Using Estimated True Safety Event Rates versus Flagged Safety Event Rates: Does It Change Hospital Profiling and Payment?

Amy K. Rosen, Qi Chen, Ann M. Borzecki, Marlena Shin, Kamal M. F. Itani, and Michael Shwartz

---

Objective. To assess whether use of the AHRQ Patient Safety Indicator (PSI) composite measure versus modified composite measures leads to changes in hospital profiles and payments.

Data Sources/Study Setting. Retrospective analysis of 2010 Veterans Health Administration discharge data.

Study Design. We used the AHRQ PSI software (v4.2) to obtain PSI-flagged events and composite scores for all 151 hospitals in the database ($n = 517,814$ hospitalizations). We compared the AHRQ PSI composite to two modified composites that estimated “true safety events” from previous chart abstraction findings: one with modified numerators based on the positive predictive value (PPV) of each PSI, and one with similarly modified numerators but whose denominators were based on the expected fraction of PSI-eligible cases that remained after removing those PSIs that were present-on-admission (POA).

Principal Findings. Although a small percentage (5–6 percent) of hospitals changed outlier status based on modified PSI composites, some of these changes were substantial; 30 and 19 percent of hospitals changed ≥ 20 ranks after adjustment for PPVs and POA flags, respectively. We estimate that 33 percent of hospitals would see a change of at least 10 percent in performance payments.

Conclusions. Changes in hospital profiles and payments would be substantial for some hospitals if the PSI composite score used weights reflecting the relative prevalence of true versus flagged events.

Key Words. Patient safety measurement, quality of care, AHRQ Patient Safety Indicators (PSIs), hospital profiling

---

There is growing emphasis on defining the quality and safety of health care through reporting of process and outcome measures. The Agency for Healthcare Research and Quality’s (AHRQ) Patient Safety Indicators (PSIs) were

designed to screen for potentially preventable complications of inpatient care using hospital discharge data (AHRQ 2003). Although they were originally intended for case finding and quality improvement (QI), they are increasingly used for comparing hospital performance. The Centers for Medicare and Medicaid Services (CMS) recently added several of the PSIs to its Hospital Compare website for the Hospital Inpatient Quality Reporting (IQR) Program (CMS 2013a). Beginning in 2015, the National Quality Forum (NQF)-endorsed PSI composite measure (PSI-90) will also be used in the Hospital Value-based Purchasing (HVBP) Program to reduce payment for Medicare hospitalizations based on their achievement and improvement performance scores relative to benchmark quality standards (CMS 2013b).

Because of these high-impacts, high-stakes programs, it is critical that the PSIs accurately reflect hospital quality. Concerns over the use of the PSIs for comparative reporting and pay-for-performance are prevalent in the scientific literature and popular press. For example, previous studies examining the criterion validity of the PSIs using explicit chart review found that flagged PSI events do not always represent true safety events (Utter et al. 2009, 2010; White et al. 2009; Sadeghi et al. 2010; Zrelak et al. 2011; Rosen et al. 2012). The positive predictive validity (PPV), the proportion of flagged cases confirmed by chart review (the “gold standard”) to have a PSI, ranged from 28 percent for Postoperative Hip Fracture to 91 percent for Accidental Puncture or Laceration. Most of the false positives were attributed to well-known coding limitations inherent in administrative data and lack of present-on-admission (POA) codes (Iezzoni 1990, 1997; Utter et al. 2009, 2010; White et al. 2009; Zrelak et al. 2011; Rosen et al. 2012).

In addition, criticisms have been raised about the specific methods used in calculating the PSI composite, as well as the moderate reliability of

---

Address correspondence to Amy K. Rosen, Ph.D., Center for Healthcare Organization and Implementation Research (CHOIR), VA Boston Healthcare System; Department of Surgery, Boston University School of Medicine, 150 S. Huntington Avenue, Boston, MA 02130; e-mail: akrosen@bu.edu. Qi Chen, M.D., Ph.D., and Marlena Shin, J.D., M.P.H., are also with the Center for Healthcare Organization and Implementation Research (CHOIR), VA Boston Healthcare System, Boston, MA. Ann Borzecki, M.D., M.P.H., is with the Center for Healthcare Organization and Implementation Research (CHOIR), Bedford VA Medical Center, Bedford MA; Department of Health Policy and Management, Boston University School of Public Health, Bedford, MA. Kamal M. F. Itani, M.D., is with the Department of Surgery, Boston University School of Medicine; Department of Surgery, VA Boston Healthcare System; Harvard Medical School, Boston, MA. Michael Schwartz, Ph.D., is with the Center for Healthcare Organization and Implementation Research (CHOIR), VA Boston Healthcare System; Department of Operations and Technology Management, Boston University School of Management, Boston, MA.

individual PSI component measures (i.e., the ability to distinguish high- from low-quality hospitals), potentially threatening the reliability and validity of the composite measure (Schone, Hubbard, and Jones 2011). Although use of a composite measure may improve reliability, it is not known whether the PSI composite does, in fact, improve reliability with respect to hospital profiling (AHRQ 2008; Schone, Hubbard, and Jones 2011). Thus, hospital rankings based on flagged PSI composite rates could be affected by limitations inherent in the use of an administrative data-based measure. An unintended consequence of the HVBP Program would be to penalize all hospitals with high rates—those whose rates reflect poor performance, as well as those whose rates might reflect better coding practices and documentation of events (Haut et al. 2007). It might also cause hospitals to misdirect QI activities and focus on the wrong areas.

