NATIONAL QUALITY FORUM

Memo

TO: NQF Members

FR: Kathryn Streeter, Senior Project Manager and Christy Skipper, Project Manager RE: Voting Draft Report: *NQF- Endorsed Measures for Surgical Procedures, 2015-2017*

DA: November 15, 2016

Background

The Surgery portfolio is one of NQF's largest measure portfolio with over 100 endorsed measures. This NQF project aimed to evaluate additional performance measures that will help guide cardiac, vascular, orthopedic, urologic, and gynecologic surgeries that include adult and pediatric population. The 25-member Surgery Standing Committee met for a 2-day in-person meeting to evaluate 24 measures: 10 new measures and 14 measures undergoing maintenance review. The Committee recommended 16 measures for endorsement and eight were not recommended for endorsement.

Comments Received

NQF solicits comments on measures undergoing review in various ways and at various times throughout the evaluation process. First, NQF solicits comments on endorsed measures on anongoing basis through the Quality Positioning System (QPS). Second, NQF solicits member and public comments prior to the evaluation of the measures via an online tool located on the project webpage. Third, NQF opens a 30-day comment period to both members and the public after measures have been evaluated by the full committee and once a report of the proceedings has been drafted.

Comments Received

NQF solicits comments on measures undergoing review in various ways and at various times throughout the evaluation process. First, NQF solicits comments on endorsed measures on anongoing basis through the Quality Positioning System (QPS). Second, NQF solicits member and public comments prior to the evaluation of the measures via an online tool located on the project webpage. Third, NQF opens a 30-day comment period to both members and the public after measures have been evaluated by the full committee and once a report of the proceedings has been drafted.

Pre-evaluation comments

The pre-evaluation comment period was open from June 1 to June 6, 2016 for all 24 measures under review. One pre-evaluation comment was received and was in favor of endorsement of a measure submitted to the portfolio. The pre-evaluation comment was provided to the Committee prior to their initial deliberations.

Post-evaluation comments

The draft report went out for the 30-day Public and Member commenting period on September 22 – October 21, 2016. During this commenting period, NQF received 71 comments from 6 member organizations and 11 comments from the public. Comments were generally supportive of the Committee's recommendations.

A complete table of comments submitted post-evaluation, along with the responses to each comment and the actions taken by the Standing Committee, is posted to the project page along with the measure submission forms.

The Committee reviewed all comments received and considered the pre-meeting comments prior to making an endorsement recommendation. The Committee also responded to all post-evaluation comments. Revisions to the draft report and the accompanying measure specifications are identified as red-lined changes.

Reconsideration Request

#0351: Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04)

The developer has requested that this measure be reconsidered by the Committee.

During the in-person meeting, the Committee did not pass the measure on Validity. In regard to Validity, the Committee was concerned that the specification that excludes patients who are transferred from the denominator could provide a method to game the measure, particularly if the patient's condition worsens. The Committee noted that transferring a patient to a higher level of care is often appropriate, however, when a patient is transferred to a higher level of care, the receiving hospital becomes responsible for the treatment provided by the transferring hospital. The Committee did not believe risk adjustment fully accounted for the fact that the receiving hospital would become responsible for the patient's outcome. The Committee also questioned whether transfer issues were addressed adequately to understand threats to validity and that handling of transfers make it difficult to validate if the appropriate efforts were made to save a patient.

During the comment period, the developer submitted a request for reconsideration on the grounds that the Committee did not appropriately review and evaluate the measure on the Validity criteria; the Committee's discussion included concerns about how the measures might be used rather than focusing solely on scientific acceptability of the measure; and a separate NQF committee reviewed a similar measure and reached a different conclusion than did the Surgery Standing Committee, e.g., inconsistent review of measures across NQF standing committees. The developer also submitted additional information on transfers, risk adjustment, and use of claims data to measure complications. Please see Appendix A for the full reconsideration request and additional data submitted).

> **Committee Response**: The Committee agreed that the additional data submitted on transfers were sufficient to address their concerns on Validity. The additional data showed that the inclusion or exclusion of

transferred patients in the model would have little effect on the outcome. The Committee also acknowledged that the measure would risk adjust for transfers and whether the patient arrived at the hospital with a complication already present. Ultimately, the Committee recommended this measure for endorsement.

"Consensus Not Reached" Measure

1543: Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Artery Stenting (CAS)

During the in-person meeting, the Committee noted that there were no published guidelines for carotid artery stenting and that this procedure is not recommended by the major medicalsocieties. Committee members also questioned whether the measure should be considered an appropriate use measure due to the increased risk of stroke or death, compared to the risk of stroke or death by surgery. Other Committee members stated that despite indication, the procedure is still being performed and therefore it would be important to measure the outcome. The Committee should review the comments that were received, and then re-discuss the measure. During the in-person meeting, consensus was not reached on the Evidence (Y-12; N-10) and Validity (H-0; M-13; L-9; I-0) subcriteria.

Thirteen comments were submitted for this measure; all were supportive comments, noting that the measure will aid in the appropriate selection of patients who receive carotid stenting. One comment stated that "review of the medical literature that was discussed…omitted significant amounts of data and that it was inaccurate in summarizing the clinical guidelines on carotid stenting." See Appendix B or the Comment table (ID 6370) for the full comment.

Developer Response: In response to Dr. Powell's comments: As Chair of the SVS Quality Performance and Measures Committee and one of the presenters of this measure that day we are not against CAS. As I stated that day CEA and CAS have similar outcomes with experience operators. We are advocating this measure be continued so we can monitor outcomes and compared them with CEA. There have been studies that suggest some subsets of patients may have different outcomes yet ongoing research will define this. I strongly encouraged the NQF Committee that day to continue this measure. Brad Johnson, MD SVS Chairman of Quality Performance and Measures Committee

Committee Response: After reviewing member and public comments, the Committee agreed that that they should not be dissuaded from recommending the measure for endorsement although there is controversy surrounding the carotid artery stenting procedure. The Committee believes that this measure provides a well-defined tool to assess the outcome of the measure and recommends it for continued endorsement.

