NATIONAL QUALITY FORUM

# Memo

TO: Consensus Standards Approval Committee (CSAC)

FR: Helen Burstin, Chief Scientific Officer

Marcia Wilson, Senior Vice President, Quality Measurement

RE: Appeal of Measure #0351 for the Surgery 2015-2017 Project

DA: March 17, 2017

**ACTION REQUIRED**

The CSAC will review the letter of appeal and this memo in consideration of the appeal. The CSAC will determine whether to uphold measure endorsement for the following measures:

- #0351: *Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) (Agency for Health Care Research and Quality)*

The following documents are appended to this memo:

1. Appendix A: Appeal Letter from Tufts Medical Center for #0351: *Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04)*
2. Appendix B: Surgery Standing Committee response
3. Appendix C: Measure evaluation summary table. Please see the full report here.

**BACKGROUND**

In accordance with the NQF Consensus Development Process (CDP), the measures recommended by the Surgery Standing Committee were released for a 30-day appeals period, which closed on February 24, 2017. The National Quality Forum (NQF) has received one appeal of its endorsement of the measure listed above.

During the in-person meeting on August 16-17, the Committee agreed that the underlying evidence for the measure remained unchanged since the previous endorsement review and also agreed there was a gap in performance. Discussion of the scientific acceptability of the measure focused on a number of concerns including:

- whether claims data can accurately capture complications reliably;
- the need for additional testing since the risk adjustment strategy includes patients transferred in with complications present on admission, thus, inappropriately penalizing institutions financially;
- the measure not including patients transferring to other facilities, thus providing a potential for “gaming” (i.e., transferring the sickest patients to other facilities); and

- the absence of testing data that demonstrates the measure assesses what it is supposed to measure.

Ultimately, the Committee did not agree the measure met the Validity criterion and the measure was not recommended for endorsement.

During the member and public commenting period, the developer submitted a request for reconsideration on the grounds that: 1) the Committee did not appropriately review and evaluate the measure on the Validity criteria; 2) the Committee's discussion included concerns about how the measure might be used rather than focusing solely on scientific acceptability of the measure; and 3) a different NQF committee reviewed a similar measure and reached a different conclusion than did the Surgery Standing Committee, i.e., there had been an inconsistent review of measures across NQF standing committees. The developer also submitted additional testing data on transfers, risk adjustment, and use of claims data to measure complications.

On the post-comment call on November 3, 2016, the Committee reviewed the reconsideration request and the additional testing data submitted by the developer. The Committee agreed to reconsider the measure for endorsement. After a review and discussion of the additional data submitted, the Committee re-voted and passed the measure on the Validity criterion. The Committee agreed the measure was feasible; the Committee did not agree that the measure met the Usability and Use criterion, noting that the measure was not specific enough to aid providers in performance improvement and in recognizing patterns. Overall, the Committee recommended the measure for continued endorsement.

**SUMMARY OF APPEAL**

The appeal focused on the following issues:

- The validity of the measure is based on outdated information that is poorly generalizable;
- The measure implies that any hospital should be able to prevent death after surgery in patients with serious complications;
- The measure rates contradict disease-specific mortality findings; and
- The measure results in unintended consequences by penalizing an institution receiving high risk transfer cases.

**COMMITTEE RESPONSE TO THE APPEAL**

The appeal letter and the response from the measure developer were shared with the Committee; the responses from the Committee members were mixed. Please see Appendix B for Committee members' responses.

**MEASURE DEVELOPER RESPONSE TO THE APPEAL**

AHRQ appreciates the opportunity to address several concerns regarding NQF #0351 (PSI 04 - Death rate among surgical inpatients with serious treatable complications) that were raised in an appeal letter to the National Quality Forum from Tufts Medical Center dated February 24, 2017.

In the paragraphs below, we respond to these concerns and describe how they have already been addressed through the endorsement maintenance process and the conscientious efforts of the Surgery Standing Committee over the past year.

AHRQ recently announced that enhancements to the risk-adjustment approach for PSI 04 (as voted on by the NQF Surgery Standing Committee) will be incorporated into a forthcoming release of the AHRQ Patient Safety Indicators software, v6.03. (https://qualityindicators.ahrq.gov/News/PSI_v6.0_SASQI_WinQI_Memo.pdf)

**1. "The validity of the measure is based on outdated information that is poorly generalizable"**
The validity of the concept of "failure to rescue" or "death after serious treatable complications" has, in fact, been a subject of ongoing attention and analysis throughout the 25 years since Silber and colleagues first described the concept (Med Care 1992;30(7):615-629), the 15 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 12 years since PSI 04 was introduced.

