER ASHER: So, validity was looked at in the context of face validity and I think that we've discussed that in the context of other STS measures with respect to the empirical validity testing.

They basically look to see if the information could be used to predict future performance and they submitted data to suggest

that that was the case.

Although, again, we have very, very few individuals in some of these extreme categories. And so you just have to keep that in mind. But it appears to me that based on the constructs that they've used that it deserves a high rating with respect to validity.

CO-CHAIR GUNNAR: Any other discussion? Let's vote.

MR. SANCHEZ: Voting will now begin for 2b validity. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

Still waiting on one more vote. Can you please cast your vote again? We have 17 for high, 5 for moderate, zero for low, zero for insufficient.

MEMBER ASHER: There's no new information here with respect to feasibility. We've all discussed this in the context of the other STS measures and I think it deserves a

moderate or high rating with respect to feasibility.

CO-CHAIR GUNNAR: Further discussion? Hearing none, let's vote.

MR. SANCHEZ: Voting will now begin for criteria 3 feasibility. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

CO-CHAIR GUNNAR: And vote one more time. It's failing to pick up one person so one more time.

MR. SANCHEZ: We have 12 for high, 10 for moderate, zero for low, zero for insufficient.

MEMBER ASHER: This measure is in current use. We've already discussed it from ways that these various measures are being used so I won't talk about that.

I do question whether or not the fact that it is part of the CABG composite score limits its usefulness as an isolated measure.

I know we're supposed to evaluate specific characteristics of this but it is part of the composite.

There is this very limited change in terms of the number of individuals who fall into these extreme categories. And so I raise that question.

I also raise a question as to whether or not there's evidence of risk-adjusted improvement in performance. In 4b1 there is a reference to 1b which references the appendix.

In looking at the odds ratios presented over two time periods it wasn't clear to me that performance improved in the various provider deciles that were presented. I discussed this with Amy. We were hoping that perhaps the STS could address that particular issue.

We've discussed this before about 4b as a part of this overall evaluation, whether or not there's evidence that there is

improvement. And again, with most people in the moderate category wondering what the relevance of that is.

MR. O'BRIEN: I think it may be difficult to assess improvement over time because back in the two thousands the time frame in the STS database for defining stroke was a 72-hour time frame. Currently it's 24. So there's only within a certain window we can look for improvement.

And I think going back for several of the outcomes considering mortality and stroke you saw kind of dramatic improvements over the nineties and early two thousands, and then more of the leveling off in recent years. So I'm not sure just across the last few years whether there's much difference. I don't recall seeing data suggesting a dramatic difference over the past two years.

DR. JACOBS: And the definitional changes I can clarify a bit. Back in the early

part of the STS database there was transient ischemic attacks which were less than 24 hours, something called a reversible ischemic neurologic deficit which was 24 to 72 hours, and then a stroke was greater than 72 hours.

Then at some point in time when the universal definition of stroke evolved the term "reversible ischemic neurologic deficit" either went away or some people would say well, that's just a subtype of stroke and thus it's dichotomized into two things, TIA and stroke.

MEMBER DUTTON: I'd comment about that too. As performance on some of these outcomes tops out I would suggest that you will continue to demonstrate improvement because you will achieve the same level of outcome in progressively sicker and sicker patients as time goes on.

And as a suggestion for STS about how to analyze it and how to show ongoing improvement that might be a way to approach it.

CO-CHAIR GUNNAR: Any other discussion regarding usability and use? Hearing none let's vote.

MR. SANCHEZ: Voting will now begin for 4 usability and use. One is for high, two is for moderate, three is for low, four is for insufficient information. Timer starts now.

Still waiting on a few if you could please resubmit. We have 13 for high, 6 for moderate, 2 for low, zero for insufficient information.

CO-CHAIR GUNNAR: Any further discussion? Please vote on whether this measure meets NQF criteria.

MR. SANCHEZ: Voting will now begin for overall suitability for endorsement. One is for yes, two is for no. Timer starts now.

CO-CHAIR GUNNAR: One more time.

MR. SANCHEZ: We have 20 for yes, 2 for no.

CO-CHAIR GUNNAR: So, the committee

votes in favor of this maintaining status as an NQF-endorsed measure.

CO-CHAIR FLEISHER: So we're moving onto 0114 so you can stay there. I assume this is another part of the composite. Who's the discussant? Great. Allan, you want to briefly tell us how other than the change in the endpoint, how this differs from the previous measure?

DR. JACOBS: Right. So the issues and dialogue for this measure are going to be very, very similar to the last measure, including the rate of occurrence and very similar to operative mortality.

This is a measure of risk-adjusted postoperative renal failure, the percentage of patients over the age of 18 undergoing isolated CABG without preexisting renal failure who develop postoperative renal failure or require dialysis.

And again, this is another extremely

morbid complication after cardiac surgery. It occurs rarely but it's a life-altering event for sure.

And rather than discuss the evidence base I think the evidence base speaks for itself. The development of the model speaks for itself. And the issues we're going to discuss are very similar to the ones we just discussed with stroke. And I'm happy to answer any questions.

MEMBER SIPERSTEIN: So yes, I think we can minimize a lot of repetition on this model. Analogous to, for example, the mortality in the stroke measure. Really talk about the outcome of a multidisciplinary team throughout the entire process of care.

CO-CHAIR FLEISHER: So are we ready to vote on evidence? Okay, please do.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for yes, two is for no. Timer starts now.

We have 22 for yes, zero for no.

CO-CHAIR FLEISHER: Okay. Can we go onto gap?

MEMBER SIPERSTEIN: So in terms of an opportunity for important the data was presented that categorized various centers as being high, mid or low performers.

And the rate ranged from 0.3 percent in the high performers, 2 percent in the mid and up to 6.7 percent in the low. So I'd interpret this as there is still a range in the measure and a continued opportunity for improvement.

CO-CHAIR FLEISHER: Questions? Comments? So it's a larger gap.

DR. JACOBS: Right. I think this one should be a little easier to vote on than mortality in stroke because the gap is larger.

MEMBER YATES: And we discussed this in the workgroup. The gap being larger is reflective of the fact that this is a true cardiac team gap analysis in that multiple

health providers are involved in terms of maintaining pressure or giving the right or wrong drugs and the like.

And that this goes -- this extends from the operating room environment all the way out through the ICU and to the floor.

CO-CHAIR FLEISHER: Let's vote. Thank you.

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

Still waiting on a few of you. Please resubmit. And one last time, please. We have 12 for high, 10 for moderate, zero for low, zero for insufficient.

CO-CHAIR FLEISHER: Okay. Next.

MEMBER SIPERSTEIN: I think the priority has been well addressed with this affecting about 2.5 percent of the patients with the high morbidity and cost. And also a

link to survival. And there are multiple publications that address this.

CO-CHAIR FLEISHER: Let's vote, unless any comments.

MR. SANCHEZ: Voting will now begin for 1c high priority. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

Still waiting on a few responses if you could please resubmit. We have 21 for high, 1 for moderate, zero for low, zero for insufficient.

MEMBER SIPERSTEIN: So in terms of the reliability the numerator statement is clear and easy to calculate. The definition of renal failure is either creatinine that's greater than or equal to 4 or a threefold increase in the creatinine. Both of those are easily numerical data. Or a new dialysis requirement.

The denominator is pretty much

all-inclusive with all patients over 18 undergoing isolated CABG.

Very minimal exclusions. Even patients who have had renal transplants previously are included. And obviously re-op CABGs are included also. So minimal exclusions.

We've talked about the databases and the clear definitions.

CO-CHAIR FLEISHER: Questions? Comments? Let's vote.

MR. SANCHEZ: Voting will now begin for 2a reliability. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

MR. LYZENGA: We're still waiting on a few if you could resubmit your vote, please.

MR. SANCHEZ: We have 17 for high, 4 for moderate, zero for low, zero for insufficient.

CO-CHAIR FLEISHER: Validity.

MEMBER SIPERSTEIN: I think we've been through all of this previously so ditto to the last.

CO-CHAIR FLEISHER: Any objections to voting? Let's vote.

MR. SANCHEZ: Voting will now begin for 2b validity. One is high, two is moderate, three is low, four is for insufficient. Timer starts now.

Still waiting on a few if you could please resubmit. We have 21 for high, 1 for moderate, zero for low, zero for insufficient.

CO-CHAIR FLEISHER: Okay. Feasibility.

MEMBER SIPERSTEIN: Has been reviewed.

CO-CHAIR FLEISHER: Thank you. Any objections? Let's vote.

MR. SANCHEZ: Voting will now begin for criteria 3 feasibility. One is for high, two is for moderate, three is for low, four is

for insufficient. Timer starts now.

Still waiting on one more if you could please resubmit. We have 15 for high, 7 for moderate, zero for low, zero for insufficient.

CO-CHAIR FLEISHER: Usability.

MEMBER SIPERSTEIN: I think all these points closely mirror the former measure also.

CO-CHAIR FLEISHER: Any comments? Let's vote.

MR. SANCHEZ: Voting will now begin for criteria 4 usability and use. One is for high, two is for moderate, three is for low, four is for insufficient information. Timer starts now.

Waiting on a few responses if you could please resubmit. One more time, please. We have 17 high, 5 moderate, zero low, zero insufficient information.

CO-CHAIR FLEISHER: And let's move onto voting for endorsement.

MR. SANCHEZ: Voting will now begin

for overall suitability for endorsement. One is for yes, two is for no. Timer starts now.

CO-CHAIR FLEISHER: And if we can switch out the developer group because we're moving to the next.

MR. SANCHEZ: Still waiting on two more if you could please resubmit. Twenty-one yes, zero no.

CO-CHAIR FLEISHER: Okay. So, as we're about to start the bariatric measures it would be great for Reva or Helen to just comment on what's in this space. And if nothing's on the space my assumption is we do or don't change the criteria from your perspective. I just want to have that out there.

MS. WINKLER: I would just refer you back to the surgical portfolio document that I gave you yesterday and you'll see that the three measures that are newly submitted for your consideration are the only measures in that space.

I mean, clearly as surgeons the other measures that more generally apply to all types of surgery would be applicable to this area, but specifically to bariatric surgery these are the first ones.

And the criteria are the same for all measures new or for maintenance as we talked about briefly yesterday. For maintenance there are a few expectations of additional information based on use, experience and that sort of stuff. But for new measures everything is the same.

CO-CHAIR FLEISHER: Yes and if I could just confirm that Dr. Morton has an open line.

DR. MORTON: Yes, this is John Morton from the American Society of Metabolic and Bariatric Surgeons. And you guys also have Dr. Matt Brengman in attendance as well representing ASMBS. Thank you.

CO-CHAIR GUNNAR: So, five minutes from the developers to provide an overview of

2556 Yearly Surgical Case Volume of Primary Stapled Bariatric Procedures for Morbid Obesity.

DR. MORTON: Sure. So this is, as mentioned before these are new measures that we're presenting on behalf of bariatric surgery. We don't have any previous measures.

This specific measure is looking at a yearly surgical case volume for primary stapled bariatric procedures in bariatric surgery.

There are three main procedures, sleeve, the band and the bypass. In looking at the current evidence when we examine it pretty closely we see the preponderance of the morbidity and mortality associated with the procedures lies with the stapled cases.

We're also pretty aware in review of the literature that volume emerges as a predictor for both mortality and morbidity.

The other thing that we're able to know is that this is a particular variable that is

reliable and easy to access.

We have had discussion about surgeon volume versus hospital volume. However, the data is more clearly available and more consistent when it comes to hospital volume.

And so the case volume, again as I mentioned before is important for morbidity and mortality but there are some data to also support its use for patient satisfaction and even some of the resource utilization measures such as return to work, length of stay and readmissions.

So as what we're offering as our first measure here total yearly primary stapled bariatric surgery cases in 18 year or older patients. And we've listed the codes for the specific procedures. These are both listed as ICD-9 and CPT codes. It should be pretty readily available whether it be through a data registry or even through an OR log.

I think that's most ofwhat we wanted

to identify. I think we've mentioned already that the volume we do is an important measure for prediction of complication. And we've outlined why hospital is more straightforward as a captured variable.

The number that we've looked at at case volume after a pretty extensive literature review is for stapled cases 50 annual cases a year.

So I think I'll probably pause here because I think the measure is fairly straightforward. And I'm happy to answer any questions.

CO-CHAIR GUNNAR: So, let the record show that Dr. Ko will recuse himself from this measure voting.

MR. BRENGMAN: So I'll just make one comment and that is that in our review of the NQF measures there are four current measures of surgical volume as it relates to different procedures, not bariatrics.

CO-CHAIR GUNNAR: Kelsey?

MEMBER MCCARTY: So, as stated this measure looks at yearly case volume of primary stapled bariatric surgical procedures performed on patients 18 years and older.

It is neither an outcome nor a process measure, but rather a structural measure. The developers are making the case that procedure volume is an easily quantifiable variable that appears to correlate to overall outcomes of morbidity and mortality, and that this variable acts as a surrogate for experience, expertise and institutional commitment.

They also provide the rationale that case volume reflects patient choice which is in turn reflective of patient satisfaction and economic factors such as time to return to work, hospital length of stay and ease of follow-up.

And finally, they also describe this measure as a surrogate for physician to physician referral patterns.

My concern with this measure is that it's not paired with an outcome measure or anything actionable. So it's not clear to me what you would do with the results.

So they have selected a threshold of 50 cases per year based on expertise of the society. So if you're less than 50 cases a year, and I guess I would ask this to the developer, what happens.

So, do you encourage people to get the case volume up? Do you encourage patients not to go there, physicians not to make referrals? Is it just patient information to decide to do with that what they will? I'm not sure what to make of that.

DR. MORTON: Well, there's a few things to consider there.

First, the recommendation is based on the presented literature, the evidence that's available. So this is above and beyond any sort of expert opinion.

There's considerable literature to support the threshold number.

In terms of what to do with the actual volume criteria I think that it's pretty well linked to both mortality and morbidity. And we can certainly amend that because I think that's a valid point. Any sort of measure that we want to look at we want to make actionable. And so we could easily marry it to morbidity and mortality.

CO-CHAIR GUNNAR: Any other discussion? Dr. Sawin?

MEMBER SAWIN: I wonder if the developers could comment on why children under the age of 18 were excluded.

DR. MORTON: I appreciate that point. And we did look at it. Unfortunately we simply don't have data for the young adults undergoing bariatric surgery. It's probably less than 1 percent of the overall case volume. There's about 180,000 cases being done annually. And

in the data that we have currently we're seeing extremely small sample sizes on an institutional level.

And there's a lot of variation, anywhere from, you know, 2 cases to up to about 50 or 60 seems to be the largest. So we just don't have enough signal to figure out what a volume threshold might be for the pediatric patients.

And we felt that the overwhelming majority of the cases being done are adult cases, not pediatric. And we felt most comfortable in recommending the adult population.

CO-CHAIR GUNNAR: Dr. Asher.

MEMBER ASHER: So, low volumes of procedures can be a safety issue. High volumes can sometimes lead to better outcomes but can also imply over-utilization. That's certainly the case for a lot of surgical procedures. So, is there anything in the

existing evidence to suggest that in some high-volume centers that perhaps it's an over-utilized procedure?

DR. MORTON: That's a great question. So, it really gets to the core of appropriateness for these patients.

In that regard bariatric surgery I think has been pretty straightforward about indications for procedure. And these indications for procedure have been set by the 1991 NIH consensus conference criteria.

And that has been a general rule for almost all -- for all the insurance companies, all the different payers. So there's little opportunity for over-utilization as long as the indications are met. So I think given the fact that we've got very clear criteria about appropriateness of indication there's less likelihood for over-utilization.

One point about utilization. There's about 18 million patients who qualify for

weight loss surgery and there's currently about 180,000 cases being done. So it's about 1 percent of the eligible population. So there's at this point little indication of over-utilization for the procedure in the population.

CO-CHAIR GUNNAR: Dr. Burstin.

DR. BURSTIN: Just one comment. I just wanted to follow up on Kelsey's point. I did check. We do in fact have four endorsed volume measures on AAA, esophageal resection, pancreatic resection and pediatric heart surgery.

Those are all attached to mortality measures. Those are considered high-risk, low-volume procedures. So I just wanted to put that out there.

So the question is going to be is there a quality signal here for the volume standing alone.

DR. MORTON: There is. There's

fairly clear indication about the effect of volume on mortality. You know, there's quite a few cases. Probably the best cases have been done with the nationwide inpatient sample and with that threshold of 50. I mean, I can try to get you the exact number here, what the differences are in mortality. I apologize, give me one moment.

It's an odds ratio of 2. So you can imply there that there's a twofold increase in mortality for these patients when they're going to a lower-volume center. So it does appear that there is adequate signal to determine that volume does have an effect.

CO-CHAIR GUNNAR: We'll go to this side of the room and then to that. So, Dr. Grover.

MEMBER GROVER: Yes, thanks to your group for putting forward these proposals.

I have a question on -- in cardiac surgery we've looked at this for many, many

years. And although there's a statistical relationship to volume if you do a scattergram and look there are in the lower-volume centers some high-performing centers, exceptions to that overall trend that you see. And in the high-volume centers there are some that don't do so well.

So how can you -- does the literature discriminate in that way to allow some type of accommodation for the centers that do well that are low-volume and the ones that don't do so well that are high-volume that are more the exceptions to the rule?

DR. MORTON: When we look at -- we've actually done some scatter plots. And I think what it reflects is where we are in the evolution of bariatric surgery vis-a-vis CT surgery, a much more established field of surgery.

We're still seeing a pretty consistent trend. And actually, the threshold we saw

around 50 continues to extend frankly once you're getting into some hospitals that do greater than 400.

We've looked at whether or not there are high-performing low-volume centers and it appears to be a really rare event in just looking at the scatter plot. So we don't see the same sort of relationship we've been able to see in thoracic.

And I think part of the reason for that is simply it's still early on in the quality improvement process for the field. And I think some of the lessons learned by the high-volume hospitals will at some point and should diffuse to the low-volume centers and you'll see more homogenization of outcome.

But we have not seen that to date. To date there's still a pretty strong indication that volume makes a big difference.

CO-CHAIR GUNNAR: Dr. Markman?

MEMBER MARKMAN: In reading your

paper and your evidence for this the centers of excellence are defined as greater than 125 cases per year. I'm just curious why you picked 50.

And then further in your paper there's a statement that when you go to a center of excellence that there was lower mortality but higher morbidity.

So, if you can address those two questions. And this was in the paper that you just sent out to us.

DR. MORTON: Okay. So the paper we just sent to you guys is in reference to accreditation. So in reference to to the volume question we have now the reason that the 125 came up is that was a historic number at this point. That was the previous volume standard when the accreditation programs started back in 2006-2007.