Comments and their Disposition

Two major themes were identified in the post-evaluation comments, as follows:

1. Comments in support of Committee's recommendations
2. Evaluation and discussion of the Sociodemographic Status trial period

Theme 1 – Comments in support of Committee's recommendations

A vast majority of the comments (68) were in support of the Committee's recommendations on all measures for which consensus was reached. Of these, 13 comments emphasized the importance of *#1534 In-hospital mortality following elective EVAR of AAAs*.

Theme 2 – Evaluation and discussion of the Sociodemographic Status trial period

Two comments expressed concern related to the "lack of rigor and robustness of the risk adjustment reviews" and suggested that other SDS factors must be considered "to understand the potential impact on a hospital's performance".

NQF Response: The SDS trial period is a temporary change to NQF's policy. During this 2-year trial period, NQF is gathering information about the feasibility, limitations and challenges of including SDS factors in the risk-adjustment approach. Consideration of sociodemographic factors in risk adjustment models is a critical issue in measurement science. The Committee was charged with evaluating the submitted measure specifications and testing by the measure developer. Given the constraints on the current, available data elements, the Committee relied on the methods used by the developer to test the conceptual and empirical relationship between SDS factors and readmissions and complications.

The developer stated there was a conceptual relationship between the SDS variables. Specifically, the developer reported socially disadvantaged patients had a higher disease burden or receive worse or disparate care, and that there is a source unrelated to hospital quality of care which could hinder patient adherence to things like post-discharge instructions. Using decomposition analysis to measure effects at the patient level and at the hospital level, the hospital effect had greater impact than patient level factors. These results were in contrast to the clinical data elements, where the patient effect tended to dominate. Due to the dominant hospital effect, the developer reported that they risk adjust away a component of hospital quality when the variables are included in the model. The developer reported that the three SDS factors were statistically significant in the model and inclusion of the variables in the model did not change the c-statistic. Ultimately, the Committee agreed that they would not recommend risk adjustment for SDS since finding disparities among groups on these measures is something that should be reported and followed.

NQF has maintained a non-prescriptive approach to the selection and testing of variables in risk adjustment models. NQF has not required that certain SDS variables be tested and does not set requirements around the inclusion of any specific variables. Similarly, NQF does not set "cut-points" for the statistical testing of a risk adjustment model. The evaluation of the model is left to the Standing Committee reviewing the measure. This approach applies to both clinical and SDS variables.

Developer Response: CMS and Yale/CORE share the FAH's concern for the scientific rigor and robustness of the risk adjustment analyses. A risk adjustment model that is scientifically sound hinges to a large part on the use of

data sources that are scrutinized, vetted, and representative of the population of interest. The process of variable selection must be done thoughtfully, as the inclusion of highly-correlated variables in a model often yields spurious results. In the context of socioeconomic status (SES) and quality reporting, there are questions that every developer and user must ask: If changes in the models result in changes in the findings, are these changes methodological artifacts? Do they alter the big picture of the overall ranking of the hospitals? Are these changes clinically meaningful? In order to identify relevant SES variables that can be used in a national measure of hospital quality, we have identified all available data sources assessing SES as patient-level variables, or proxies for patient-level variables, and can be linked to Medicare Fee-for-Service claims for all, or nearly all, over 65 year-old Medicare patients. We also performed a thorough review of relevant literature to identify SES variables that had a conceptual relationship with the measures' outcomes. The only SES variables that met the criteria above and were supported by evidence linking the variable to measure outcomes including mortality, readmission and complications were:

- Dual eligible status (meaning enrolled in both Medicare and Medicaid)
- Agency for Healthcare Research and Quality (AHRQ)-validated SES Index score (composite of 7 different variables found in census data: percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th-grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room)

In selecting variables for analyses across all measures, our intent was to be responsive to the NQF guidelines for measure developers in the context of the SES Trial Period, and to identify variables that are feasible to test and use in the near term. We examined patient-level indicators of both SES and race or ethnicity that are reliably available for all Medicare beneficiaries. We aimed to select those variables that are most valid and available. The variables used are aligned with what the National Academy of Medicine committee identified as available for use in outcome measures.

Committee Response: After a full discussion on SDS risk adjustment, the Committee accepted the developer's rationale not to include the SDS variables in the risk adjustment model. However, concerns were raised that the lack of effect might be lost in the variability of the overall low C statistic in the measure, and that the methodology used might not be capturing a community effect seen as a source of performance gaps in the relevant literature. The Committee recognizes that risk adjustment for SDS factors is a rapidly progressing area and that more work is needed to appreciate the effects of social risk, understand the most relevant patient and community level factors, collect data on these risk factors, and determine the best methods to incorporate these risk factors into performance measures. The Committee looks forward to continued deliberations on these issues and to reexamining these measures as better data emerge.

NQF Member Voting

Information for electronic voting has been sent to NQF Member organization primary contacts. Accompanying comments must be submitted via the online voting tool.

Please note that voting concludes on December 5, 2016 at 6:00 PM ET – no exceptions.

Appendix A

To: National Quality Forum (NQF) Surgery Standing Committee Members

From: Pam Owens and Mamatha Pancholi, AHRQ Quality Indicators Project
Agency for Healthcare Research and Quality
On behalf of the entire AHRQ QI Team

Re: Additional Information regarding Maintenance Measure NQF 0351: Death Rate Among Surgical Inpatients with Serious Treatable Complications (PSI 04)

Date: October 21, 2016

AHRQ appreciated the opportunity to have the maintenance measure NQF 0351: Death Rate Among Surgical Inpatients reviewed by the NQF Surgery Standing Committee. Reviews by external stakeholders allows AHRQ to inform multiple experts about measure refinements, discuss and understand concerns of the measures, and be able to address those concerns and refine to make the measures even better.

As a follow-up to the NQF Surgery Standing Committee review on August 16, 2016, AHRQ would like to share with the Committee additional information that was requested during the review and present additional proposed refinements based on reviewers' concerns.