Specifically, the NQF Evidence form under Section 1a.8.2 includes an extensive and detailed 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. A systematic review published in 2015 (Johnston MJ, et al. Surgery 2015; 157(4):752-63) identified 42 separate studies on measures of "failure to rescue," including several that highlighted the importance of "delayed escalation" related to "hierarchy and failures in communication." Based on this literature, 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.

In addition, the NQF 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. To further address this question, we used the publicly available HospitalCompare data from the Centers for Medicare & Medicaid Services (CMS) to estimate the direction and strength of associations between risk-standardized PSI 04 rates and structural measures of hospital quality reported to CMS. Among 1,362 hospitals that reported all of the relevant measures through June 30, 2013, those that participated in a surgery registry had 3.3 fewer events per 1,000 persons (p=0.006) and those that implemented safe surgery checklists had 3.6 fewer events per 1,000 persons (p=0.093). Participation in nursing and cardiac registries was not associated with risk-standardized PSI 04 rates among Medicare fee-for-service patients.

**2. "The assumption of the 'preventable death' is dubious for patients transferred to tertiary referral centers…"**

PSI 04 does not imply "that any hospital should be able to prevent death after surgery in patients with these serious complications." The title of the indicator simply describes these complications as "treatable," but there is no implication that all of these deaths are preventable – at tertiary referral centers or anywhere else. If that were the implication behind PSI 04, then there would be no need for risk-adjustment; individual deaths could simply be tallied and observed mortality rates could be reported, without further statistical analysis.

The health services research literature and AHRQ's own analyses confirm that preventability is a relative or proportional concept, not an absolute concept. Very few inpatient deaths are 100% preventable, and only some are 0% preventable. A substantial portion of the observed variation in hospital outcomes can be attributed to patients' severity of illness and pre-hospital care, another portion can be attributed to random noise, and only what is left over (after risk-standardization and reliability-adjustment or smoothing) can reasonably be attributed to quality of care. This conceptual rationale applies to all risk-standardized mortality measures, including PSI 04. Although one might challenge Tufts' assertion that "tertiary and quaternary medical centers" are "alone" in their ability "to provide evidence-based advanced medical and surgical care in critically ill patients," there is no doubt that risk-adjustment and carefully drawn exclusion criteria help to identify the true hospital quality signal.

For this reason (with feedback from the NQF Surgery Committee members), AHRQ has revised its PSI 04 risk-adjustment model for version 6.03 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. 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 reduces any residual bias resulting from the transfer of patients with severe complications.

AHRQ's updated models for v6.03 have significantly higher performance than the previous models. Specifically, the re-estimated c statistics, based on the AHRQ 34-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 the surgical outcome measures reviewed in 2016, 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."
With respect to the performance gap, AHRQ's analysis of all-payer data indicates that major teaching hospitals, defined as hospitals with 0.250-0.599 residents per licensed bed, have an aggregate observed-to-expected ratio of 0.969-0.972. In other words, major teaching hospitals have about 3% fewer deaths than expected, based on the five risk models described above. By comparison, non-teaching hospitals (with 45.3% of eligible patients) have an aggregate observed-to-expected ratio of 1.002-1.007. In other words, non-teaching hospitals have more deaths than expected, whereas major teaching hospitals have fewer deaths than expected (on average), demonstrating that there is indeed a "performance gap" that favors patients treated at major teaching hospitals, after risk-adjustment.

**"Unintended consequences … whether this measure has been adequately vetted."**
Tufts challenges AHRQ's assessment that "no evidence has been identified suggesting unintended consequences for this measure." AHRQ completed the NQF forms, presenting all relevant evidence about its measures, without regard to whether the evidence may be favorable or unfavorable. In this process, AHRQ follows NQF guidance in distinguishing between "evidence" and anecdotes. Although AHRQ is aware of and has been provided anecdotes suggesting that the measure lacks face validity (such as the four case examples described in the Appeal Letter), the current (version 3.0, September 2016) Committee Guidebook for the NQF

Measure Development Process states (page 52) that anecdotes do not qualify as evidence of unintended consequences:

"The Usability and Use criterion also reflects the need for consideration of unintended negative consequences of the measure to individuals or populations (if any). This consideration should not center on theoretical negative consequences but instead should be those that are supported by evidence (e.g., the nature of the unintended negative consequence, the affected party, the number of people affected, and the severity of the impact)."
See also NQF's Guidance for Evaluating Usability and Use of Performance Measures, which came out of the work of NQF's Usability Task Force in 2012:

"The Task Force agreed that unintended negative consequences should be considered under Usability along with the evidence of use and influence on quality. Negative consequences should be to individuals or populations. (Issues regarding fair comparisons among the entities whose performance is being measured would relate to the measure's validity.) It would not be feasible to request evidence that no adverse consequences occurred; however, the potential for unintended negative consequences should be considered in measure development, and this information should be solicited from users of endorsed measures. Reports of negative consequences should be accompanied by evidence including the nature of the consequence, the affected party, the number of people affected, and the severity of the impact."

