© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2007.00760.x
RESEARCH ARTICLE

# Rare Adverse Medical Events in VA Inpatient Care: Reliability Limits to Using Patient Safety Indicators as Performance Measures

*Alan N. West, William B. Weeks, and James P. Bagian*

**Objective.** To assess Agency for Healthcare Research and Quality's Patient Safety Indicators (PSIs) as performance measures using Veterans Administration hospitalization data.

**Data Sources Study Setting.** Nine years (1997–2005) of all Veterans Health Administration (VA) administrative hospital discharge data.

**Study Design.** Retrospective analysis using diagnoses and procedures to derive annual rates and standard errors for 13 PSIs.

**Data Collection/Extraction Methods.** For either hospitals or hospital networks (Veterans Integrated Service Networks [VISNs]), we calculated the percentages whose PSI rates were consistently high or low across years, as well as 1-year lagged correlations, for each PSI. We related our findings to the average annual number of adverse events that each PSI represents. We also assessed time trends for the entire VA, by VISN, and by hospital.

**Principal Findings.** PSI rates are more stable for VISNs than for individual hospitals, but only for those PSIs that reflect the most frequent adverse events. Only the most frequent PSIs yield significant time trends, and only for larger systems.

**Conclusions.** Because they are so rare, PSIs are not reliable performance measures to compare individual hospitals. The most frequent PSIs are more stable when applied to hospital networks, but needing large patient samples nullifies their potential value to managers seeking to improve quality locally or to patients seeking optimal care.

**Key Words.** Patient Safety, veterans, administrative data, reliability

Patient safety has become an issue of focal interest in the evaluation of health care quality (Kohn, Corrigan, and Donaldson 1999). Recently the federal Agency for Healthcare Research and Quality (AHRQ) has developed a comprehensive set of standardized patient safety measures, derived from hospital discharge data, which potentially might be used to compare performance within and between health care systems. In collaboration with the University of California-Stanford Evidence-based Practice Center, AHRQ has put forth 23 Patient Safety Indicators (PSIs) that use administratively obtained discharge data to calculate risk-adjusted rates of medical/surgical adverse events in hospitalized patients (AHRQ 2003a; Zhan and Miller 2003a). Hospitalizations involving adverse events identified by the PSIs have been associated with worse outcomes and higher costs in both the private sector (Zhan and Miller 2003b) and the Veterans Health Administration (VA; Rosen et al. 2005). AHRQ also has produced statistical software that calculates PSI rates and variances from administrative hospital discharge data, using comorbidities to risk-adjust them for cross-system comparisons. The intention is that the PSIs eventually may help health care managers improve quality and medical consumers choose higher quality services. Potentially, the PSIs may serve dual purposes, including (a) case-finding to explore the local root causes of adverse events for corrective intervention, and (b) establishing normative rates against which a system's overall performance can be compared. An implication of the second purpose, however, is that the reliabilities of the PSI rates must be demonstrated.

Studies of PSI rates in the VA (Rosen et al. 2005, 2006a, 2006b), the nation's largest health care system, have generated interest in their potential application as performance measures. Because PSI rates are derived from administrative data, they have been promoted as easily obtained, inexpensive, and objective measures of medical quality. But if they are to serve as adequate performance measures they also must be both reliable and discriminating. As some health care systems consistently perform better than others, if PSI rates reflect underlying quality, they should be sufficiently sensitive to distinguish systems from one another, and their rank orders should show some stability over time. Rosen et al. (2006) showed that hospital-level PSI rates correlated after a 3-year lag, but the proportions of variance accounted for were low, ranging from 0.06 to 0.29; a shorter lag may have revealed better reliability.

Fortunately for health care consumers, the adverse medical events that the PSIs identify are very uncommon occurrences. Unfortunately for health

Address correspondence to Alan N. West, Ph.D., VA Outcomes Group REAP, VA Medical Center, White River Junction, VT 05009. William B. Weeks, M.D., MBA, is with the VA Outcomes Group REAP, VA National Quality Scholars Fellowship Program, & Field Office of VA National Center for Patient Safety, VA Medical Center, White River Junction, VT, and the Departments of Psychiatry and Community & Family Medicine, Dartmouth Medical School, Hanover, NH. James P. Bagian, M.D., PE, is with the VA National Center for Patient Safety, Ann Arbor, MI.

care analysts, the rareness of adverse events may limit the reliability of PSI rates as performance indicators. PSIs vary widely in terms of how uncommon the underlying events are. Some, such as Decubitus Ulcer, occur each year in nearly every hospital, but others, such as Death in Low Mortality DRGs, or Foreign Body Left during Surgical Procedure, are so rare that AHRQ's current algorithms do not even provide variance estimates for them from which risk-adjusted rates and confidence intervals can be derived. As PSI rates reflect rare adverse events, how well do they reflect changes in health care quality over time or differences among providers? To how large a health care system should they be applied to achieve reliable quality measures?

To address these questions we calculated annual PSI rates from 9 years (1997–2005) of VA hospital discharge data. For each PSI, we assessed stability/change across time for the VA health care system overall, for each of the 22 regional Veterans Integrated Service Networks (VISNs) into which the VA system is organized (in 2002, two VISNs were merged into one, but we kept their data separate for these analyses), and for each VA Medical Center belonging to these VISNs. We anticipated that from year to year, PSIs would be more stable if they are calculated for larger health care systems, i.e., VISNs, than at the hospital level. We expected, as well, that those PSIs that are based on higher numbers of underlying adverse events would be more stable over time.

