# The Pitfalls of Benchmarking ICUs*

**Dylan W. de Lange, MD, PhD**
Department of Intensive Care and Emergency Medicine
University Medical Center Utrecht
University of Utrecht
Utrecht, The Netherlands

Almost 50 years ago Donabedian (1) suggested to evaluate health-related quality of care on the basis of three different components: structure, process, and outcome. Structure indicators are related to rather fixed resources (e.g., number of rooms and number of ventilators). Process indicators refer to the activities related to treatment and care (e.g., time-to-first-antibiotic and prevention bundles). Outcome is defined as changes in the state of health of a patient that can be attributed to an intervention or to the absence of an intervention (e.g., hospital mortality and health-related quality of life) (2).

Undeniably, the most important outcome measure is survival, but it is difficult to set the standard. A 100% survival, although ultimate, is probably unattainable in critical care (3). Often, survival is described as crude mortality figures, but these cannot be compared between different ICUs without appropriate correction for case-mix and severity of illness. Some ICUs treat patients who are more severely ill than other ICUs. As a consequence, they will encounter higher crude mortality ratios than the ICUs caring for less ill patients.

A popular way to deal with these imbalances between ICUs is to adjust for severity of illness using a prognostic model. Such models designate predicted mortalities to groups of patients. If the observed mortality is lower than the predicted mortality, then the ICU is performing better than the model suggests. Such a ratio between observed and predicted mortality is called the "standardized mortality ratio" (SMR). A genuine gold standard is lacking, and usually, the SMR of the reference population is used instead. Already six countries have made the SMR an obligatory quality indicator (3).

In this issue of *Critical Care Medicine*, Kramer et al (4) use two popular prognostic models to adjust for severity of illness: the Acute Physiology and Chronic Health Evaluation (APACHE) IV score and the National Quality Forum (NQF) prognostic model (5).

*See also p. 261.

**Key Words:** benchmark; intensive care; prediction model; standardized mortality ratio

Dr. de Lange served as a board member for the National Intensive Care Evaluation Foundation (which benchmarks the Dutch ICUs).



**DOI: 10.1097/CCM.0000000000000732**

When a model "ages," its accuracy wanes. The last update of both models was based on patient cohorts from before 2009 (4, 5). New treatment modalities that have been introduced after updating these models are, obviously, not been incorporated in these prognostic models but can influence overall survival. Prognostic models should be periodically retested, customized, and updated (6). For that reason, Kramer et al (4, 5) have customized (or recalibrated) both models to repair a model's tendency to under- or overpredict death in their particular ICU population.

When a model is deemed "best fitted for a particular ICU population," it can be used to adjust severity of illness and case-mix. The next step is to compare the observed mortality of that particular ICU to the predicted mortality according to the prognostic model (the benchmark). ICUs with an SMR more than 1.0 are considered "poor performers." However, the current trend of openly publishing performance-based data and rank-order listings should be carefully appraised. Such publications might have profound implications for hospital organizations. Therefore, the reliability of procedures for performance assessment and performance comparison is extremely relevant (7). Kramer et al (4, 5) show that the SMR (and therefore the position in a ranking list) is very much dependent on the prognostic model that is used (Fig. 2 in [4]). Most of the ICUs in this study have APACHE IV SMRs that are, at least in CIs, overlapping with the NQF SMRs. These SMRs are, although numerically not identical, statistically not differing from each other. However, 11 of 47 ICUs (23%) have SMRs without overlapping CIs, suggesting a statistically significant difference in SMR between the two models. Even after recalibration, the diversity in SMRs remained (eFig. 7 in [4]): still eight of the 47 ICUs did have nonoverlapping SMR CIs. Table 3 in (4) shows that only 21 of 47 ICUs (44%) are actually in agreement on the direction of the SMR (4). This means that the direction of the SMR (below, equal, or above 1.0) is the same for both models. The most extreme difference was seen in four ICUs that appear to have a SMR less than 1.0 according to the APACHE IVa model but a SMR more than 1.0 according to the NQF. Clearly, such dispersing SMRs hamper proper interpretation of outcome indicators and highlight that there is substantial uncertainty if the SMRs were to be used in a ranking list.