**Case Examples – "Concern about the Face Validity of PSI 04"**
Tufts Medical Center describes four specific cases, all of which presumably triggered PSI 04. These cases are difficult to evaluate, because the case summaries provided to NQF are abbreviated by necessity and do not provide important information about the triggering complication, when and how that complication started, and the circumstances surrounding the patient's transfer to Tufts. However, we used the information provided to estimate the risk of death for each of these patients, according to the "new" v6.03 risk-adjustment model that was reviewed by the NQF Surgery Standing Committee. Please recognize that these estimates do not consider comorbid conditions that may slightly decrease or slightly increase the risk of PSI 04.

Case 1 describes a relatively young man (54 years old) who underwent successful cardioversion in the field, and was transported to a tertiary care center, where he was apparently diagnosed with coronary artery disease justifying intervention, but died for unclear reasons after suffering acute kidney injury (AKI) from muscle breakdown. His risk of PSI 04 is estimated at 41%-52%, depending on which "operating room procedure" drove the patient's MS-DRG assignment. These estimates have face validity, based on the summary provided. AKI due to rhabdomyolysis is generally considered to be a treatable condition, so this case (as summarized) may plausibly offer opportunities for improved care.

Case 2 describes a relatively young man (57 years old) with sepsis who developed cardiogenic shock and did not respond to appropriately aggressive efforts to support his cardiovascular function. His risk of PSI 04 is estimated at 50%-61%, depending on which "operating room procedure" drove the patient's MS-DRG assignment. These estimates have face validity based

on the patient's age and condition at presentation; his failure to respond to aggressive interventions did not become apparent until later in his hospital course.

Case 3 describes a slightly older patient (65 years old) who suffered a massive intracerebral hemorrhage, presumably due to an undiagnosed aneurysm or hypertensive emergency, and died after unsuccessful neurosurgical intervention (following transfer from a community hospital). This patient's risk of PSI 04 is estimated at 96%, indicating that the receiving hospital takes virtually no "penalty" for such an event.

Case 4 describes a transplant-eligible young man (40 years old) with severe chronic heart failure who was transferred, for unclear reasons, from one tertiary care center to another. He sustained "massive intracranial hemorrhage" for unclear reasons, and died shortly thereafter. His risk of PSI 04 is estimated at 46%-58%, depending on which "operating room procedure" drove the patient's MS-DRG assignment. These estimates have face validity, based on the summary provided. Intracranial hemorrhage of this magnitude is an unusual complication of heart failure, suggesting iatrogenic factors (e.g., anticoagulation, fall or other trauma) that may plausibly offer opportunities for improved care.

In summary, all four of these cases were very high risk transfers, and the receiving institution receives appropriate credit for that risk through enhanced risk-adjustment. However, at least two of these cases still warrant evaluation for potential opportunities to improve care at the receiving hospital.

**3. "PSI 04 rates contradict disease-specific mortality findings"**
Tufts Medical Center notes that its organization "consistently performs at the national benchmark or better in the CMS mortality measures" but it still has "a PSI 04 rate worse than the national rate." The authors of the appeal letter posit that this may be "true for other hospitals as well."

First, it is important to note that CMS mortality measures have focused on medical conditions such as acute myocardial infarction (AMI), heart failure, pneumonia, stroke, and chronic obstructive pulmonary disease (COPD), with the recent addition of one surgical procedure (coronary artery bypass graft surgery). By comparison, PSI 04 focuses on a broad spectrum of operative procedures across nearly all subspecialty domains and organ systems. Therefore, a strong relationship between PSI 04 and CMS mortality measures is not essential, and may not even be expected.

However, it is quite possible that some aspects of hospital quality affect all service lines, to at least some degree, so a positive correlation between these metrics at the hospital level would be reassuring. To address this question, we used the publicly available HospitalCompare data from CMS to estimate the direction and strength of correlations among these risk-standardized measures. We estimated Pearson correlation coefficients because hospital-specific, risk-standardized mortality rates are normally distributed. In fact, PSI 04 (reported for the period 7/1/2011-6/30/2013) is positively correlated with AMI mortality at r=0.18, COPD mortality at

r=0.15, heart failure mortality at r=0.13, pneumonia mortality at r=0.20, stroke mortality at r=0.14, and a composite of the CMS mortality measures at r=0.23 (reported for the period 7/1/2010-6/30/2013). In other words, on average, hospitals with lower risk-standardized mortality rates on CMS measures also tend to have lower risk-standardized mortality rates on PSI 04 (all comparisons $p<0.0001$).