## METHODS

We applied AHRQ's algorithms for finding adverse events and calculating PSI rates to all VA hospital discharge data for federal fiscal years 1997–2005 inclusive. VA's administrative database for hospitalizations is its annual Patient Treatment File (PTF; maintained by VA's Austin Automation Center), which includes demographics as well as admission and discharge dates, and ICD-9-CM diagnoses and procedures, for each hospitalization and any specialty service portion thereof.

PTF data include both acute and nonacute (e.g., long-term care) admissions, and As the PSIs were developed for acute admissions only, we eliminated all nonacute admissions as well as the nonacute portions of mixed admissions, following the procedures developed by Rosen, Rivard, and their colleagues (Rivard et al. 2005; Rosen et al. 2005). The method involves identifying each bedsection (admission to a particular specialty service during the hospitalization) in a mixed admission as either acute or nonacute based on the definitions of VA's Health Economics Resource Center, and then splitting these admissions into chronologically contiguous sections of acute care only to create new hospitalization records. Any new record may have a new admission or discharge date, a new principal diagnosis derived from its first bedsection, or a new DRG defined as the highest cost-weighted DRG for the new admission. The method also entails an algorithm for identifying the principal procedure for an admission (not specified in PTF data) by searching for valid operating room procedures (based on a DRG grouping list) and selecting the chronologically first procedure in either the Procedure or Surgery file. Exploiting the richness of the PTF data, the method eliminates nonelective hospitalizations by screening out nonelective DRG codes, surgeries that were performed on the third day of the admission or later, and procedures done at times other than weekdays between 5 A.M. and 5 P.M. From the resulting dataset for each year, we then derived the input variables to submit to the PSI algorithms, following AHRQ's specifications (AHRQ 2003b); these variables include patient demographics, date of admission, DRG, diagnoses, procedures and their dates, and VISN or hospital code.

AHRQ has made available on its website (www.ahrq.gov) a set of SAS programs to be applied to prepared data to compute its PSIs; we used version 3.0, which was released in February 2006. The programs implement algorithms that find patients at risk for each PSI as well as those who actually experience the adverse event. Observed rates are calculated for all PSIs, and most PSIs are then risk-adjusted (using comorbidity weights derived from nationally normative hospitalization data) and given confidence intervals. Some PSIs do not yield risk-adjusted rates or confidence intervals because AHRQ considers them too rare for reliable estimates. Several PSIs also pertain to obstetrical DRGs, which are extremely rare among VA's predominantly male clientele. Therefore, we limited this study to the following 13 PSIs, which all yield risk-adjusted rates:

Complications of Anesthesia (1)
Decubitus Ulcer (3)
Failure to Rescue (4)
Iatrogenic Pneumothorax (6)
Selected Infections Due to Medical Care (7)
Postoperative Hip Fracture (8)
Postoperative Hemorrhage or Hematoma (9)
Postoperative Physiologic & Metabolic Derangement (10)
Postoperative Respiratory Failure (11)

Postoperative Pulmonary Embolism or Deep Vein Thrombosis (12)
Postoperative Sepsis (13)
Postoperative Wound Dehiscence (14)
Accidental Puncture or Laceration (15)

The specific diagnoses and procedures defining these PSIs have been well described elsewhere (AHRQ 2003a; Rosen et al. 2005); it suffices here to note that these 13 PSIs assess a wide range of medical/surgical procedures and adverse events.

From the risk-adjusted PSI rates AHRQ's SAS programs also calculate smoothed rates with standard errors, using multivariate signal extraction (AHRQ 2003b, pp. 38–39). Smoothing removes "noise" in the risk-adjusted rates, which may be greater for smaller health care systems or very rare adverse events, by adjusting rates toward the overall mean. For this purpose, AHRQ calculated "shrinkage" estimates from HCUP Year 2002 SID data from 35 states. For clarity, in this study we report results for smoothed PSI data only, as the smoothed rates reduce outlying values due to smaller Ns and therefore should yield better reliability, but we note that our findings were similar when we submitted the intermediate risk-adjusted rates to the same analyses.

Data analyses assessed the stability of these PSIs across the 9 years (1) for the entire VA system, (2) at the level of VISNs, and (3) at the level of individual hospitals that had any patients at risk for the given PSI. Findings were related to the annual frequencies at which the underlying adverse medical events occur. Averaging the annual smoothed PSI rates and standard errors produced by the AHRQ software across the 9 years, 1997–2005, we calculated overall rate/SE ratios, whose magnitudes reflect the power of a PSI to yield reliable differences in statistical comparisons (rather like the inverse of a coefficient of variation). PSI stability was tested with intraclass correlations of PSI rates. Additionally, for each PSI we identified those VISNs or hospitals whose rates were among the highest, middle, or lowest third in 2001, the central year in our range. Using 2001 as our reference, we then calculated the percentages of VISNs or hospitals that also were among the highest, middle, or lowest thirds in other years (a measure we call "retention" whether before or after) to determine whether indicators or health care systems of different size differ with respect to stability over time. As stability should also help general trends emerge, we considered how well PSIs of different size reveal time trends for the VA as a whole and within VISNs or hospitals.