And since then there's been an evolution in the types of procedures we do.

There's much less banding being performed now and more bypasses and sleeves. And as a result the stapled procedures are the ones that garner more attention.

In regards to the specific paper about the accreditation this is a paper that's being published in the Annals of Surgery.

And in it the big differences were around mortality as well as failure to rescue. If you look at the individual complications we looked at roughly about 20. And there were -- the vast majority had improved outcomes at accredited centers. There were a handful that did not have improvement. But the majority did show improvement for those specific complications.

MEMBER MARKMAN: So what is the percentage of complications. And I guess leakage is probably the greatest one. But what is the percentage of this complication?

DR. MORTON: So, I have here just

looking at the overall. And this is around accreditation on volume. So, those were two different effects if you will. And volume exerts a pretty big effect but so does accreditation. In fact, we've seen that accreditation exerts an effect above and beyond volume. But volume is a place to start.

And when we look at the actual complication rates in that particular paper the accredited hospitals had roughly about an 11.3 percent complication rate of any sort in the inpatient stay. And the unaccredited hospitals were at 12.3 percent.

MEMBER MARKMAN: So, the question now comes down is that we have several measures that we're reviewing. And we will review each one. But don't you have a measure also on accreditation?

DR. MORTON: We indeed do. And the three measures that we have are around accreditation, around surgical volume and

about 30-day readmissions.

If we view this in any sense of priority, if we really view accreditation as being the most important measure because from those a lot of things emanate. It's a platform for benchmarking. It gives the centers the opportunity for quality improvement.

It should be mentioned that volume is a subset for accreditation and in many ways accreditation is a composite measure of many different things that are going on all at once. It can render significant advantage for the patient entering an accredited center. Everything from the data registry that allows for benchmarking to having resources in place, the quality improvement requirement and the standards that we've already mentioned.

MEMBER MARKMAN: So, basically the evidence on this particular measure of 50 cases, most of your evidence is on 125 cases.

DR. MORTON: I wouldn't say most of

it. I would say the historic data is around the 125, but the most contemporary data that we have indicates that 50 for stapled cases is the best measure.

That 125 number came out back in 2004-2005. And at that point it was a very different landscape for bariatric surgery. Banding was a much, much more common procedure. And that has now changed where we're seeing more and more stapled procedures.

And it doesn't take as many stapled procedures to indicate changes in morbidity and mortality because the rates are higher. And with the most contemporary data that we have from the University of California Irvine the 50 stapled cases make a difference.

Not to get too technical, but when we went back to look at the 125 threshold we found that about half of the procedures being performed were banding. And if we excluded the banded procedures it really had no impact on the

overall morbidity and mortality because the band procedure was so safe.

But we saw a lot of the morbidity and mortality was associated with the stapled cases. And that's what led to the review of the data to make it most contemporary. And that's where we came up with the 50 stapled cases.

CO-CHAIR GUNNAR: Collette?

MEMBER PITZEN: I just have a process question I guess in terms of -- and maybe everyone is already there and I'm not. But what kind of evidence are we looking for for a structural standalone volume measure in order for it to go forward in the process?

DR. BURSTIN: Yes, this was part of when we did the Evidence Task Force Report a couple of years ago it very clearly said that the requirements for structural measures are identical to process measures.

So it's still -- that's why I asked the question earlier. It still is the quality and

the quantity, the consistency of the evidence that the structural element has an impact on outcomes.

CO-CHAIR GUNNAR: So, just before we vote I have a comment. A question. I know that NIH has seven studies out there currently active which may change the guidance and recommendations going forward.

If the criteria for bariatric surgery is modified, it becomes more metabolic-based as opposed to weight-based going forward, and the volume as a result of that goes down how does a strictly freestanding volume-related measurement stand up?

DR. MORTON: Well, we do view the procedures as being important in consideration that they're above and beyond weight as you mentioned. It's a very powerful metabolic operation with high remission rates for diabetes.

We don't anticipate even if there were

changes in any sort of indications that there would be large changes in the number of procedures being done. Because this has been a longstanding indication that probably will not change with any sort of recommendations.

It should be mentioned that NIH did look at this in terms of guidelines and perhaps revisions of them. They deferred it. They decided guidelines are not NIH business and they deferred the guidelines to both the American Heart Association and the American Academy for Clinical Endocrinologists as well as the Obesity Society.

And when they came out with their revised guidelines they were essentially consistent with what we've done in the past but with even a broadening of indication for some of the diabetics.

So I don't anticipate that any sort of recommendations coming forward will decrease volume. Frankly, I think that it will most

likely increase volume as there's better awareness of the procedure and its efficacy and safety.

CO-CHAIR GUNNAR: Any further discussion? I think we're ready to vote on the evidence for this particular measure.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for high, two is for moderate, three is for low, four is for insufficient evidence. Timer starts now.

We have zero for high, 11 for moderate, 9 for low, 2 for insufficient evidence.

MS. WINKLER: This is a gray zone result. In other words, there's no consensus of the committee at this point in time. So we'll continue evaluating the measure but realizing that there isn't a decision out of the committee on this criteria so far.

MEMBER MCCARTY: Okay, so for performance gap the developers say that the American Society of Metabolic and Bariatric

Surgery in conjunction with the American College of Surgeons have created a joint quality improvement program again using this volume threshold of 50 stapled bariatric cases per year as the requirement.

So this is a little bit different from what you just cited in the paper because in the -- what was put before us in the measure description says that 50 cases per year is actually the highest requirement for level of certification. So, clarification on that would be helpful.

And there was also no data supplied in terms of how many hospitals today that perform bariatric surgery meet this 50 threshold. So it's difficult to evaluate what the opportunity to improve is without that baseline data.

CO-CHAIR GUNNAR: Any additional comments? I think we're ready to vote.

DR. BURSTIN: Do they want to respond? You asked a question.

CO-CHAIR GUNNAR: Sorry. I went to the vote inappropriately. So, the developers, please.

DR. MORTON: Well, yes, there is a performance gap. And in some -- they're within the literature cited. There is at least 25 percent of centers that do bariatric surgery are below that threshold of the 50 stapled cases annually. So that's a pretty substantial performance gap.

And we have been able to substantiate that by looking at nationwide and patient samples and incorporates both -- that incorporates over 1,000 sampled hospitals that may or may not be accredited. So we do see a pretty significant performance gap there.

And I do want to emphasize that there is a very, very clear relationship between volume and mortality which is a significant outcome.

CO-CHAIR GUNNAR: Dr. Fleisher?

CO-CHAIR FLEISHER: So, I'm actually confused in how to rate this. Because you actually -- you keep saying 50 but your measure is number. So be very specific. This is not do you make 50 which is a cutoff. You say how many have you done. So how you rate a performance gap if you are 51?

So I'd actually turn to staff for some recommendation. Because I don't understand how to analyze this.

DR. BURSTIN: Well and again, just to read what's up there, it's that there's variation or gap. So I guess the question would be is there sufficient variation across institutions and volume would be the way I would read that.

CO-CHAIR FLEISHER: But that, okay, volume that it's important enough to --

DR. BURSTIN: You've already made the assessment on your first vote sort of. I take that back, you didn't make the assessment on

your first vote of whether evidence is important or not.

But assuming for now, assuming that's the case you're now at the point where you're discussing whether there's sufficient variation or a gap in performance. So I would base this pretty much on variation I would think.

MEMBER SIPERSTEIN: So are we trying to say then -- so we're basically just saying someone has to tell you how many cases you do.

But implicit in all this is this number 50 keeps on coming up in the measure that we're saying 50, 50, 50. But that's not really what the measure is saying. It's just saying numbers.

So, as far -- if we exclude the whole idea of 50 then there isn't a performance gap because we don't know what that means. We just know that hospitals are doing different numbers.

And we could define a quality measure for every procedure then and it wouldn't have any real value.

DR. BURSTIN: Right. And I guess the issue here is really this question from the first criterion of whether you believe that there's evidence that volume affects the outcome in this case.

In this particular one you're really just saying is there variation. It's not based on 50. The 50 is more their evidence was really my assumption. It's not the construction of the measure. The measure itself is volume as are other volume indicators.

CO-CHAIR FLEISHER: So Helen, just we're at near the end because I'm thinking usability versus performance gap.

DR. BURSTIN: Yes.

CO-CHAIR FLEISHER: And that's where I think there's some confusion. Because how it's used.

DR. BURSTIN: So, as an example, I pulled up from -- and again, this is more on evidence, but from the Evidence Task Force Report. The measure they used as the example was nurse staffing hours which in some ways doesn't have a threshold either. It is more of a continuous variable. But in that instance the evidence was that higher nursing hours resulted in lower morbidity and mortality.

So I think, I mean I'm just following the same argument. I think the argument here is they're saying that volume as a structural measure has an impact, has a relationship to outcome. And so I think what you're looking at here is really whether you believe there is sufficient variation across hospitals to push it forward on performance gap.

I'm not indicating whether I think that's a good or bad idea. Just again, more so from our precedence of looking at other structural measures.

MR. BRENGMAN: I will just make a comment on that and that is that's exactly what made it very difficult actually to sort of write this measure. There are only very few volume measures and it is the number.

The evidence piece is pretty definitive that there's a continuum of higher quality over a number. We presented 50 as part of the accreditation process but not as the basis of this particular measure.

And so there is a high degree of variability of centers doing as low as 15 cases and as high as greater than 400. And that's presented within the literature base as well.

DR. MORTON: I just wanted to follow up on Matt's point. There's actually a lot of data around the volume effect and how 50 stapled cases makes a difference.

There's a study from Procolis survey 2003, Wellwer and Kax 2007, Kells, Obesity Surgery 2009, Flung 2004, Jack, Sword, Smith in

2010. I mean I can go on. You guys have that list there. There's considerable evidence demonstrating a very strong relationship between volume and outcome.

And even though some of the data were historic where we saw different numbers, the most contemporary data are quite clear about 50 stapled cases having significant impact on patient mortality.

And the mortality difference between that threshold of 50 greater than, less than is quite high. It's stated here as being greater than twofold.

CO-CHAIR GUNNAR: Okay, we're going to go on this side and then there. So, Kelsey, did you have a comment?

MEMBER MCCARTY: Well, just that I feel like where I'm really struggling where to evaluate this measure, and this is the problem with all of these types of measures I suppose, is that yes, there's research that shows

there's a strong correlation, but it is on this continuum basis. So the science is there to support it.

But when you adopt measures like this you're trying to fit it in this administrative framework that doesn't always work. Sometimes we don't -- it might need more administrative definition to figure out what to do with this.

So, is there good research to support a correlation? Yes. Is there good research to support a constructive way to make this an administratively run metric? I don't think so.

DR. MORTON: Well, it's a fairly straightforward measure. It's 50 stapled cases. So I think getting that --

CO-CHAIR GUNNAR: No, no, no, just to get focus back in on performance gap I think let's stay there for the moment. And so, Dr. Saigal.

MEMBER SAIGAL: So it sounds like it's

conceptually not possible to have a performance gap the way it's written, right? Because it basically is reporting the number that's happening. So everyone meets it and there's no gap.

CO-CHAIR GUNNAR: Except there's variation.

MEMBER SAIGAL: Depends on how we interpret that variation.

CO-CHAIR GUNNAR: Dr. Temple?

MEMBER TEMPLE: I hope this fits in the performance gap section. But what I'm hearing is it's not just the volume but it's also the case mix. So I'm hearing that -- so if somebody has been, based on how this is written if you're doing 50 bands instead of 50 stapled and you should be doing the 50 stapled, that's variability that you would want to capture in a measure. It's not captured in this performance gap the way the measure is written. So I don't see the volume piece in

this.

And then I see that the whole issue about case mix and choice, you know, the appropriateness of a band versus a staple is not addressed in this. And that to me may be where you see the variability and could find a performance gap measure to measure.

CO-CHAIR GUNNAR: I think we take the submission the way it's written. We must and let's vote unless there's further discussion. So I think we're ready.

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We've got 2 for high, 3 for moderate, 7 for low, 10 for insufficient.

CO-CHAIR GUNNAR: So I believe that stops this analysis. So, for 2556 the recommendation of the committee is that it not be endorsed as an NQF measure.

So we will go onto 2557.

(Whereupon, the foregoing matter went off the record at 12:14 p.m. and went back on the record at 12:25 p.m.)

CO-CHAIR FLEISHER: Can you open the phones for public comment?

OPERATOR: At this time if you would like to make a comment please press * and then the number 1. At this time there are no public comments.

CO-CHAIR FLEISHER: Okay, anybody in the room, short public comment? Okay. Thank you, Lisa.

CO-CHAIR FLEISHER: Okay, let's restart. We won't let you eat. We'll let you talk first.

We are on the readmission 2557. And just checking, anyone leaving before 3:30 today? One. Two. What time? 1:30. About 3. Do you have any measures that you have reviewed? Any measures? Which is which one?

Okay. So we'll make it by 3. Great. Yes.

Okay, readmission. Want to give us a two-minute to three-minute overview.

DR. MORTON: Sure.

CO-CHAIR FLEISHER: Yes, we hear you.

DR. MORTON: Hi. This is John Morton again. This is the NQF measure 2557 and it's Hospital-level 30-day All-cause Readmission Rates after Elective Primary Bariatric Surgery Procedures.

The American Society for Metabolic and Bariatric Surgery is the measure steward. And we're examining hospital-level 30-day all-cause readmission rates following elective primary bariatric surgery. And the ages 18 to 65.

We've included the specific bariatric procedures that are listed there. We're looking at the outcome as being defined as a readmission for any cause within 30 days of the discharge date of the index procedure.

We wanted to make sure we have a good population of homogeneity by excluding some of the other procedures, mainly open or revisional procedures or extremes of age where we don't have good data about the level of readmissions for those patients.

Our rationale for this is that we view the NQF's previous statements on all-cause 30-day readmission measures as being important. And we view this particular measure as being a standard for quality monitoring and looking at the accountability of care for patients.

There have been great strides made in bariatric surgery around decreasing mortality and even individual complication rates are relatively low. DVT, leaks are less than 1 percent.

There is still quite a bit of variation around readmission. National average for bariatric surgery is still about anywhere from

6 to 8 percent. And again, we view it as an important measure because it is a composite of different components of care, complications, patient and physician satisfaction and resource utilization.

And in sum we feel this all-cause readmission measure would provide an ample opportunity to improve hospital performance.

CO-CHAIR FLEISHER: Thank you. And who's reviewing? Collette.

MEMBER PITZEN: Thanks. As stated this is measure 2557 Hospital-level 30-day All-cause Readmission after Elective Bariatric Procedures which include gastroscopic rho and y gastric bypass, sleeve gastrectomy, biliopancreatic diversion and lab-adjustable gastric banding.

It's a new measure. The numerator is readmission to an acute care hospital with a stay that is at least 24 hours for any reason within 30 days of the index hospital discharge

date.

The intended level of analysis for this measure is a hospital-reported measure. However, no performance results were provided, no analysis and no data provided in this application for that consideration.

Before I go into each of the selection criteria I want to share that I do think this measure has potential but I do have some concerns. So I'd just like to briefly talk about that and then we can go into the criteria.

On the plus side obesity is at epidemic level in our nation. This is an important related procedure for the population at significant risk for morbidity.

Many surgical procedures have a fairly low readmission rate. For example, total knee, less than 1 to 2 percent which may or may not make them able to be used for measures for accountability and public reporting.

Granted, though this is not as high as

readmission rates for chronic conditions in the 18 to 22 percent ranges in the literature provided by the developer demonstrate a range between 1.7 and 9.4 percent based on procedure type which may demonstrate opportunity for improvement.

I applaud the developer's decision to include any reason for readmission which is appropriate for this population.

Concerns. Again, no performance data provided so we're unable to understand the true opportunity.

Concerns about the measure specifications overall. The intent of the exclusion, for one example, the intent of the exclusions are to exclude procedures performed related to gastric cancer which is an appropriate exclusion. However, listed diagnosis codes are malignant neoplasm of the esophagus and do not include malignant neoplasm of the stomach.

There's no reliability or validity testing on the data for this measure. There's no current risk adjustment plan or model for this outcome-based measure. The developer shares a struggle with the lack of administrative data that could be used in a risk adjustment model. But they have plans for future development around that.

And my last concern is the ability of a registry-based data system to reliably capture readmission data. Medicare data demonstrates that as many as 22 percent of readmissions occur to a facility other than the index hospital.

During our workgroup call the developers shared that part of the registry database process includes a follow-up with the patient at 30 days. In the registry participants the follow-up success rate is around 90 percent. But this mechanism would need to be tested and included for

understanding the measure performance and impact of missing data. Claims data-based may be a more reliable and complete data source.

CO-CHAIR FLEISHER: So, can we focus on -- because you've identified a lot of the future issues.

MEMBER PITZEN: I'm going to start walking through the criteria for evidence.

This is an outcome measure. The developer at one point in time stated it was an outcome measure. However, in the actual application on criteria 1a1 they talk about this as being an intermediate outcome measure.

I would disagree a little bit. It's an outcome. So, they provided a lot of information, a lot of literature that might not have been necessary. However, they did not provide that direct link of the processes to the outcome.

However, we can glean that from the literature that was provided. Examples of

related processes include fluid and electrolyte balance, surgical technique, prevention of infection, deep vein thrombosis and coordination of care. So I think this would suggest that the link is present, it's just not part of the application.

CO-CHAIR FLEISHER: So purely for evidence, any comment on whether outcome, and I agree with you, this is an outcome measure, that there's -- the rationale supports the relation of the outcome to at least one process as Collette has nicely outlined. Comments? Let's vote.

DR. MORTON: We do agree it is an outcome measure.

CO-CHAIR FLEISHER: Okay, thank you. Thank you.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for yes, two is for no. Voting timer starts now.

We're needing one more. Can everyone

resubmit please? We've got 20 for yes, 2 for no.

CO-CHAIR FLEISHER: Okay, gap.

MEMBER PITZEN: No performance data was provided. Unable to understand the true opportunity. However, literature provided demonstrated some potential opportunity for improvement with literature citing readmission rates of laparoscopic bypass 6.5 percent, open gastric 9.4, sleeve gastrectomy 5.4, and adjustable gastric banding at 1.7. No data was presented or discussed about disparities.

MS. WINKLER: I would just mention that for new measures that it is not -- because there frequently may not have been a great deal of use to generate data it is acceptable to reference literature data for this criteria for a new measure.

CO-CHAIR FLEISHER: So if I could ask you stated the rates. Do we know how much variability between hospitals?