Incidentally, we find the same correlation at the state level. Across the 34 states with hospitals that had Hospital Compare rates reported both PSI 04 and CMS risk-standardized mortality measures, PSI 04 rates are correlated with AMI mortality at r=0.41, COPD mortality at r=0.15, heart failure mortality at r=0.16, pneumonia mortality at r=0.40, and stroke mortality at r=0.004. In other words, geographic disparities for all of these measures, with the exception of stroke mortality, are mildly to moderately consistent.

**Conclusion**

AHRQ appreciates the opportunity to respond to the concerns raised in the appeal letter from Dr. Weingart and colleagues, dated February 24, 2017. AHRQ acknowledges the vital role that academic medical centers and other tertiary referral centers play in our health care system, and continually strives to improve the reliability, validity, and usefulness of its measures for multiple stakeholders. In fact, some of the same questions were raised in the NQF Surgery Standing Committee over the past year. In response, AHRQ implemented significant enhancements to its risk-adjustment models, and tested the impact of those enhancements, as described above. Based on these findings, the NQF Surgery Standing Committee recommended re-endorsement of NQF # 0351, just as a separate NQF committee (Patient Safety) recommended re-endorsement of the 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)." Of note, these re-endorsed measures #0353 and #0352 include ALL 30-day deaths and ALL inpatient deaths, respectively, without excluding patients who seek or require "advanced medical care."

AHRQ welcomes additional suggestions and input from users of PSI 04, and will continue to work with NQF on evidence-based enhancements to the measure specifications and risk-adjustment methods.

**Appendix A Appeal Letter from Tufts Medical Center for 0351: Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04)**

February 24, 2017

To Whom it May Concern,

We would like to submit an appeal for your consideration regarding PSI-04 and recommend that NQF reconsider this measure for endorsement.
PSI 04 - Death rate among surgical inpatients with serious treatable complications (shock/cardiac arrest, sepsis, pneumonia, deep vein thrombosis/ pulmonary embolism or gastrointestinal hemorrhage/acute ulcer).
Our objection to the measure is based on several concerns:

*1.* The validity of the measure is based on outdated information that is poorly generalizable.
.Section 1c.3 of the #0351 NQF measure rationale says "the underlying premise was that better hospitals are distinguished not by having fewer adverse occurrences but by more successfully averting death among (i.e., rescuing) patients who experience such complications. The original definition used by Silber et al was based on key clinical findings abstracted from the medical records of 2,831 cholecystectomy patients and 3,141 transurethral prostatectomy patients admitted to 531 hospitals in 1985." The surgical cases reviewed by Silber in 1985 could be cared for at community hospitals as well as academic medical centers. Given the current payer restrictions and acuity of patients receiving advanced therapies available today, the patients that are presently cared for at academic medical centers would not have survived long enough to make it to surgery in 1985. Taking a premise that is over 30 years old and applying it to today's patient population exposes this measure to the need for reconsideration.

*2.* The assumption of the "preventable death" is dubious for patients transferred to tertiary referral centers.
The implication of PSI 04 is that any hospital should be able to prevent death after surgery in patients with these serious complications. The reality is that there is a class of hospitals that recognize that these conditions in seriously compromised patients cannot be treated without extraordinary care and that sometimes, the care provided will not be enough to save the patient. These patients are then transferred to tertiary and quaternary medical centers, owing to the expertise and proficiency found in these institutions. They alone are able to provide evidence-based advanced medical and surgical in critically ill patients. Performing highly skilled procedures in a critically ill population as well as providing the clinical support needed for these complex patients should be proof that there is no "performance gap" as noted in section 1b of the #0351 NQF measure rationale. We would recommend consideration that these patients, inherently requiring very advanced therapies and interventions, be removed from the population of "treatable complications". We would strongly advocate for a definition that is more specific, with consideration of expanded risk adjustment variables, and specific exclusion criteria inclusive of those patients who seek advanced medical care as an option, given the severity of their underlying diagnoses and complications.

NQF Performance Gap statement on unintended consequences in section 4c brings into question whether this measure has been adequately vetted. "No evidence has been identified suggesting unintended consequences for this measure." Our experience, particularly in the 4$^{th}$ case example, shows evidence that we receive transfers outside of Silber's original population, and therefore exist outside of the "normal" expectation for "averting death". These high risk transfer cases should be excluded from the receiving institution's PSI 04 measure count. These extremely high risk cases are transferred to our institution owing to our expert capabilities in caring for the sickest of patients. Including these cases in the measure and then penalizing institutions for "harm" based on the otherwise inevitable patient outcome is incongruous with the ultimate goal of this very measure.