Given the lack of knowledge about the ability of the PSI composite measure to accurately detect differences between high- and low-quality hospitals, and the potentially serious implications of this, we conducted a study in the Veterans Health Administration (VA) to assess whether use of PSI-flagged events obtained from the AHRQ PSI composite measure versus estimates of true safety event rates derived from two modified composite measures based on previous chart abstraction findings led to changes in hospital ranks, hospital categorization, and hospital payments under a pay-for-performance program.

Because the VA has fewer financial incentives for the upcoding of diagnoses and lacks penalties or rewards tied to PSI events compared to the private sector (Rosen et al. 2012), it represents an ideal setting in which to study the impact of using routinely collected hospital discharge data on hospital performance. However, due to increasing financial imperatives to reduce the excess costs associated with patient safety events, as well as an already constrained global budget, it is likely that the VA will adopt similar payment policies to control costs in the near future. Thus, evaluating any changes that occur from using modified composites versus the original AHRQ PSI composite measure on hospital profiling in the VA will hopefully yield important insights into areas that are modifiable, as well as opportunities for improving the measure.

METHODS

Data Source

We conducted a retrospective study using VA Fiscal Year 2010 (October 1, 2009–September 30, 2010) inpatient administrative data. Hospital discharge

information included demographics, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis and procedure codes, utilization, and summary information related to each episode of care from the VA's National Patient Care Database. The database contained a total of 151 VA hospitals (VA Information Resource Center 2002). Based on prior work, we eliminated nonacute care stays from the data as the PSIs were designed to screen for patient safety events in the acute care setting (Rivard et al. 2005; Rosen et al. 2005). We also used a previously developed algorithm to distinguish elective from nonelective admissions because VA data lack an admission type that is required for the denominator of three surgical PSIs included in the AHRQ PSI composite (Rivard et al. 2005; Rosen et al. 2005). Our final database included a total of 517,814 hospitalizations from which PSI composite rates were calculated.

AHRQ PSI Composite Measure

We selected the AHRQ PSI composite as the outcome measure, rather than individual PSIs, because of the relatively low frequencies of several of the PSIs, as well as mounting concerns about the reliability of individual indicators (Schone, Hubbard, and Jones 2011). As with other composite measures, it provides both a broad picture of quality across multiple domains, as well as allows one to determine which individual components contribute most to assessed performance. Further, use of a composite measure may improve reliability (AHRQ 2008).

The PSI composite is calculated using the weighted average of the scaled and reliability-adjusted ratios for each of the component indicators (AHRQ 2008). The composite measure includes all the provider-level indicators with the exception of Death Among Surgical Inpatients with Serious Treatable Complications (PSI #4) (considered a composite itself), and Foreign Body Left During Procedure (PSI #5) and Transfusion Reaction (PSI #16), which are reported as counts. The software calculates rates for three PSIs (Postoperative Hemorrhage or Hematoma, Physiologic and Metabolic Derangement, and Respiratory Failure); however, as they were not officially endorsed by the NQF at the time the composite was developed, it excludes them from the composite by assigning them zero weights.

To calculate the PSI composite for this study, we applied the AHRQ PSI software (version 4.2) to the data to obtain observed ("flagged") rates and estimate 95 percent confidence intervals (CIs) for

the risk-adjusted rates of each of the hospitals (AHRQ 2003). Risk adjustment was conducted first at the individual level by taking into account specific patient characteristics at discharge: age, gender, DRGs, and 27 comorbidities based on the AHRQ Comorbidity Software (Healthcare Cost and Utilization Project [HCUP] 2011). A hospital's risk-adjusted rate (i.e., the rate the hospital would have if it had an average case mix) was then calculated by dividing its observed PSI rate by its expected rate and multiplying the O/E ratio by the indicator rate for the reference population (AHRQ 2003). Of note, VA databases do not yet contain POA flags; thus, the risk-adjusted rates included PSI events that may have been POA.

Hospital Profiling and Payment

Similar to the approach used on the CMS Hospital Compare website (CMS 2013a) to profile hospitals, we computed a 95 percent CI for each hospital's PSI composite score, and then assigned hospitals to performance categories based on a comparison of each hospital's 95 percent CI with the national VA PSI composite score. If the hospital's CI included the national VA composite score, then the hospital's performance was considered to be "no different from" the national VA performance. However, if the hospital's CI was below or above the overall VA's composite score, the hospital was considered performing "better than" or "worse than" the national average, respectively.

We also examined the impact in payment that each hospital would receive under a pay-for-performance (P4P) program using an approach that is conceptually similar to that used by the CMS. We considered only the achievement part of the performance score and not the improvement part because of the lack of longitudinal data. As such, CMS defines M as the median PSI composite score, B as the benchmark score (mean of the top 10 percent of hospitals), and P as the performance of an individual hospital. A hospital's performance score is then calculated as a linear transformation of $(P-M)/(B-M)$ and their payment calculated as an additional linear transformation of the score combined with scores from other dimensions (CMS 2013b). We calculated a hospital's score as $(\text{max} - P)/(\text{max} - \text{min})$, where "max" is the highest (worst) PSI composite score across our hospitals and "min" the lowest (best). The percentage of the payment pool awarded to each hospital was their performance score divided by the sum of all of the performance scores.