MEMBER PITZEN: We don't. It was literature cited that those are what the rates are in the application. I don't have any information about variability.

CO-CHAIR FLEISHER: This is a facility-level.

MEMBER PITZEN: It's a facility-level measure.

CO-CHAIR FLEISHER: Does the developer have any literature?

DR. MORTON: We do have data about that and it is listed in the evidence area. But there is a tremendous amount of variation, anywhere from 1 percent up to about 20 percent depending on center.

CO-CHAIR FLEISHER: Further comments? You have a comment?

MEMBER MARKMAN: How are we going to collect the data? I mean is this claims-based, or is this going to be --

CO-CHAIR FLEISHER: That's

specification.

MEMBER MARKMAN: Oh, okay.

CO-CHAIR FLEISHER: So let's stay on gap.

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

We have 7 high, 13 moderate, 1 low, zero insufficient.

CO-CHAIR FLEISHER: Okay, next.

MEMBER PITZEN: Next is priority. Although not explicitly stated but filtered throughout the literature provided the measure does relate and reflect to a high priority of aspect of care related to obesity.

CO-CHAIR FLEISHER: Okay. Comments? Let's vote.

MR. SANCHEZ: Voting will now begin for 1c high priority. One is high, two is moderate, three is low, four is insufficient.

Timer starts now.

Still waiting on a few if you could please resubmit. We have 13 high, 9 moderate, zero low, zero insufficient.

CO-CHAIR FLEISHER: Next. Reliability.

MEMBER PITZEN: Reliability. The numerator again is readmission to an acute care hospital with a stay that is at least 24 hours for any reason within 30 days of hospital discharge.

The actual numerator details are not well specified in the application and simply repeat the procedures in the denominator.

I guess I just wanted one overarching comment. Even as an experienced measure person I could not take these specs as written today and implement them with any certainty or reliability. So I think that there's some work that needs to be done around that area.

The denominator statement is

incorrect. It's stated as all hospitals performing bariatric surgery but the implied intent of that is it's all patients aged 18 to 65 undergoing -- I'm sorry, elective primary bariatric procedures.

The denominator is well specified with CPT and ICD-9 pd codes listed.

Again, I talked about the intent of the exclusion and I would recommend that the developer really look carefully at the exclusion codes that are related to malignant neoplasms because that is significant for the measure.

The data sources indicated include electronic health record and registry data. However, no data on results from these systems were provided and no testing occurred.

A small thing. The type of score is listed as a continuous variable. This should be labeled as an outcome proportionate rate type of score.

Testing. No reliability testing was completed or provided on actual data. A few minutes before we broke for lunch we had gotten some additional information but that is around published literature studies. So we have no reliability and validity testing on the actual data for the measure.

So I would rate the reliability of the measure and the specifications to be low.

CO-CHAIR FLEISHER: Barry, you want to now?

MEMBER MARKMAN: Yes. And so how are you going to collect the data? You see electronic records. Can you be more specific on how you're going to determine readmission?

DR. MORTON: Well, readmission as we've listed is through the data registry. With the data registries we've already discussed we do have the ability to have follow-up at 30 days that's based on from report from patient.

So, it's similar to many other programs. And that way we're able to capture if the patient got readmitted to a different hospital which I think is an important component.

So, where exactly the readmission data can come from is a variety of sources including the electronic medical record. But ultimately the registry is the one that's going to give the very best results and that's what we're advocating for to be as complete as possible.

Because if we rely simply on individual hospitals we may miss that patients got admitted elsewhere. So we are stating it would be through the registry.

I think it's already been mentioned before. Given that this is a new measure some of those reliability issues may not come up. We may not have data for that.

We have looked at the exclusion criteria and the idea was to exclude all GI

malignancies which is generally ICD-9 code 150 subset X. But we will make sure that that's absolutely clarified.

CO-CHAIR FLEISHER: Any other comments regarding Collette's concerns?

MEMBER YATES: Who's the central collector of the data? Where does the data from -- I'm looking at S24 and it's "Each facility will maintain a registry." There's various numerous registries suggested. But who's the central data sorter if you will? Where does the data go?

DR. MORTON: The central data will be stored in the Metabolic and Bariatric Surgery Accreditation Quality Improvement Program.

Again, that's the one that has the highest reliability of data in terms of it being clinically derived. And we have that 30-day patient follow-up. So we're able to account for readmissions that might have occurred at a different hospital.

MEMBER YATES: But the actual collection process at the individual hospitals will be individual registries or databases as well as -- or statewide registries that they then are voluntarily sending up to MBSAQIP? Or are we assuming that all hospitals that are involved with that program are the ones that are sending data?

DR. MORTON: All the hospitals that are involved with MBSAQIP will be the ones sending data. There's pretty clear-cut criteria and training for how to obtain this data within the data registry itself.

The nurse reviewers who are the people collecting the data have very strict specifications about how to obtain that data. And so the sample sizes there were just to indicate that there are other means of doing this.

But we're advocating for the best way of doing this which is through the MBSAQIP where

we have to find definitions and we also have standards in place to make sure that the data is reliable.

CO-CHAIR FLEISHER: Barry?

MEMBER MARKMAN: Yes. Is the data that you collect, I mean, is that also going to be used in your accreditation of the facility?

DR. MORTON: To this date we have not decided about specific thresholds for readmission. All we are asserting in this measure is that 30-day all-cause readmissions for bariatric surgery primary procedures should be collected.

In the future there may be some role that readmissions play around re-accreditation. But at this point in time what we're asking for is enforcement of the collection of the 30-day readmission rates.

CO-CHAIR FLEISHER: So let's stay on reliability. That fits in usability which is --

MEMBER SIPERSTEIN: Just to clarify because I was over there. Does everybody participate in the registry?

We had this discussion that the penetrance of STS was 95 percent. So, is this -- and I know John kept on saying we would prefer to use this.

I just want to know how many hospitals. Because it goes to the validity. It goes to whether or not you're going to be able to get the data. And then it goes later on to data burden.

MR. BRENGMAN: It's currently 75 percent of the programs in the country, 750 total programs right now.

There is an increasing number as the new program comes online. For those who don't know two major databases have combined and is going online right now to become the MBSAQIP which you see in the application.

Those are two similar databases, not

totally exactly the same. Now they're unified as a single data set. They have defined data collectors who are trained. There's no inter-reliability, inter-observer reliability, but there's then validation of the data ongoing and then at site visits where all things like complications and readmissions are audited. So, it's a very reliable database structure over time.

CO-CHAIR FLEISHER: Just to follow up on that. STS was able to tell us not only programs but cases and there's a strong correlation. What defines a program versus what defines a non-program surgeon who does bariatric surgery? Do you know how many people just do bariatric surgery out there, the number of procedures?

DR. MORTON: This is a question of how many bariatric surgeons there are?

CO-CHAIR FLEISHER: Well, it's actually -- the data is regarding programs in

bariatric surgery, but that does not necessarily define the universe of bariatric surgery per se. STS can comment on that.

Not bariatric surgeons. What percentage of the bariatric surgery that CPT codes included in this measure would this cover in total? Do you have any idea?

DR. MORTON: As Matt mentioned, about 75 percent of hospitals performing bariatric surgery are accredited.

That does not translate into the number of procedures because the lion's share of procedures are done by accredited hospitals. So, it's probably closer to about 85 percent of the total cases are done at accredited centers. That's the distinction between number of cases and number of hospitals.

MEMBER PITZEN: This is Collette. I'm sorry if I missed this but what I'm having a hard time grappling with is is there a registry already built? Are the data fields --

DR. MORTON: Yes.

MEMBER PITZEN: Okay. It would have been nice had you provided a data dictionary to us for us to understand that there was some structure present.

DR. MORTON: We're happy to do that. And there was actually a web link to the data elements and how they're collected and the criteria around them. But that's in the document, but we can make it even more explicit by sending the entire data dictionary if that's helpful.

CO-CHAIR FLEISHER: So let's vote on reliability unless anyone has a comment specific on reliability? Okay.

MR. SANCHEZ: Voting will now begin for 2a reliability. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

We have 1 high, 11 moderate, 8 low, 2 insufficient.

CO-CHAIR FLEISHER: So we are in the gray area. So we will continue, correct? Okay, next.

MEMBER PITZEN: Again, there was no validity testing of the data that was submitted with the application. Literature again was provided today right before the meeting. However, that is not the actual testing of the measure. Therefore no meaningful differences were demonstrated because no data was submitted.

Again, the exclusion coding which can be corrected.

And then unsure of the potential impact of missing data. Again. So I still view this in the low area.

CO-CHAIR FLEISHER: Comments? Amy?

MEMBER MOYER: I was concerned about the lack of information regarding the risk adjustment. It sounded like there's nothing formal yet and some might submit un-risk

adjusted. And it seems like that would have an impact on the readmission rate.

MR. BRENGMAN: Risk adjustment is being built into the new data set but we don't have it at this time.

And part of the problem with having two separate database collecting all of this data was it's not uniform enough to do that kind of work on.

And so having a unifying data set now, that can go on going forward. But we have to say that's going to be something in the future.

Validation of the data. I think there's a description in one of the measures of how the data is validated for the MBSAQIP database. I know we supplied it as an adjunct perhaps to one of the other measures. But it doesn't do any testing on it. I hear what you're saying, about how the data is collected.

CO-CHAIR FLEISHER: So, I would actually just like to get clarification from

staff in that if risk adjustment is felt to be important given this criteria and they don't have it yet how do we interpret that, or how does NQF suggest?

MS. WINKLER: Well, risk adjustment is, you know, not always necessary and sometimes it is. It really impacts the validity of the measure results. And that's -- certainly how you handle case mix adjustment is a potential threat to validity that you should consider in your evaluation.

DR. MORTON: If I could comment. We are in the process --

CO-CHAIR FLEISHER: If we could just let -- Helen will comment and then we'll be happy to hear the developer comments.

DR. BURSTIN: Yes, so just briefly it's risk adjustment or justification for the lack thereof is essentially the way you should view it.

We do have other outcome measures that

are not risk-adjusted and in fact are done through exclusions by, for example, taking the lowest risk cases or the lowest risk pregnancies and pulling those together. It's kind of a bit of a poor man's risk adjustment but that's acceptable as well. So the real issue is whether the absence of risk adjustment is justifiable or is it just that it's not ready. Sorry, John, go ahead.

CO-CHAIR FLEISHER: So, John, do you want to comment now?

DR. MORTON: Yes, I sure do. I think that's exactly what I was going to mention is that we have done some means of risk adjustment here by excluding revisional cases that we all know as surgeons are tougher cases, higher rates of readmissions.

We've also excluded the open cases because we know those cases tend to be done for specific reasons. They may be tougher patients to do.

And we also excluded the extremes of age. So we have done some exclusions. And I do think it goes a long way towards addressing the issues that have come up where we are able to get a more reliable measure.

So we did exclude open, revisional and extremes of age. And we are working towards risk adjustment but we already have done some work in the area to get a homogenous population by excluding those populations.

CO-CHAIR FLEISHER: So essentially you've done segmentation as opposed to risk adjustment.

DR. MORTON: Exactly.

MEMBER MARKMAN: If you're talking about patient risk adjustment I can add that to have the procedure itself, to go for bariatric surgery and to have it approved through a prior auth process, it is extensive. And the clearances that you need to get the surgery include cardiac, pulmonary, psychological.

You have to have other morbidities with that.

And I'm trying to help you here a little bit in terms of the risk assessment because they are scrutinized if they're under the insurance umbrella.

CO-CHAIR FLEISHER: Okay. Collette, did you -- or Rick?

MEMBER DUTTON: Just to add to that though the most important risk which would be the patient's weight really does need to be accounted for and it doesn't sound like they've done that yet or they have data for that yet.

MEMBER MARKMAN: To do the surgery you have to have certain BMIs which --

DR. MORTON: Right.

MEMBER MARKMAN: -- which you start at 35 with a comorbidity. If you're above 40 you're kind of put a pass on it.

So in doing -- I mean, in terms of a risk assessment I don't --

CO-CHAIR FLEISHER: But what I think

Rick's saying is a lot of ours are above 60 at one of our hospitals and you know, that accounting for it at the high end. I think one of my hospitals -- yes.

Collette, any other comments or are we ready to vote?

MEMBER PITZEN: I just want to open it up to Keith who was the secondary reviewer to see if he had any additional comments.

MEMBER OLSEN: I don't. I agree with your assessments on the measure.

MS. WINKLER: One comment. I know I probably should have said something to reliability, but being an outcome measure for endorsement these -- all measures do need to be tested for reliability with formal reliability testing as well as testing for validity as explained in the criteria.

So Collette, could you review just to be sure we know what the status of that is?

MEMBER PITZEN: I'll try. So,

because no actual data was submitted with this measure the actual reliability testing between, you know, assessing the performance between sites, that was not present.

There was literature that was provided that talked about some of the reliability about capturing a readmission rate. However, we don't have that testing present.

And again, I have some concerns about the specifications as they're written today and moving that into implementation will be difficult.

CO-CHAIR FLEISHER: Okay, please vote.

MR. SANCHEZ: Voting will now begin for 2b validity. One is for high, two is for moderate, three is for low, four is for insufficient. Voting starts now.

Please resubmit. One more time, please. We have zero for high, 9 for moderate, 12 for low and 2 for insufficient.

CO-CHAIR FLEISHER: So it actually does not -- right. We're at 61 percent. So it does not pass reliability. Validity. Reliability. Right.

So I think you've gotten a lot of guidance on some of the issues here. And this committee will be intact for two to three years as far as the members. So I think there will be a chance to improve the measure. Okay.

DR. MORTON: Well, we will certainly take it to heart.

CO-CHAIR GUNNAR: So, the next is 2559 Bariatric Surgery Hospital Accreditation. Developers, would you like to provide an overview?

DR. MORTON: Yes. So this is measure 2559 and the measure title is Bariatric Surgery Hospital Accreditation. We are the stewards.

I want to point out that bariatric surgery is a new surgical specialty. Accreditation has only been in place as a

concept for about seven years now.

We do know that bariatric surgery has improved over time and one of the reasons for it is hospital accreditation.

As mentioned at the beginning we have had no measures submitted to NQF before and we view this particular measure of clear importance.

We're also aware that accreditation is not at all uniform. There is a performance gap here with only about 75 to 80 hospitals, 80 percent of hospitals being accredited for bariatric surgery.

There are also some opportunities for harmonization amongst the different accrediting bodies, whether they be American College of Surgeons and ASMBS, MBSAQIP as well as some of the insurance payers.

But we do know that there are particular key elements of accreditation that are important, namely the standards involved.

The data registry is absolutely critical to allow for benchmarking for quality improvement efforts.

In addition, the requirement for quality improvement as a condition for accreditation.

And finally, looking at the data provided there's 10 studies that have been performed. Seven of the studies are well in support of accreditation across the board for complication. Mortality and failure to rescue, resource utilization. And all of the papers utilize the same data set so there's not any difference in terms of where they're obtained from. And there are some particular issues with the papers against accreditation.

But we really view this as being critical for continued quality improvement for bariatric surgery. Without accreditation there will not be opportunity for collection of data as this is an additional resource for

hospitals to provide.

And the data registry is absolutely critical to have the ability to benchmark, compare, move forward with your quality improvement efforts.

So, in sum we view accreditation as being absolutely critical for our future endeavors in quality improvement in bariatric surgery. And without it a lot of these efforts will be in peril.

And there is a considerable performance gap here with approximately 25 percent of hospitals not being accredited in the United States. And I think there's a clear preponderance of evidence in support of accreditation as listed in the literature cited.

And this is a particular measure that is fairly easy to obtain and define and has been reliable to date. And we've had all the citations there listed.

So I do appreciate the committee reviewing this and all the measures provided and appreciate the support they'll provide in bariatric surgery moving forward in terms of quality improvement and supporting the accreditation measure. Thank you.

CO-CHAIR GUNNAR: Dr. Roth.

MEMBER ROTH: This is Gary Roth. This is measure 2559 Bariatric Surgery Hospital Accreditation sponsored by the American Society of Metabolic and Bariatric Surgery.

Due to the inherent delays in my operating room during the workgroup call which I'm sure is unique to my operating room I missed the call. But I did review the transcript. It appeared to be quite an interesting discussion.

There was a comment made that there hasn't been endorsement of an accreditation measure like this. Was that relative to the NQF, or was that just relative to the bariatric?

MS. WINKLER: That was relative to

NQF.

MEMBER ROTH: Okay. So in that respect this may be a little bit different. I'm not sure how that's going to affect the discussion.

But bariatric surgery is as described relatively speaking a new field. The premise that there's going to be a favorable impact that accreditation will have upon surgical outcomes versus those institutions that are non-accredited.

One of the demonstrated drivers for accreditation for bariatric surgery is, as mentioned, is safety and effectiveness.

Also as mentioned accreditation for bariatric surgery programs is not uniform, about 75 to 80 percent. It described 730 hospitals that were part of the registry and 250 that are also doing bariatric surgery that are not accredited.

There is multiple accrediting bodies

reportedly which is an opportunity for harmonization among those.

Accreditation -- well, the measure is accreditation versus non-accreditation of course.

It's described as a process measure but there's also components of outcomes with what's in the discussion. Possibly that this could be some type of composite measure.

The process measures of course are such things as patient selection, level of critical care support, continuous quality improvement. But there's also outcome discussions too including case volumes and data collection --

CO-CHAIR FLEISHER: Just a clarification. I believe this would actually be considered a structure.

MEMBER ROTH: Okay. What I was reading described it as process in the application. Structure most certainly would

fit more appropriately.

The accreditation of course incorporates many different processes as mentioned. Quality improvement requirements, multidisciplinary team.

And one of the premises of course is mortality and morbidity after bariatric surgery's influenced by center's accreditation status including such things as accredited centers having a reduction in failure to rescue.

The numerator is the number of hospitals that are accredited. The denominator of course all those that are performing bariatric surgery. And the cases were identified through ICD-9 procedure codes.

The key elements of the accreditation include such things as case volume, patient selection, and approved procedures, commitment to quality care standards, appropriate equipment and instrumentation, critical care

support, continuation of care, data collection and continuous quality improvement.

Under evidence there was an extensive citation of the literature but there were no level 1 type studies. Of course, all the studies in this case are observational cohort type studies.

The literature though is really not homogenous in the sense that when you look at the studies 4 of the 10 that were described here actually do not support the concept of accreditation including morbidity, mortality issues, length of stay and cost.

So, within the literature the -- I would say the literature supports it you know at a moderate level at best.

CO-CHAIR FLEISHER: Comments? Comments from the developer?

DR. MORTON: Yes, I think it's important to take a close look at the literature because if you review the articles that are

against accreditation there's numerous flaws.