These 4 cases from our institution illustrate our concern about the face validity of the PSI-04 measure.

| Case | Summary |
|---|---|
| Case 1 | A 54 yr old collapsed at the airport, was cardioverted and intubated in the field. He was transferred to a tertiary care center where he had an urgent cardiac catheterization. He developed acute renal failure due to rhabdomyolysis. Dialysis was attempted but patient decompensated and died within 24 hrs of admission. |
| Case 2 | A 57 yr old treated for sepsis, intubated, and transferred to tertiary center in cardiogenic shock for advanced therapies. Unstable hemodynamically and in respiratory failure despite intubation, he required a heart pump, an additional machine to oxygenate his blood, and continuous hemofiltration for renal failure because he could not tolerate dialysis. He continued to decline and expired within 3 days of admission. |
| Case 3 | 65 yr old found confused and disoriented by EMS. At a community hospital found to be unresponsive, intubated and transferred. At the tertiary center found to have a large cerebral hemorrhage. Despite surgery to relieve cerebral pressure, the neurologic evaluation showed no purposeful movement and no improvement. The patient expired 20 days after admission. |
| Case 4 | 40 yr old waiting for a heart transplant and receiving advanced therapy for chronic heart failure was inpatient at a tertiary care center and transferred to another tertiary care center for care and management. He sustained massive intracranial hemorrhage. His condition progressed to grave and recovery unlikely given the extent of brain damage. He expired within 6 days after admission |

*3.* PSI-04 rates contradict disease-specific mortality findings.

While our organization consistently performs at the national benchmark or better in the CMS Mortality measures we have a PSI 04 rate worse than the national rate. Is this true for other hospitals as well?

We believe that PSI-04 is a flawed measure that does not justify NQF endorsement. We support the required research that would refine this measure with appropriate exclusion criteria so that tertiary centers are not penalized for providing the advanced care for which the community relies upon them.

Thank you for your consideration of our appeal,

Saul Weingart, MD CMO Tufts Medical Center
Catherine Feleppa Camenga, MS, RN Director of Quality and Patient Safety
Alex Pavoll, Manager of Quality Decision Support

**Appendix B Surgery Standing Committee Responses to Appeal for Measure #0351**

| Committee Member | Response |
|---|---|
| Alan Siperstein | Many of the concerns of the Tufts group were discussed at our committee meeting. I believe that the attribution question is a real one, where community hospitals are referring patients to more resourced centers. Although it is clearly desirable from a patient perspective for a smaller hospital to send deteriorating patients to a tertiary care center, a potential adverse consequence would be for those tertiary centers to resist such transfers. The Tufts group also gave specific examples where they were penalized on this metric and review of the case specifics raised questions whether those types of cases should be included. Although my own institution performs well on this metric, in discussion with our quality officers, they also cite clinical examples that did not seem aligned with the intention of this measure.<br><br>One of my issues with the PSI-4 maintenance submission (and this holds true for many of the maintenance measures) is the limited amount of information regarding how the measure has performed since it was last reviewed. A specific concern that was not addressed in the submission was the reliability of the administrative data used to flag the complications. A concern was that this may vary from institution to institution, particularly because this measure is used to compare different hospitals. (If a given hospital used consistent tracking mechanisms over time, then this metric would be helpful in assessing "rescue" longitudinally at that institution.) |
| Lee Fleisher | Thanks for pulling this together. For transparency, I continue to collaborate and I am a co-I on several grants with Jeff Silber which include failure to rescue measures. The measures continue to show value in measuring quality and variation in rates between hospitals. |
| Karl Bilimoria | There remain a number of issues with PSI-4, and I do not believe it should be endorsed by the NQF.<br><br>There is considerable national pushback regarding this particular PSI. It would be worthwhile to look at the comments sent to CMS regarding PSI-4 in during the last comment period for the Proposed Rule. A number of institutions, including mine, sent very detailed complaints about PSI-4 that totaled several pages of criticisms.<br><br>Several major issues remain from our in-person meeting that were not adequately addressed on the follow-up conference call:<br>1. The underlying issues with the validity of the data are particularly problematic. The data are not audited, and there are widely accepted institutional differences in coding practices (and overt gaming by hospitals). Data validity issues are supported by Cliff Ko's studies demonstrating a considerable lack of agreement between the admin data and clinical data. The admin data misses a considerable number of postoperative complications.<br>2. The adequacy of the risk adjustment is still a concern given the underlying data available, particularly adjustment for the complexity of the operation. |