## RESULTS

We applied the AHRQ programs to calculate, for each PSI and for each year, (a) the number of adverse events detected, (b) the number of patients at risk for them, (c) the risk-adjusted, smoothed PSI rate, and (d) the standard error for this rate. These calculations were performed at each of three levels of analysis: (1) for the VA system nationwide, (2) at the VISN level, and (3) at the hospital level. Within each year, VISN-level data were averaged across the 22 VISNs; similarly, each year's hospital-level data were averaged across those hospitals with patients at risk, which numbered from 103 to 118 hospitals, depending on the PSI. For each PSI and level of analysis, we then averaged across the 9 years, 1997–2005, so as to yield a single value for each measure. Table 1 shows the resulting annual averages (PSI rates and standard errors are expressed as cases per 10,000 patients at risk). The ratio of average annual PSI rate to average standard error is given also, as this measure reflects the potential strength of the PSI to detect differences statistically, much like the inverse of a coefficient of variation. The PSIs are listed in order of frequency of adverse events.

VA-wide there are a few thousand Failure to Rescue or Decubitus Ulcer events each year. Accidental Punctures and Postoperative Deep Vein Thrombosis each account for more than a thousand events annually, as well. The other PSIs are considerably less frequent. Failure to Rescue also is much more likely to occur among patients at risk for it, with nearly 15 percent suffering death. Many more patients are at risk for Decubitus Ulcer, but fewer than 2 percent of them experience it. Other PSIs are much lower risks. At the level of VISNs, and particularly hospitals, most PSI events are very rare. Furthermore, at either the VISN or hospital level, correlations between different PSIs are low, rarely exceeding $r = 0.40$ (principal components analyses yield several components, each accounting for little variance), which suggests that the PSIs vary independently of one another.

Table 1 also shows that the average standard errors of per hospital PSI rates are considerably greater than those for per VISN rates, which in turn are greater than those for the entire VA system. Higher standard errors reduce the likelihood that comparisons among hospitals will yield statistically reliable differences, whereas comparisons among VISNs or repeated measures on national data might. The ratio of average PSI rate to average standard error, which reflects the potential statistical strength of comparisons, drops considerably from the most to least frequent indicators. When we looked at the ratios of PSI rate to standard error for individual VISNs or hospitals in each year

Table 1: Mean Annual Adverse Medical Events, Cases at Risk, Smoothed Patient Safety Indicators (PSI) Rates, and Standard Errors for 1997–2005

| *PSIs (in Order of Frequency):* | *(4) Failure to Rescue* | *(3) Decubitus Ulcer* | *(15) Accidental Puncture or Laceration* | *(12) Post-operative Pulmonary Embolism or Deep Vein Thrombosis* | *(7) Selected Infections Due to Medical Care* | *(11) Post-operative Respiratory Failure* | *(6) Iatrogenic Pneumo-thorax* | *(9) Post-operative Hemor-rhage or Hematoma* | *(14) Post-operative Wound Dehiscence* | *(13) Post-operative Sepsis* | *(10) Post-operative Physiologic or Metabolic Derangement* | *(1) Compli-cations of Anesthesia* | *(8) Post-operative Hip Fracture* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| For the entire VA* | | | | | | | | | | | | | |
| Adverse events | 3,322.8 | 2,925.0 | 1,210.7 | 1,032.3 | 627.6 | 472.3 | 372.4 | 306.0 | 133.8 | 122.3 | 91.6 | 60.2 | 29.8 |
| *Cases at risk* | *22,090* | *20,9838* | *45,3532* | *98,100* | *336,662* | *34,844* | *427,209* | *98,321* | *19,889* | *18,834* | *46,265* | *98,881* | *71,293* |
| PSI rate | 1,498.9 | 157.0 | 39.0 | 87.4 | 14.7 | 99.0 | 10.7 | 26.4 | 33.1 | 58.2 | 17.8 | 6.0 | 5.6 |
| *Standard error* | *23.0* | *3.2* | *0.9* | *3.1* | *0.8* | *5.0* | *0.4* | *1.5* | *3.1* | *7.3* | *1.4* | *0.9* | *0.6* |
| PSI rate/SE | 65.2 | 49.1 | 43.3 | 28.2 | 18.4 | 19.8 | 26.8 | 17.6 | 10.7 | 8.0 | 12.7 | 6.7 | 9.3 |
| Per VISN† | | | | | | | | | | | | | |
| Adverse events | 151.0 | 133.0 | 55.0 | 46.9 | 28.5 | 21.5 | 16.9 | 13.9 | 6.1 | 5.6 | 4.2 | 2.7 | 1.4 |
| *Cases at risk* | *1,004* | *9,538* | *20,615* | *4,459* | *15,302* | *1,584* | *19,419* | *4,469* | *904* | *856* | *2,103* | *4,495* | *3,241* |
| PSI rate | 1,446.6 | 148.3 | 39.2 | 87.6 | 15.1 | 97.4 | 9.6 | 24.5 | 24.4 | 62.6 | 14.2 | 6.5 | 4.9 |
| *Standard error* | *105.3* | *15.9* | *4.3* | *14.3* | *3.8* | *20.3* | *1.6* | *5.4* | *7.8* | *33.9* | *5.2* | *4.0* | *2.7* |
| PSI rate/SE | 13.7 | 9.3 | 9.1 | 6.1 | 4.0 | 4.8 | 6.0 | 4.5 | 3.1 | 1.8 | 2.7 | 1.6 | 1.8 |