Simulating PSI Composite Measure Based on “Estimated True Rates”

We simulated hospital profiles using “estimated true events” based on the PPVs obtained from prior VA research (Rosen et al. 2012). These PPVs included all false-positive cases (i.e., those that were POA or due to other reasons, such as miscoding or lack of coding specificity) (see Appendix SA1 for individual PPVs). The first simulation was designed to represent “estimated” true PSI composite rates for those hospitals (i.e., all VA hospitals and some private sector hospitals) that do not yet have POA flags incorporated into administrative data (herein called “VA-PPV-adjustment”). To estimate true PSI composite rates, we first recalculated the observed PSI rates by including only estimated true cases (i.e., the flagged rate for $PSI_i \times PPV_i$); next, we modified the numerator weights used to create the PSI composite measure by including only estimated true cases [i.e., the new weight for $PSI_i = (\text{original weight for } PSI_i \times PPV_i) / \sum (\text{original weight for } PSI_i \times PPV_i)$]. These steps created a “VA-PPV-adjusted PSI composite score.” We used numerator weights, as recommended by AHRQ, to reflect the impact of the number of adverse events for each individual PSI in the composite.

The second simulation of estimated true PSI composite rates was designed for those hospitals covered under the CMS Inpatient Prospective Payment System that are mandated to incorporate POA flags into their data (Department of Health and Human Services [DHHS] 2010) (herein called “CMS-PPV-adjustment excluding POA”). We modified the algorithms described above by excluding those flagged PSIs that were false positives due to POA, and recalculated the PPVs of each PSI to create a “CMS-PPV-adjusted PSI composite score excluding POA.” These modifications were guided by previous VA chart abstraction findings that showed that eliminating cases with POA improved PPVs, particularly for those PSIs with the lowest PPVs, such as Pressure Ulcer, where 59 percent of flagged cases were POA (Rosen et al. 2012) (see footnote added to Appendix SA1 for more details). Specifically, we recalculated the observed PSI rates [i.e., the flagged rate for $PSI_i \times (PPV \text{ excluding } POA_i)$]. We assumed that these PSIs would not have been flagged by the original PSI composite because they occurred prior to index hospitalization. We then recalculated the numerator weights for the PSIs [i.e., the new weight for $PSI_i = (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i) / \sum (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i)$].

Analyses

We first examined changes in hospital ranks when the original PSI composite score was modified using PPV flags and also POA exclusions. Next, we assessed changes in hospital rankings, first in categorization (high, average, low performing), and second in payment under P4P program using the original PSI composite, the VA-PPV-adjusted PSI composite, and the CMS-PPV-adjusted PSI composite excluding POA. All analyses were performed using SAS software (version 9.1) (Statistical Analysis System [SAS] Version 9.1 2004).

RESULTS

Numerator Weights of Individual PSIs in the Original AHRQ PSI Composite and Modified PSI Composites

As shown in Table 1, there were changes in the weights based on PPVs and lack of POA. For example, when adjusted for PPV (i.e., estimated true rate), the weight for Pressure Ulcer (PSI #3) decreased from 0.2404 to 0.1334, whereas the weight for Accidental Puncture or Laceration (PSI #15) increased from 0.2983 to 0.4691. Compared to the sum of the weights of the three least prevalent PSIs (Postoperative Hip Fracture, Sepsis, and Wound Dehiscence), which was about 0.05 percent, the sum of the weights of the three most prevalent PSIs (Pressure Ulcer, Postoperative Pulmonary Embolism/Deep Vein Thrombosis, and Accidental Puncture or Laceration) comprised about 80 percent of the total composite score. This occurred in both the original AHRQ PSI composite as well as the two modified PSI composites.

Change in Hospital Ranks

As highlighted in Figure 1A, modifying the AHRQ PSI composite score by PPV flags resulted in a change in rank for most hospitals. Using the VA-PPV-adjusted PSI composite scores, the median absolute change in rank position among hospitals was 11 places (interquartile range [IQR], 3–23). In 98 hospitals, the rank changed by >5, in 77 by >10, and in 46 by >20.

Similarly, when using the CMS-PPV-adjusted PSI composite scores excluding POA (Figure 1B), the median absolute change in hospital rank position was 10 (IQR, 5–18). Specifically, in 109 hospitals, the rank changed by >5, in 74 by >10, and in 28 by >20. Appendix SA2 illustrates in detail how two hospitals can change ranks (decrease or increase rank) based on the specific

Table 1: Numerator Weights of Individual Patient Safety Indicators (PSIs) in AHRQ PSI Composite and Modified PSI Composites

<i>PSIs in AHRQ PSI Composite Measure</i>	<i>Original Weights*</i>	<i>VA-PPV-Adjusted Weights†</i>	<i>CMS-PPV-Adjusted Weights Excluding Present-on-Admission (POA)‡</i>
#3 Pressure ulcer	0.2403	0.1334	0.2557
#6 Iatrogenic pneumothorax	0.0457	0.0617	0.0539
#7 Central venous catheter-related blood stream infection	0.1280	0.0900	0.0847
#8 Postoperative hip fracture	0.0011	0.0006	0.001
#9 Postoperative hemorrhage or hematoma	0§	0	0
#10 Postoperative physiologic and metabolic derangement	0§	0	0
#11 Postoperative respiratory failure	0§	0	0
#12 Postoperative pulmonary embolism/ Deep vein thrombosis	0.2360	0.1877	0.1697
#13 Postoperative sepsis	0.0383	0.0376	0.0336
#14 Postoperative wound dehiscence	0.0124	0.0200	0.0155
#15 Accidental puncture or laceration	0.2983	0.4691	0.386