| Committee Member | Response |
|---|---|
| | 3. The issues of VTE surveillance bias end up pulling more patients into the denominator for PSI 4 at VTE high-surveillance rate hospitals, and these patients frequently have many other underlying issues, so their death is unrelated to the subclinical VTE. Death from VTE is fairly infrequent, so more likely the death is due to a different cause.<br>4. Sepsis has a poorly defined definition that differs from institution to institution – the rules used to code sepsis leave a lot of discretion to the coders at the hospital and to the physicians. This was demonstrated last week in a query reported by Vizient members last week that showed hospitals defined PSI sepsis in very different ways.<br>5. There remain tremendous national concerns surrounding how PSI 4 inadequately handles transfers and conditions that are present on admission. These issues severely limit the ability to compare hospitals on this metric.<br>6. There are evidently mapping issues related to the transition to ICD-10 that need to be addressed.<br><br>As we discussed at the in-person meeting, all of these issues combined really destroy any confidence the public, hospitals, or the NQF should have in this metric.<br><br>On the follow up call that this committee had with Patrick Romano, there were a number of committee members who could not attend in the middle of a work day, and thus the vote may have been different if we had the whole committee present to review all these issues again from the in-person meeting. |
| Robert Cima | I agree with Karl. As an enterprise, Mayo, has been looking at this issue across our 20 plus hospitals and have found many of the concerns raised by Karl. I can assure you that these are not theoretical problems with the measure. |
| Adolph Yates | I concur. The diagnostic intensity has dramatically increased over the last two decades (e.g. CTA), which may explain the sense of the measure not being "fair". |
| Mark Jarrett | I am also in agreement. The measure has real issues - and we have enough measures to deal with already. Becomes hard to push staff when they have concerns about the measure itself. |
| Lawrence Moss | The link between quality and performance on this measure does not meet the NQF bar in my opinion. |

## Appendix C Measure Evaluation Summary Table

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

**Submission | Specifications**

**Description**: In-hospital deaths per 1,000 surgical discharges, among patients ages 18 through 89 years or obstetric patients, with serious treatable complications (shock/cardiac arrest, sepsis, pneumonia, deep vein thrombosis/ pulmonary embolism or gastrointestinal hemorrhage/acute ulcer). Includes metrics for the number of discharges for each type of complication. Excludes cases transferred to an acute care facility. A risk-adjusted rate is available. The risk-adjusted rate of PSI 04 relies on stratum-specific risk models. The stratum-specific models are combined to calculate an overall risk-adjusted rate.

**Numerator Statement**: Number of deaths (DISP=20) among cases meeting the inclusion and exclusion rules for the denominator.

**Denominator Statement**: Surgical discharges, for patients ages 18 through 89 years or MDC 14 (pregnancy, childbirth, and puerperium), with all of the following:

- any-listed ICD-9-CM or ICD-10-PCS procedure codes for an operating room procedure; and
- the principal procedure occ

**Exclusions**: Exclude cases:

- transferred to an acute care facility (DISP = 2)
- with missing discharge disposition (DISP=missing), gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing)

**Adjustment/Stratification**: Statistical risk model

"The predicted value for each case is computed using a hierarchical model (logistic regression with hospital random effect) and covariates for gender, age (in 5-year age groups, except for the youngest age range), Modified Diagnosis Related Groups (ie. MS-DRGs without any distinction for “comorbidity and complications” (CC/MCC), Elixhauser Comorbidity Index (https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp), Major Diagnosis Categories (MDC) based on the principal diagnosis, and transfer in from another acute care hospital. A parsimonious model was identified using a backward stepwise selection procedure with bootstrapping. The expected rate is computed as the sum of the predicted value for each case divided by the number of cases for the unit of analysis of interest (i.e., hospital). The risk-adjusted rate for the overall PSI 04 is calculated as the observed to expected ratio multiplied by the reference population rate, where the observed and expected values are summed across five strata (categories) of PSI 04 risk. This approach differs from other AHRQ Patient Safety Indicators without strata, in that each discharge-record’s expected value is computed using one of five distinct stratum-specific risk adjustment models that correspond to an assigned PSI 04 stratum. The five PSI 04 strata group records together based on secondary diagnoses that represent complications of care, and place the patient at risk of death (which is the numerator of PSI 04).

Additional information on methodology can be found in the Empirical Methods document on the AHRQ Quality Indicator website (www.qualityindicators.ahrq.gov). The Empirical Methods are also attached in the supplemental materials.

The specific covariates for this measure are provided for each Stratum as part of the Technical Specifications attached to section S.2b.

Source: http://www.qualityindicators.ahrq.gov/Modules/psi_resources.aspx"

**Level of Analysis:** Facility

**Setting of Care:** Hospital/Acute Care Facility

**Type of Measure**: Outcome

**Data Source**: Administrative claims

**Measure Steward**: Agency for Healthcare Research and Quality

**STANDING COMMITTEE MEETING 08/16 - 08/17/16**

**1. Importance to Measure and Report: <u>The measure meets the Importance criteria</u>**

(1a. Evidence, 1b. Performance Gap)

1a. Evidence: **Accepted previous evaluation**; 1b. Performance Gap: **H-6**; **M-16**; **L-0**; **I-0**

<u>Rationale</u>:

- The Committee noted that evidence presented with the recent submission is directionally the same as when last considered, at which time the measure passed on evidence, thus the Committee accepted the previous evaluation of evidence without vote.
- A member observed that the performance gap has improved by about 6% per year; however, significant gap remains in that there are some 43,000 deaths/year in 34 states as measured in all payer datasets. Further there are variations in the deaths by age, insurance status and other groupings. The Committee agreed that there is an actionable gap.
- The Committee noted that consideration should be given to including the pediatric population in this measure going forward.