I agree that the level 1 evidence is not going to be present there and it never will be. This does not lend itself to level 1 evidence in terms of randomization for accreditation.

One study against accreditation was from the Livingston and Good study should be dismissed frankly because it predated accreditation. It utilized the data set from 2005 and accreditation did not start until 2006.

The Michigan paper stated they did not see differences in accreditation. However, every center participating in the Michigan collaborative had elements of accreditation, namely volume standard, data registry, quality improvement, site visits. So they were all virtually the same.

The JAMA paper had significant methodological flaws in the sense that it

compared it to a control group that was stated not to have accreditation. Unfortunately all of those control population were private payers who also had accreditation in place.

So it was not a very accurate assessment of a control group. So I think the preponderance of the evidence is for accreditation.

And one final point to make is I agree this is more of a composite measure and I'll readily grant that it has strong elements of a structural measure. Thank you.

MEMBER ROTH: And relative to the literature some of the support of literature also predates the accreditation as far as the dates that the articles were published.

DR. MORTON: I'm not sure which ones those might be because I was -- I'm just looking at them now and there's none that I can see that predate 2006 other than the Livingston article.

MEMBER ROTH: I'd have to go back to

the references but I looked at them date-wise also and characterized them by the date they were published.

CO-CHAIR FLEISHER: Any other comments? Dr. Cima?

MEMBER SIPERSTEIN: John, can you just clarify one thing about the accrediting bodies?

So, there's the bariatric NSQIP but then there's five other accrediting bodies. Do they all use the same sort of -- are they all going to be measuring the same thing? Are we just saying accrediting is accrediting and they're all equal?

DR. MORTON: Well, the main accrediting body, the national accrediting body is the MBSAQIP.

The other ones that are listed there are national accrediting bodies that are represented by different payers, Cigna, Aetna, United as well as Blue Cross.

Blue Cross in particular has cited the MBSAQIP standards and accreditation program as being consonant with what they're pursuing to the point that if you are an MBSAQIP-accredited hospital no further review of those standards are required. So that's one step forward already around harmonization with that particular payer.

Had discussions with the other payers to make sure that they're also moving in that direction.

That being said, there is very strong consistency along most of the criteria with those private payers.

MEMBER JARRETT: Hi, this is Mark. I've been kind of quiet all morning but I have to ask this question philosophically.

If we start saying that a measure is going to be are you accredited, where do we stop that? Do we start -- then we start saying well, do you have some Joint Commission accredited to

be a stroke center? Do you have to be so-and-so accredited to be a patient-centered home?

I mean, accreditation is wonderful and I think everybody should do it, but the question is does that become a measure that people are going to say oh, here's another quality measure that we have to say yes or no on. That's my concern.

DR. MORTON: Well, I think bariatric surgery is a little bit different in the sense that it is new and there's not a consistent application of it. Unlike say the example cited there, Joint Commission, there's pretty good uniformity around that type of accreditation.

But for bariatric surgery there's not uniformity for that. There's still quite a gap in terms of who's accredited and who's not.

And I do think it bears a lot of elements that are going to improve both hospital performance and patient care.

MEMBER JARRETT: Well, I beg to differ because Joint Commission for stroke is relatively recent in the last couple of years and I would say there's more variability in the treatment of strokes throughout the country which is a much more frequent event.

So if you were going to go somewhere towards that then I would say well then we should mandate everybody who takes care of a stroke has to be an accredited stroke center. And that's going to be -- that becomes a whole -- that's just the road I'm afraid we're going to go down.

That's, again, not knocking the fact that this is very critical, very important and I think everybody should participate in. The question is does it become an NQF measure.

CO-CHAIR GUNNAR: I think we have one other comment here. We appreciate that line of discussion but we'll move back to the evidence category. And any other discussion? Hearing

none I think we're ready to vote.

MR. SANCHEZ: Voting will now begin for 1a evidence. One is for high, two is for moderate, three is for low, four is for insufficient evidence. Timer starts now.

We have 1 high, 9 moderate, 12 low, 1 insufficient evidence.

CO-CHAIR GUNNAR: So I believe that's gray. So we will carry on.

MEMBER ROTH: Okay, under opportunities for improvement, performance gap, there most certainly are described worse outcomes especially in the older patients and ethnic minority patients. But not necessarily based purely on accredited or non-accredited centers using accreditation as the discussion point.

So there is a need to improve the quality in outcomes. I don't know that we see a specific performance gap associated --

CO-CHAIR GUNNAR: If I read this

correctly, not to interrupt, there is a gap. It's 20 to 25 percent of the facilities performing bariatric surgery are not accredited.

MEMBER ROTH: So is that affecting the outcomes.

CO-CHAIR GUNNAR: Okay. Any other discussion? Hearing none we'll go to voting.

MR. SANCHEZ: Voting will now begin for 1b performance gap. One is high, two is moderate, three is low, four is insufficient. Timer starts now.

We have 4 for high, 16 for moderate, 3 for low, zero for insufficient.

MEMBER ROTH: On high priority 180,000 bariatric surgeries are performed annually. Again the concept that 25 percent of the centers are not accredited.

There is a high priority associated with the number of surgeries that are performed per year and the known outcomes.

CO-CHAIR GUNNAR: Any other discussion? Dr. Fleisher.

CO-CHAIR FLEISHER: Yes, just a question for clarification. High priority of accreditation? Versus high priority of the health of bariatric patients? The disease or the --

MS. WINKLER: Yes, it's more the condition of the patient.

CO-CHAIR GUNNAR: Ready to vote?

MR. SANCHEZ: Voting will now begin for 1c high priority. One is for high, two is for moderate, three is for low, four is for insufficient. Timer starts now.

We have 11 for high, 10 for moderate, 2 for low, zero for insufficient.

MEMBER ROTH: Okay under 2, reliability, scientific acceptability of measure properties. Reliability testing, it was listed as not applicable I believe in the application.

But the article that was just submitted today described reliability testing, demonstrated that the measure data elements are repeatable, reproducing the same results in a high proportion of the time when assessed in the same population at the same time period. So that was again the article that was just distributed today.

MS. WINKLER: Specifications come under reliability and I think we need to be very clear what this measure is. And so, let's be sure we're all thinking about it in terms of the same thing.

As a structural measure where the question is is this hospital or facility accredited and it's a yes or no dichotomous answer I just want to verify. Because some of the things that Dr. Morton has been saying confuse me just a little bit. So I want to be sure we're all on the same page that that is the measure.

DR. MORTON: Accredited, yes. Accreditation is the measure. I'm happy to clarify any points of confusion. But accreditation is the measure.

MS. WINKLER: Okay. I think the secondary question to that is is it accreditation by any of those accrediting bodies? Or is it accreditation by your particular group?

DR. MORTON: What we're looking for is accreditation based on those components that we've described there. And we're looking at it specifically at national registries that allow benchmarking. And some of those payers do have some of that, but most of them do not.

And what we're looking forward to is having a single accreditation process and body where all of those that are mentioned there are going to incorporate the same measures.

MS. WINKLER: Okay. So trying to clarify again. So you're defining the data

element of accreditation by whether they assess those elements.

DR. MORTON: Right. The best expression of accreditation is going to be through participation with the MBSAQIP. That is going to be the absolute best expression of the accreditation measure that we're putting forward.

We cite some of those other accrediting bodies as examples of other national accrediting agencies. And they incorporate some of the elements of MBSAQIP. But, what we are asking for enforcement for is accreditation through its best expression which is the MBSAQIP program. That incorporates national clinically derived data that allows for benchmarking, has a requirement for quality improvement, has the resources in place, multidisciplinary team. So, to be clear, that's what we're asking for.

CO-CHAIR FLEISHER: So, to

re-clarify, you're not asking about your database specifically, correct?

MR. BRENGMAN: To clarify, we are asking for accreditation that encompasses the seven components that are described in the application. So it has to have those seven components. If it has all of those seven then it qualifies as an accreditation program under this measure. Is that right, John?

DR. MORTON: That's correct, Matt. And the best expression of those seven components is through MBSAQIP where if there were another national organization that were able to supply those seven elements then that too would be -- that would be compatible with this accreditation measure.

MR. BRENGMAN: To give you an example of that, currently the Cigna program encompasses MBSAQIP with some additional volume data. So they would qualify because they have a program that has the seven

components and then something else. So that would be fine.

CO-CHAIR FLEISHER: So who would actually be in charge of deeming this accrediting organization as meeting the measure to meet the measure?

I mean, in other words there has to be somebody who -- STS is in charge of the database. So we know who's in charge of the database.

Who deems an accrediting organization as stamping the approval to meet this measure if you join them?

DR. MORTON: Well, to be clear it would be MBSAQIP if that can be as clear as possible. It would be MBSAQIP.

We cite those other ones to point out that there are other accrediting bodies out there, but we're looking to harmonize all of these.

And MBSAQIP in the current

configuration is the one best example of it that we feel should be the leader and the certifier and the best expression of accreditation.

CO-CHAIR FLEISHER: Collette, can you?

MEMBER PITZEN: I'll try to add. It is confusing because if I try to take it in a development example someone has to be the owner and the steward of the measure that you're measuring, the numerator and the denominator. So somebody would have to be gathering all that data and trying to make the decision if those seven key criteria were met in terms of how this measure is specified.

If the intent is accreditation for MBSAQIP, I mean that is a different statement in your numerator.

CO-CHAIR GUNNAR: Dr. Grover?

MEMBER GROVER: I had two questions kind of along the same lines. Because you're kind of in a way putting the response, all your

faith in the accreditation system and you aren't really controlling the data elements that go into that yourself.

The other question I had, and maybe, Reva or Helen, you can help me here, is that I thought historically, and maybe you've changed your policy. I thought when you brought measures before the NQF you generally had had them in place for a year or two or something and collecting data to show that you do have the reliability and the validity of collecting that data. Is that passé?

MS. WINKLER: No. We don't necessarily expect a great deal of data from new measures because they haven't been widely used. But we certainly expect reliability and validity testing for all measures that are submitted.

So that will generate some data. It may not be large amounts.

CO-CHAIR FLEISHER: Collette?

MEMBER PITZEN: So, if I could just carry that theme along in terms of the criteria. Then we don't have the information to validate that that is occurring.

CO-CHAIR GUNNAR: Dr. Yates?

MEMBER YATES: The application as given, there are multiple different potential accrediting bodies. It's not clear how many of them carry your organization's blessing. It says that they may include the following organizations as opposed to dropping the "may" and just saying that they do include.

And the concern would be is that a moving target and within each of those accrediting bodies, are they moving targets in the sense that they could be changing one of the seven criteria that you've accepted one year but have dropped one or added one a year after.

DR. MORTON: Hello?

CO-CHAIR GUNNAR: We hear you.

MR. MORGAN: I'm sorry, I missed that

last question. There was some sort of noise. Sorry.

MEMBER YATES: To reiterate, the application shows multiple different potential accrediting bodies. And the actual statement is, is that the other accredited bodies may include, but it's not clear that they have your organization's blessing for each of them. And that it's been vetted, that that process has been vetted.

So it's kind of hard to know what the real numerator is for those hospitals that are accredited.

MR. MORGAN: Well the numerator and denominator for hospitals that are accredited is actually pretty clear. We're going by the MBSAQIP of accreditation status. And based on that, about 75 percent of bariatric surgery hospitals are accredited.

To be complete, we included those additional payers and their accreditor

strategies in order to be complete. To be very clear we view MBSAQIP as the appropriate accrediting body and ASMBS as a steward for this measure.

And so we're very clear about that. We're looking to harmonize what we're doing along with those other payors. And as I mentioned a little bit earlier, Blue Cross and Blue Shield have endorsed that and they're moving forward in that direction.

And we would like these other accrediting agencies to do the same. And they've given every indication that they're looking at a single standard.

MEMBER YATES: So in fact the statement would be other accrediting bodies might include the following organizations if they indeed meet the requirements as you've put them forth.

MR. MORGAN: Yes, I think that would be accurate. Yes. And again, the main

reason that was listed was just to be you know, complete about what's the landscape currently. But we do view MBSAQIP as the leading model for accreditation.

CO-CHAIR GUNNAR: Dr. Dutton.

MEMBER DUTTON: We've had this discussion before at the NQF around for instance participating and registry. And it will come up again in a few minutes when we talk about the STS measure.

I think traditionally how this has gone here is the steward puts forward the criteria, i.e., participation and registry or accreditation. And here are the conditions of that. The measure would then be open to reporting by anybody who can meet those criteria.

And if we asked MBAS here if another registry appears with the same requirements or another accrediting agency appears that meets these criteria, would that be acceptable?

MR. MORGAN: Yes. I think that would be acceptable as long as they met those criteria listed. And that's why we specify them because we view those criteria as being bedrock to the accreditation process. And to ensuring better outcomes.

And so I think it is in some ways very similar to what you've outlined. And participation in the data registry like the next measure is a component of this composite measure if you will, of accreditation. As we do view that as being perhaps the single most important aspect of accreditation is participation in the data registry.

CO-CHAIR GUNNAR: Collette.

MEMBER PITZEN: So just to clarify process. So if we treat this like all other measures, there really hasn't been any testing of the reliability of collecting and capturing that information from the other sites.

MR. MORGAN: So if you're referring

to reliability of trying to determine if center is accredited or not, there are you know, years of data of demonstrating which centers were accredited. I'm not sure if that's exactly what you're asking, but we have -- we do have multiple years of the accrediting body as having listed who's accredited.

MEMBER PITZEN: This is Collette. I'll try again. When you're listing seven key elements of this is defining accreditation, if I translated that into a clinical measure that we're used to working with for diabetes, we have five components that we expect patients to meet in order to meet that numerator criteria. So I'm getting confused, is it as simple yes/no, you're accredited by someone, or is it you're meeting these key elements and then you would be considered to be accredited?

MR. MORGAN: I see. So in order to be accredited, a yes/no, you have to meet those seven. And there can be no exceptions around

those seven. So it's an all or none. Does that make sense?

Hello?

MEMBER SIPERSTEIN: But if we -- John, if we say if you're saying then another body can then accredit, assuming they meet those seven, we have good data from MBSAQIP that you can meet those standards, but do we have any reliable data from other accrediting bodies that they also would meet those standards?

Because you either have to make the -- I think you have to make the statement, in order to be accredited, you have to be part of the metabolic MBSAQIP, or these are the standards that everyone needs to meet, and what data do we have that other accrediting bodies will provide us that data?

I mean is that sort of what you're getting at Collette?

MEMBER PITZEN: Yes.

MR. MORGAN: Yes, so where we have

the data for this, and that those individual centers have met those individual components is around MBSAQIP, we don't have data for the other accrediting bodies. So again, just to reinforce, MBSAQIP is the leading and best example for accreditation that incorporates those seven different elements.

And we do have data from the different site reviews that have been performed. So as part of a site review, all of those elements are investigated for and accounted for. And the inability to meet those components, would not result in accreditation. So you have to meet all those component in order to be accredited.

Where we have good data and reliability around that is with MBSAQIP. The other ones that we mentioned there were simply to be complete and to give you examples and opportunities for further harmonization. But we do have very good data about centers meeting

those components through the site reviews.

And again, it is an all or none. They have to meet all of those criteria in order to be accredited.

CO-CHAIR GUNNAR: Collette.

MEMBER PITZEN: I'll try one more time. So we don't have any information on the testing of that hypothesis?

CO-CHAIR GUNNAR: Dr. Yates? Dr. Fleisher?

CO-CHAIR FLEISHER: Yes, just to go back to the criteria, because I've looked at the criteria again. So it says appropriate instruments, am I correct. The seven elements, the appropriate equipment and instruments can count commitment to quality care standards. Those are nebulous enough in this specification, unless I missed somewhere in the specification that you define it that other organizations know what that means.

Am I just -- not like we do -- John

is, do you in this document define so other organizations can meet it, what appropriate equipment and instruments, and commitment to quality care standards mean, and what critical care support means?

MR. MORGAN: We most certainly do. And it's an 82 page document. And it is the standards associated with MBSAQIP. And it's absolutely detailed. In the interest of brevity in the application, it wasn't incorporated entirely there because it is an 82 page document.

However this is a web citation, reference. But there is absolute clear data, specifications about exactly what that means. It's an 82 page document that can very easily be supplied.

CO-CHAIR FLEISHER: But that's to me your criteria, right? For your accreditation?

MR. MORGAN: Correct. Right.

CO-CHAIR FLEISHER: Not for any accreditation, correct?

MR. MORGAN: That's correct.

CO-CHAIR FLEISHER: Thank you.

CO-CHAIR GUNNAR: Amy?

MEMBER MOYER: I just -- I'm looking through the algorithm for evaluating reliability. And the first boxes are the specifications precise and ambiguous and complete. And we spent a lot of time discussing this. It feels like it maybe doesn't meet that first box.

CO-CHAIR GUNNAR: Any other comments? All right, I think we're ready to vote.

MR. SANCHEZ: Voting will now begin for 2A, reliability. 1 is high, 2 is moderate, 3 is low, 4 is insufficient. Timer starts now.

Zero for high, 4 for moderate, 12 for low, 6 for insufficient.

CO-CHAIR GUNNAR: So the measure

fails to pass. And therefor is not recommended for endorsement.

MS. WINKLER: Would it be fair to say that as a committee, because you considered this a high priority condition, you would certainly like to see you know, improved measures to come back on these because the topic area is an important one.

MEMBER GROVER: Yes, I feel strongly about that. I think what you all have heard here as a committee, a large committee here that spent probably an hour on this. So we obviously would like to see you succeed. And we just need more information and more experience. At least I'm speaking for myself on that issue.

CO-CHAIR FLEISHER:: Just -- you're within the college, correct? I mean the organization sits within the college, so perhaps Cliff, you can help them as they come back.

MR. MORGAN: Well we appreciate it. And what we would greatly appreciate.

MEMBER KO: So I just had a question about the structural measures. Because this is a, you know there was -- two of them are structure, volume and participating in accreditation program. And then the outcome. Outcome measure is easy enough.

But the structural measure is very interesting as we kind of go through our different eras of how we do things. You know, first it was structural volume. And then it was the process measures with Mark McClellan's piece in CMS and what not.

And then we've gone to outcomes. And as we run this program, outcomes are not enough. So maybe we go to composites, or maybe we kind of even go back to a -- maybe the biggest composite is an accreditation program, where it has those pieces, a structural piece of process and an outcome.

And it even has the piece of putting it into play. Because then there's the accreditation itself. Is this, I mean is this worthwhile. Or is this something that's in the vision of this group of that piece? Because that's a huge step forward. It's a lot of work obviously to put the application forward.