Even if two ICUs are performing identically for each patient type and the prediction of the risk of a poor outcome in the reference population is perfect, the ICUs are very likely to have different values for their SMR. When two ICUs have identical performances for low-risk patients and high-risk patients but different proportions of these patient groups, these ICUs might prove to have different SMRs. This counterintuitive phenomenon is called the "Simpson paradox" and is eloquently explained in a recent study in pediatric intensive cares (7). This

shows that the SMR does not point out which ICU is best for an individual patient. It merely shows what the performance (or outcome) of an ICU is, given its particular case-mix of patients in comparison to a reference population (8).

Another major drawback of comparing SMRs is that discharge policies are ignored. ICUs that transfer their patients to other facilities have a good outcome (discharged alive), even when these patients die in the next hospital. Therefore, it preferred to focus on long-term outcome (e.g., 1-year survival) (9).

If the process of benchmarking is difficult and the interpretation of SMRs is hampered by pitfalls, then the obvious question is: should we continue benchmarking? Let's go back to Donabedian (1). Why did we start benchmarking in the first place? Because we wanted to learn which ICU processes are associated with a better outcome (so-called best practices) and should be implemented in all ICUs. This is the proper way to quality improvement. However, the unintelligent translation of SMRs to rank-order listings for ICUs should only be scored as "a negative quality indicator."

## REFERENCES


1. Donabedian A: Evaluating the quality of medical care. 1966. *Milbank Q* 2005; 83:691–729
2. de Vos M, Graafmans W, Keesman E, et al: Quality measurement at intensive care units: Which indicators should we use? *J Crit Care* 2007; 22:267–274
3. Flaatten H: The present use of quality indicators in the intensive care unit. *Acta Anaesthesiol Scand* 2012; 56:1078–1083
4. Kramer AA, Higgins TL, Zimmerman JE: Comparing Observed and Predicted Mortality Among ICUs Using Different Prognostic Systems: Why Do Performance Assessments Differ? *Crit Care Med* 2015; 43:261–269
5. Kramer AA, Higgins TL, Zimmerman JE: Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: Implications for national benchmarking. *Crit Care Med* 2014;42:544–553
6. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
7. Bakhshi-Raiez F, Peek N, Bosman RJ, et al: The impact of different prognostic models and their customization on institutional comparison of intensive care units. *Crit Care Med* 2007; 35:2553–2560
8. Manktelow BN, Evans TA, Draper ES: Differences in case-mix can influence the comparison of standardised mortality ratios even with optimal risk adjustment: An analysis of data from paediatric intensive care. *BMJ Qual Saf* 2014; 23:782–788
9. Brinkman S, Abu-Hanna A, de Jonge E, et al: Prediction of long-term mortality in ICU patients: Model validation and assessing the effect of using in-hospital versus long-term mortality on benchmarking. *Intensive Care Med* 2013; 39:1925–1931


# Nonbeneficial Care: We Have Got to Do Something?*


**Amaya D. George, DO**
**Christopher J. Colombo, MD**
Critical Care Section
Department of Medicine
Dwight David Eisenhower Army Medical Center
Fort Gordon, GA


Concerns regarding the effectiveness of communication and the provision of end-of-life care are apparent in the medical literature, beginning largely with the publication of the Support Trial in 1995 which demonstrated a chasm between patient autonomy and physician awareness and implementation of the same. At the time, the commonsense intervention of a communication specialist had no meaningful impact (1). The subsequent 2 decades of research, public policy, and physician education have been marked by multiple attempts to improve communication among the patient, the family, and a team of providers.