**2. Scientific Acceptability of Measure Properties: <u>The measure does not meet the Scientific Acceptability criteria</u>** <u>(See Section 6 for the results of the revote on Validity following a reconsideration of the measure.)</u>

(2a. Reliability - precise specifications, testing; 2b. Validity - testing, threats to validity)

2a. Reliability: **H-1**; **M-15**; **L-5**; **I-2;** 2b. Validity: **H-0**; **M-9**; **L-10**; **I-4**

<u>Rationale</u>:

- In discussing inclusion of conditions that are present on admission (POA), AHRQ staff stated that analyses had shown that excluding patients with conditions POA did not improve validity of the measure but did reduce the number of cases that could be captured.
- The Committee discussed the specification that excludes patients from the denominator who are transferred to an acute care hospital in terms of potential for "gaming" the measure by transferring patients, particularly if patient condition worsens. The developer representative agreed there is a small window for gaming but stated there is not a way to assess the outcome of interest in such cases since hospitalizations cannot now be linked.
- The Committee raised several concerns about transfers, specifically:
    - In addressing the effect of cases where hospitals receive patients in transfer, with complications of interest who then die, the developer stated that these cases are not excluded from the measure because they contribute to detectable signal; rather they are handled with risk adjustment. They further noted that patients received in transfer have lower rates of death.
    - The Committee noted that it did not see specific testing data that the measure assesses what it is supposed to be measuring. Members also noted that, based on the data provided the number of patients transferred out and excluded is not a high number (3% of 300,000).
    - The Committee noted that transferring patients to higher levels of care is often the right thing to do but expressed concern that risk adjustment to handle patients transferred in cannot fully address the issue that the receiving hospital becomes responsible for events it cannot control. Further, the Committee stated that retaining these patients to improve signal is concerning and penalizes the receiving hospital.
    - The Committee also questioned whether the transfer issues were addressed adequately to understand threats to validity and, separately, that the handling of transfers makes it impossible to validate that appropriate effort was made to save the patient while in-hospital analysis over time could provide useful information.

- The Committee suggested that the developers provide sensitivity data around transfers out including facility variability analyzed in terms of such things as rural/urban, high technology/low technology, large/small as well as impact of transfers by looking at hospitals with and without that data. The developers stated they could provide this information.

- The Committee expressed concern that while claims data are a reliable way to identify a population of interest and will provide patient death, it has limitations in its ability to accurately capture complications.
- Members noted that studies comparing clinical to administrative data, false negative and high false positive rates have been found. Committee members acknowledged that coding variability among institutions can occur with clinical as well as administrative data and further noted that, particularly for multifactorial complications, significant discrepancies using administrative data have been found.
- In its discussion of SDS, the Committee agreed that there is no conceptual basis for inclusion of SDS factors in risk adjustment model.

**3. Feasibility: H-6; M-10; L-0; I-0**

*(3a. Clinical data generated during care delivery; 3b. Electronic sources; 3c. Susceptibility to inaccuracies/ unintended consequences identified 3d. Data collection strategy can be implemented)*

Rationale:

- On the post-comment call, the Committee agreed the measure was feasible, noting that the measure was straightforward and data sets are readily available.

**4. Usability and Use: H-2; M-4; L-9; I-1**

*(Used and useful to the intended audiences for 4a. Accountability and Transparency; 4b. Improvement; and 4c. Benefits outweigh evidence of unintended consequences)*

Rationale:

- On the post-comment call, the Committee discussed that the measure was not specific enough to aid providers in performance improvement and may not be useful in comparing hospital quality. The developer stated the measure should be used to track rates over time and not tracked by individual cases.

**Standing Committee Recommendation for Endorsement**: **Y-10; N-5**

Rationale:

Since the measure failed on Validity during the in-person meeting, the Committee did not take a vote for overall suitability for endorsement. During the post-comment call, the Committee re-voted and passed the measure on Validity and voted on the remaining criteria. The Committee then voted on overall suitability for endorsement.