The *NQF-Endorsed Measures for Surgical Procedures 2015-2017: Draft Report for Comment* (September 22, 2016) noted that the reviewers wanted additional information and/or had concerns about the following areas:

- "Risk-adjustment strategy includes patients transferred in with complications present on admission"
- "Claims data cannot accurately capture complications reliably"
- "Does not include the transfers out thus providing a potential for 'gaming'"
- "Absence of testing data that demonstrates the measure assesses what it is supposed to measure"

In the paragraphs below, we respond to these concerns and describe enhancements to the risk-adjustment approach that AHRQ is planning to implement to PSI 04 in the next release of the Patient Safety Indicators, v6.0.2. These enhancements will carry forward into the ICD-10 PSI software, as soon as an adequate period of ICD-10 coded data is available for predictive modeling.

"Risk-adjustment strategy includes patients transferred in with complications present on admission"

This question has been a subject of considerable attention and analysis over the 12 years since PSI 04 was introduced, the 14 years since Needleman and Buerhaus (*N Engl J Med* 2002;346(22):1715-22) first described this approach to operationalizing "failure to rescue," and the 24 years since Silber and colleagues (*Med Care* 1992;30(7):615-629) first described the concept of "failure to rescue".

First, it is important to recognize that NQF recently re-endorsed the related (not competing) Silber/Children's Hospital of Philadelphia (CHOP) version of this measure, with two different specifications of the outcome variable, "0353 Failure to Rescue 30-Day Mortality (risk adjusted)" and "0352 Failure to Rescue In-Hospital Mortality (risk adjusted)." Both of these specifications of the "failure to rescue" concept include ALL 30-day deaths and ALL inpatient deaths, respectively. In other words,

not only do these existing NQF-endorsed measures include deaths among patients transferred in with complications present on admission, but they even include deaths among patients who never experienced any reported complication (see http://www.qualityforum.org/QPS/0352).

In previous discussions with the Patient Safety Standing Committee, the 0352/0353 measure stewards from CHOP provided extensive support for the concept of NOT limiting the denominator to patients who experienced a complication acquired in the same hospital. There are two important conceptual arguments supporting this decision:

1. PSI 04 and other measures of "failure to rescue" focus on the progression from complication to death (and the hospital's ability to influence that progression); whether the hospital was responsible for causing the complication during the same admission, during a prior admission, or not at all (i.e., the complication originated in ambulatory surgery), is not material to that focus. PSI04 is a risk-adjusted inpatient surgical mortality measure in which the denominator is limited to patients with certain complications that often arise in association with surgical care. Prior studies have suggested that high-quality care can prevent similar proportions of these deaths, whether the complication happened to originate during the same hospital stay or not.
2. AHRQ and other measure developers attempt to design measures in a manner that minimizes unintended consequences and reduces opportunities for "gaming." Limiting the denominator for PSI 04 to patients with complications reported as hospital-acquired (i.e., not present on admission) would encourage hospitals to shift blame for complications, which is an unproductive exercise. In other words, a serious complication may lead to death if it is not recognized and treated in a timely manner; this downward trajectory can often be interrupted with aggressive treatment. The goal of PSI 04 and other measures of "failure to rescue" is to reward hospitals for providing such treatment, regardless whether the complication started within that hospital's walls or outside its walls. AHRQ does not wish to encourage hospitals to waste effort on determining exactly when a complication started, given that the evidence-based focus of this indicator is on treating complications, not preventing them.

Thus, this critique of PSI 04 was already discussed extensively, and set aside, during last year's successful re-endorsement process for "0353 Failure to Rescue 30-Day Mortality (risk adjusted)" and "0352 Failure to Rescue In-Hospital Mortality (risk adjusted)." AHRQ encourages the Surgery 2015-2017 Standing Committee to consider the previous work of the Patient Safety Standing Committee, as consistent evaluation processes benefit all of NQF's stakeholders.

Second, AHRQ is revising its PSI 04 risk-adjustment model for the next release, v6.0.2, by adding adjustors for whether the denominator-triggering complication was present on admission, and whether it was relatively mild or severe. This addition is superimposed on stratified risk-adjustment models, in which patients with different complications are allowed to have different predictors of death, and different relationships between age, gender, and transfer status and the risk of death. In other words, the v6.0.2 risk models will explicitly account for whether a denominator-triggering complication was present on admission, and whether it was mild or severe at the time of presentation. For example:

- Stratum A includes deep vein thrombosis (DVT) and pulmonary embolism (PE); the latter diagnosis is now considered to be more severe than the former.
- Stratum B includes pneumonia; staphylococcal, gram negative, anaerobic, and aspiration pneumonias are now considered to be more severe than other types of pneumonia.
- Stratum C includes sepsis; sepsis with septic shock or acute organ system dysfunction is now considered to be more severe than uncomplicated sepsis.

- Stratum D includes shock and cardiorespiratory arrest; the latter diagnosis is now considered to be more severe than the former.
- Stratum E includes all types of gastrointestinal (GI) hemorrhage; GI bleeding with perforation is now considered to be more severe than GI bleeding without perforation.

Conceptually, this change will reduce any residual bias resulting from the transfer of patients with severe complications.

AHRQ's updated models for v6.02 have significantly higher c statistics than the current v6.01 models (see details in Appendix A). Specifically, the re-estimated c statistics, based on the AHRQ 37-state reference population, increased from 0.780 to 0.797 for Stratum A, from 0.771 to 0.782 for Stratum B, from 0.726 to 0.776 for Stratum C, from 0.715 to 0.818 for Stratum D, and from 0.860 to 0.878 for Stratum E.