In this issue of *Critical Care Medicine*, Downar et al (2) investigate an aspect of end-of-life care they term nonbeneficial treatment (NBT). Through a survey used at multiple facilities in Canada, they sought to arrive at a healthcare provider's functional definition of NBT and explore perceptions of NBT as well as causes, impacts, and mitigation strategies for NBT. Strengths of the study include the exceptional agreement with the two NBT definitions that include quality of life and patient self-determination as well as demonstrating providers comfort with being able to differentiate NBT from beneficial treatment with fairly high certainty. Not surprisingly, unrealistic expectations on the part of the patients and their surrogates were felt to be a driving force behind continuing NBT, and there were several variables that demonstrate different perceptions on the part of physicians and nurses; themes that were present 20 years ago (1) and still evident in recent literature (3–5). Lastly, the proposed solutions are to have better communication, training, and strategies and increased use of advanced care planning. Lack of advanced care planning as a perception by providers is congruent with current literature, showing only

***See also p. 270.**

**Key Words:** communication; ethics; healthcare cost; medical futility; rationing

Dr. George disclosed government work. Dr. Colombo is employed by the United States Army, received support for travel from Georgia Regents University for the 2012 and 2013 critical care symposium (lodging reimbursed during days lectures given for critical care conference and meeting registration fee waived), and disclosed government work. His institution received grant support from AMEDD Advanced Medical Technology Initiative (military research grant for collaborative research with industry and civilian academia on medical technology).



**DOI: 10.1097/CCM.0000000000000759**

# Comparing Observed and Predicted Mortality Among ICUs Using Different Prognostic Systems: Why Do Performance Assessments Differ?*

Andrew A. Kramer, PhD[1,2]; Thomas L. Higgins, MD, MBA, MCCM[3,4]; Jack E. Zimmerman, MD, FCCM[1,5]

**Objectives:** To compare ICU performance using standardized mortality ratios generated by the Acute Physiology and Chronic Health Evaluation IVa and a National Quality Forum–endorsed methodology and examine potential reasons for model-based standardized mortality ratio differences.

**Design:** Retrospective analysis of day 1 hospital mortality predictions at the ICU level using Acute Physiology and Chronic Health Evaluation IVa and National Quality Forum models on the same patient cohort.

**Setting:** Forty-seven ICUs at 36 U.S. hospitals from January 2008 to May 2013.

**Patients:** Eighty-nine thousand three hundred fifty-three consecutive unselected ICU admissions.

**Interventions:** None.

**Measurements and Main Results:** We assessed standardized mortality ratios for each ICU using data for patients eligible for Acute Physiology and Chronic Health Evaluation IVa and National Quality Forum predictions in order to compare unit-level model performance, differences in ICU rankings, and how case-mix adjustment might explain standardized mortality ratio differences. Hospital mortality was 11.5%. Overall standardized mortality ratio was 0.89 using Acute Physiology and Chronic Health Evaluation IVa and 1.07 using National Quality Forum, the latter having a widely dispersed and multimodal standardized mortality ratio distribution. Model exclusion criteria eliminated mortality predictions for 10.6% of patients for Acute Physiology and Chronic Health Evaluation IVa and 27.9% for National Quality Forum. The two models agreed on the significance and direction of standardized mortality ratio only 45% of the time. Four ICUs had standardized mortality ratios significantly less than 1.0 using Acute Physiology and Chronic Health Evaluation IVa, but significantly greater than 1.0 using National Quality Forum. Two ICUs had standardized mortality ratios exceeding 1.75 using National Quality Forum, but nonsignificant performance using Acute Physiology and Chronic Health Evaluation IVa. Stratification by patient and institutional characteristics indicated that units caring for more severely ill patients and those with a higher percentage of patients on mechanical ventilation had the most discordant standardized mortality ratios between the two predictive models.

**Conclusions:** Acute Physiology and Chronic Health Evaluation IVa and National Quality Forum models yield different ICU performance assessments due to differences in case-mix adjustment. Given the growing role of outcomes in driving prospective payment patient referral and public reporting, performance should be assessed by models with fewer exclusions, superior accuracy, and better case-mix adjustment. (*Crit Care Med* 2015; 43:261–269)

***See also p. 473.**

[1]Cerner Corporation, Vienna, VA.