**6. Public and Member Comment**

- The developer submitted a request for reconsideration during the member and public commenting period:

We are writing to request that the National Quality Forum (NQF) Surgery Standing Committee reconsider the decision to remove endorsement of Death Rate Among Surgical Inpatients with Serious Treatable Complications (PSI 04), (NQF 0351). This long-standing Patient Safety Indicator (PSI) has been endorsed by NQF since 2008. Our request for reconsideration is based on concern that NQF's standard review process was not applied properly during the in-person meeting on August 16, 2016, particularly with respect to the following:

1) Appropriate review and evaluation of the measure for Criteria 2. Scientific Acceptability Sub-criteria 2a. Validity
2) Discussion of the use case of the measure prior to full discussion of the scientific acceptability for the measure
3) Consistent evaluation of related (not competing) measures across NQF standing committees

First, according to the NQF's Guidance for Evaluating Validity and as noted by Karen Johnson, Senior Director, during the review, measure developers need only submit validity testing with respect to computed performance measures scores, not data element validity. AHRQ submitted information about construct validity, which should have been the focus of the validity discussion, not the detailed discuss of claims data and data element validity.

Second, although AHRQ acknowledges the difficultly of conducting reviews that are use-agnostic, the reviewers brought up concerns about the use of the measure by CMS during scientific acceptability discussions. It is AHRQ's understanding the NQF seeks to endorse measures that are deemed scientifically rigorous and suitable for not just quality improvement but also general accountability purposes (not specific accountability purposes). The NQF review process is intended to be use-agnostic. Specific use cases of the measure, particularly the appropriate use of the measures in CMS programs, are to be discussed during NQF's Measure Application Project committee meetings.

Third, while acknowledged in the introduction of the measure, NQF's re-endorsement of a related measure by the Patient Safety Standing Committee was not emphasized during the review discussions. In particular, in the course of that re-endorsement discussion for NQF 0352 (Failure to Rescue In-Hospital Mortality, risk adjusted), which was developed and is stewarded by the Children's Hospital of Philadelphia, the Patient Safety Committee carefully evaluated the design of "failure to rescue" measures. This Committee discussed and accepted the developer's evidence-based arguments in favor of including patients who had reported complications present on admission in the measure denominator. When different NQF Standing Committees fail to evaluate similar measures, with similar design features, in a consistent manner, the consequences include confusion across the stakeholder community and mixed messages to measure developers, stewards, and users.

In addition, as noted in the NQF-Endorsed Measures for Surgical Procedures 2015-2017: Draft Report for Comment (September 22, 2016), reviewers wanted additional information about transfers, risk adjustment and use of claims data to measure complications.

AHRQ respectfully requests that NQF ask that the Committee exercise the option to re-vote on the validity of the measure during the post-comment call to preserve the integrity of the NQF process, and consider the additional information being submitted by AHRQ.

<u>NQF Post Comment Call</u>

- On the post draft report comment call, the Committee reviewed the reconsideration request and the additional testing data submitted by the developer. Ultimately, the Committee agreed to reconsider the measure for endorsement.
    - The Committee noted that the issue of transfers was addressed through the additional sensitivity analysis showing that including or excluding transfers would have little effect on the outcome. The developer confirmed that the measure would risk adjust for transfers and whether the patient arrived at the hospital with a complication already present.
    - The Committee also questioned the potential surveillance bias of including deep vein thrombosis (DVT) and pulmonary embolism (PE), since hospitals that detect more DVT or PE will have more cases in the denominator. The developer stated that studies have shown high performing hospitals with effective multi-disciplinary teams can intervene early on and prevent an adverse outcome.
    - In addressing the Committee's concern that some hospitals may game the measure by transferring patients out before they die, the developer acknowledged that the issue was inherent among smaller or rural hospitals that transfer patients to larger, teaching hospitals. The developer also stated events such as post-operative complications that are counted in the denominator for this measure, are also identified in the numerator in other patient safety measures. The developer also stated they have tried to create a severity flag with the

| |
|---|
| administrative data to be able to detect the severity of the patient's condition when transferred to the receiving hospital.<br>o The Committee again raised that while administrative data is more useful to track individual hospitals, there are still concerns in terms of hospitals' ability to compare their performance to others, based on how well administrative data are collected. Ultimately, the Committee re-voted and passed the measure on the Validity criterion<br>o The Committee agreed the measure was feasible, and in discussion of usability, did not agree that the measure met this criterion, noting that the measure was not specific enough to aid providers in performance improvement and in recognizing patterns. Overall, the Committee recommended the measure for continued endorsement.<br><br>**Vote Following Consideration of Public and Member Comments:**<br>**Validity: H-4; M-10; L-2; I-1** |
| **7. Consensus Standards Approval Committee (CSAC) Vote (December 21, 2016): Y-16; N-0; NR-1**<br>• Decision: Approved for continued endorsement |
| **8. Board of Directors Vote (January 25, 2017): Yes**<br>• Decision: Ratified for continued endorsement |
| **9. Appeals** |