*Continued*

Table 1. *Continued*

| *PSIs (in Order of Frequency):* | *(4) Failure to Rescue* | *(3) Decubitus Ulcer* | *(15) Accidental Puncture or Laceration* | *(12) Post-operative Pulmonary Embolism or Deep Vein Thrombosis* | *(7) Selected Infections Due to Medical Care* | *(11) Post-operative Respiratory Failure* | *(6) Iatrogenic Pneumo-thorax* | *(9) Post-operative Hemor-rhage or Hematoma* | *(14) Post-operative Wound Dehiscence* | *(13) Post-operative Sepsis* | *(10) Post-operative Physiologic or Metabolic Derangement* | *(1) Compli-cations of Anesthesia* | *(8) Post-operative Hip Fracture* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Per hospital[‡] | | | | | | | | | | | | | |
| Adverse events | 26.2 | 23.0 | 9.4 | 8.4 | 4.9 | 4.1 | 2.9 | 2.5 | 1.2 | 1.1 | 0.8 | 0.5 | 0.2 |
| *Cases at risk* | *172* | *1,632* | *3,526* | *762* | *2,618* | *271* | *3,322* | *763* | *155* | *146* | *359* | *766* | *554* |
| PSI rate | 1,410.6 | 155.6 | 33.5 | 88.5 | 14.9 | 91.8 | 7.4 | 22.5 | 21.7 | 78.2 | 11.2 | 7.0 | 3.9 |
| *Standard error* | *192.4* | *43.3* | *9.8* | *30.3* | *9.0* | *31.1* | *2.6* | *7.4* | *8.8* | *70.9* | *6.9* | *7.3* | *4.7* |
| PSI rate/SE | 7.3 | 3.6 | 3.4 | 2.9 | 1.7 | 3.0 | 2.8 | 3.0 | 2.5 | 1.1 | 1.6 | 1.0 | 0.8 |

*Note*: Mean PSI rates and standard errors are expressed as cases per 10,000 patients at risk.

*Averaged across all 9 years, 1997–2005.

[†]Averaged across all 22 VISNs within each year, then across all 9 years.

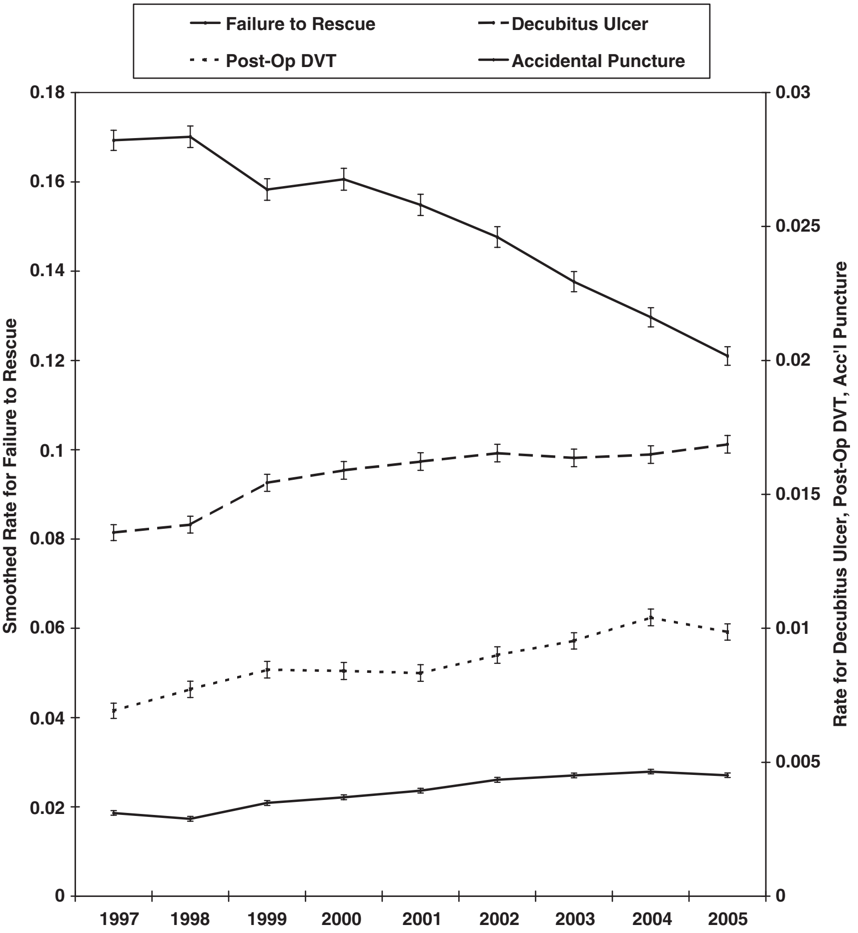
[‡]Averaged across all hospitals within each year, then across all 9 years.

(data not shown in the table), there were numerous instances, particularly for the less frequent PSIs, where the ratio was $<2$; as the ratio is comparable to a *t* statistic, in these instances the PSI rate would be statistically indistinguishable from zero. Consequently, comparisons among VISNs or time series analyses of national data using the most frequent PSIs are more likely to yield reliable differences than comparisons among hospitals on the less common PSIs.