Note. Definition of the PSIs: PSI #3 Pressure Ulcer: Cases of pressure ulcer per 1,000 discharges with a length of stay >4 days; PSI #6 Iatrogenic Pneumothorax: Cases of iatrogenic pneumothorax per 1,000 discharges; PSI #7 Central Venous Catheter-related Bloodstream Infections: Cases of ICD-9-CM codes 99662 or 9993 or 99931 per 1,000 discharges; PSI #8 Postoperative Hip Fracture: Cases of in-hospital hip fracture per 1,000 surgical discharges with an operating room procedure; PSI #9 Postoperative Hemorrhage or Hematoma: Cases of hematoma or hemorrhage requiring a procedure per 1,000 surgical discharges with an operating room procedure; PSI #10 Postoperative Physiologic and Metabolic Derangement: Cases of specified physiologic or metabolic derangement per 1,000 elective surgical discharges with an operating room procedure; PSI #11 Postoperative Respiratory Failure: Cases of acute respiratory failure per 1,000 elective surgical discharges with an operating room procedure; PSI #12 Postoperative Pulmonary Embolism or Deep Vein Thrombosis: Cases of deep vein thrombosis (DVT) or pulmonary embolism (PE) per 1,000 surgical discharges with an operating room procedure; PSI #13 Postoperative Sepsis: Cases of sepsis per 1,000 elective surgery patients with an operating room procedure and a length of stay >4 days; PSI #14 Postoperative Wound Dehiscence: Cases of reclosure of postoperative disruption of abdominal wall per 1,000 cases of abdominopelvic surgery; PSI #15 Accidental Puncture or Laceration: Cases of technical difficulty (e.g., accidental cut or laceration during procedure) per 1,000 discharges.

*Original numerator weights of individual PSIs built in the PSI software version 4.2. (AHRQ quality indicators).

†VA-PPV-Adjusted numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV_i) / \sum (\text{original weight for } PSI_i \times PPV_i)$. This simulates the current VA scenario, in that the VA has not yet incorporated POA flags into administrative data.

‡CMS-PPV-Adjusted excluding POA numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV_{\text{excluding POA}_i}) / \sum (\text{original weight for } PSI_i \times PPV_{\text{excluding POA}_i})$. This simulates the current CMS scenario, because CMS has incorporated POA flags into Medicare claims data.

§These three PSIs do not have any weights assigned to them. When the AHRQ PSI composite measure was developed, these PSIs were not yet endorsed by the National Quality Forum. Only PSIs that were officially endorsed were given weights in the composite measure.

calculations used in calculating the original AHRQ PSI composite, the VA-PPV-adjusted PSI composite, and the CMS-PPV-adjusted PSI composite excluding POA.

Agreement in Hospital Categorizations between the Original AHRQ PSI Composite and Modified PSI Composites

We found differences in hospital categorization depending on whether the original or VA-PPV-adjusted PSI composite was used. Overall, a total of 9 of 151 hospitals changed categorization (6 percent). Of the four hospitals identified as performing “better than average” using the original PSI composite, three (75 percent) were identified as “average” using the VA-PPV-adjusted PSI composite. Of the 131 average-performing hospitals, two (2 percent) were reclassified to “worse than the national average”; among the 16 hospitals that performed worse than the national average, four (25 percent) improved and were reclassified as average-performing hospitals. Thus, about 4 percent of average and worse than average performers changed hospital categorization when the VA-PPV-adjusted PSI composite was used (see Table 2.)

Using CMS-PPV-adjustment excluding POA, overall, a total of eight hospitals changed categorization (5 percent) (Table 2). This included three of four (75 percent) hospitals that were identified as performing “better than average” using the original PSI composite but changed to “average” using the CMS-PPV-adjusted PSI composite excluding POA. Focusing specifically on the “average” and “worse than average” performers, one hospital went from average to worse than the national average, and 4 of 16 were reclassified as average performers from worse than the national average performers, resulting in a change in categorization of 3 percent.

Changes in the Percentage of the Payment Pool under a P4P Program

Figure 2A shows the percentage of the payment pool each hospital would receive using the original PSI composite (monotonically declining line) and the VA-PPV-adjusted PSI composite (jagged line). Notably, for about the first 50 hospitals (arrayed in the graph from highest to lowest in terms of payment under baseline rates), the majority of hospitals would receive lower payments based on VA-PPV-adjustment. In a noticeable number of these cases, the payments would decrease by 10 percent. The effects on the average-performing hospitals were more symmetrical. However, there were some hospitals whose payments would increase or decrease as much as 20 percent from payments

Figure 1A: Comparison of Hospital Ranks Based on the Original AHRQ Patient Safety Composite Score§ and the VA-PPV-Adjusted PSI Composite Score**

Note. §Original numerator weights of individual PSIs built in the PSI software version 4.2. (AHRQ quality indicators); **VA-PPV-Adjusted numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV_i) / \sum (\text{original weight for } PSI_i \times PPV_i)$. This simulates the current VA scenario, in that the VA has not yet incorporated POA flags into administrative data. Hospitals are arrayed on the horizontal axis from best performer (rank 1) to worst performer using the original PSI composite score.

based on the VA-PPV-adjusted PSI composite. The large fluctuations for some of the lower performing hospitals reflect the instability of percentage changes when the percentage of payment pool based on original PSI composite is very low. The same pattern is seen in Figure 2B, which compares the percentage of payment based on the original PSI composite and the CMS-PPV-adjusted PSI composite excluding POA, although the impact was not as large.

DISCUSSION

Patient safety events, although relatively infrequent, can have serious consequences, including excess mortality, length of stay, cost, and readmission

Figure 1B: Comparison of Hospital Ranks Based on the Original AHRQ Patient Safety Composite§ and the CMS-PPV-Adjusted PSI Composite Excluding POA §§

Note. §Original numerator weights of individual PSIs built in the PSI software version 4.2. (AHRQ quality indicators); §§CMS-PPV-Adjusted excluding POA numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i) / \sum (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i)$. This simulates the current CMS scenario, because CMS has incorporated POA flags into Medicare claims data. Hospitals are arrayed on the horizontal axis from best performer (rank 1) to worst performer using the original PSI composite score.