But it's a huge step. And it might be in the right direction, but if it's the feeling of the group that accreditation like what this is, or the stroke program, or you know, any of those types of programs, is not in -- is not the direction of what NQF wants -- of a measure, then that's really helpful for us.

CO-CHAIR FLEISHER: Collette?

MEMBER PITZEN: I just wanted to make some additional comments. Again, I think we're excited about that outcome measure. And I'd like to see that come back through again.

And I had a suggestion. People do this to me all the time. I think a really cool

measure would be looking at the weight loss of patients undergoing this procedure, and their maintenance of that weight loss over time.

I think consumers, providers, purchases would be interested in that kind of outcome measures. Thanks.

CO-CHAIR FLEISHER: Amy?

MEMBER MOYER: From my perspective, I really struggle with accreditation as something that NQF would endorse. And I guess part of what I struggle to understand about it is I've never met a hospital that won't tell me they're accredited.

They won't tell me their mortality rate or their complication rate or you know, all these other things. But they're not shy about saying we're accredited. And so I struggle with what adding the NQF endorsement to that means, unless we're endorsing something specific. And then that feels like a direction we may not want to go in either. That was an

issue I had.

MR. BRENGMAN: Can I ask a question about the accreditation thing. Because I'm also a little puzzled because it is, this particular accreditation is quite lengthy.

It's got elements of structure, you have to buy all this junk for your hospital. You have to have nursing pathways which are clearly process. And then it's also outcomes because if you don't meet certain outcome requirements you can't be accredited either.

So you have to do all of those things, and then you have to be inspected by an onsite inspector. And they have to sign off on you to meet these measures, which are all scientifically based. And the committee for MBSAQIP anyway, it's like 40 surgeons who have met for two years to come up with the criteria, who are knowledgeable in the field.

And so you have this accreditation, and it seemed to me we got derailed on sort of

-- we tried to include too many people honestly in my opinion. But we could come back to that.

But I agree with Dr. Ko. I mean if you can't get that through as a single measure, do you want to see more of that? Did you want to see more of what the details of the structure were? Of what we do to make accreditation? I think the other two were a little easier.

MEMBER HANDY: So one of the problems that I had which was why I asked the question. Is there's Blue Cross/Blue Shield accreditation. There are other accreditations. And the way you open it up for the measure, made it unmanageable.

So I don't know what the committee would think of, you know if there was one accrediting body, that would change in my mind how it's organized, who owns it, who maintains it.

Now whether that would get through NQF is other people's opinions. That was the

biggest problem I had.

MEMBER SIPERSTEIN: I mean there is a point that Amy brings up. And that is you know, every hospital in the country that gets paid by CMS is accredited by the joint commission. But we wouldn't be having this meeting if every hospital was doing great.

So accreditation in and of itself, just because they got the stamp of approval from some organization, and I know that's not the case with this. I mean but that's the point, should NQF be in the business of saying who's accreditation is more important than someone else accreditation.

I'm not sure if this is the right forum for this. I think the problem with this measure, in my personal line, was there was confusion about are we picking a winner, saying that you have to use a specific tool, you know we had this -- I remember being very vocal about this when we had, when I was on this three years

ago or four years ago, about STS. Are you picking a winner.

I'm not sure we're in the business of picking winners. We shouldn't be. We're in the business of writing a very good design. And I think what was the problem about it for me was it kept on coming back and John kept on saying we're going to use MBSAQIP as the benchmark.

Well, the thing STS went away -- always came back at me with, these are measures that anybody can collect the data on and do. And if you keep on saying well we're the racehorse, we're the winner, and that you need to -- we're hoping to get to us, that's not what NQF should be doing.

MEMBER LEVY: You know I think a way to think about this is accreditation is a lot like board certification for us. It's something, it's a benchmark, it's something we'd like everybody to have

But then we don't say that all board certified surgeons are the same. And we continue to hold board certified surgeons to standards and to performance review and all those things.

So I think accreditation is not what we do at NQF. And it's not a measure, an outcome measure, it is just a benchmark. It's something that payers may clearly want to look at. But that's not what we're -- it doesn't describe quality in the way that we need to do that.

So I think if we look at the analogy between board certification and accreditation in these different kinds of programs, that might help us.

CO-CHAIR GUNNAR: A.J.?

MEMBER YATES: I'll take a contrary point, and to answer Dr. Ko's question. I mean there's accreditation in the board sense. But then you may help if NQF were to define as a

structure process say had a paradigm for looking at this issue. Because it's come up again apparently more than once.

And the paradigm I would suggest is, is there a patient problem, or is there a patient population that is unique. Is that patient population helped by a hospital going through all the steps to reach accreditation and to maintain accreditation that includes data that comes -- goes out, comes back and causes quality improvement on a regular basis.

And is there data to show that that process helps. And lastly fourth, is there a critical mass of hospitals involved with it that you can maybe say that this is the winner. That this is a good thing to be involved with.

STS as an example, has evolved to that 95 percent winner status if you will. I would suggest that the bariatric group is very close to this, having I'm assuming ACS and the other group merged to create MSG.

And your data's there in this unpublished paper. Your data, your information's there. And it may be that you sold yourself short on this by opening it up to any other accrediting body, which is a black -- a whole bunch of black boxes.

Whereas you really ought to be shooting at the problem are these obese patients going through surgery. There's a performance gap we can show. We can show that if people do go through the accrediting process and as a team we have to create all this information, and we have feedback loops in place that we're going to improve outcomes, I would definitely say that it is part of NQF's business.

You know it's -- and in terms of board accreditation, we're way, way away from anything like that as individual physicians. But the whole move for master you know, mastering certification, or recertifying and

going through those processes, is mean to address that issue.

But at this level, for specific programs, for specific patient populations, I think it is within NQF's -- I think it's within their bounds to work with this. But I think that some guidelines should be made.

MEMBER PITZEN: Yes, I would just like to share an opinion. I think the QI movement has vastly changed in the last 20 years. NQF has a hierarchical preference for measures. And there's nothing wrong with having accreditation and those kinds of bodies. But it's a means to the end of the outcomes that you desire.

So there may be less value in a simple yes/no, we're all accredited, we all have joint commission accreditation. But what are the pieces and components, what are the outcomes underneath that that are going to support population health and improvement and

quality.

MEMBER MARKMAN: Yes, I'd just like to commend the two on your outcome. And I agree with Collette, that to take on a 30 day readmission initiative, and I've looked through the entire portfolio of our -- and we don't really have one, other than the CMF. I think there's a cardiac one.

But I really commend you guys for coming out and trying to accumulate the data on readmissions, which is very important to everybody. So kudos.

MEMBER SIPERSTEIN: I think the context of this measure's very important. Because obviously this is an evolving specialty. It's coming from their professional organization. And they are striving to improve the quality of their patient care and obviously have to commend them for putting these forward.

And in that context, I think even

the issue of being accredited or not is an important move towards improving quality in a fledgling areas. The implication for example in a more mature area like cardiac surgery are different.

But I think definitely in bariatrics in this context, it's a very important group of measures, and I would encourage the group to refine them and resubmit them.

MEMBER KO: I have a comment and then maybe a question. So the comment is to share just our quickly our experience in MBSAQIP in that we, for all the criteria that we vote on within getting the data and the statistics and reliability and doing the audits and whatnot, we perseverate on that.

We spend hours and hours fighting and figuring out what's the best way to do this. And we could probably submit a ton of measures that would -- I mean I'm not trying to be -- that

would probably pass our rigor here, with our rigor in the last couple of days.

What we have found in MBSAQIP is that the data are essential, but it's not enough. So even though we -- the participants of MBSAQIP have these measures basically at their fingertips, more is needed. And I think that all of us who have done QI realize that. There's much more needed than just whatever the mortality rate, or this risk adjusted readmission rate, or this whatever SSI, UTI rate.

And there's going to -- and I think that that's what STS is figuring out, that it's going to be a composite of more than just one thing. It's going to be a group of things. Like theirs is 11 things.

And these accreditation things are getting us potentially to that next level of getting all these pieces together. So whether it's these metrics that we have, or it's some

kind of structural thing of, I don't know, some communication thing. Or an experience, that's the potential here.

And so I'm asking maybe Dr. Gunnar and Dr. Fleisher, that is this something that we can do in this committee, if we're a standing committee for two or three years, to identify that as a potential gap in our scan of the environment of this is a way to potentially raise that next level, so a next level within surgery.

CO-CHAIR FLEISHER: So I'm going to actually ask the NQF to comment. Because if you remember as we were charged yesterday, we were specifically told that the reason we went to standing committees, is to identify gaps and to go out and solicit developers.

And what you're hearing, I think very clearly, and I think that was said very nicely is, this is an area where there is a huge gap. And I think uniformly, the committee

believes that improving bariatric care, can I say that for the committee, would be a thing that would be outstanding.

MS. WINKLER: Speaking more to Cliff's question, which I think is a little bit more global, and that is the fact that you're putting out an idea, and you're really like to get feedback on what people think of where measurement could or should go. You've gotten it from this group.

And one of the things I'm going to do, is in the report that this goes out, and you can help me write the section if you'd like, is to actually highlight that as a discussion area that you guys have considered and are asking.

And we can specifically call it out as something you're really like to see member and public comment on, in terms of how they see the next generation of measurement evolving and perhaps get some useful feedback. So we can use the process to help along that way.

CO-CHAIR FLEISHER: And remember, that's a really a really important of member and public comment will be critical. Because -- and that will hopefully come back to this committee when we get those comments. Not just up the chain within.

DR. BURSTIN: So yes, I just want to add that completely, Cliff, that's the reason we made you guys a standing committee. It doesn't just have to be measure you know, gaps, but also information, measurement science gaps.

What do we need to know to understand how benchmarking relates to measurement, or whatever the case may be. We can raise it up beyond the surgery committee, or we can have it be a surgery specific kind of initiative. All that would be exciting.

Just a couple of thoughts on the issue of structure measures because I think this is a really important question, really

back. We have not to date endorsed measures of accreditation for largely many of the reasons that Barbara outlined, of whether they're within scope or not.

I will tell you that we've had many discussions over the years about for example should maintenance and certification potentially be a composite that has different elements within it. Those are the kinds of things that have come forward.

So when this measure came forward, my feeling was I really wanted this Committee to take a look at it because I thought it was an important discussion for all of you to have. I will say that again, as part of the work we've done to date as was pointed out by Collette, we do have a hierarchical preference for outcomes.

Structural measures and actually went back to the evidence task force. But again, specifically says structural measures are appropriate primarily when they are very

well established structural process outcome relationships. And when it is not feasible to directly measure the outcome or processes.

So it still feels like that is a, okay we'll do that if we can't get the other one. So I think in general, the preference would certainly be for the outcome to bariatric surgery complications, pass through you know, whether there's leaks or other really important issues that should come forward. Those would be absolutely welcome I'm sure by the broadest array of multiple stakeholders at NQF.

And I think if you wanted to pursue the one around accreditation, I think the key is really to the specifics of what are those seven elements? Can they be reliably measured? And is that really something you think would have a strong relationship to outcome?

CO-CHAIR GUNNAR: Barbara?

MEMBER LEVY: Just one other point

about accreditation, and that's for how long is it now. You know a surgeon leaves one group and goes someplace else. The team changes.

I just have problems with that as a measure because it's not a static thing at all. It requires a lot of validation in terms of audit. And I just think looking at the outcomes is going to get us closer to where we want to be.

And Cliff, I totally agree with you, that in terms of implementation, of quality improvement, we've got a long way to go. And benchmarking is step one, that we know we need to do. But I'm not sure that the accreditation gets us where we want to be.

CO-CHAIR GUNNAR: Dr. Moss.

DR. MOSS: So I agreed strongly with what Cliff articulated. There is a significant gap, and being able to close that would be a step forward in surgical quality improvement. I just want to advocate that we

need to do something more sophisticated then just a yes/no answer.

I mean if we're going to put these composites together with the elements of accreditation, we've got to put them together in a way that provides something more to the public then just yes or no. There has to be some sort of graded or comprehensive assessment to quality that people can use to compare one institution against another.

MR. MORGAN: If I could make a comment.

CO-CHAIR FLEISHER: John, did you want to say something before we move on?

MR. MORGAN: Yes, I just you know, I appreciate all the comments. And I just want to emphasize again, that bariatric surgery is very early on in its quality improvement process.

And I love all the different thoughts involved here. But we have to walk

before we can run. And to that end, what we know about accreditation is currently that it does make a difference for those outcomes.

Which one of those elements makes the biggest difference? That's area for further investigation. We're not as far along as other fields.

What we do know, is that the accreditation process has helped in terms of having better outcomes. And I would love to have more opportunity to refine the measures, particularly around accreditation, as well as readmission and move forward.

I am concerned about focusing in on specific outcomes. For leak for example. That's less than one percent of an occurrence. If we're really looking to move the needle of quality improvement, having the elements and tools in place to perform the quality improvement, is bedrock.

And that's where we view

accreditation as being so important to accomplish those tasks. Future inquiry will let us know what elements of inquiry -- of accreditation will make a difference. But right now, the only thing we do know is that accreditation makes a big difference in outcomes for patients.

But we really do appreciate the opportunity to present to you all. And we will be resilient and we will see what we can do to move the measures forward. Thank you very much.

CO-CHAIR FLEISHER: Thank you. Okay, we are going to attempt to get through some -- at least the composite measures.

MS. WINKLER: No, no, no. We need --

CO-CHAIR FLEISHER: No, we have the improvement and the status of surgical wounds.

MS. WINKLER: Right. If you'll recall, yesterday we didn't have the lead

reviewer here for the meeting, and we delayed this measure. So is Debra Deitz on the line?

MS. DEITZ: I am.

MS. WINKLER: Hi Debra. Thank you for bearing with us and coming back again today. So we've had overnight members of the committee review the measure, so that we can have a more fair discussion of it at this point in time.

So we'll give you a couple minutes to introduce the measure, and then we'll proceed with the discussion.

MS. DEITZ: Okay, can I just -- we weren't sure that you were going to deal with us right first thing. And I just would like to see what other members of the team are also here, of the measure developer team.

So Caroline, are you on the line?

MS. GALLAHER: Yes.

MS. DEITZ: Okay.

MS. GALLAHER: Hi, this is Caroline Gallaher. I'm the CMS lead for the Home Health

Quality Reporting Program.

MS. DEITZ: Okay. And hopefully members of the Acumen team will also be joining us. So we can get started, and we'll expect them to join us.

So the measure that you're looking at today is as someone said yesterday, you needed to put on your home health hats. And I think that's really important.

This is looking at the improvement in status of surgical wounds in the home health setting. Specifically how many episodes of care were there in which the patient had a better status of surgical wounds at the end of their home health stays, then they did at the beginning of their home health stay.

And if calculated as most of the home health measures are, on data that comes from the OASIS data set, which is the standard data set that home health agencies need to collect. It's part of the conditions of

participation, that they collect these data as part of their comprehensive patient assessment.

And it's the information on the surgical wounds -- so they are actually recording the status of the wound as healing, or partially granulating, or not healing, as part of the measure. And while they are recording the patient's status as part of their normal clinical practice.

The CMS, Medicare, Home Health Compare website, currently reports a total of 22 outcome in process measures for Medicare and Medicaid patients, so that consumers can review and compare agency performance. And this outcome measure is the only measure related to surgical wounds that is publically reported.

According to the date that we've -- the most recent data that we've looked at, about 25 percent of all the home health patients had a surgical wound. And about 13 percent of

patients showed an improvement in their surgical wound during their home health episode.

And there are lots of home health wound care treatments that are known to improve wound healing, such as keeping the wound clean and dry. And avoiding activities that will cause problems for the wound, skin torsion or tension near the wound. Educating patients about lifting restrictions, and nutritional intake, and what's the signs of wound problems that they need to report.

The -- we looked at the data from the 2011 measurement period, and the 2013, or 12/13 measurement period, and we saw a change in the main risk adjusted performance rate on this measure. And increase from 86.2 percent to 87.9 percent for the agencies in which there were at least 20 valid episodes, which is what we report.

And because of the high prevalence

of surgical wounds in home health patients, and because there are agency practices that are associated with improving that outcome, CMS believes it is important to continue publically reporting the measure.

And that's my presentation.

CO-CHAIR GUNNAR: So this is a discussion of an established metric and it's maintenance, so outcome measures. Dr. Markman.

MEMBER MARKMAN: I just put this together since last night. But this is a significant measure. And it's -- it hasn't been reviewed since 2009. This is the first time it's come up. And just -- I want to read, I'm going to ask you to help me with the numerator and the denominator because we're going to look at the rationale on our first vote.

And it says that the numerator is the number of home health episodes of care where

the patient has a better status of surgical wound to discharge, compared to the start of care. And the denominator is all home health episodes of care in which the patient was eligible to improve.

So I mean my -- I'm not really sure what you mean by eligible, but then most wounds will heal and the what I'm concerned about, or I want you to explain, is why only 13 percent of those wounds have gotten better. Because most of them -- wounds heal. I would say would be 100 percent with the right care.

So that's my first question. And I'll give you an opportunity to answer. Why from the body of evidence, that we came up with only 13 percent?

MS. DEITZ: Well I think that the first thing is your question about what does it mean to be eligible to improve. And that means that when a patient has a surgical wound at the time of admission, and it is already starting

to heal, or al -- you know fully granulating, then we would not say that they had an opportunity to improve, because the -- they could no long -- they could not be rated any higher at discharge.

So I believe that we submitted some data on a attachment that shows the number of wounds and how they fall out, but in terms of how many are actually eligible to improve.

MEMBER MARKMAN: For the committee, we're talking about millions of wounds. We're talking about four, on the episodes, there's 4.8 million episodes, and almost 3.9 wounds.

So it's -- I'm mean, I'm just looking for you know, I mean I see -- listen, I think that the -- that as part of the wound care team, you contribute -- your contribution is significant.

But in terms of finding the rational for the healing, which is what we're going to

vote on, I see you know, and I have that graph in front of me. I just -- I see 13 percent. And this has been ongoing, so --

CO-CHAIR GUNNAR: Dr. Dutton?

MEMBER DUTTON: With an unfortunate amount of experience in the wound business, I think it would help the committee to understand what kind of wounds we're talking about. Obviously a closed surgical wound where it's sew up, the skin is closed and it's granulating, would not be included in this.

But are we talking about an open surgical wound, for example where it was superficially opened because of a superficial infection and we are now waiting closure. Or are we talking about decubitus diabetic foot ulcers and at that kind of wound?

MEMBER MARKMAN: That's the question because it's not defined, you know the types of wounds.