To illustrate this point, we tested the annual PSI rates for the entire VA system nationwide for linear or higher degree trends from 1997 through 2005. Only four PSIs yielded statistically significant time trends, which we show (with standard errors) in Figure 1. All four PSIs represent strongly significant linear trends (tested via a *z* score calculated as linear slope divided by the ratio of mean annual standard error to the standard deviation of number of years). For the most frequent PSI, Failure to Rescue, the smoothed rate dropped with time, $z = -8.04$, $p < .0001$. But for the other three, there were strong increases with time: Decubitus Ulcer, $z = 3.29$, $p < .001$; Postoperative DVT, $z = 3.37$, $p < .001$; and Accidental Puncture or Laceration, $z = 7.14$, $p < .0001$. For Failure to Rescue, there was a sharp drop in numbers of adverse events (20 percent) from 1997 to 1999, and then steady annual counts in the context of increasing numbers of patients at risk (data not shown). For the latter three PSIs, annual numbers of adverse events rose linearly with time, while patients at risk remained fairly constant. The strong linearity of these trends suggests the influence of some long-standing, steady process(es) active throughout the 9 years that may relate to care provision or alternatively to different coding practices. However, these processes have not been reflected in the rates of the other, less common, PSIs. Though their greater variability made the linear trends insignificant, rates for two other PSIs, Postoperative Respiratory Failure and Postoperative Sepsis, also rose more than 20 percent during this time; for each, patients at risk rose across years, but adverse events per year rose faster. Rates for the other PSIs show no clear time trends.

When we looked for time effects within individual VISNs, only the most frequent PSIs yielded significant trends: At $p < .05$ or better, six (of 22) VISNs showed significant linear trends for Accidental Puncture, with one of these also having linear trends for Decubitus Ulcer and Selected Infections Due to Medical Care, and another joining five other VISNs in showing linear trends for Failure to Rescue; several other VISNs trended nonsignificantly in the same directions on these PSIs. On the other hand, significant time trends for individual hospitals were slightly fewer than would be expected by chance. In short, time trends emerged only for the more frequent PSIs, at either the national or VISN level, but not at the individual hospital level above chance.

Figure 1: Annual Smoothed Rates (and Standard Errors), for the Four Most Frequent Patient Safety Indicators (PSIs) in the Veterans Health Administration Health Care System, Which Are the Only PSIs to Yield Significant Time Trends at the National Level.

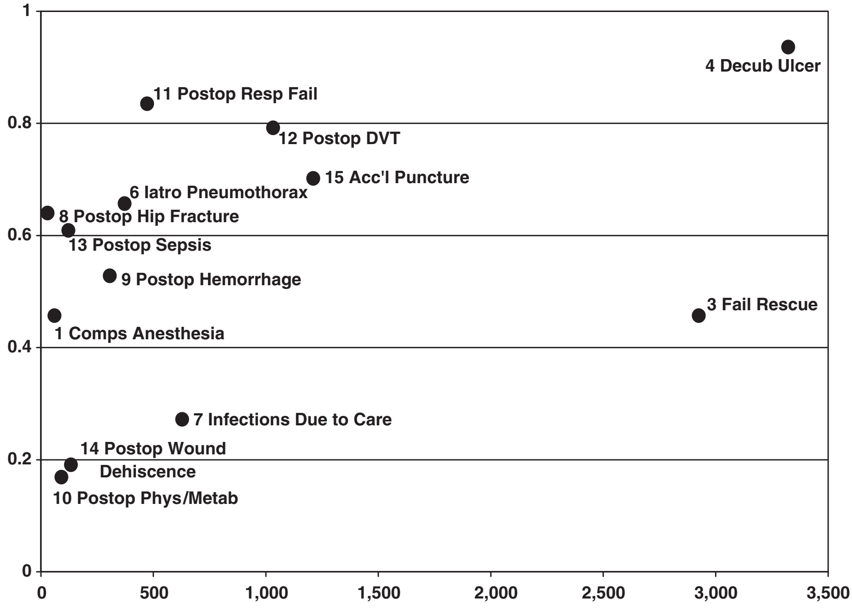


*Note*: Left *y*-axis is smoothed rate for Failure to Rescue; right *y*-axis is for Decubitus Ulcer, Post-Op DVT, and Accidental Puncture.

Variances at the hospital level, or for the less frequent PSIs at any level, are too great relative to rates to yield significant trends.

One way to assess a PSI's reliability is to consider how well one year's rate predicts rates in other years. For each PSI we calculated an intraclass

Figure 2: IntraClass Correlations of Smoothed Patient Safety Indicator (PSI) Rates, for VISNs across 9 Years, 1997–2005, Plotted against Average Annual Number (throughout the VA) of PSI Adverse Events (Rate Explains > 50 percent of Variance in Other Years' Rates for PSIs 4, 11, and 12 Only).