[2]Department of Biostatistics, Kansas University Medical Center, Kansas City, MO.

[3]Critical Care Division, Department of Medicine, Baystate Medical Center, Springfield, MA.

[4]Tufts University School of Medicine, Boston, MA.

[5]Department of Anesthesiology and Critical Care Medicine, George Washington University, Washington, DC.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Supported, in part, by Cerner Corporation, which owns the registered trademark for Acute Physiology and Chronic Health Evaluation (APACHE) and holds the marketing rights to APACHE and Mortality Probability Model ($MPM_0$-III).

Dr. Kramer is employed by and has stock options with Cerner Corporation (Cerner Corporation owns the marketing rights to Acute Physiology and Chronic Health Evaluation [APACHE]). Dr. Higgins served as Chair of the Project IMPACT research committee (2003–2007) and had access to Project IMPACT data used to develop $MPM_0$-III during that time. He owns Cerner stock, has received travel support to speak at the Cerner Critical Care Outcomes forum, and has previously collaborated with Drs. Zimmerman and Kramer on other MPM and APACHE articles. Dr. Higgins received support for travel from and lectured for Cerner (travel support at Cerner Critical Care Conference January 2014 and in previous years) and has stock options with Cerner (not related to the article or any work with Cerner in the past). Dr. Zimmerman consulted for, received support for travel from, lectured for, and received support for article preparation from Cerner Corporation.

For information regarding this article, E-mail: akramer@cerner.com



**DOI: 10.1097/CCM.0000000000000694**

**Key Words:** Acute Physiology and Chronic Health Evaluation; benchmarking; health quality indicators; hospital mortality; intensive care; outcome assessment

The quality of intensive care is commonly assessed using the standardized mortality ratio (SMR), that is, the ratio of observed to predicted hospital mortality (1, 2). Mortality is predicted using contemporary prognostic systems to control for severity of illness and other risk factors (3–5). SMR is the most commonly used ICU quality indicator in Western Europe and is mandated by six countries (6).

In the United States, SMRs have been used as ICU performance benchmarks in Veterans Administration ICUs (7), Project IMPACT participants (8), Acute Physiology and Chronic Health Evaluation (APACHE) users (9), and the eICU Research Institute (10). SMR-based performance measures are used voluntarily in 10–15% of U.S. ICUs (11, 12). SMRs are easily calculated and support mortality comparisons (9, 13), assessment of outcomes over time (14), and evaluation of associations between outcome and care processes (15, 16).

As a quality metric, the SMR has multiple limitations. Variations in SMR can be due to differences in case-mix adjustment (17, 18), model obsolescence (3–5), type and extent of severity measurement (19, 20), and differences between the population originally used for model development versus an external population (21). At the ICU level, SMRs require data for large numbers of patients to reduce the impact of random variation (22) and provide meaningful CIs (23). SMRs can be influenced by transfers from other acute care hospitals (3, 24) and differences in hospital discharge practices, particularly the use of postacute care facilities (25, 26). Better than average mortality for an ICU's low-risk patients can be obscured by worse than average mortality among high-risk patients because more of these patients die and contribute disproportionately to the SMR (12). Finally, it is possible for an ICU to look bad or good statistically using one model while the opposite result is found with another model (20).

Predictive models have been compared for accuracy at the patient level (18, 21, 24, 27) and among ICUs (9, 20, 27–29), but fewer studies have examined the use of two or more predictive models to compare ICU-level SMRs (20, 29). We are not aware of previous studies that have explored how the components of different prognostic models might result in disparate performance evaluation at the ICU level.