MS. DEITZ: I'm sorry, but it does

not -- may not appear to be well defined in the actual measure. However, we have specific other measures that refer to and measure the healing of decubitus pressure ulcers, stasis ulcers, et cetera. And this is restricted to surgical wounds that are either healing by primary or secondary intent.

And there's a good deal of guidance that is provided to the clinician based on WOCN guidance as to assist them in determining what the status is of the surgical wound.

MEMBER MARKMAN: So these don't include Wagner's foot ulcers or -- these are just postoperative -- I mean are they open wounds? Or are they infected wounds? Or are they closed wounds?

And I brought this up during the discussion. I'm not really sure about your data set you know. And if you have additional information about it, I mean I actually review a lot of OASIS forms.

But I you know, I just don't have a good grasp of what types of wounds and why such the low rate of healing.

MS. DEITZ: Well I think that it has to do with the number of wounds that are not moving from one status to the next at the time that the patient is discharged from home health. So the status that they can record is either newly epithelialized, fully granulating, early partial granulation, or not healing.

So and again, we provide the agency clinicians with a good deal of guidance about how they determine which of those -- which of those criteria apply. Definitions apply.

MEMBER MARKMAN: I mean this measure has been in place for five years. And you know, we have to show the relationship has a -- did it fluctuate from the 13 percent? Was it better one year? Or has it been kind of stable at 13 percent for five years?

MS. DEITZ: Well I see this is discussed. I apologize because the team members who have been doing the analysis of the responses, and the healing are -- appear to not have joined us yet on the call.

CO-CHAIR GUNNAR: Well let's move forward.

MS. WINKLER: Yes, we can postpone it to another.

CO-CHAIR GUNNAR: Yes, Dr. Yates?

MEMBER YATES: Well I --

MS. GALLAHER: Yes, this is Caroline from CMS. I believe our other team members should be getting on the call in just a few minutes. So if the committee would just bear with us and you know --

MS. WINKLER: Actually we really need to you know, if we need to let you get yourselves together. If that's the case, if you could just hang there and let us talk about another couple of measures, we'll get back to

you when you have your team all together.

MS. GALLAHER: Okay. We weren't sure what time you were going to call us. We weren't expecting to be quite this early. So I apologize for us not being quite ready for our presentation.

But yes, I would appreciate that if you would give us as little -- a few minutes.

CO-CHAIR FLEISHER: Okay, so thank you. We will at the latest, 3:15, because that's when we're essentially finishing.

So Jeff? So we are going to go on to 2561. And who's the discussant? Do you want to take care of this one?

CO-CHAIR GUNNAR: Yes, that's fine. This is measure 2561 STS, aortic valve replacement composite score. This is a new measure being introduced.

DR. JACOBS: Well good afternoon, and we'll do our best to get through as many of the remaining STS measures in the time

available.

The first two measures that we have on the table now are two composite measures. And just to briefly give a history, the STS currently has a composite measure in place for isolated coronary artery bypass grafting.

And this measure is a four domain composite that is composed of 11 national quality forum endorsed measures that are grouped into the following domains. Absence of operative mortality, so that's risk adjusted mortality; absence of major morbidity, and then high quality interoperative care by using the internal mammary artery and appropriate perioperative medication usage. So that's the current CABG composite, the four domain composite.

What we're going to talk about today is an AVR composite and then an AVR CABG composite. The AVR composite is a two domain composite. The absence of operative mortality

and the absence of operative morbidity.

And for all of the composite measures we've brought forward so far, including this one, the absence of morbidity is an any or none, for stroke, external wound infection, renal failure, reoperation and prolonged ventilation. So any of those events would qualify for morbidity as opposed to absence of morbidity.

And we've created a composite for aortic valve replacement surgery based on the structure just described, the two domain composite. And it's captured over a three year time interval because aortic valve replacement is done less commonly then coronary bypass grafting, and to have sufficient sample size, we used three years.

I think I'll stop there, that's a pretty good opening statement about it.

CO-CHAIR GUNNAR: So the question from the front table is, do you -- Jeff do you

have a problem if we take these next two, both 61 and what's the other one, the next one?

DR. JACOBS: 63

CO-CHAIR GUNNAR: 63 together? So AVR and AVR CABG together?

DR. JACOBS: Yes, I think that would make sense from a time efficiency point of view, because the issues are essentially identical.

CO-CHAIR GUNNAR: So Lynn Reede is also discussant on the 63. So we'll -- so a couple of just to frame this I think for others. So there's three parts of this. There is the morbid -- the absence of mortality, so within 30 days, correct? Or is it the whole, any hospitalization?

DR. JACOBS: It's the same operative mortality measure we discussed earlier. So it's the union of 30 day plus discharge.

CO-CHAIR GUNNAR: All right, so

it's the absence of mortality, yes or no, that's one. And then the absence of morbidity in one of those --

DR. JACOBS: Five morbidity domains.

CO-CHAIR GUNNAR: Five morbidity domains, which are all NQF endorsed.

DR. JACOBS: Correct.

CO-CHAIR GUNNAR: So you could have -- you could -- now here's my question. This was one that was left from our phone call. Would you be can -- if you died, so you had checked off on that, would you also be included in the morbidity realm as well?

MR. O'BRIEN: Not automatically. So that morbidity endpoint does not in -- you know, it's a --

CO-CHAIR GUNNAR: Stroke, died.

MR. O'BRIEN: Did any one or more of the following things happen.

CO-CHAIR GUNNAR: Stoke, died.

MR. O'BRIEN: So your death would show up in the death --

CO-CHAIR GUNNAR: Right.

MR. O'BRIEN: Domain and then the stroke would show up in the stroke domain. And they'd both be counted, but the death would only be counted in the death domain, it wouldn't be recounted in the morbidity domain.

CO-CHAIR GUNNAR: Exactly. I just wanted to clarify. And then there's the cumulation, there's the domain which is the additive of those two, is the true or not?

MR. O'BRIEN: That is correct.

CO-CHAIR GUNNAR: So there is actually -- and then those are converted to the star composite for each of those. The absence of mortality, the absence of morbidity and then

DR. JACOBS: And then an overall star and a composite staring.

CO-CHAIR GUNNAR: And then the composite star, okay. And each.

DR. JACOBS: Right, so when it's publically reported, one can go to the website and get the star rating for the whole composite. The star rating for any individual domain of the composite and can drill down even further and get the percentages of the different events as well that make up the composite.

CO-CHAIR GUNNAR: So this applies to both of these and Lynn did you want to say anything else?

MEMBER YATES: No, keep going.

CO-CHAIR GUNNAR: Okay. Dr. Yates -- or who's up, Dr. Dutton.

MEMBER DUTTON: Are the morbidities weighted in some way? Or do they occur at approximately equal incidents, so are they exerting the same weight on the measure?

MR. O'BRIEN: So just to step back, the composite consists of two domains and each domain has its own score. So in a sense it's a composite of composites, because we're taking

a composite of mortality and morbidity, but morbidity itself is a composite.

So within the morbidity domain, that's constructed as an any one or more yes or no. It's a composite end point like you'd have in a clinical trial if the patient died or had a heart attack, yes or no.

So it's any one or more of the morbidities that occurred. So there's no explicit waiting there, it's just basically any one or more happened.

In terms of the weighting, sure --

MEMBER DUTTON: So if one of those occurs ten times more often than the other three, for instance the prolonged ventilation, does that end up driving this method?

MR. O'BRIEN: No, it's a great question. The end points that occur more frequently will tend to have a little bit more statistical influence.

But when we've looked at you know,

different ways of weighting the individual morbidity components, and looked at item total correlations to kind of assess to what extent are they being driven by individual items, this was the approach that a surgeon committee felt you know, although it doesn't have explicit weighting, they looked at kind of the implicit weightings that went along with having it's any or none.

And you know, although certain items did contribute more, for example the prolonged ventilation, it had face validity with the panel that developed the measure. So that's an inherit feature of any kind of composite, and I think you've identified you know, one of the features of the approach.

But it was an approach that was very transparent during the development process. And had to face validity with the development team and stakeholders.

CO-CHAIR GUNNAR: Dr. Fleisher.

CO-CHAIR FLEISHER: I'm still a little confused. Can a patient for the composite score be counted twice or once for -- in other words, if a patient had morbidity and mortality, it's back to yours, does that count once or twice against them, the hospital in the composite score?

MR. O'BRIEN: Well I mean, there are two different domains. So each domain is -- what's the percentage -- each hospital has estimated percentage of patients that die, okay. Now you're also going to estimate the percentage of patients that experience morbidity. Yes the same patient can contribute information to the estimation of both of those by percentages.

CO-CHAIR FLEISHER: But in other words, I'm basically asking almost a survival curve of death or what percentage of patients have death or am I at the composite score, death or a complication, is the composite score -- or

really doesn't have an accurate score?

MR. O'BRIEN: On the morbidity there's not -- so the -- on the morbidity domain, that's the composite of did any one or more of these things happen. In that list, there's five things on the list, death is not one of the things on that list.

DR. JACOBS: In other words this is not an estimate of morbidity in survivors, it's an estimate of all patients that have experienced morbidity. And I think, if for face validity, it makes sense to me because one can die after a CABG with no complications, or an aortic valve replacement with no complications. Or one can die with a stroke and renal failure and prolonged intubation and five reoperations.

So I think those are different forms of death. It still ends up in a horrible situation with death. But it makes sense to track both of those things. But it's clearly

not -- the second domain is not morbidity in survivors, it's morbidity in anybody.

CO-CHAIR FLEISHER: So just, is there any measure that says how many people walked out of the hospital with zero compli -- walked out of that hospital without having a complication or die?

DR. JACOBS: I don't think that's the -- that would be easy enough for us to do, but I think that's not the way this composite is structured.

CO-CHAIR FLEISHER: But that's not -- okay, that's, that was my, okay.

CO-CHAIR GUNNAR: So let me frame it a little bit, because it's -- the way that plays in the -- it comes into play would be in the fringes, right. If I'm a two almost one in mortality, I am a two almost one in morbidity, I could be some two-two, but I could be a one statistically overall because the two added together would put me in the lowest category.

True or not true?

DR. JACOBS: I think that's a Sean O'Brien question.

MR. O'BRIEN: That sounds true. That just --

CO-CHAIR GUNNAR: It's critical math, you can't do that.

DR. JACOBS: I'm a big picture guy. If you want to get to the mathematical statistics, you've got to talk to the PhD over here. I don't want him doing any operations, and I certainly don't want to be talking about statistics.

MR. O'BRIEN: I just say as maybe as an aside, STS does report a measure that's a positive six items, which is any of those four measures of morbidity or mortality where you know, the word composite can have different meanings and be confusing.

So here we're talking about a composite where we're taking a domain score

that exists for mortality, a domain score each participant gets a score for morbidity, and we're basically averaging them together in a -- you know after standardizing them, they kind of have a similar variance as denotes under the age, and they're going to get averaged together. Like you know, one goes into the other, divided two -- separately, separately.

CO-CHAIR GUNNAR: But each domain is rated separately. The mortality domain, the morbidity domain and then the composite domain.

MR. O'BRIEN: The composite is an average of the two domain specific scores, or it's literally an average, one plus the other divided by two.

DR. JACOBS: That's the math that I was getting at.

MR. O'BRIEN: And just to you know, they're all estimated together in a big multi-variant model where all the end points

are analyzed and estimated simultaneously, and combined in fancy ways. But the math part is simple, it's just what we're estimating.

CO-CHAIR GUNNAR: So the second part of the question, the question we asked on the phone call, which you were great to take, had to do with volume number as to the application of this composite with regard to AVR and AVR CABG.

And they have -- given you cite that report, CABG numbers are significant, right? They're almost -- they're triple digited, it's 100 I think is the average that --

DR. JACOBS: Yes, most common operation we do.

CO-CHAIR GUNNAR: Right, it's a common operation. The question here from a statistical point of view, and again, Sean, what you know, when we're looking at the average number of AVRs per facility at less than 30, it's 28-30. And the number of AVR CABGs coming

in around 18, and not all facilities reporting.

So the -- it's interesting, there's a group that does AVRs, and then there's a group that does not do, small number, that actually doesn't do AVR CABG. The question is --

DR. JACOBS: You mean publically report?

CO-CHAIR GUNNAR: Publically -- no, well reported to your registry.

DR. JACOBS: Well, so I don't mean to interrupt, but --

CO-CHAIR GUNNAR: No, please do.

DR. JACOBS: First of all, anybody who does any of these and participates in the data base reports all of these operations. The STS data base works by every cardiac operation being done at a hospital is reported. And that's verified at the audit process to comparison with the operative logs.

So if a center is performing AVRs or performing AVR CABGs, those show up in the data

base. That's one fact. Another fact is that the sample size issue is a real sample size issue. Because even though AVR and AVR CABG are you know, both within the top five operations done by adult cardiac surgeons by volume, to get a significant sample size -- to get a large enough sample size in order to truly differentiate outliers in a meaningful fashion, we aggregate three years of data for this composite as opposed to one year of data for the CABG composite. That's the way we address the small sample size.

CO-CHAIR GUNNAR: So just to finish up that, just the point I was trying to make was that you had 970 in the last reporting cycle that reported having in -- of your reporting sites, 970 reported doing an AVR, but only 933, if I have the numbers correct, reported doing an AVR CABG. That was the point I was making.

Is that not all sites did an AVR CABG. 40 more sites, or rough 30 some, did an

AVR, but not a combined procedure.

DR. JACOBS: Yes, I think that's believable.

CO-CHAIR GUNNAR: Right.

DR. JACOBS: Some smaller programs may not have a higher claim of valve replacement surgery, and certainly of valve CABGs. They may just be focusing on you know CABGs alone, and referring those other more complex operations to other hospitals nearby.

CO-CHAIR GUNNAR: So I'm getting beyond the questions, because I hear Reva in my left ear. And so we want to go and vote -- is there any more discussion on this first, just the rationale support, the health outcome?

And so, let's -- and this goes to -- remember, we're voting on both of these together.

CO-CHAIR FLEISHER: Well actually, you probably need to separately vote, or can you --

CO-CHAIR GUNNAR: Yes, how can we do this? No, no, no, we're just. Can we do 61 and 63 Helen, together? That was the question.

DR. BURSTIN: Yes, I mean if people feel comfortable that whatever they would vote on, would be for the other, let's just go ahead and we'll put them together in the report like that.

CO-CHAIR GUNNAR: The developers are fine with that, so.

DR. BURSTIN: Okay, that sounds fine.

DR. JACOBS: We're fine with that.

CO-CHAIR GUNNAR: Okay, so we're voting on both 61 and 63, and go ahead.

MR. SANCHEZ: Voting will now begin for 1A evidence. 1 is for yes, 2 is for no. The timer starts now.

CO-CHAIR GUNNAR: Yes, Dr. Grover is abstaining, and we lost Dr. Saigal and Dr. Asher. So we're at 21.

MR. SANCHEZ: 21 for yes. Zero for no.

CO-CHAIR GUNNAR: So next, and I think Dr. Reede do you want to talk about-- well, okay go -- yes, please.

MEMBER REEDE: For the performance gap, the issue looking at this particular measure was that some of the lower performing organizations didn't show up in the data. So there's still opportunity to improve there as a performance gap.

DR. JACOBS: I think if we look at the distribution of star ratings for first of all the AVR composite, we have a distribution of about three percent one star, 90 percent two star, and six or seven percent three star. So that's all comers.

Now unlike the isolated CABG composite where we have a very nice spectrum of publically reporting one star, two star and three star programs, I think as of now, the AVR

public reporting is a little bit more skewed. But that I think is in part because it's a newer measure that we've just started to publically report, and it's a few years behind on the adoption curve of public reporting.

So I think ultimately, the same phenomenon will happen with a little lag to what happened with CABG. We'll continue to get more and more sites to publically report, and then the publically reported website will have a distribution very similar to what's the distribution in the actual data base.

Does that answer your question? Thanks.

CO-CHAIR GUNNAR: Yes, I mean I might take a different perspective, which is the entire composite measure is put together to establish a performance gap with regard to the outcomes, the collective outcomes, correct?

Because before the numbers were so small on mortality, and each and every one of

the individual morbidity fields, that in fact the cumulative aggregate is meant to establish a performance gap. The question about how that's applied is really a threshold that is made -- is determined by the society, not -- and your aspirational goal was that everybody be a two or three star at the get go.

So the measure really I don't think is about -- or this domain of performance gap is actually I think, for me is high to moderate, because it actually expands on the cumulation of a small points of light, right?

So let me see if I've got that correct. Do I have that correct Reva? The very first thing?

MS. WINKLER: Yes, that's fine. I mean one of the advantages to composites, is that it does give you an opportunity to look at the data differently and perhaps open up variation in gaps that didn't exist in the individual components.

CO-CHAIR GUNNAR: Any other discussion? So I think we can vote.

MR. SANCHEZ: Voting will now begin for 1B, performance gap. 1 is high, 2 is moderate, 3 is low, 4 is insufficient. Timer starts now.

CO-CHAIR GUNNAR: Vote again. We need one more. Oh, there it is, okay.

MR. SANCHEZ: We have 12 high, eight moderate, one low, zero insufficient.

CO-CHAIR GUNNAR: So I think we've -- I think this has been answered in its components, but I think -- any other discussion on the priority of this? I think we can go ahead and vote.

MR. SANCHEZ: Voting now will begin for 1C, high priority. 1 is high, 2 is moderate, 3 is low, 4 is insufficient. The timer starts now.

21 high, zero moderate, zero low, zero insufficient.

MS. WINKLER: I just want to point out that because this is a composite measure, there's an additional category -- criteria under 1, and that is for composite specifically. And that really speaks to is the composite structuring, what the components that are put together, are they -- does it make sense? Is it logical? Does the quality construct have meaning as opposed to just sort of random throwing things into the pot.

CO-CHAIR GUNNAR: So my perspective is, is that it in fact is thoughtful and speaks to the significant outcomes that one would look to in evaluating the overall quality of the surgical program under the evaluation.

My issue here and what I have mentioned numerous times I think, is the issue regarding the amount of noise that can exist when you have low numbers where an event can cause a big swing in impact on your overall score. And you know, that gets smoothed away

substantially when you have the higher CABG numbers.

And fundamentally I don't know why this wouldn't all be in one. It's about the program, and why shouldn't AVR or AVR CABG and CABG just be put together. That's the first question.

And then the second question is, and we were reassured on the phone about -- during a conference call, about the movement between one category to the next, that has -- but the data on that is relatively small, for particularly AVR CABG.

So I don't know that there's, from my impression, that there's great data to argue against the thought that I could in fact in one cycle be in reporting if I was a lone small program, that I could do my 18 or 30 AVR, AVR CABGs, have no mortality or morbidity associated with that on one cycle, yet the next cycle be absolutely terrible but really have

the same program, and no difference other then the fact that I had some events accumulate in one reporting period because I'm a low number.