These revised c statistics, representing the ability of the model to discriminate between patients who survived and patients who died, are very consistent with other NQF-endorsed mortality measures, including postoperative morbidity and mortality measures based on registry data. For example, among measures reviewed in 2016 for the same project (Surgery 2015-2017), developers reported c statistics of 0.716-0.719 for the American College of Surgeons' (ACS) "Risk Adjusted Colon Surgery Outcome Measure," 0.65-0.70 for CMS' "Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)," 0.708-0.738 for the STS "Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score," 0.708-0.807 for the STS "Individual Surgeon Composite Measure for Adult Cardiac Surgery," and 0.758-0.772 for the ACS "Risk Adjusted Case Mix Adjusted Elderly Surgery Outcomes Measure." (http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80864)

Therefore, the argument that the risk-adjustment strategy for PSI 04 is inadequate lacks empirical foundation. AHRQ believes that the risk-adjustment upgrade to be implemented in v6.0.2 will address the remaining concerns of hospitals that receive transfers from other hospitals.

"Claims data cannot accurately capture complications reliably"

As described above, the concept of "failure to rescue" or "death given a complication" has never hinged on the precise level of coding accuracy. The intent of "failure to rescue" measures is not to "ding" hospitals for complications, but to identify hospitals that are able to recognize complications early and treat them aggressively to reduce the risk of death.

This criticism is also puzzling because there is no clear connection between inaccurate coding and PSI 04 performance. Inpatient death—the numerator or outcome event for PSI 04—is widely understood to be accurately reported. Thus, the only ways a hospital could artificially reduce its PSI 04 rate through inaccurate coding would involve manipulation of the *denominator* by either: (1) preferentially decreasing the coding of denominator complications among patients who subsequently died; or (2) preferentially increasing the coding of denominator complications among patients who did *not* subsequently die. (In epidemiologic language, clinically significant information bias would require *nondifferential* misclassification of complications, with respect to an outcome that was unknown when the complication occurred and was diagnosed by the treating physician.) Numerous factors argue against the likelihood of such manipulation: (1) when complications are diagnosed, it is frequently unpredictable which patients will go on to die as inpatients; (2) even if it were possible to predict which

patients will die, it would be implausible that providers would suppress the accurate description and treatment of a complication for the sake of PSI 04 performance; and (3) the most plausible means of gaming PSI 04—systematic over-coding of complications such that the denominator might be inflated with milder cases unlikely to result in death—would be counter-intuitive to providers and an inefficient means of influencing PSI 04 rates. For example, the most common evidence-based critique of PSI 12 – that it is susceptible to "surveillance bias" or "overdiagnosis bias" from inter-hospital variation in how postoperative patients are screened for venous thrombosis – would lead hospitals with falsely high PSI 12 rates to have falsely low PSI 04 rates. In other words, a hospital attempting to suppress its PSI 04 rate could do so by exaggerating its PSI 12 rate – an implausible scenario on its face.

Finally, it should be noted that the complications captured in the denominator of PSI 04 are serious, often life-threatening complications (i.e., DVT, PE, pneumonia or aspiration, sepsis, shock or cardiac or respiratory arrest, GI hemorrhage) that affect MS-DRG assignment in the Inpatient Prospective Payment System, so failing to code them would trigger underpayment, whereas consistently overcoding them would trigger auditing and financial penalties. The underlying ICD-coded data are widely accepted as sufficiently accurate to determine hospital payment (through MS-DRG complication or "CC/MCC" assignments), so they should logically be accepted as sufficiently accurate to identify patients who had complications that placed them at risk of death.

"Does not include the transfers out thus providing a potential for 'gaming'"

Patients transferred-out are excluded because the outcome of the hospital episode is unknown for these patients, in the absence of linked data across multiple hospitals. In other words, the numerator for PSI 04 is in-hospital death, but if the patient is appropriately transferred to a regional referral center for more advanced care, the patient's outcome is unknown to the referring hospital.

Conceptually, it may be preferable to attribute the outcomes of transferred patients proportionally to both referring hospitals and receiving hospitals, since both hospitals were involved in providing care. In some cases, referring hospitals may transfer patients "too late" for receiving hospitals to provide effective treatment, and it would be appropriate to place most of the "blame" on the referring hospital. In other cases, referring hospitals may transfer patients immediately, but receiving hospitals may either fail to deliver effective and timely treatment, or may be powerless to do so due to the patients underlying condition(s). This idea of "split attribution" merits methodologic attention, but would be difficult to operationalize. NQF has a separate process currently underway to promote more transparent, reproducible, and valid approaches to attribution. AHRQ is not aware of any currently endorsed quality measures, whether based on claims data, registry data, or electronic health records, that "split attribution" across multiple hospitals.

Given that there is currently no validated methodology for "split attribution" between referring and receiving hospitals, AHRQ continues to favor exclusion of these patients. Empirically, this exclusion has virtually no effect on the distribution of risk-adjusted, smoothed PSI 04 rates across hospitals, because the hospitals that transfer out more than 10% of their eligible patients are usually small hospitals. These small hospitals have smoothed observed/expected ratios close to one, indicating that their observed results (without smoothing) are not very reliable. Approaches of this type have become standard practice among measure developers, and have been extensively reviewed by NQF committees.

"Absence of testing data that demonstrates the measure assesses what it is supposed to measure"

This concern is demonstrably false. AHRQ believes that the NQF Surgery Standing Committee may have misinterpreted NQF's 2016 Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement. This document (Algorithm 3) specifically emphasizes "validity testing computed with performance measure scores for each measure entity" over "validity testing conducted with patient-level data elements." Recommended approaches to "validity testing computed with performance measure scores" include "(1) correlation of the performance measure score on this measure and other performance measures; (2) differences in performance scores between groups known to differ on quality; and (3) other accepted method…" AHRQ presented substantial and unrefuted evidence of "construct validity" based on these approaches. Such findings are the only "testing data" requested by the NQF for measures of this type. Evidence of "patient-level data element" validity is only required – indeed, only helpful – if evidence of validity of "performance measure scores" is unavailable or conflicting, which does not apply in the case of PSI 04.