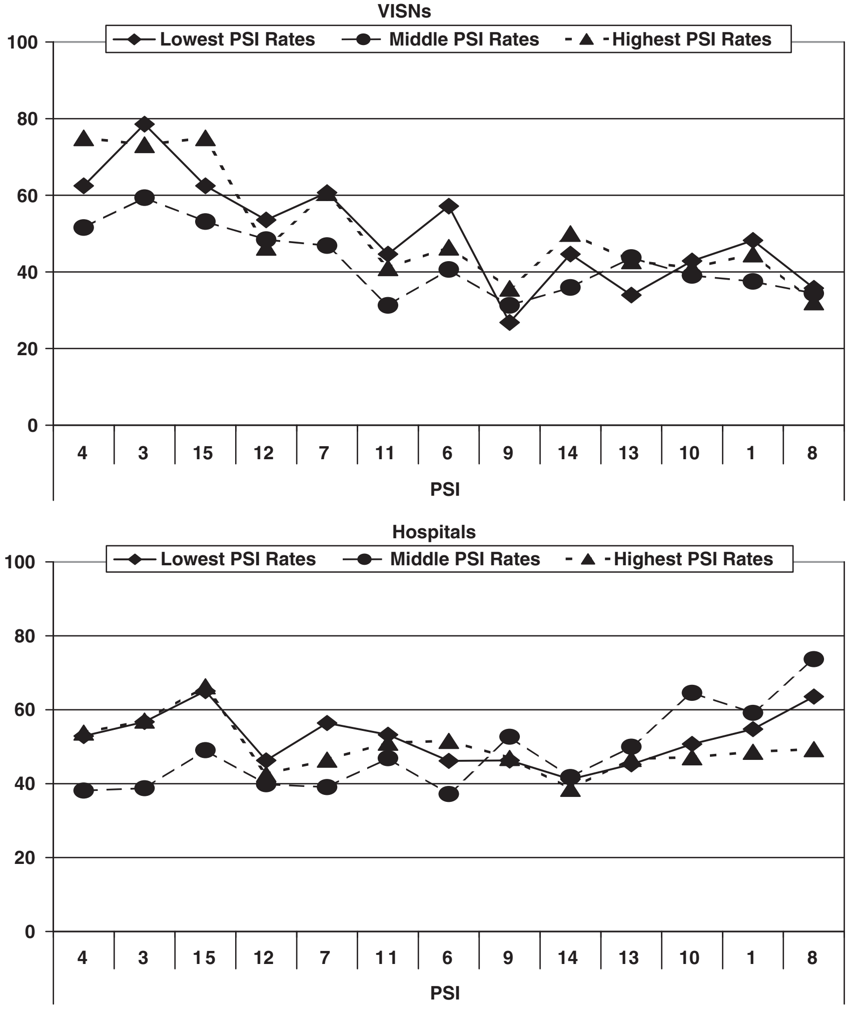
correlation from a general linear models analysis of PSI rates between 1997 and 2005; we calculated each PSI's intraclass correlation twice, once using VISNs and once using hospitals as the units of analysis. Figure 2 shows these correlations for VISN-level data plotted against average annual number of PSI events throughout the VA. Generally, reliability over time is better for those PSIs that are based on the most adverse events, although for Failure to Rescue reliability is low despite its comparatively large number of underlying events. For only three PSIs, Decubitus Ulcer, Postoperative Respiratory Failure, and Postoperative Pulmonary Embolism or Deep Vein Thrombosis, did annual rates predict more than 50 percent of the variance in other years. Hospital-level correlations (not shown) are consistently lower across all PSIs, so that the variance explained for any of the most common PSIs is only somewhat greater than 30 percent. On other PSIs, VISNs and hospitals changed their rank orders considerably from year to year.

There were, however, several VISNs, as well as hospitals, whose PSI rates were consistently in each year's lowest or highest third of scores. For each PSI, we rank-ordered the VISNs on their smoothed rates for 2001, the central year in our 9-year time span, and apportioned them to the highest, middle, or lowest thirds for that year. We then searched the PSI rates for each of the other 8 years, 1997–2000 and 2002–2005, to determine whether each VISN remained in the same third; we calculated the percentage that did so in each year, under the assumption that if this "retention" rate is higher, the PSI is more stable over time. We repeated these calculations using hospitals as the units of analysis, as well. Preliminary analyses showed no systematic effects for years, e.g., retention rates were not higher for years closer to 2001, nor did pre- and post-2001 years differ. Figure 3 shows these percentages averaged across the 8 years (VISN-level data: upper plot; hospital-level: lower).

The figure suggests that for VISN-level data retention in the same third across years was more likely for the more frequent PSIs; in other words, VISNs were more likely to maintain their relative positions on these measures than on the less frequent PSIs. To test this notion, we submitted the annual retention rates to general linear model analysis of variance, using the 8 years, 1997–2000 and 2002–2005, as units of analysis, and including PSI and 2001 ranking (highest, middle, or lowest third) as crossed factors. Retention rate varied significantly across PSIs, $F(12, 311) = 18.94$, $p < .0001$. Retention also was greater overall for VISNs with the lowest third of PSI rates (50 percent) or the highest rates (51 percent) than for VISNs in the middle third (43 percent), $F(2, 311) = 12.40$, $p < .0001$. The interaction approached significance, $F(24, 311) = 1.46$, $p < .08$, as differences among thirds occurred primarily for the more frequent PSIs. We confirmed this in a follow-up analysis in which we combined the three most frequent PSIs, Failure to Rescue, Decubitus Ulcer, and Accidental Puncture, and compared them to the ten other PSIs combined, finding a significant interaction, $F(2, 311) = 4.43$, $p < .05$. VISNs whose rates are among the highest or lowest thirds on the most common PSIs are the most likely to remain at the same levels over time, whereas stability across years on the most infrequent PSIs is much closer to chance.

The hospital-level data in Figure 3 show much different patterns, with less retention overall than for VISNs, though again average retention varied significantly across PSIs, $F(12, 311) = 23.43$, $p < .0001$. Quality thirds also differed in their average retention rates, $F(2, 311) = 9.03$, $p < .001$, but the differences were not great (lowest rates: 52 percent, middle: 49 percent, highest rates: 50 percent). For hospitals, however, there was a significant interaction, $F(24, 311) = 11.07$, $p < .0001$, due to an anomalous finding that

Figure 3: Average Percentages of VISNs, or Hospitals, That Were, among the Lowest, Middle, or Highest Third of Smoothed Patient Safety Indicator (PSI) Rates in 2001, and Also Were in the Same Third in Other Years. Percentages in the Same Third in Each of the Other 8 Years Were Averaged. PSIs Are Listed in Order of Frequency.



middle-level hospitals had higher retention rates when quality rankings were based on the rarest PSIs. Yet when we again compared the three most frequent PSIs to all others, there also was a significant interaction, $F(2, 311) = 21.08$, *p*<.0001, revealing that only the most frequent PSIs yielded higher retention rates for hospitals in the highest or lowest thirds of PSI rate rankings. In summary, rankings based on the more common PSIs are more stable, particularly for larger health care systems, and for extreme rather than middle rankings. Most PSIs are too unstable to yield reliable comparisons among health care systems, particularly at the hospital level.