So I'll stop there.

DR. JACOBS: Yes, so I'll try to address a few of these points. I think the issue of sample size is an important issue, and that is the reason why this is reported over a three year time interval instead of a one year time interval. And that is why confidence intervals are use.

And with the confidence intervals that we used and with the three year reporting interval, the sample size is large enough that we can say that there's a 97 and a half percent Bayesian probability that a one star provider or a three star provider is different from a two star provider. And that's regardless of programmatic volume when one uses three years.

Second, your discussion about creating a composite of composites. And

that's the next step in all of this. This is an exercise in building blocks.

So first we created all of the individual measures and put them through NQF endorsement. That allowed us to then create composite scores for each of the lesions.

The next step in this is a composite of composites where one can have a programmatic composite that would take into account mitral valve surgery, aortic valve surgery, AVR CABG and CABG. And that composite of composites would have the added value of potentially creating sample size large enough for individual provider public reporting to complement the programmatic public reporting.

MR. SHAHIAN: And this is Dave Shahian. The only thing I would add to that is that although we are building that composite of composites, in fact patients are having one procedure.

So I think from a patient

perspective, they not only want the overall assessment of a program, but they want to know how did that program do with my planned procedure? Because there are in fact programs that do proportionally a lot more valve surgery for example.

And it is not necessarily true that a program that excels in CABG surgery is going to excel in valve surgery. So the ability to drill down to the level of individual procedures we think is important for the consumer.

DR. JACOBS: Yes, there's published literature that shows that a program that's excellent with operation A in cardiac surgery, may or not be excellent with operation B. And I think Dave's point is a good one, that an individual patient is going to be having an aortic valve replacement, and he wants to go to a place that he knows is good at that operation.

CO-CHAIR GUNNAR: I just want to

make sure we touched on all the arguments. Any -- Dr. Reede anything else from you?

Oh, I was just asking if -- I just wanted to make sure we touched on all of the -- we put all of the arguments out there.

MEMBER REEDE: We did, thank you.

CO-CHAIR GUNNAR: Any -- Dr. Temple?

MEMBER TEMPLE: Just sort of a -- I guess it's a phase validity and it fits this composite thing, it -- structure area. So as a surgeon it seems like if you have a -- if you have two complications, your score is the same as you have with death and a complication. Is that right kind of for the composite?

MR. O'BRIEN: No, so it doesn't work out that way.

MEMBER TEMPLE: Okay, good.

MR. O'BRIEN: So in terms of the weighting between mortality and morbidity endpoints, there's not any really great

rational objective way to determine the weighting.

MEMBER TEMPLE: Right.

MR. O'BRIEN: So the approach that was adopted by the committee, was to take the two domains and weigh them equally, recognizing that we wanted to avoid a situation where one of the two domains dominated because of having one of the scores that has a very large standard deviation. You take a measure that has a very large standard deviation and average it with something that has very little variation, the overall average is being dominated by the one with the larger variation.

So they were first standardized to have the same standard deviation, then averaged together so you could say that they were being weighted equally. Of course when you do that, those weights have implications and it turns out that so, you know if you said what's the impact on that score if I increased my avoidance

morbidity by one percentage point, how much is that going to change my composite score? Would that change it as much improving my mortality by one percentage point?

And the answer is no. The impact of improving on mortality has a substantially larger impact on improving the overall composite then the one percentage -- the equivalent percentage point difference in morbidity. And I don't recall the exact numbers, but it's a ratio that I can pull off this.

MEMBER TEMPLE: No, no. I just -- when I read, I just wanted to make sure that that was the case.

CO-CHAIR GUNNAR: Any other discussion? Get ready to vote.

MR. SANCHEZ: Voting will now begin for 1D, composite. 1 for high, 2 for moderate, 3 for low, 4 for insufficient. Timer starts now.

We have 13 for high, 8 for moderate, zero for low, zero for insufficient.

CO-CHAIR GUNNAR: So we'll run through reliability and by validity and feasibility fairly quickly. We've been through this before. If there's any discussion, just stop and raise your hand.

So reliability? Time for a vote.

MR. SANCHEZ: Voting will now begin for 2A, reliability. 1 for high, 2 for moderate, 3 for low, 4 for insufficient. The timer starts now.

We have 17 for high, four for moderate, zero for low, zero for insufficient.

CO-CHAIR GUNNAR: Validity, any discussion? None, go for vote.

MR. SANCHEZ: Voting will now begin for 2B, validity. 1 is for high, 2 for moderate, 3 for low, 4 for insufficient. Timer starts now.

Still waiting on two votes, so if

you could please resubmit.

CO-CHAIR GUNNAR: Yes, vote again. We might have missed -- Dr. Temple stepped out, so.

MR. SANCHEZ: We have 15 for high, five for moderate, zero for low, zero for insufficient.

MS. WINKLER: And because it's a composite, there's one more criteria under scientific acceptability. And this is really empirical analysis that supports the composite construction. Really responds to some of your questions about the relative frequencies of the various contribution to the composite.

CO-CHAIR GUNNAR: So there are two publications, one for each of these measures in the Annals of Thoracic Surgery, which is a society, STS journal, that walks through the construction and the validation of the composite measure for both of those. For both 61 and 63.

There is no other evidence outside of that to my knowledge, other than the extensive amount of evidence that exists for each of the components.

DR. JACOBS: Yes, I would agree with that. I would just clarify by saying that the Annals of Thoracic Surgery is a free institution from STS. They're separate institutions. STS uses the Annals as their official journal, as does another surgical organization, the Southern Thoracic Surgical Association.

And papers that are presented at the STS meeting are submitted for publication to the Annals, but not all of them actually get published. They go through a separate, very rigorous peer review process, which is completely separate from STS.

So these papers have certainly gone through a separate peer review process that's outside the scope of the Society of Thoracic

Surgeons. And it is correct that these are the papers that have been used to publish and provide the scientific basis of these composites, because they are new tools, and these are the first publications about them. But I think they provide ample evidence.

CO-CHAIR GUNNAR: Dr. Reede, any additional comments?

MEMBER REEDE: Just that correlation was done really at the participant level, so that they looked at on the Pearson and the Spearman, they did that sort of analysis of the correlation between the components.

MR. O'BRIEN: I'm sorry, I didn't hear the end of your comment, and I'd like to respond if it's something I should have picked up.

MEMBER REEDE: Oh, there's really nothing to respond. It's just saying that you actually did the correlation between the different measures to bring the composite to --

to validate the composite.

MR. GUNNAR: Any other discussion? Then we can go ahead and vote on 2D.

MR. SANCHEZ: Voting will now begin for 2D, composite. 1 for high, 2 for moderate, 3 for low, 4 for insufficient. Timer starts now.

We have 13 high, seven moderate, zero low, zero insufficient.

CO-CHAIR GUNNAR: Feasibility, any further discussion? Ready for a vote.

MR. SANCHEZ: Voting now will begin for criteria 3, feasibility. 1 for high, 2 for moderate, 3 for low, 4 for insufficient. Timer starts now.

CO-CHAIR GUNNAR: We need -- are we good? Okay.

MR. SANCHEZ: We've got 12 for high, eight for moderate, three for low, four -- I'm sorry, zero for low, zero insufficient.

CO-CHAIR GUNNAR: And usability

and use. Any further discussion? Let's carry on, vote.

MR. SANCHEZ: Voting will now begin for criteria 4, usability and use. 1 for high, 2 for moderate, 3 for low, 4 for insufficient information. Timer starts now.

We have 17 for high, three for moderate, zero for low, zero for insufficient information.

CO-CHAIR GUNNAR: And so the overall suitability for endorsement, any further discussion on these two measures? We'll vote on these collectively then.

MR. SANCHEZ: Voting will now begin for overall suitability for endorsement. 1 for yes, 2 for no. Timer starts now.

CO-CHAIR GUNNAR: Please re-vote.

MR. SANCHEZ: 20 for yes, zero for no.

CO-CHAIR GUNNAR: So resounding yes for both 61 and 63.

CO-CHAIR FLEISHER: Okay, we're going to go to 0734, 0113 and 0456. If we can get this done in 15 minutes.

DR. JACOBS: Sure.

CO-CHAIR FLEISHER: And does anybody have an objection to doing all three together, the databases? We can vote independently for endorsement if anybody feels they need to be separated out.

DR. JACOBS: All right, so what I'd like to do first is talk a little about the penetration of each of these databases. So these three measures successively are participation in national database for pediatric congenital heart surgery, then for adult cardiac surgery and then for general thoracic surgery.

We know for pediatric cardiac surgery, that there's 125 programs in the United States that do pediatric heart surgery. And currently 108, or 86 percent of those

programs, participate in the STS database.

Those 86 percent of programs represent over 90 percent of the cases because all of the large volume programs participate. However, 14 percent of the lower volume programs do not participate in the STS database.

In general thoracic surgery, -- well first, in adult cardiac surgery, we've talked about multiple times already, that we have 90 to 95 percent of the programs and over 95 percent of the operations. The programs that we don't have in the majority are VA hospitals, military hospitals or Kaiser hospitals.

And finally, in general thoracic surgery, the denominator is a little bit more allusive, because general thoracic surgery is performed by thoracic surgeons and also by general surgeons. The STS database has welcomed the participation of general surgeons

in the thoracic database along with board certified thoracic surgeons.

And there's no actual source of data that can provide a true denominator, but my best estimate of penetrants of the thoracic database would be that we capture 30 to 40 percent of the thoracic surgery done in the country.

CO-CHAIR FLEISHER: So --

DR. JACOBS: The only other thing I would say about these measures, and the topped out concept, because clearly at the concept of topped out's going to come up at least for the adult cardiac one, and maybe for the pediatric one, although I think at 86 percent is probably not topped out.

One should not underestimate the value of having this measure exist when one goes to meet with a middle manger in the hospital and asks for allocation of resources to support the existence of this database and the personnel to collect the data for the database.

I've had to go to hospital administrators in my own hospital and in other hospitals and explain to them why this is important. And when we can say that it's an NQF endorse measure simply to participate, that is a tool that gets a middle manager to authorize the writing of the check to pay for the salary of the person to enter the data. So --

CO-CHAIR FLEISHER: So the key question is, you have multiple outcome, which are endorsed by NQF, and then which could argue for the same thing. So why do you need the database itself?

DR. JACOBS: So -- well I think that question has a different answer for adult cardiac, general thoracic and congenital. Because in congenital there's really only two NQF endorsed measures. This and reporting mortality stratified by STAT categories. Those are the only two congenital measures that are endorsed that involve the STS database.

In thoracic there's somewhere between that and what we have for adult cardiac. And I think it makes a cleaner argument with the hospital administrator to have these measures there then just to say we need this to capture the data for this endpoint measure. Sometimes the act of participation being an endorsed measure just gets the guy to write the check easier.

And there's data that shows that the very act of participation in and of itself improves quality. And we've published that. There are papers that have published that show that database participation alone leads to improvement in quality.

CO-CHAIR FLEISHER: Why don't -- the question is going to come up whether we're going to have sufficient time. So why don't while they're here, we may have to continue this on the call.

DR. JACOBS: Okay.

CO-CHAIR FLEISHER: But why don't we use the next, if it's okay, 10 minutes, to have a discussion with the developers while we're all in the room, which is very different then on the call potentially. And then we could potentially vote on the call about this.

MEMBER HANDY: Well I just wanted to make the point that many of the foregoing outcomes that we had talked about, especially with regard to the burden of reporting, is predicated on the presence of these databases. Without these databases are essential to the execution of those measures.

CO-CHAIR FLEISHER: True, but do you need the databases to be endorsed to actually -- there are other databases utilized for outcomes, correct? In which the database participation is not endorsed. So that's not a definite linkage to my knowledge. Helen?

DR. BURSTIN: That's correct. I will say these issues, these measures have come

up all the way to the level of the NQF Board of Directors because this became an issue a couple of years ago. Actually right around the time Jeff and his group submitted the pediatric surgery one.

We also had one from ACR, on participation and radiation dose registry as well. Those are the two new ones. And there was a sense by the board, and we've really been trying to follow this very much along the lines of the comment I read earlier from the evidence task force report, that you know, these kind of structural measures should really only be there if you in fact don't have other measures that get at the outcomes.

And so I think you know, the question of participation and had drive participation and how it relates to the use of an endorsed measure, I think is something I would ask you to consider, fully noting I'll point out last, it just seems like there's one

of these a week, and there is.

But just a few weeks ago, the committee, the safety committee, in fact did not put forward the endorsement of the radiology, radiation and safety measure. Feeling like it was time to move forward to get to measures of radiation safety rather than measures of participation and registry. I get that right around radiation safety.

CO-CHAIR FLEISHER: So, Robert?

MEMBER CIMA: I just wanted to ask and get a clarification from the developer because we're grouping these. I know on the pediatric one, there isn't a national one. But in the other ones you define national, regional and local participation.

So how do we define that? I mean how is that distinguished then if you know, a new -- if the State of New Hampshire decided to form their own regional -- I just wanted to -- I know in the pediatric one such a thing doesn't

exist, but how do you define these? And so it doesn't necessary need to be STS then, it could be some other regional thing.

DR. JACOBS: I think it's designed by the creation of a multi-institutional registry with broad enough participation that the registry would be useful for benchmarking performance improvement, quality improvement and public reporting.

So it doesn't have to be the STS database that can provide those functions. But a certain number of institutions would be necessary in order to be able to provide those functions. I don't think there's any standard definition of what that number is. I think it's probably a little like pornography, you know it when you see it.

MEMBER CIMA: And how about for specifications of a measure? We need to know it.

DR. JACOBS: But I don't think --

that's a random number. I don't think anybody could say that a multi-institutional database is useful when it has 10 or 11 or 14 participants. I think that number doesn't exist. There would be no database to come up with it.

I understand that we need numbers to endorse things here whenever they exist. But the best you can do there is have a panel of experts look at the database and say does this database really meet the requirements of being able to perform these functions.

CO-CHAIR FLEISHER: Thank you. We're not going to vote today and we're probably going to have to take these separately on the call. Rick?

MEMBER DUTTON: Two quick questions. The missing practices in the adult registry are systematically missing in the DoD/VA. Do they have a registry?

DR. JACOBS: The VA does have an

adult cardiac registry. Dr. Grover could speak to that because he was one of the founders and I mean he kind of created that ball game. So yes, they do have their own.

MEMBER DUTTON: And is the Northern New England Collaborative still functioning? And is that another competing registry? Or not competing, but collaborative registry?

DR. JACOBS: No, I honestly don't know the level of function of the Northern New England Registry right now myself. I couldn't comment on that. Maybe Dr. Grover or Dr. Shahian could.

MEMBER DUTTON: And then the second question, how many of these registries of these three have been nominated as QCDRs?

DR. JACOBS: The adult cardiac one.

CO-CHAIR FLEISHER: A.J.?

MEMBER YATES: Yes, and to follow up that question, and I had brought this up in the workgroup. The repository of this data is

going to be STS. I mean you're going to the be the steward of this measure.

So the question is can you leave room for the reporting by VA hospitals that they are participating in the VA registry. Do you leave room for perhaps hospital systems that choose not to pay the money to belong to STS, but run their own registry, such as the Kaiser system or something to that effect? And I have a follow up question.

DR. JACOBS: Well I think the VA database would meet this requirement.

MEMBER YATES: But will you leave room for them to report to STS so that if you're the steward of this -- who keeps score, is the question.

DR. JACOBS: I don't understand what you mean by leave room.

MEMBER YATES: Because when we talked about this on the conference call, the -- being involved with the STS registry is

adjudged by the fact that people send in their scores. That's how you know people are involved.

DR. JACOBS: Right.

MEMBER YATES: But there isn't some questionnaire that goes out to the universe of hospitals doing cardiac, thoracic and pediatric surgery -- cardiac surgery, asking them if they belong to a registry.

DR. JACOBS: Correct. I don't think that that exists. I think that in --

MEMBER YATES: It's self defining that it has to be your registry if it's only recorded by them putting it out.

DR. JACOBS: No, I don't think so. I think that the VA registry could similarly say that this is a list of hospitals that participate in our registry and meet this requirement should they choose to do so. And that would be --

MEMBER YATES: But they could be

self-defining then is what I'm saying.

DR. JACOBS: Sure.

CO-CHAIR FLEISHER: Rick did you have a comment that you think can help clarify?

MEMBER DUTTON: Yes, specific to that, Dr. Yates, so the score would be kept by the user of the measure. So CMS would determine that DoD registry meets the requirements of this measure.

MEMBER YATES: That would satisfy me. Because the way it's written --

DR. JACOBS: That's a good answer.

MEMBER YATES: The way it's written, it's written as your group seeing that people have put in the data, then they get a score saying that they participate.

CO-CHAIR FLEISHER: Thanks. Larry.

MEMBER MOSS: So I understand that we're looking at all three of these measures together. I just want to make the comment that

I think the argument is most compelling on congenital heart surgery because there are not readily available outcome measures.

And the participation outcome link is very, very strong. And so we have the most at stake to not authorize that one.

DR. JACOBS: I would agree.

CO-CHAIR FLEISHER: Great. We'll discuss that on the call. John?

MEMBER HANDY: So I wanted to make a point about the general thoracic database which I had quoted some literature on yesterday, saying that it's actually not very representative of the national experience, and has far superior outcomes to the national experience. So it is linked to better outcomes.

DR. JACOBS: Yes, you got that.

CO-CHAIR GUNNAR: Yes, for that reason I wasn't going to push against including them three together. I think you have to look

at the thoracic one separately, it's a different animal.

As far as the VA's concerned, I mean we were in fact, we were -- Dr. Grover can attest to the fact that we began before the STS, but they look remarkably similar somehow. And you can say the same about ACS and NSQIP.

So I think Dr. Yates' point is well taken. Is this is about having a clinical database with certain features, the way I look at it, that answers certain questions and collects certain data.

And from CMS' point of view, which is one of the things in a time frame, which is the other part of the discussion I would like to have at some point, you know, what that time frame would be, because in relationship to quality improvement, the closer you can get the time frame from the events that are occurring at the facility, the tighter-knit the afferent and efferent loop of this are.

So those are my points.

CO-CHAIR FLEISHER: Great. I think -- thank you for allowing us to defer the vote, because I think it would take a lot more discussion. But I think you've at least heard some of the concerns and perhaps the two individuals who are reviewing this, or three individuals that reviewed this, if you have other questions based upon this, or who can help define that even better on the call, would that be acceptable that the reviewers continue a dialog with you?