Specifically, the Evidence form under Section 1a.8.2 includes an extensive and detailed environmental scan of the literature, demonstrating that measures of "failure to rescue," including PSI 04, are consistently associated with many hospital-level measures of high quality care, including higher nurse-to-bed ratios, better nurse skill mix ratios (i.e., baccalaureate-trained nurses), higher US-trained nurse ratios, Magnet designation by the American Nurses Credentialing Center, and the Practice Environment Scale of the Nursing Work Index. Higher hospital volume was associated with lower rates of "failure to rescue," based on multiple specifications of the concept, in at least six studies. The Measure Testing form under Section 2b.2.3 reports confirmatory analyses using PSI 04, as specified by AHRQ. The analyses reported included expected differences in performance scores between teaching and non-teaching hospitals, between high-technology and low-technology hospitals, between large and small hospitals, between hospitals with high and low nurse staffing levels, and between hospitals with better and poorer nursing skill mix.

Additional evidence of face validity was also reported under Section 2b.2.3, although evidence of face validity is not required for a "high" or "moderate" rating of validity under NQF's 2016 Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement.

AHRQ has confirmed some of these findings in updated analyses after the NQF Standing Committee discussion this summer. For example, AHRQ's analysis of both the current v6.01 and proposed v6.02 risk-adjusted measures indicates that major teaching hospitals, defined as hospitals with 0.2500 to 0.5999 residents per licensed bed, have an aggregate observed-to-expected ratio of 0.969 to 0.972. In other words, major teaching hospitals have about 3% fewer deaths than expected, based on the five risk-adjustment models described above. By comparison, non-teaching hospitals (with 45.3% of all eligible patients) have aggregate observed-to-expected ratios of 1.002 to 1.007. Hospitals with unknown teaching status in the American Hospital Association's survey database (with 2.3% of all eligible patients) have aggregate observed-to-expected ratios of 1.032-1.046. In other words, non-teaching hospitals consistently demonstrate more deaths than expected, whereas major teaching hospitals consistently demonstrate fewer deaths than expected.

In summary, PSI 04 and other measures of "failure to rescue" have demonstrated validity at the hospital-score level because of their correlation with structural and process measures of quality, particularly those related to the nursing workforce. There is no evidence that these correlations are appreciably affected by the details of how denominator-triggering complications are specified. For

example, Mattke et al (*Med Care* 2004; 42(2 Suppl):II21-33) found that hospital rankings of failure-to-rescue for surgical patients (of which AHRQ's PSI 04 is a widely used specification) were not influenced appreciably by denominator definitions: "the indicator's estimate of relative hospital performance was fairly robust, even if specific patient cases were misclassified in terms of the timing of complications. This finding comports with the original FTR definition, which did not require that complications occur during the hospital stay." Silber et al's analysis (*Med Care* 2007; 45(10):918-25) also supports having as broad a denominator definition as possible, including conditions that were present on admission. Construct validity relationships were consistent across all tested specifications of failure-to-rescue, including PSI 04. Absent any empirical evidence that restricting the denominator to conditions that arose during the same hospital stay would increase the validity of the indicator, AHRQ has chosen to retain fidelity with the original concept of failure to rescue, as it was developed by Silber et al. (*Med Care* 1992; 30:615-29) and adapted by Needleman et al. (*N Engl J Med* 2002; 346(22):1715-22). As described by Needleman and Buerhaus (*Med Care* 2007; 45(10:913-5), "We have argued elsewhere that with adequate risk adjustment and use of methods that compare actual to expected rates of complications, inclusion of cases in which complications were present on admission can be controlled… FTR-N was developed with some sensitivity to these concerns… which contributed to its adaptation by AHRQ as FTR-A. As noted above, FTR-N was developed to be a nursing sensitive measure. It was constructed ex ante around complications that nursing was believed to influence and for which early identification and intervention by nurses might be central to rescue."

Additional, more recent literature is summarized in Appendix B of the document (see also the National Health Institute for Health Research review at: http://www.crd.york.ac.uk/prospero/display_record.asp?ID=CRD42013004080) . This literature again highlights that PSI 04 should be viewed principally as a measure of hospital team performance, due to its particular sensitivity to aspects of nursing skill mix and the nursing work environment. It should not be viewed as a measure of quality for individual surgeons.

Appendix A. Revised PSI 04 Risk Models

Source: Agency for Healthcare Research and Quality, Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases, 2013

Table 1. PSI 04A Risk Model: Pulmonary Embolism / Deep Vein Thrombosis Stratu

Appendix B. Literature Review – Preventing Failure to Rescue

Recent research targeting preventing failure to rescue (FTR) events in the surgical population continues to demonstrate a strong association with structure and processes of care, which are often intermingled.

Structures and processes of care

Residents and hospital

The first set of articles looked at the impact of residents. Use of residents can represent a process of care, but can also be a surrogate for structural measures or components. Performing a retrospective cohort analysis using the National Surgical Quality Improvement Program (NSQIP) Participant User Files for 2005-2009, Castleberry et al. (2013) found that resident trainee participation in complex, oncologic surgery was associated with significantly higher rates of 30-day postoperative complications in NSQIP-participating hospitals; however, this effect was countered by overall lower 30-day mortality and improved rescue rate in preventing death among patients suffering complications (5.9 vs. 7.6%, AOR 0.79, 95% CI 0.68-0.90). Navathe, A. S., et al. (2013) found no differences in FTR in teaching hospitals when comparing rates prior to and after ACGME duty hour reform, suggesting that resident fatigue was not a major factor contributing to mortality given complications.

Ferraris et al (2014), also using the NSQIP database to evaluate failure to rescue among 200 hospitals over 5-years, found that FTR was lower with resident involvement (9.4% vs 12.4% for attending alone; $P < .001$), despite significantly increased operative morbidity (11.4% vs 7.8% with attending only; $P < .001$) and prolonged operative time (127 minutes vs 93 minutes for attending only; $P < .001$). The most serious complications occurred 5 to 10 days before death, suggesting that there is a window for intervention to rescue patients with early aggressive treatments (especially of the initial sentinel complication).