## DISCUSSION

AHRQ has proposed multiple PSIs to broadly assess medical/surgical quality in inpatient settings. But as AHRQ acknowledges (AHRQ 2003a), the different PSIs cannot be given equal weight as the annual numbers of adverse medical events they represent vary by orders of magnitude. The most frequent PSI, Failure to Rescue, represents more than 3,000 deaths in VA care per year; though relatively few patients are at risk for it, among these, 15 percent die. The VA can congratulate itself that during these recent 9 years the smoothed rate for this most common PSI has dropped by more than 25 percent nationally. The next most common PSIs, Decubitus Ulcer, Accidental Puncture or Laceration, and Postoperative Pulmonary Embolism or Deep Vein Thrombosis, also represent thousands of adverse events per year, and for these there have been steady increases in reported incidence over the same time, which may indicate a need for case-finding and focused interventions to improve quality, changes in coding practices during a time period when VA markedly increased commercial insurance billings, or simply a trend of increased reporting of adverse events. Other PSIs reflect many fewer adverse events annually, and their smoothed rates are sufficiently variable from year to year that time trends do not emerge. PSIs vary independently of one another, within and across years, so that global conclusions about patient safety within VISNs or hospitals are not possible.

Because the PSIs vary so widely in the annual numbers of adverse events underlying them, we sought to assess their stability as health care quality measures in systems of different size, i.e., the VA as a whole; the health care networks, VISNs, into which the VA has been organized; and the individual hospitals. We found that intraclass correlations of PSI rates are substantially higher at the VISN level than for hospitals, but primarily for the most common PSIs; uncommon PSIs had relatively low correlations, regardless of the size of the system. Similarly, when we considered how well PSI-defined groups of VISNs or hospitals remained stable over time, we found that rankings based on the more common PSIs are more stable for the VISNs, and for extreme rather than middle rankings, and are considerably less stable at the hospital level and for the less common PSIs. Our findings suggest that a health care system (hospital or network) cannot expect a given PSI rate to be a reliable performance indicator unless its patients at risk number in the thousands and adverse events number in the hundreds. Consequently, only the largest systems might reasonably consider using PSI rates to assess quality of care. But the need for such large patient samples must necessarily limit the usefulness of PSI rates to managers seeking to improve health care quality or patients seeking optimal care.

This study is limited, in part, because AHRQ PSIs are calculated from administrative data. To the degree that important data elements are not recorded in discharge summaries, or there are variations in coding across providers within the VA, our analyses and conclusions may be inaccurate. Another limitation may stem from our assumption that efforts to reduce adverse events measured by the PSIs likely have been randomly dispersed in time among VISNs or hospitals during the time period examined. This is not to say that the VA has not made a concerted effort to improve patient safety during this time period; it has (Weeks and Bagian 2000; Bagian et al. 2001, 2002; Best et al. 2002; Heget et al. 2002; Neily et al. 2003; Mills et al. 2004, 2006; Eldridge et al. 2006). We simply have assumed that these efforts were not systematically distributed in time across VA hospitals and VISNs.

Most importantly, AHRQ PSIs have not yet been strongly validated within the VA or other health care systems. At the least, a large-scale study reviewing clinical records will be needed to assess how well the PSIs represent the documented incidence of preventable adverse events. Although prior studies (Zhan and Miller 2003b; Rosen et al. 2005) found higher mortality, lengths of stay, and costs associated with admissions involving PSI events, there is some evidence that the sensitivity/specificity of AHRQ PSIs is less than ideal. A comparison of adverse events identified from five surgical PSIs to data collected through VA's National Surgical Quality Improvement Project (Best et al. 2002) found the PSIs had sensitivities of between 37 and 67 percent (Rosen et al. 2006a). In unpublished work, we have examined confirmed cases of "foreign bodies left in" that were identified through VA's root cause analysis system (Bagian et al. 2002) and found that only 47 percent were detected by the AHRQ PSI software. Additionally, an AHRQ researcher (A. Elixhauser, personal communication) told us of recent evidence that roughly eight of every nine Decubitus Ulcer cases may have been present on admission to the hospital, suggesting a need for additional data elements to detect true adverse events.

Because they do not reflect health care system performance consistently over time, it is quite premature to use PSI rates as performance measures; for now, their use should be limited to research that will test their validity. If they are presented as reliable performance measures before validation, hospital or network managers who are confronted with poor PSI rates may be expected to respond rapidly to correct them, which may be more detrimental to than supportive of improved operations. Given the rareness of these adverse events, a much better approach is to use root cause analysis to understand and address system vulnerabilities, perhaps using adverse events identified by the PSI algorithms and validated through chart review as the triggering mechanism for focused inquiry. Consistency in rates for the most common PSIs may potentially serve to identify those hospital networks that are exemplars of superlative care, or those that may require special attention for improvement efforts. But at this time, PSI rate comparisons should not provide the basis for guiding managers in local quality improvement efforts or health care consumers in selecting high quality providers.