DR. JACOBS: Absolutely. Yes, feel free to email me. Jeffjacobs at msn.com. Send me an email and we can set up a phone conference and involve Dave Shahian if we wanted to. Or we can have an email dialog or whatever's easier.

MS. WINKLER: You have a conference call set up as a post-call for this committee to meet on June 9 from 2:00 to 4:00 Eastern. So

we'll see what we can do. Although that agenda's starting to look kind of packed.

MEMBER JARRETT: It would be good though if you could combine it with that conference call. If not, we'll do it separately.

CO-CHAIR FLEISHER: The call is June 10, which is Tuesday. Okay. Why don't we go back to the -- thank you.

DR. JACOBS: Yes, I just wanted to thank everybody in this room. This has been an enjoyable experience. The work being done by every -- a surprisingly, I would say a surprisingly enjoyable experience.

I think that the work that your group is doing is commendable. It's been educational for me. And it's been a little intellectually challenging, but I've enjoyed it. And thank you. Thank you very much for all your work.

CO-CHAIR FLEISHER: Thank you.

Where were we?

MS. DEITZ: Back at home health.

DR. BURSTIN: Are you all with us Debra?

Ms. Deitz: Yes.

DR. BURSTIN: Wonderful. Okay, we will relaunch.

MEMBER MARKMAN: So then the issue is -- so I mean, this is an important measure because it entails 11,000 plus agencies, and millions of wounds. So then the question is, you know, I would submit to you that there is a relationship. I don't know about the numerator and the denominator, but I would submit to the committee --

CO-CHAIR FLEISHER: Well the first question is evidence. Is there anything further one evidence that you think, that's specifications, the numerator. That's an outcome.

MEMBER MARKMAN: Yes.

CO-CHAIR FLEISHER: Is there a sufficient evidence? It's an outcome measure.

MEMBER MARKMAN: I'm going to say yes.

CO-CHAIR FLEISHER: Okay, so let's vote.

MR. SANCHEZ: Voting will now begin for 1A, evidence. 1 is yes, 2 is no. Timer starts now.

We have 19 for yes, three for no.

CO-CHAIR FLEISHER: Next. We are in gap again.

MEMBER MARKMAN: It looks like in terms of the reporting, over the last three years that have been reported, that there's actually been a decrease in the number of valid episodes that have been reported.

CO-CHAIR FLEISHER: But is there variability in outcome between sites?

MEMBER MARKHAM: I'm going to ask the developer.

CO-CHAIR FLEISHER: Okay, for Debra. Debra?

MS. DEITZ: Yes.

CO-CHAIR FLEISHER: Are you on the phone?

MS. DEITZ: Yes.

CO-CHAIR FLEISHER: Is there variability on the -- between home health groups and the outcome of this measure, the results of this measure.

MS. DEITZ: Yes, and I'm going to let, Keziah Cook of Acumen address this.

MS. COOK: Sure. Can you guys hear me?

CO-CHAIR FLEISHER: Yes.

MS. COOK: Okay. Apologies for earlier, my line wasn't open. You know I think probably the easiest way to see the opportunity for improvement in this measure is to look at the interquartile range and the inter-decile range.

So we're basically seeing in the most recent year of data, the difference between the 25th percentile and the 75th percentile is about nine percentage points. So agencies sort of you know, the best agencies, nine percent more of their patients improve in surgical wounds then at the worst agencies.

And it's even more extreme if we compare the 10th percentile to the 90th percentile. So that difference in the most recent year is nearly 17 percentage points.

CO-CHAIR FLEISHER: Comments? Questions? Let's vote.

MR. SANCHEZ: Voting will now begin for 1B, performance gap. 1 is high, 2 is moderate, 3 is low, 4 is insufficient. Timer starts now.

CO-CHAIR FLEISHER: Are we okay?

MR. SANCHEZ: We have nine for high, 11 for moderate, zero for low, zero for insufficient.

CO-CHAIR FLEISHER: Next. Priority. Does this address an important issue?

MEMBER MARKMAN: Yes.

CO-CHAIR FLEISHER: Okay. Any comments? Questions?

MR. SANCHEZ: Voting will begin now for 1C, high priority. 1 is high, 2 is moderate, 3 is low, 4 is insufficient. Timer starts now.

You have 13 for high, six for moderate, one for low, zero for insufficient.

CO-CHAIR FLEISHER: Okay. Next, it's not a composite. Reliability.

MEMBER MARKMAN: Now the issue is the 13 percent improvement of only -- I mean they're not doing decubitus and they're not doing chronic diabetic wounds. These are post-op infections. And they've only seen a 13 percent improvement in the wound. And I would submit that --

MS. COOK: I'm sorry, this is Keziah. Can I please clarify. I know this came up earlier, and we actually have a typo on our form.

CO-CHAIR FLEISHER: Yes, and can you tell us -- specifications is really the key issue because we know it's been tested. Go ahead.

MS. COOK: Well, so first just to clarify that 13 percent number. That sentence should have read that it's 13 percent of patients had a surgical wound as a find, and their surgical wound was capable of improvement.

So it's 13 percent of the patients are eligible for the measure. And it's not that only 13 percent improved. It's the 13 percent of home health patients overall are these post-surgical patients who have a wound that is able to improve. Many of them have already fully epithelialized.

MEMBER PITZEN: This is Collette, I just wanted to comment. So the average overall performance rate is 89 percent, but probably some room for improvement.

But just to clarify, that's on all of those 4,000,000 home care visits. 13 percent of that large population had the opportunity to have -- that had a surgical wound that needed healing. Right?

MS. COOK: That's not correct.

MEMBER PITZEN: No?

MS. COOK: The population numbers are calculated only for those patients who are eligible for the measure. So there were about, I think about 500,000 patients who were eligible for the measure.

MEMBER PITZEN: Correct.

MS. COOK: So that 88 percent rate is off of those approximately 500 thousand patients. Not off of the 4,000,000 total home health population.

MEMBER MARKMAN: So in the mix of wounds, how many were closed wounds, and how many were open wounds, do you know?

CO-CHAIR FLEISHER: The question is to precisely specify how they actually did the calculation, and that's the question -- that's this question of reliability. Do you specify precisely?

MS. COOK: Right. So the item itself actually -- actually Deb would you have the item in front of you? I'm still trying to get it pulled up.

MS. DEITZ: Yes. And the response is zero, is that it's newly epithelialized. And that would be the one that would not be counted at start of care because they would not be eligible for improvement.

And then 1 is fully granulating, 2 is early partial granulation and 3 is not healing.

MS. COOK: And so what the measure

actually captures is if those are sort of in order of severity, so if a patient moves from having a wound that is not healing, which I think was number 3, to a number 2, which -- Deb which one was the number 2?

MS. DEITZ: Yes, early partial granulation.

MS. COOK: Okay, so if they move from not healing to early partial granulation, that's considered an improvement. Or if they move from early partial granulation all the way up to fully epithelialized, that's improvement.

So if they move upwards on a --

MS. DEITZ: Or if they no longer are considered to have a surgical wound at discharge because it is fully epithelialized.

CO-CHAIR FLEISHER: Thank you. Collette.

MEMBER PITZEN: I just wanted to comment. We talked about this during the

workgroup call. Initially I was a little bit confused. I thought it was simply a binary yes/no, did the wound improve? And I thought maybe that was a bit subjective.

But it was clarified that they're using WOCN guidelines for describing those wound characteristics. So there is a standardized process for determining that improvement.

MEMBER MARKMAN: Is there any time lines that you have incorporated in these wound healing things? I mean is it -- do you take a week, or? I mean when you took the episode, do you define a time element?

MS. COOK: So the typical episode length is just under 60 days. So an episode is the time period between the start of care and the patient's discharge. That's typically about 60 days.

Although, you know, there are some patients who are in home care for longer than

that. But the 60 days is the typical length.

CO-CHAIR FLEISHER: A.J.?

MEMBER YATES: Yes, I'm just going to point out that we need to look at this from the perspective of nursing, as opposed to the perspective of say surgeons. And what we're looking at is the value given to the patient and the value given to the payers, that's primarily CMS in this case, for the intervention of the home health nurse.

And whether there was observation performed, which is captured. And whether or not there was some positive influence by the nurse being there by keeping the dressing dry, keeping the patient out of trouble, doing some local wound care.

And that's going to be a different level of reliability testing then, it's in a sense observational, and I think that it's in effect, a composite of all the things they do.

MEMBER MARKMAN: Yes. They

actually do more than that. I mean you know.

MEMBER YATES: Right.

MEMBER MARKMAN: I mean yes, it's not just the wound. They don't just take care of the wound.

MEMBER YATES: Right. But this one measure of that that I think is reasonably measured this way.

MEMBER MARKMAN: Yes.

CO-CHAIR FLEISHER: Larissa?

MEMBER TEMPLE: One of my concerns is when the VNS go out to assess wounds, it's often a different person who assess the wound on the various episodes of the visits. And so I was wondering if there is any data to show the interobserver agreement?

I know that there are objective criteria, but I've certainly seen myself, disagreement. And so I'm curious to know if there are -- if they've ever looked at the interobserver agreement, and/or if they've

seen a wound get better and then get worse and then look to see if it was the same assessor.

Because I think that's an important piece of data because as clinicians we get these reports that wounds are getting better. We feel a whole lot better when it's the same nurse versus different ones. And so I think that that's an important piece to this. And I'm curious to know this.

CO-CHAIR FLEISHER: That is reliability, so can we get a comment from the developer on that aspect of reliability?

MS. COOK: So it sounds like you're talking about the item level reliability, rather than the measure reliability. Deb, I know the item level reliability was done quite some time ago, I think in the early 2000's. Do you happen to recall any of the findings? Or else we could round them up.

MS. DEITZ: I need to locate the item level reliability. I know that it was

done. And I know that it was considered to be acceptable as part of the development of the OASIS data set. And it was felt to be sufficiently reliable to be used in the data set for measure -- quality measurement and payment.

CO-CHAIR FLEISHER: Kelsey?

MEMBER McCARTY: My question was kind of similar, but I guess on the opposite side. If you do have continuity of care, it's the same person, is that person then reporting on their own ability to heal the wound, and do they have to acknowledge if that's not going -- like what's the reliability there of getting, I mean is the person there grading themselves I guess is what I'm asking?

MS. DEITZ: And I would say that the person is using very standardized criteria, the WOCN criteria to assess the wound at the time that the patient is discharged.

CO-CHAIR FLEISHER: Still, but you don't have any reliability testing of that,

correct?

MS. DEITZ: We have reliability. This item was -- did undergo reliability testing, yes.

CO-CHAIR FLEISHER: So yes. Collette?

MEMBER PITZEN: Just maybe to help clarify. I'm assuming that you're capturing and storing that initial observation, and what that wound status was and then your discharge, that it's like two separate fields that you're comparing, would that be a correct assessment?

MS. DEITZ: Yes.

MEMBER KO: Would this type of ongoing measure require an audit rather than a reliability testing ten years ago, just like you know the 10 percent STS audit making sure that it's continues to be high quality?

MEMBER DUTTON: I'll add a thought for the developers. This is 2014, we have flying cars. You can take a picture of the

wound at each visit and actually compare them independently then, or send them to the physician to compare.

MEMBER MARKMAN: I definitely agree. I -- if that, you know, I mean we're taking the whole subjective and making it an objective, and many wound care clinics do that, it's standard of care.

So that in terms of the reliability, a picture's worth a thousand words. And it's something that then can be communicated back to the physicians when the -- because these patients should be homebound. And if they can't see the doctor, then.

CO-CHAIR FLEISHER: Yes, Amy?

MEMBER MOYER: Isn't this in essence the chart that they're taking the data from? Isn't it the documentation of the care giver? I mean it sounds kind of like it's almost auditing the chart in some of the other measures we've looked at are I mean, I'm not

familiar with OASIS, but.

DR. BURSTIN: It's basically a database they're entering the information into. But again, they are using standardized approaches to do it. And again, I think that's similar things could be said about many things. We enter our own data in about our own care, so.

CO-CHAIR FLEISHER: Collette?

MEMBER PITZEN: Just a recommendation. Maybe you would want to repeat that kind of on data element reliability studies in the future submissions, but.

CO-CHAIR FLEISHER: Any other comments, Dr. Grover?

MEMBER GROVER: I think the picture aspect not only of reliability, but I would hope that if you don't, I would hope you audit these. I mean just a sample audits and then you would have a before and after picture to make it objective. And make sure that the observer or the treater that's taking care of this wound is

reporting it accurately.

CO-CHAIR GUNNAR: Okay. We'll go ahead and vote.

MR. SANCHEZ: Voting will now --

CO-CHAIR GUNNAR: Do we have -- we're close to -- do we have 18? We do not have our -- let's go reliability. Let's vote, I think we have 18, so go ahead.

MR. SANCHEZ: Voting will now begin for 2A, reliability. 1 is for high, 2 for moderate, 3 for low, 4 for insufficient. Timer starts now.

We have one for high, 11 for moderate, six for low, zero for insufficient.

CO-CHAIR GUNNAR: We have quorum for the remainder. So validity.

MEMBER MARKMAN: The last comments in terms of reliability and validity is that possibly the subjective aspect, but I think that from the standardization of the OASIS, that it's valid.

MS. WINKLER: Plus I think there's significant information -- testing information submitted in the submission that should cover it.

MEMBER MARKMAN: Yes.

CO-CHAIR GUNNAR: All right, any other comments? Let's vote.

MR. SANCHEZ: Voting will now begin for 2B, validity. 1 is high, 2 moderate, 3 low, 4 insufficient. Timer starts now.

We have one for high, 12 for moderate, one for low, zero for insufficient.

CO-CHAIR GUNNAR: Feasibility?

MEMBER MARKMAN: Feasibility, it's electronic, it's mandated, so yes.

CO-CHAIR GUNNAR: Any other comments? Let's vote.

MR. SANCHEZ: Voting will now begin for criteria 3, feasibility. 1 is high, 2 moderate, 3 is low, 4 is insufficient. Timer starts now.

Waiting on one more vote, please resubmit.

CO-CHAIR GUNNAR: Please submit again. Okay, we've got it.

MR. SANCHEZ: We have 13 high, five moderate, zero low, zero insufficient.

CO-CHAIR GUNNAR: Usability and use.

MEMBER MARKMAN: Absolutely.

CO-CHAIR GUNNAR: It's been in place, doing the job. Go ahead, no other discussion, we'll vote.

MR. SANCHEZ: Voting will now begin for criteria 4, usability and use. 1 is high, 2 is moderate, 3 is low, 4 is insufficient information. Timer starts now.

CO-CHAIR GUNNAR: We need one more. Everybody up again. And vote. There we go, got it.

MR. SANCHEZ: 10 high, eight moderate, zero low, zero insufficient

information.

CO-CHAIR GUNNAR: So overall suitability for endorsement. Any other comments, discussion? Hearing none, time to vote.

MR. SANCHEZ: Voting will now begin for overall suitability for endorsement. 1 is yes, 2 is no. The timer starts now.

CO-CHAIR GUNNAR: Everyone please vote. There we go.

MR. SANCHEZ: We have 17 yes. One no.

DR. BURSTIN: Thank you for everybody's patience on the phone. We appreciate it, so it's passed.

MS. DEITZ: Thank you.

MS. COOK: Thank you very much.

MS. WINKLER: Yes, operator, is anybody on the line for public comment? Anybody in the room want to make a comment?

OPERATOR: Okay, it is time to make

a public comment. Please press star then the number 1. There are no public comments at this time.

MS. WINKLER: Okay. Next steps briefly before you all run to catch your planes, trains and whatevers. Thank you all very much for your time. We realize this is very intense.

We do have a few left over stragglers that we will deal with. As I mentioned there was a -- there is a conference call scheduled, it should be on your calendars. Somehow I -- it's on mine too, but I somehow can't read calendars or something.

But we will be sending out a discussion agenda ahead of time. But we do have, we'll talk with the folks from ASA about the tabled measure 269. Also we'll want to do the three database measures. We need to think about how we want to do those efficiently. The call is only two hours.

Plus we need to look at related and competing measures. Given the results of your evaluation on the antibiotic prophylaxis measures, there might not be a whole lot there. But certainly we do want to look at the two CABG mortality measures side by side and have the conversation about competing measures.

We also want to take a look at the entire results of what you've done. And you know, see if it all makes sense. Have you, you know, do some last comments before we start wrapping it up and start writing up a report that reflects this that will go out for public comment.

So those are our next steps. We are anticipating, or was scheduled to go out for public comment like July 3. So this will be quickly moving through the month of June.

And so if you've got anybody you want to share this with, they should be looking for it to be available for public comment in

July. So that's kind of the time frame going forward.

This committee will regroup by conference call in August, to talk about those comments that we receive. And perhaps act on them depending on what the comments may indicate.

So we're going to be moving briskly through the process steps through the summer. So Bill, any questions from you all? Andrew anything I've forgotten? Any comments from anybody?

I know I've had sidebar suggestions from all sorts of folks on, you know, process improvements, and suggestions and all sorts of things. We're open ears. So feel free to keep those suggestions coming, and don't be surprised if we call you back and say okay you're on, we really want to hear the details.

So, any last --

CO-CHAIR FLEISHER: Reva?

MS. WINKLER: Yes?

CO-CHAIR FLEISHER: Yes, we have Lee on the phone now, and I'm particularly --

MS. WINKLER: Hi Lee.

CO-CHAIR FLEISHER: How people -- hi -- I'm particularly interested in how people think the pre-meeting conference calls could be improved and utilized in a different way. Because I'm not sure they were always as effective.

MS. WINKLER: Okay. We'll certainly see if we can figure out the best way to get your feedback and suggestions for how we can make this work. Since we're going to be working together going forward for the next couple of years.

Also, one thing we didn't do, we will need to do, is somehow randomly get you to two year or three year terms. We did it? Oh good, I missed it. Okay.

You know if you're staying for two

or three years. So renewable for another term if you so desire. So we really want to make this easy for everybody.

Dr. Yates?

MEMBER YATES: Yes, I was just going to make a comment to the question on the phone. And I had already said this to you Reva, but I think that if there could be an emphasis on the conference calls, the workgroup calls to ask -- to determine the level of evidence, and bring out the level of evidence on the process measures during the workgroup calls so that that's better assessed.

Because I think some measures failed today from an inadequate discussion of the level of evidence. And I think some of the measures -- I just think we lost a lot of time debating evidence that was clearly presented that wasn't discussed well enough to say level one, level two, level three.

MS. WINKLER: But anyway. Thank

you all very much. Travel safely.

(Whereupon, the above-entitled proceeding was concluded at 3:28 p.m.)