Hospital level variables

A systemic review of 42 relevant papers (1980-2012) by Johnston et al found that the overall incidence of FTR ranged between 8.0 and 16.9%. Two studies demonstrated that FTR discriminated high- and low-volume hospitals better than morbidity measures. Greater hospital volume was associated with lower FTR rates in four studies. Two studies that analyzed the effect of the Safe Practices of the National Quality Forum revealed that FTR was less frequent in those with greater compliance. A higher level of nurse staffing was associated with lower FTR rates in two studies, with no significant association in another study. One study analyzed several hospital characteristics and found that the following were associated with lower FTR rates: teaching status, hospital size >200 beds, daily census >50%, increased nurse-to-patient ratios, and use of technology. Lower FTR rates were associated with patient age <70 years, absence of malignancy, and gastrointestinal complications. Higher FTR rates were found in patients with medical complications (compared with surgical complications), surgical site infections, deep vein thrombosis, pneumonia, sepsis, and non-white ethnicity. Mortality rates were higher among patients with escalation delay (defined as the lack of timely recognition or action when caring for

a deteriorating patient) compared with no delay in three studies. One study found that a rapid transfer to the ICU decreased mortality.

Using the 2007 to 2011 Nationwide Inpatient Sample (NIS), Hyder et. al., looked at high-mortality hospitals compared with low-mortality hospitals and found a difference in FTR rates (22.4% vs 20.2%, p = 0.0020). They then used Monte Carlo models to estimate the potential overall mortality reduction that could be achieved by focusing on 5 target subpopulations, assuming that these target subpopulations in high-mortality hospitals could achieve the same mortality rates as analogous patients in low-mortality hospitals. Approximately 50% absolute reduction in baseline total mortality (5.22%) was demonstrated when targeting FTR (2.73%; 95% CI 2.61 to 2.87) and reducing the FTR mortality gap by nearly 75%.

A Japanese study using PSI4 (v4.2) (Kitazaqa et al, 2014), found that low-volume hospitals had more deaths among surgical inpatients with serious treatable complications (38.5%, 95% CI, 33.7% to 43.2%) than high-volume hospitals (21.4%, 95% CI, 19.0% to 23.9%). In multiple linear regression analysis, each additional surgical patient per month was associated with 0.2 fewer cases ($p < 0.01$) of PSI04.

Nursing and nurse staffing

FTR is a considered a nurse sensitive indicator. Many of the recent research articles looked at the association between nursing staffing and FTR with mixed results.

Mark et al studied the impact of nurse staffing regulations in CA on FTR using PSI04. Hospitals were divided into quartiles based on how their pre-regulation RN ratios compared to other States. California quartile 1 hospitals had lower RN staffing compared to other states whereas quartiles 3 and 4 hospitals had higher staffing. Although there was a 35% post-regulation increase in RN staffing in quartile 1 hospitals, the associated change in FTR was not significantly different than the change in the comparison hospitals. However, the researchers saw a statistically significant differential decrease in FTR in California Quartile 4 hospitals, where the differential increase in RNs was slightly greater than in any other quartile. These mixed findings suggest other factors, such as culture or other staffing enhancements, influence FTR rates. Besides staffing, the type of nurse has been shown to make a difference. Using robust regression with clustering by UHC hospital type, Blegen et al (2013) found that as RN education increased in University affiliated hospitals, FTR decreased ($r = -0.399$; $p < 0.05$). This is supported by work by Kutney-Lee et al. (2015), who found that between 1999 and 2007, 11 Magnet recognized hospitals in the State of Pennsylvania had 6.1 fewer FTR deaths per 1000 patients (P=0.02) than the 125 non-Magnet Pennsylvania comparison hospitals.

Supporting the preventability of FTR events, Brooke, et al. (2012) found that hospitals that complied fully with the 27 National Quality Forum (NQF) safe practices had an increased likelihood of diagnosing a complication after any of six high-risk operations (odds ratio [OR], 1.13; 95% confidence interval [CI], 1.03-1.25), but had a decreased likelihood of failure to rescue (OR, 0.82; 95% CI, 0.71-0.96), and a decreased odds of mortality (OR, 0.80; 95% CI, 0.71-0.91).

Park et al examined the relationship between RN staffing and FTR and evaluated the effect of patient turnover on that relationship in 42 UHC hospitals. In general, they found that more RN hours per patient day were associated with lower rates of FTR, controlling for non-RN staffing and hospital characteristics. They also found that patient turnover rates differed by unit type, with higher turnover on non-ICUs (56.1%) than ICUs (45.4%), but with no direct effect of patient turnover on FTR rates in either non-ICUs or ICUs. There was an interaction effect between patient turnover and RN staffing on non-ICUs, however, indicating that the association between RN staffing and FTR differed significantly depending on the level of patient turnover. They also found that higher technological complexity in hospitals was related to lower FTR ratios (Mark & Harless, 2010; Mark et al., 2004).

Risk adjustment

Ferris et al reported that more than two-thirds of patients with failure to rescue have multiple complications. A risk-scoring system based on preoperative variables predicted patients in the highest-risk category of failure to rescue with good accuracy. Cardiac events were associated with the highest failure-to-rescue rates, but stroke, renal failure, and pulmonary failure had nearly as high FTR risk. The following table includes the percent mortality with and without the complication using propensity matching on preoperative variables.

Sheetz and colleagues evaluated whether increased hospital care intensity (HCI) is associated with improved outcomes following seven major cardiovascular, orthopedic, or general surgical operations in the Medicare population. High-HCI hospitals had greater rates of major complications than low-HCI centers (risk ratio, 1.04; 95% CI, 1.03-1.05) and there was a decrease in failure to rescue at high compared with low-HCI hospitals (risk ratio, 0.95; 95% CI, 0.94-0.97). Using multilevel-models, HCI reduced the variation in failure-to-rescue rates between hospitals by 2.7% after accounting for patient comorbidities and hospital resources.

Table 2. Effect of Complication Type on Failure-to-Rescue Rates

Type of Complication	No. of Complications	Mortality Rate, %	
		Without Complication	With Complication
Cardiac	13 629	1.2	45.5
Renal	13 338	1.3	31.0
CNS	6130	1.5	28.0
Pulmonary	56 546	0.8	24.5
Sepsis/SIRS	45 464	1.1	17.8
Reoperation for bleeding	102 034	1.1	8.7
DVT/PE	17 631	1.5	8.0
Wound	41 844	1.5	4.8
Any serious complication	207 236	0.5	10.1

Abbreviations: CNS, central nervous system; DVT, deep vein thrombosis; PE, pulmonary embolism; SIRS, systemic inflammatory response syndrome.