(Zhan and Miller 2003; Rivard et al. 2008; Friedman et al. 2009). However, formidable barriers to reducing safety events remain, a major one being the lack of reliable and valid measures of patient safety (Leape and Berwick 2005). Although development of the PSIs by AHRQ represents a significant advance in standardizing patient safety measurement, similar to other measures that rely on administrative data, they are subject to coding inaccuracies. Herein, we explored how the PSI score weights and facility performance assessments might change if “true PSI rates” were available instead of the currently available flagged rates. Even though a small percentage (5–6 percent) of hospitals changed categorization when the CMS public reporting approach was used, some of the changes in ranks were substantial as 30 and 19 percent of hospitals

Table 2: Hospital Ranking Agreement Based on Original AHRQ Patient Safety Indicator (PSI) Composite, VA-PPV-Adjusted PSI Composite, and CMS-PPV-Adjusted PSI Composite Excluding Present-on-Admission (POA)

<i>Rank Based on the Original PSI Composite*</i>	<i>Better Than</i>	<i>Average-Performing</i>	<i>Worse Than</i>	<i>Total</i>
Rank based on VA-PPV-adjusted PSI composite [†]				
Better than [§]	1	3	0	4
Average-performing [¶]	0	129	2	131
Worse than ^{**}	0	4	12	16
Total	1	136	14	151
Rank based on CMS-PPV-adjusted PSI composite excluding POA [‡]				
Better than [§]	1	3	0	4
Average-performing [¶]	0	130	1	131
Worse than ^{**}	0	4	12	16
Total	1	137	13	151

*Original numerator weights of individual PSIs built in the PSI software version 4.2. (AHRQ 2003).

[†]VA-PPV-Adjusted numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV_i) / \sum (\text{original weight for } PSI_i \times PPV_i)$. This simulates the current VA scenario, in that the VA has not yet incorporated POA flags into administrative data.

[‡]CMS-PPV-Adjusted excluding POA numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i) / \sum (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i)$. This simulates the current CMS scenario, because CMS has incorporated POA flags into Medicare claims data.

[§]Better than: upper 95% confidence interval (CI) of hospital PSI composite is lower than national average composite.

[¶]Average-performing: 95% CI of hospital PSI composite overlaps with national composite.

^{**}Worse than: lower 95% CI of hospital PSI composite is higher than national average composite.

changed at least 20 ranks after adjustment for PPVs and POA flags, respectively. Further, the effects on the percentage of a payment pool that hospitals would receive based on the modified composites were, in about one-third of the hospitals, considerable. Excluding PSIs that were presumably POA similarly affected perceptions of hospital performance and, to a lesser degree, hospital payments.

These results have important implications for the use of administrative data in identifying hospitals as good or bad performers. For example, to identify best practices, one might examine processes in the four “better than average” hospitals flagged using the original PSI composite; however, three of the hospitals were not flagged as “better than average” using either of the two modified composites, and thus QI efforts based on the four hospitals might be misguided. Hospitals with the poorest safety performance should be of greatest concern; these are the ones that have the greatest opportunity to improve (i.e., those identified as “worse than average” performers). If improvement incentives were specifically focused on those hospitals identified as “below

Figure 2A: Comparison of the Percentage of Payment Pool Based on the Original AHRQ Patient Safety Composite§ and the VA-PPV-Adjusted PSI Composite**

Note. §Original numerator weights of individual PSIs built in the PSI software version 4.2. (AHRQ quality indicators); **VA-PPV-Adjusted numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV_i) / \sum (\text{original weight for } PSI_i \times PPV_i)$. This simulates the current VA scenario, in that the VA has not yet incorporated POA flags into administrative data. Hospitals are arrayed on the horizontal axis from best performer (rank 1) to worst performer using the original PSI composite score.

average” using baseline rates, 25 percent (4 of 16) of the hospitals identified by the AHRQ PSI composite would have been incorrectly targeted, and 12 percent (2 of 12) of the estimated true “below average” hospitals would have been missed.

These results also highlight the importance and necessity of using accurate measures to assess hospital safety performance, as well as the urgent need to improve the accuracy of administrative data. These data are commonly used for hospital comparisons and serve as the infrastructure for quality reporting and QI initiatives (The Physician Consortium of Performance Improvement [PCPI] 2009). However, numerous challenges to coding accuracy may affect the misclassification of hospitals that use administrative data for these purposes. First, current coding methods rely heavily on physicians’ clinical documentation, including their ability to accurately capture and document diagnostic information and outcomes. Second, providers must be able to

Figure 2B: Comparison of the Percentage of Payment Pool Based on the Original AHRQ Patient Safety Composite§ and the CMS-PPV-Adjusted PSI Composite Excluding POA§§

Note. §Original numerator weights of individual PSIs built in the PSI software version 4.2. (AHRQ quality indicators); §§CMS-PPV-Adjusted excluding POA numerator weight for $PSI_i = (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i) / \sum (\text{original weight for } PSI_i \times PPV \text{ excluding } POA_i)$. This simulates the current CMS scenario, because CMS has incorporated POA flags into Medicare claims data. Hospitals are arrayed on the horizontal axis from best performer (rank 1) to worst performer using the original PSI composite score.

identify outcomes that are complications of care or safety events, as well as differentiate these from events that are POA. Third, because of the ambiguity of clinical phenomena, diagnosis codes, even when adequately captured, may lack the specificity required in defining safety events. Fourth, coding errors may arise if coders do not query physicians on incomplete or confusing documentation, or if they lack knowledge on coding adverse events. Recent improvements in the quality of discharge data, such as the CMS mandate to collect POA codes (DHHS 2010) help to address some of these issues. In addition, the introduction of ICD-10-CM codes in October 2014 should increase coding detail and specificity, resulting in improved ability to measure patient safety using administrative data (CMS 2010). Nonetheless, our findings corroborate those of others, suggesting that, to some extent, differences in performance are likely due to variations in coding accuracy across hospitals rather than to true differences in safety performance (Glance et al. 2008a,b).