This study reports an assessment of ICU performance using SMRs generated by two contemporary predictive models in a large multi-institutional clinical database. The two models are APACHE IVa and the National Quality Forum (NQF)-endorsed ICU Outcomes Mortality Model (ICOMmort). Our specific aims are to: 1) evaluate the extent of differences in ICU performance rankings using the two models and 2) examine how model differences might explain variations in performance assessment.

## METHODS

Study data were obtained from the APACHE database (Cerner Corporation, Kansas City, MO) for admissions between January 1, 2008, and May 31, 2013, using software that supports automated and computer-based manual entry of APACHE and Mortality Probability Model ($MPM_0$-III) variables ($MPM_0$-III variables are used in the ICOMmort model). Previous studies have described the predictor variables, development, and validation of APACHE IV (3), $MPM_0$-III (4), and the ICOMmort (30), which represents a modification of $MPM_0$-III. The reliability, accuracy, and use in the field of the APACHE IV and $MPM_0$-III models are described elsewhere (3, 4, 30).

The ICOMmort model has been endorsed by the NQF for use as an internal or external benchmark for quality improvement and is available upon request at http://www.qualityforum.org: Measure #703. Data were stripped of patient identifiers in compliance with the Health Insurance Portability and Accountability Act. The Institutional Review Board at Baystate Medical Center deemed this study exempt from review under federal regulation. For simplicity and to avoid confusion, the APACHE IV models will subsequently be referred to as "APACHE" and the ICOMmort modification of $MPM_0$-III as "NQF."

### Institutional and Patient Data

Characteristics of each hospital and ICU were self-reported. Patient data were generated as a result of medical care. All ICUs collected ICU day 1 APACHE data for consecutive unselected ICU admissions. Each ICU chose whether or not to collect $MPM_0$-III data, which supports predictions using the NQF model. If an ICU chose to collect $MPM_0$-III data, it was obtained within 1 hour before or after ICU admission. The demographic, clinical, and outcome data collected are shown in **eTable 1** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106) and described in detail elsewhere (3, 4, 31).

### Patients and ICUs Excluded From Hospital Mortality Prediction

We excluded ICU readmissions to avoid reporting two outcomes for the same patient. We did not collect data for and thereby excluded patients with burns, ICU admission for less than 4 hours, patients less than 18 years old, patients with missing outcomes, and those with diagnoses or characteristics excluded by each model. NQF predictions excluded patients admitted for cardiac surgery, trauma, and to rule out acute myocardial infarction with no infarction at 24 hours (31); we did not have data to enforce the last exclusion. APACHE predictions excluded patients admitted from another ICU and patients undergoing coronary artery bypass graft surgery (3). ICUs with data for less than 300 patients (prior to enforcing model-specific exclusions) were excluded because these units were still in the start-up phase of data collection and to ensure that SMRs had a narrow CI (23).

### Hospital Mortality Prediction

Hospital mortality was predicted for each eligible patient using the methods prescribed by the most recent APACHE and NQF

versions. The APACHE IV model was developed and validated using 2001–2003 patient data (3) and updated (APACHE IVa) using 2006–2008 patient data (18). The NQF model uses $MPM_0$-III variables (4) for admissions from 2001 to 2004 (17). It was subsequently modified using 2009 patient data (NQF #0703) (31). Mortality was predicted using APACHE and NQF in their published form because our primary goal was to compare SMRs based on models in current use. As an adjunct analysis, we also derived mortality predictions using APACHE and NQF models that were recalibrated using the study database (first level customization), a procedure that optimizes calibration for a different set of patients (20, 21, 27, 32).

***Assessment of Model Accuracy at the Patient Level.*** We first compared the accuracy of APACHE and NQF mortality predictions at the patient level.

The following statistics were used: 1) the difference between observed and mean predicted hospital mortality; 2) area under the receiver operating characteristic curve (AU-ROC) (33, 34); and 3) Brier score (35), modified because the raw Brier score is affected by the incidence of mortality. The modified Brier score adjusts for mortality differences and represents the percent reduction in deviation when using a specific predictive model as opposed to assigning everyone a probability equal to the incidence rate (18). A higher percentage reduction in adjusted Brier score indicates better model accuracy. We did not use the Hosmer-Lemeshow test because it is highly sensitive to differences in sample size (36). Further details about the performance of APACHE and NQF mortality predictions at the patient level are reported elsewhere (18).