Healthcare Disparities

Silber et al using Medicare data from six States found that when matching on age, sex, year, state, and the exact type of procedure, blacks had higher failure-to-rescue rates (6.1% vs. 5.1%, $P<0.001$) than whites. When preoperative medical risk factors were added to this matching algorithm, there was no significant racial difference in FTR. These findings confirm that race/ethnicity should not be included in PSI04 risk-adjustment.

Indicator specifications

Needleman et al, using PSI04 v 3.1, examined whether the accuracy of PSI04 (v3.1) could be improved by testing three exclusion rules using California HCUP Data. All POA-informed specifications of exclusion rules improved the C-statistic of the failure-to-rescue measure and its sensitivity, with modest losses of specificity. For the entire failure-to-rescue pool, the mortality rate was 22% among patients with hospital-acquired complications compared to 13% for patients with POA complications, and the risk-adjusted difference in mortality was 6.6% (from 1.9% for pneumonia to 7.1% for sepsis). Lengths of stay were longer for patients with hospital-acquired complications than for patients with POA complications: 14.1 versus 8.4 days, with a risk-adjusted difference of 4.3 days (from 2.2 days for shock/cardiac arrest to 5.4 days for sepsis and GI bleeding). For all tested specifications, higher licensed hours and proportions of registered nurses were statistically significant predictors of lower FTR rates.

Summary

1. Better processes and structures of care are associated with lower FTR rates.
2. Nursing type, higher nurse staffing, and lower unit turn-over are associated with lower FTR rates.
3. Use of residents is associated with higher complication rates, but lower mortality and better rescue efforts.
4. There is no new evidence to support treating academic hospitals differently than non-academic facilities in risk adjustment models.
5. Based on the work by Needleman et al, a patients' risk for death and extended (and more expensive) treatment appear to be higher when their secondary diagnoses are hospital-acquired complications versus being POA.
6. The work by Silber emphasis the need to risk-adjust for preoperative medical factors to account for healthcare disparities.

Appendix C. References

Blegen MA, Goode CJ, Park SH, Vaughn T, Spetz J. Baccalaureate education in nursing and patient outcomes. *J Nurs Adm.* 2013;43(2):89-94.

Blegen MA, Goode CJ, Park SH, Vaughn T, Spetz J. Baccalaureate education in nursing and patient outcomes. *J Nurs Adm.* 2013;43(2):89-94.

Brooke BS, Dominici F, Pronovost PJ, Makary MA, Schneider E, Pawlik TM. Variations in surgical outcomes associated with hospital compliance with safety practices. *Surgery.* 2012;151(5):651-659.

Brooke BS, Dominici F, Pronovost PJ, Makary MA, Schneider E, Pawlik TM. Variations in surgical outcomes associated with hospital compliance with safety practices. *Surgery.* 2012;151(5):651-659.

Brooke BS, Dominici F, Pronovost PJ, Makary MA, Schneider E, Pawlik TM. Variations in surgical outcomes associated with hospital compliance with safety practices. *Surgery.* 2012;151(5):651-659.

Castleberry AW, Clary BM, Migaly J, et al. Resident education in the era of patient safety: a nationwide analysis of outcomes and complications in resident-assisted oncologic surgery. *Ann Surg Oncol.* 2013;20(12):3715-3724.

Ferraris VA, Bolanos M, Martin JT, Mahan A, Saha SP. Identification of patients with postoperative complications who are at risk for failure to rescue. *JAMA Surg.* 2014;149(11):1103-1108.

Hyder JA, Wakeam E, Adler JT, DeBord Smith A, Lipsitz SR, Nguyen LL. Comparing Preoperative Targets to Failure-to-Rescue for Surgical Mortality Improvement. *J Am Coll Surg.* 2015;220(6):1096-1106.

Johnston MJ, Arora S, King D, et al. A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. *Surgery.* 2015;157(4):752-763.

Kutney-Lee A, Stimpfel AW, Sloane DM, Cimiotti JP, Quinn LW, Aiken LH. Changes in patient and nurse outcomes associated with magnet hospital recognition. *Med Care.* 2015;53(6):550-557.

Mark BA, Harless DW, McCue M, Xu Y. A longitudinal examination of hosptial registered nurse staffing and quality of care. Health Serv Res. 2004; 39(5):1629.

Mark BA, Harless DW, Spetz J, Reiter KL, Pink GH. California's minimum nurse staffing legislation: results from a natural experiment. *Health Serv Res.* 2013;48(2 Pt 1):435-454.

Mark BA, Harless DW, Spetz J, Reiter KL, Pink GH. California's minimum nurse staffing legislation: results from a natural experiment. *Health Serv Res.* 2013;48(2 Pt 1):435-454.

Mark BA, Harless DW. Nurse staffing and post-surgical complicatins using the present on admission indicator. Res Nurs Health. 2010; 33(1):35-47.

Mattke S, Needleman J, Buerhaus P, Stewart M, Zelevinsky K. Evaluating the role of patient sample defintiions for quality indicators senstive to nurse staffing patterns. Med Care. 2004; 42(2 Suppl):II21-II33.

Navathe AS, Silber JH, Small DS, et al. Teaching hospital financial status and patient outcomes following ACGME duty hour reform. *Health Serv Res.* 2013;48(2 Pt 1):476-498.

Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K. Nurse-staffing levels and the quality of care in hospitals. *N Engl J Med.* 2002;346(22):1715-1722.

Needleman J, Buerhaus PI, Vanderboom C, Harris M. Using present-on-admission coding to improve exclusion rules for quality metrics: the case of failure-to-rescue. *Med Care.* 2013;51(8):722-730.

Needleman J, Buerhaus PI. Failure-to-rescue: Comparing definitions to measure quality of care. *Medical Care.* 2007;45(10):913-915.