Despite the issues related to coding, some patient safety experts suggest that public reporting of performance data and feedback to providers is a

valuable and promising method to motivate hospitals to improve patient safety (Leape 2010). This belief is based, in part, on experiences such as the VA's National Surgical Quality Improvement Program, where feedback to hospitals on their risk-adjusted complication rates led to substantial decreases in mortality and morbidity (Khuri et al. 1998). Informing hospitals on the strengths and weaknesses in their safety performance may also facilitate identification and attention to underlying system failures (Leape 2010). On the other hand, if hospitals do not "trust" or "believe in" the validity of the measures being used, they will be less likely to use them for QI initiatives.

There are other issues we need to consider in interpreting our findings. First, although the PSI composite measure was developed through clinical input and rigorous empirical testing (AHRQ 2008), our results suggest that its reliability was relatively poor, raising questions and concerns about its suitability for hospital reporting (Schone, Hubbard, and Jones 2011). Second, while AHRQ recommends using numerator weights in calculating the PSI composite, our findings suggest that the current AHRQ PSI composite numerators overweight specific PSIs with lower PPVs (e.g., Pressure Ulcers) relative to others with higher PPVs (e.g., Post-operative Wound Dehiscence). It is also notable that changes in the numerator weights may lead to important disparate findings. Use of a modified composite measure, in particular one that is conceptually more consistent with the intended goal of numerator-event-based weighting, should be considered as a modification to the current AHRQ PSI composite measure. While we have no gold standard for knowing which hospitals are the "true" low-performing ones, to the extent that numerator-based weights are desired for composite scores, PPV-adjusted weights are likely to be closer to that ideal than weights based on total number of flags.

Third, we assumed that PPV rates were the same across hospitals. In our previous studies, we did not have enough cases per hospital to examine variability in rates across hospitals or the extent to which this variation may depend on specific hospital characteristics (Borzecki et al. 2011a,b,c; Cevasco et al. 2011a,b,c; Chen et al. 2011; Kaafarani et al. 2011; Rosen et al. 2012). However, because both our PPV and POA results are comparable to non-VA settings, we assume that our findings are generally robust (Utter et al. 2009, 2010; White et al. 2009; Sadeghi et al. 2010; Zrelak et al. 2011). However, this may not be the case for all settings, suggesting that it is important to replicate analysis of PPV rates outside the VA to determine the generalizability of our results. Finally, we used only 1 year

of data for analyses; thus, the number of hospitals with differences in ranks or payment penalties based on the modified composites may have been subject to random noise. Had we used multiple years of data, this may have eliminated some of the random noise, resulting in changes to our results.

Our study also has several strengths. Findings were based on a comprehensive, clinically integrated electronic medical record that helped to streamline the abstraction process and facilitate accurate information retrieval (Rosen et al. 2012). We were also able to account for PSIs that occurred prior to admission using previous results on POA as a guide for eliminating these false positives. Further, because the VA has a fixed budget with relatively few, if any, financial incentives, it was a relatively neutral setting in which to examine differences between flagged PSI rates and estimated true PSI rates on hospital rankings.

Nonetheless, even within a setting such as the VA, perceptions of hospital safety performance and hospital rank changed moderately because of the limitations of an administrative data-based measure. For some hospitals, differences in categorization or payment were substantial. Thus, we recommend that the PSIs should be used as they were originally intended—as tools for screening, case finding, and QI—rather than for public reporting or pay-for-performance. However, given that the AHRQ PSI composite measure is currently being used for public reporting, hospitals should carefully evaluate PSI-flagged cases to distinguish those cases that arise from coding issues from those that are due to underlying safety lapses. Until coding of administrative data improves, and revisions of the PSIs occur, we need to be cautious in acting on “apparent findings.” The current measures appear to be moderately sensitive to relatively arbitrary decisions about how to weight the relative contributions of individual PSIs in creating a composite measure, and to uncertainties as to which flagged events are true quality problems.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This research was supported by funding from the VA Health Services Research and Development Service grant number SDR 07-002 (P.I. Rosen).

Disclosures: None.

Disclaimers: None.