***Comparisons of Model Performance at the ICU Level.*** To assess and compare the performance of each ICU, we calculated its SMR by dividing observed hospital mortality by the aggregate mean predicted value using the APACHE and NQF models for all eligible patients. We calculated 99% rather than 95% CIs to measure dispersion as it decreases type I error and generates fewer false outliers (37, 38).

Graphics showing mean SMR and 99% CIs were used to display ICU ranking based on the APACHE and NQF models. We also constructed $3 \times 3$ tables showing the aggregate assessment of each ICU's ranking using the two models: significantly less than 1.00, not significantly different from 1.00, and significantly greater than 1.00. Agreement was measured using the $\kappa$ statistic and Bowker test of symmetry.

The latter effectively tests whether APACHE and NQF are correlated in their assessment of the significance and direction (below 1.00, not significantly different from 1.00, above 1.00) of the ICUs' rankings.

To assess the potential reasons for differences in ICU performance ranking, we examined the impact of institutional and patient case-mix characteristics on each unit's APACHE and NQF-based SMR. The characteristics included the following: 1) teaching status (Council of Teaching Hospitals [COTH] vs non-COTH teaching and nonteaching hospitals); 2) ICU type (mixed medical-surgical vs medical, surgical, cardiac, and neurological specialty units); 3) severity of illness by whether an ICU's aggregate day 1 acute physiology score (APS) was below or above the median; 4) whether an ICU's median patient age was below or above the median; and 5) whether an ICU's median percentage of patients placed on mechanical ventilation was below or above the median.

For predictions from the recalibrated models, we calculated the SMR for each ICU using APACHE and NQF, respectively. The SMRs were plotted and compared for similarity to a normal distribution, and then used to graphically display each ICU's ranking by mean SMR and 99% CI for both recalibrated models.

## RESULTS

Data were collected for 175,585 patients admitted to 61 ICUs at 40 hospitals. $MPM_0$-III data were collected for 92,168 (52.5%) of the 175,585 patients. Removing ICUs with less than 300 admissions resulted in a cohort of 89,353 patients in 47 ICUs at 36 hospitals. As shown in **Table 1**, exclusions common to both models eliminated 7,082 patients (8.0%); model-specific exclusions resulted in the elimination of 2,354 patients (2.6%) for APACHE and 17,823 patients (19.9%) for NQF. Model exclusions left 79,917 patients (89.4% of 92,168) available for APACHE predictions; model exclusions and data not collected for $MPM_0$-III variables resulted in 64,448 patients (72.1% of 92,168) available for NQF predictions.

The median number of admissions to each ICU was 1,129 (interquartile range [IQR], 526; 2,763). Model-specific exclusions resulted in a median of 964 (IQR, 497; 2,525) patients eligible for APACHE predictions and a median of 920 (IQR, 466; 2,134) for NQF predictions at the unit level. NQF-specific exclusions left two ICUs with less than 200 observations.

The characteristics of the ICUs and hospitals eligible for performance comparison are shown in **eTable 2** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106). Hospitals were diverse in regard to geographic distribution, teaching status, and hospital bed size. Among the 47 ICUs, 59.6% were medical-surgical, 27.7% medical, and 8.5% surgical. **Table 2** shows the patient characteristics, resource use, and outcomes of the 82,271 patients after exclusions common to both models were applied, but before enacting the model-specific exclusions. Except for hospital mortality (11.5%), these results are similar to previously reported APACHE cohorts (13.6%).