## ACKNOWLEDGMENTS

This research was funded by two Veterans Health Administration grants: The Research Enhancement Award Program (REAP; grant # REA 03-098) and a grant from VHA's Health Services Research & Development Service (grant # IIR 04-306-3).

*Disclosures:* The authors have no conflicts of interest relating to this manuscript.

*Disclaimers:* The views expressed in this article do not necessarily represent the views of the Department of Veterans Affairs or of the United States government.

## REFERENCES

Agency for Healthcare Research and Quality (AHRQ). 2003a. *AHRQ Quality Indicators–Guide to Patient Safety Indicators. Version 2.1, Revision 2.* Rockville, MD: October 22, 2004. AHRQ Pub.03-R203.

———. 2003b. *AHRQ Quality Indicators–Patient Safety Indicators: Software Documentation. Version 2.1–SAS, Revision 3a.* Rockville, MD: February 15, 2005. AHRQ Pub.03-R204.

Bagian, J. P., J. Gosbee, C. Z. Lee, L. Williams, S. D. McKnight, and D. M. Mannos. 2002. "The Veterans Affairs Root Cause Analysis System in Action." *Joint Commission Journal on Quality Improvement* 28 (10): 531–45.

Bagian, J. P., C. Lee, J. Gosbee, J. DeRosier, E. Stalhandske, N. Eldridge, R. Williams, and M. Burkhardt. 2001. "Developing and Deploying a Patient Safety Program in a Large Health Care Delivery System: You Can't Fix What You Don't Know About." *Joint Commission Journal on Quality Improvement* 27 (10): 522–32.

Best, W. R., S. F. Khuri, M. Phelan, K. Hur, W. G. Henderson, J. G. Demakis, and J. Daley. 2002. "Identifying Patient Preoperative Risk Factors and Postoperative Adverse Events in Administrative Databases: Results from the Department of Veterans Affairs National Surgical Quality Improvement Program." *Journal of the American College of Surgeons* 194: 257–66.

Eldridge, N. E., S. S. Woods, R. S. Bonello, K. Clutter, L. Ellingson, M. A. Harris, B. K. Livingston, J. P. Bagian, L. H. Danko, E. J. Dunn, R. L. Parlier, C. Pederson, K. J. Reichling, G. A. Roselle, and S. M. Wright. 2006. "Using the Six Sigma Process to Implement the Centers for Disease Control and Prevention Guideline for Hand Hygiene in 4 Intensive Care Units." *Journal of General Internal Medicine* 21 (Suppl. 2): S35–42.

Heget, J. R., J. P. Bagian, C. Z. Lee, and J. W. Gosbee. 2002. "John M. Eisenberg Patient Safety Awards. System Innovation: Veterans Health Administration National Center for Patient Safety." *Joint Commission Journal on Quality Improvement* 289 (12): 660–5.

Kohn, L., J. Corrigan, M. Donaldson. 1999. *To Err Is Human: Building a Safer Health System*, edited by the Institute of Medicine. Washington, DC: National Academy Press.

Mills, P. D., J. M. DeRosier, J. Neily, S. D. McKnight, W. B. Weeks, and J. P. Bagian. 2004. "A Cognitive Aid for Cardiac Arrest: You Can't Use It If You Don't Know about It." *Joint Commission Journal on Quality and Safety* 30 (9): 488–96.

Mills, P. D., J. Neily, E. Mims, M. E. Burkhardt, and J. P. Bagian. 2006. "Improving the Bar-Coded Medication Administration System at the Department of Veterans Affairs." *American Journal of Health-System Pharmacy* 63 (15): 1442–7.

Neily, J., G. Ogrinc, P. Mills, R. Williams, E. Stalhandske, J. Bagian, and W. B. Weeks. 2003. "Using Aggregate Root Cause Analysis to Improve Patient Safety." *Joint Commission Journal on Quality and Safety* 29 (8): 434–9.

Rivard, P. E., A. R. Elwy, S. Loveland, S. Zhao, D. Tsilimingras, A. Elixhauser, P. S. Romano, and A. K. Rosen. 2005. "Applying Patient Safety Indicators (PSIs) across Healthcare 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, pp. 7–25. Rockville, MD: Agency for Healthcare Research and Quality.

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., P. Rivard, S. Zhao, D. Tsilimingas, S. Loveland, C. Christiansen, W. Henderson, S. F. Khuri, A. Elixhauser, and P. Romano. 2006a. "Identification of Patient Safety Events from VA Administrative Data: Is It Valid?" Twenty-Fourth Annual HSR&D Service Meeting, Washington, DC.

Rosen, A. K., S. Zhao, P. Rivard, S. Loveland, M. E. Montez-Rath, A. Elixhauser, and P. S. Romano. 2006b. "Tracking Rates of Patient Safety Indicators over Time: Lessons from the Veterans Administration." *Medical Care* 44 (9): 850–61.

Weeks, W. B., and J. P. Bagian. 2000. "Developing a Culture of Safety in the Veterans Health Administration." *Effective Clinical Practice* 3: 270–6.

Zhan, C., and M. Miller. 2003a. "Administrative Data Based Patient Safety Research: A Critical Review." *Quality and Safety in Health Care* 12 (Suppl. 2): 58ii–63ii.

———. 2003b. "Excess Length of Stay, Charges, and Mortality Attributable to Medical Injuries during Hospitalization." *Journal of the American Medical Association* 290 (14): 1868–74.