REFERENCES

- Agency for Healthcare Research and Quality [AHRQ]. 2003. "AHRQ Quality Indicators: Guide to Patient Safety Indicators." Rockville, MD. [accessed on January 31, 2013]. Available at http://www.qualityindicators.ahrq.gov/downloads/software/sas/v42/psi_sas_documentation_v42.pdf
- Agency for Healthcare Research and Quality [AHRQ]. 2008. *Quality Indicators (QI) and Patient Safety Indicators (PSI) Composite Measure Workgroup Final Report*. Rockville, MD: Agency for Healthcare Research and Quality.
- Borzecki, A. M., M. Cevalco, Q. Chen, M. Shin, K. M. Itani, and A. K. Rosen. 2011a. "How Valid Is the AHRQ Patient Safety Indicator "Postoperative Physiologic and Metabolic Derangement"?" *Journal of the American College of Surgeons* 212 (6): 968–76.
- Borzecki, A. M., H. Kaafarani, M. Cevalco, K. Hickson, S. Macdonald, M. Shin, K. M. Itani, and A. K. Rosen. 2011b. "How Valid Is the AHRQ Patient Safety Indicator "Postoperative Hemorrhage or Hematoma"?" *Journal of the American College of Surgeons* 212 (6): 946–53.
- Borzecki, A. M., H. M. Kaafarani, G. H. Utter, P. S. Romano, M. H. Shin, Q. Chen, K. M. Itani, and A. K. Rosen. 2011c. "How Valid Is the AHRQ Patient Safety Indicator "Postoperative Respiratory Failure"?" *Journal of the American College of Surgeons* 212 (6): 935–45.
- Centers for Medicare and Medicaid Services [CMS]. 2010. "ICD-10-CM/PCS: An Introduction. Official CMS Industry Resource for the ICD-10 Transition. Department of Health and Human Services (DHHS). April 2010" [accessed on January 31, 2013]. Available at <http://www.cms.gov/Medicare/Coding/ICD10/downloads/ICD-10Overview.pdf>
- Centers for Medicare and Medicaid Services [CMS]. 2013a. "CMS Hospital Compare Website" [accessed on January 31, 2013]. Available at <http://www.medicare.gov/hospitalcompare/?AspxAutoDetectCookieSupport=1>
- Centers for Medicare and Medicaid Services [CMS]. 2013b. "FY 2013 IPPS Final Rule Home Page. Department of Health and Human Services (DHHS)" [accessed on January 31, 2013]. Available at <http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY-2013-IPPS-Final-Rule-Home-Page.html>
- Cevalco, M., A. M. Borzecki, Q. Chen, P. A. Zrelak, M. Shin, P. S. Romano, K. M. Itani, and A. K. Rosen. 2011a. "Positive Predictive Value of the AHRQ Patient Safety Indicator 'Postoperative Sepsis': Implications for Practice and Policy." *Journal of the American College of Surgeons* 212 (6): 954–61.
- Cevalco, M., A. M. Borzecki, D. A. McClusky 3rd, Q. Chen, M. H. Shin, K. M. Itani, and A. K. Rosen. 2011b. "Positive Predictive Value of the AHRQ Patient Safety Indicator 'Postoperative Wound Dehiscence.'" *Journal of the American College of Surgeons* 212 (6): 962–7.
- Cevalco, M., A. M. Borzecki, W. J. O'Brien, Q. Chen, M. H. Shin, K. M. Itani, and A. K. Rosen. 2011c. "Validity of the AHRQ Patient Safety Indicator 'Central

- Venous Catheter-Related Bloodstream Infections.” *Journal of the American College of Surgeons* 212 (6): 984–90.
- Chen, Q., A. K. Rosen, M. Cevasco, M. Shin, K. M. Itani, and A. M. Borzecki. 2011. “Detecting Patient Safety Indicators: How Valid Is ‘Foreign Body Left during Procedure’ in the Veterans Health Administration?” *Journal of the American College of Surgeons* 212 (6): 977–83.
- Department of Health and Human Services [DHHS]. 2010. “Hospital-Acquired Conditions (HAC) in Acute Inpatient Prospective Payment System (IPPS) Hospitals” [accessed on January 31, 2013]. Available at <http://www.cms.gov/hospitalacqcond>
- Friedman, B., W. Encinosa, H. J. Jiang, and R. Mutter. 2009. “Do Patient Safety Events Increase Readmissions?” *Medical Care* 47 (5): 583–90.
- Glance, L. G., Y. Li, T. M. Osler, D. B. Mukamel, and A. W. Dick. 2008a. “Impact of Date Stamping on Patient Safety Measurement in Patients Undergoing CABG: Experience with the AHRQ Patient Safety Indicators.” *BMC Health Services Research* 8: 176.
- Glance, L. G., T. M. Osler, D. B. Mukamel, and A. W. Dick. 2008b. “Impact of the Present-on-Admission Indicator on Hospital Quality Measurement: Experience with the Agency for Healthcare Research and Quality (AHRQ) Inpatient Quality Indicators.” *Medical Care* 46 (2): 112–9.
- Haut, E. R., K. Noll, D. T. Efron, S. M. Berenholz, A. Haider, E. E. Cornwell 3rd, and P. J. Pronovost. 2007. “Can Increased Incidence of Deep Vein Thrombosis (DVT) Be Used as a Marker of Quality of Care in the Absence of Standardized Screening? The Potential Effect of Surveillance Bias on Reported DVT Rates after Trauma.” *Journal of Trauma* 63 (5): 1132–5.
- Healthcare Cost and Utilization Project [HCUP]. 2011. “HCUP Comorbidity Software” [accessed on January 31, 2013]. Available at <http://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp>
- Iezzoni, L. I. 1990. “Using Administrative Diagnostic Data to Assess the Quality of Hospital Care. Pitfalls and Potential of ICD-9-CM.” *International Journal of Technology Assessment in Health Care* 6 (2): 272–81.
- . 1997. “Assessing Quality Using Administrative Data.” *Annals of Internal Medicine* 127 (8 Pt 2): 666–74.
- Kaafarani, H. M., A. M. Borzecki, K. M. Itani, S. Loveland, H. J. Mull, K. Hickson, S. Macdonald, M. Shin, and A. K. Rosen. 2011. “Validity of Selected Patient Safety Indicators: Opportunities and Concerns.” *Journal of the American College of Surgeons* 212 (6): 924–34.
- Khuri, S. F., J. Daley, W. Henderson, K. Hur, J. Demakis, J. B. Aust, V. Chong, P. J. Fabri, J. O. Gibbs, F. Grover, K. Hammermeister, G. Irvin 3rd, G. McDonald, E. Passaro Jr, L. Phillips, F. Scamman, J. Spencer, and J. F. Strengle. 1998. “The Department of Veterans Affairs’ NSQIP: The First National, Validated, Outcome-Based, Risk-Adjusted, and Peer-Controlled Program for the Measurement and Enhancement of the Quality of Surgical Care. National VA Surgical Quality Improvement Program.” *Annals of Surgery* 228 (4): 491–507.