### Accuracy of Mortality Predictions

At the patient level, AU-ROC was 0.883 for APACHE and 0.808 for NQF ($p < 0.001$). The percentage reduction in prediction error (adjusted Brier score) was 30.9% for APACHE and 18.0% for NQF. Observed and mean predicted hospital mortality for APACHE-eligible patients were 11.2% and 12.5%, respectively (difference = –1.3%, $p < 0.01$). For NQF-eligible patients observed and mean predicted hospital mortality were 12.4% and 11.4%, respectively (difference = 1.0%, $p < 0.01$).

At the ICU level, the median SMR was 0.89 (IQR, 0.76; 1.01) for APACHE and 1.07 (IQR, 0.92; 1.34) for NQF. **Figure 1** shows the distribution of SMRs across all ICUs for both models. Median SMR for APACHE is further from 1.00 than the median SMR for NQF, but the distribution of SMRs

**TABLE 1. Patient Exclusions Among 89,353 Eligible Admissions to 47 ICUs at 36 Hospitals**

| Exclusions | Predictive Model | | | |
|---|---|---|---|---|
| | Acute Physiology and Chronic Health Evaluation | % | National Quality Forum | % |
| Total eligible admissions | 89,353 | 100.0 | 89,353 | 100.0 |
| Exclusions applicable to both models | | | | |
| Coronary artery bypass graft | 2,382 | 2.7 | 2,382 | 2.7 |
| Readmission | 4,452 | 5.0 | 4,452 | 5.0 |
| Age < 18 yr | 248 | 0.3 | 248 | 0.3 |
| Exclusions applicable to one model | | | | |
| Trauma | 0 | 0.0 | 6,001 | 6.7 |
| Cardiac surgery | 0 | 0.0 | 5,657 | 6.3 |
| ICU transfer | 2,338 | 2.6 | 0 | 0.0 |
| Missing data or not collected[a] | 16 | 0.0 | 6,165 | 6.9 |
| Total excluded | 9,436 | 10.6 | 24,905 | 27.9 |
| Total included | 79,917 | 89.4 | 64,448 | 72.1 |

[a]Data not collected occurred when a site did not want to manually enter Mortality Probability Model data.

for APACHE is more symmetric than for NQF, which is widely dispersed and multimodal. **Figure 2** shows each ICU's performance ranking based on its SMR (99% CI) using APACHE and NQF. **Table 3** shows the 47 ICUs' aggregate performance rankings.

SMRs generated by the two models agreed on significance and direction 21 times (44.7%), which was not significantly different from expected ($\kappa = 0.126$; $p = 0.22$). Bowker test of symmetry was highly significant (chi-square = 14.27; 3 *df*; $p = 0.003$), indicating APACHE and NQF did not assign significance similarly. SMR was significantly greater than 1.0 for four ICUs using APACHE and 15 ICUs using NQF. SMR was significantly less than 1.0 for 18 ICUs using APACHE IVa and seven ICUs using NQF. There were four ICUs in which the SMR was significantly less than 1.0 using APACHE IVa but significantly greater than 1.0 using NQF. Three ICUs had SMRs exceeding 1.75 using NQF; two of these ICUs' SMR were not significantly different than expected using APACHE.

### ICU Characteristics Associated With Different SMR Rankings

Performance assessments for each ICU after stratifying for patient and institutional characteristics are shown in **Table 4**. ICU SMRs stratified by hospital teaching status are shown in **eFigure 1** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106). At ICUs in hospitals that were COTH members, seven of 24 (29.2%) had APACHE and NQF 99% CIs that did not overlap. For non-COTH teaching and nonteaching hospitals, four out of 23 (17.4%) had nonoverlapping 99% CIs (difference between strata $p = 0.18$). ICU SMRs stratified by unit type are shown in **eFigure 2** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106). There was little difference between mixed medical-surgical ICUs, in which six of 28 (21.4%) units had nonoverlapping 99% CIs, and specialty units, in which five of 19 (26.3%) units had nonoverlapping 99% CIs (difference between strata $p = 0.25$).