Park SH, Blegen MA, Spetz J, Chapman SA, De Groot H. Patient turnover and the relationship between nurse staffing and patient outcomes. *Res Nurs Health.* 2012;35(3):277-288.

Sheetz KH, Dimick JB, Ghaferi AA. The association between hospital care intensity and surgical outcomes in medicare patients. *JAMA Surg.* 2014;149(12):1254-1259.

Sheetz KH, Dimick JB, Ghaferi AA. The association between hospital care intensity and surgical outcomes in medicare patients. *JAMA Surg.* 2014;149(12):1254-1259.

Silber JH, Romano PS, Rosen AK, Wang Y, Even-Shoshan O, Volpp KG. Failure-to-rescue: comparing definitions to measure quality of care. *Med Care.* 2007;45(10):918-925.

Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. *Med Care.* 1992;30(7):615-629.

Appendix B

Comment ID 6370 on #1543: Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Artery Stenting (CAS)

Thank you for the opportunity to comment on the Society for Vascular Surgery's quality measure #1543 - Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Artery Stenting (CAS). The Society for Cardiovascular Angiography and Interventions (SCAI) is a non-profit professional association representing the majority of interventional cardiologists and cardiac cath lab teams in the United States. We believe that the review of the medical literature that was discussed while this Committee was deliberating omitted significant amounts of data and that it was inaccurate in summarizing the clinical guidelines on carotid stenting. The medical literature surrounding the NQF's review should be balanced and free from bias or self-promotion. We are concerned that comments offered by SVS did not encompass this ideal completely, instead focusing on highly selected subsets of data in favor of their position, but not including the totality of substantial evidence. As noted in the discussion below, we find conflict in their recommendation of support for carotid stenting in asymptomatic high risk surgical patients via surgical cut down based on one small trial, yet recommend against carotid stenting via percutaneous access in the same patients despite an overwhelming rigorous data base.

SCAI specific position stated below incorporates the following beliefs. We believe that both carotid endarterectomy and carotid stenting offer safe and effective treatment alternatives for patients with obstructive carotid disease, provided that the procedure is performed by an experienced operator with proven results and in carefully selected patients. Furthermore, we believe that these therapies are complementary, not competitive. There are clearly patients who are at high risk for endarterectomy who are at low risk for carotid stenting and those who are high stent risk but low surgical risk, and some patients who are better served by medical therapy without either form of revascularization. Furthermore, as we believe that the two therapies are equally effective at preventing strokes and maintaining carotid patency (discussed below), and that the vast majority of strokes occur in asymptomatic patients, there is a role for both interventions in carefully selected asymptomatic patients.

SCAI's position on carotid artery stenting in asymptomatic patients: Carotid stenting in asymptomatic patients can be recommended if:
1. Patient has a documented carotid stenosis of > or = 80% (2 modalities), AND
2. Patient has a life expectancy (of quality) of > or = 5 years, AND
3. Is performed by an experienced, certified operator with documented acceptable event rates, AND
4. The patient has pre- and post- procedural independent neurological evaluation AND is enrolled in a prospective data base AND
5. That the patient's risk/benefit for CAS individualized to their medical condition and carotid anatomy AND to the operator/institution results is at least as good as that of the same patient's individualized risk/benefit ration for carotid endarterectomy or medical therapy.

Finally this statement and our response to the SVS's statements are based on extremely

rigorous data bases including all 3 North American randomized comparisons of CEA vs. CAS (SAPHIRE, CREST, and ACT 1), the 10 year follow-up data on the CREST randomized trial (all four of these published in the New England Journal of Medicine), extremely large real world carefully regulated registries (representing over 40,000 patients prospectively studied with independent neurological adjudication), multiple IDE trials, and meta-analyses.

Based on extensive data, we believe that the totality of extensive data should be included and considered. The comprehensive data shows:

1. We acknowledge statistically increased risk of minor strokes for CAS over CEA in CREST. However, that alone cannot be used as the only determinant in decisions, and must be considered in the context of all information.

a. First despite increased peri-procedural minor strokes, there is NO difference in neurological outcomes in these patients. Careful neurological assessments in CREST showed no difference between CAS and CEA patients, and multiple trials have shown return to neurologic baseline in patients with small peri-procedural strokes b. These studies were done relatively early in CAS experience AND did not utilize proximal protection. Operator experience, improved case selection, and proximal protection have all been shown to reduce peri-procedural events.

2. There is no difference in major strokes or death between the two therapies in any of the North American randomized trials.

3. There is a consistent increase in cranial nerve injuries of CEA over CAS. While some resolve, others persist. If these were included to the same extent that minor peri-procedural strokes (which resolve as above), there would be no advantage of CEA over CAS.

4. There is NO difference between the 2 therapies for preventing long term ipsilateral strokes. BOTH therapies are excellent and durable.

5. There is either no difference or a slight advantage to carotid stenting in maintaining long term carotid patency. Again, both are excellent.

6. There is significant increase for peri-procedural myocardial infartions in CEA patients over CAS.

We believe this is very important to consider and monitor, as post procedural MI's do correlate with increased mortality, and patients with carotid artery disease frequently have concomitant coronary artery disease. In summary, we believe that there is compelling evidence (including the recent ACT 1 and 10 year CREST randomized trials) that carotid artery stenting is a viable treatment option for patient with severe carotid artery disease, especially those at high risk for carotid endarterectomy. To exclude it as a treatment option based on a small increased rate of peri-procedural minor strokes in earlier trials would be every bit as unfair as rejecting endarterectomy because of an increased rate of cranial nerve injuries or peri-procedural myocardial infarctions. As above, we believe that both therapies when performed by experienced operators in appropriately selected patients are excellent complementary procedures. We advocate for all societies to strive for established standards for carotid revascularization in a collaborative fashion, emphasizing the optimal benefits for patients independent of specialty bias.

I would like to thank D. Chris Metzger, MD, FSCAI, the Chair of our Carotid Stenting Committee

for developing these public comments.

Sincerely,

/s/

Kenneth Rosenfield, MD, MHCDS, MSCAI