- Leape, L. 2010. "Transparency and Public Reporting Are Essential for a Safe Health Care System." *The Commonwealth Fund* 4: 1–4.
- Leape, L. L., and D. M. Berwick. 2005. "Five Years after to Err Is Human: What Have We Learned?" *Journal of the American Medical Association* 293 (19): 2384–90.
- Rivard, P., R. Elwy, S. Loveland, S. Zhao, D. Tsilimingras, A. Elixhauser, P. Romano, and A. Rosen. 2005. "Applying Patient Safety Indicators (PSIs) Across Health-care Systems: Achieving Data Comparability." In *Advances in Patient Safety: From Research to Implementation*, edited by K. Henriksen, J. B. Battles, E. Marks, and D. I. Lewin, Rockville, MD: Agency for Healthcare Research and Quality.
- Rivard, P., S. Luther, C. Christiansen, Z. Shibe, S. Loveland, A. Elixhauser, P. Romano, and A. Rosen. 2008. "Using Patient Safety Indicators to Estimate the Impact of Potential Adverse Events on Outcomes." *Medical Care Research and Review* 65 (1): 67–87.
- Rosen, A. K., P. Rivard, S. Zhao, S. Loveland, D. Tsilimingras, C. L. Christiansen, A. Elixhauser, and P. S. Romano. 2005. "Evaluating the Patient Safety Indicators: How Well Do They Perform on Veterans Health Administration Data?" *Medical Care* 43 (9): 873–84.
- Rosen, A. K., K. M. Itani, M. Cevasco, H. M. Kaafarani, A. Hanchate, M. Shin, M. Schwartz, S. Loveland, Q. Chen, and A. Borzecki. 2012. "Validating the Patient Safety Indicators in the Veterans Health Administration: Do They Accurately Identify True Safety Events?" *Medical Care* 50 (1): 74–85.
- Sadeghi, B., R. Baron, P. Zrelak, G. H. Utter, J. J. Geppert, D. J. Tancredi, and P. S. Romano. 2010. "Cases of Iatrogenic Pneumothorax Can Be Identified from ICD-9-CM Coded Data." *American Journal of Medical Quality* 25 (3): 218–24.
- Schone, E., M. Hubbard, and D. Jones. 2011. "Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures Revised." Mathematica, Inc. November 2011 [accessed on January 31, 2013]. Available at http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability.pdf
- Statistical Analysis System [SAS] Version 9.1. 2004. *Version 9.1*. Cary, NC: SAS Institute Inc.
- The Physician Consortium of Performance Improvement [PCPI]. 2009. "Work Group Charge. American Medical Association (AMA)" [accessed on January 31, 2013]. Available at <http://www.ama-assn.org/resources/doc/cqi/pcpi-work-group-charge.pdf>
- Utter, G. H., P. A. Zrelak, R. Baron, D. J. Tancredi, B. Sadeghi, J. J. Geppert, and P. S. Romano. 2009. "Positive Predictive Value of the AHRQ Accidental Puncture or Laceration Patient Safety Indicator." *Annals of Surgery* 250 (6): 1041–5.
- Utter, G. H., J. Cuny, P. Sama, M. R. Silver, P. A. Zrelak, R. Baron, S. E. Drosler, and P. S. Romano. 2010. "Detection of Postoperative Respiratory Failure: How Predictive is the Agency for Healthcare Research and Quality's Patient Safety Indicator?" *Journal of the American College of Surgeons* 211 (3): 347–54.
- VA Information Resource Center. 2002. "VIREC Research User Guide: FY2002 VHA Medical SAS Inpatient Datasets" [accessed on January 31, 2013]. Available at <http://www.virec.research.va.gov/References/RUG/RUG-Inpatient02.pdf>

- White, R. H., B. Sadeghi, D. J. Tancredi, P. Zrelak, J. Cuny, P. Sama, G. H. Utter, J. J. Geppert, and P. S. Romano. 2009. "How Valid Is the ICD-9-CM Based AHRQ Patient Safety Indicator for Postoperative Venous Thromboembolism?" *Medical Care* 47 (12): 1237–43.
- Zhan, C., and M. R. Miller. 2003. "Excess Length of Stay, Charges, and Mortality Attributable to Medical Injuries during Hospitalization." *Journal of the American Medical Association* 290 (14): 1868–74.
- Zrelak, P., B. Sadeghi, G. Utter, R. Baron, D. Tancredi, J. Geppert, and P. Romano. 2011. "Positive Predictive Value of the Agency for Healthcare Research and Quality Patient Safety Indicator for Central Line-related Bloodstream Infection ("Selected Infections Due to Medical Care")." *Journal for Healthcare Quality* 33 (2): 29–36.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Positive Predictive Values (PPVs) among All Flagged Cases by Selected Patient Safety Indicators (PSIs) in the VA with and without Present-on-Admission (POA) Data.

Appendix SA2: Examples of Specific Calculations for Hospital Rank Changes.

Appendix SA3: Author Matrix.