ICU's SMRs stratified by age, severity of illness, and frequency of mechanical ventilation are shown in **eFigures 3–5** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106). For ICUs with patients having a mean age below the median (< 62.7 yr), six of 24 (25.0%) had nonoverlapping 99% CIs, similar in magnitude to five of 23 (21.7%) with a mean age above the median (difference between strata $p = 0.26$). ICUs with a mean severity of illness (APS) below the median (< 39.9), four of 24 (16.7%) had nonoverlapping CIs, as opposed to seven of 23 ICUs (30.4%) with a mean APS above the median (difference between strata $p = 0.15$). For ICUs with a percentage of patients receiving mechanical ventilation below the median (≤ 35%), there were three of 24 (12.5%) units with nonoverlapping 99% CIs. Conversely, eight of 23 ICUs (34.8%) with a percentage of mechanically ventilated patients above the median had nonoverlapping CIs (difference between strata $p = 0.06$).

The distribution of SMRs after model recalibration is shown in **eFigure 6** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106). SMRs based on APACHE were symmetrical and had a median of 0.998. The recalibrated NQF predictions produced SMRs that were highly skewed to the right and had

**TABLE 2. Patient Characteristics and Outcomes Prior to Applying Model-Specific Exclusions[a]**

| Categorical Variables | No. of Admissions | % |
|---|---|---|
| Gender = male | 43,841 | 53.3 |
| Location prior to admission | | |
| Emergency room | 36,747 | 44.7 |
| Operating room, recovery room | 17,758 | 21.6 |
| Floor | 9,568 | 11.6 |
| Other hospital | 8,273 | 10.1 |
| Step down unit | 4,096 | 5.0 |
| Telemetry | 2,400 | 2.9 |
| Other ICU | 2,338 | 2.8 |
| Direct admission | 882 | 1.1 |
| Other, unknown | 194 | 0.2 |
| Patient type | | |
| Medical | 64,358 | 78.2 |
| Elective surgery | 13,309 | 16.2 |
| Emergency surgery | 4,604 | 5.6 |
| Received active therapy on day 1 (other than mechanical ventilation) | 48,828 | 59.4 |
| Ventilated at any time during day 1 | 28,022 | 34.1 |
| Sedated, unable to assess Glasgow Coma Scale on day 1 | 8,361 | 10.2 |
| One or more chronic health conditions | 10,924 | 13.3 |
| ICU mortality | 6,468 | 7.9 |
| Hospital mortality | 9,502 | 11.5 |
| **Continuous Variables** | **Median** | **IQR** |
| Age (yr) | 64 | 51, 76 |
| Acute physiology score on day 1 | 35 | 23, 51 |
| Hospital stay prior to ICU admission (d) | 0.36 | 0.11, 0.70 |
| ICU length of stay (d) | 1.93 | 1.04, 3.79 |
| Hospital length of stay (d) | 6.09 | 3.17, 11.11 |
| Length of mechanical ventilation (d) | 2 | 1, 5 |
| **Five Most Frequent Medical Diagnoses** | **No. of Admissions** | **%** |
| Drug overdose | 3,972 | 4.8 |
| Pulmonary sepsis | 3,755 | 4.6 |
| Gastrointestinal bleeding, upper | 2,502 | 3.0 |
| Sepsis arising in urinary tract | 2,450 | 3.0 |
| Chronic obstructive pulmonary disease | 2,362 | 2.9 |
| **Five Most Frequent Postoperative Diagnoses** | **No. of Admissions** | **%** |
| Valvular heart surgery | 1,392 | 1.6 |
| Surgery for multiple trauma (excluding the head) | 1,055 | 1.3 |
| Surgery for gastrointestinal malignancy | 1,021 | 1.2 |
| Surgery for gastrointestinal perforation | 742 | 0.9 |
| Aortic aneurysm, elective repair | 736 | 0.9 |

IQR = interquartile range.
[a]Total number of eligible admissions = 82,271.

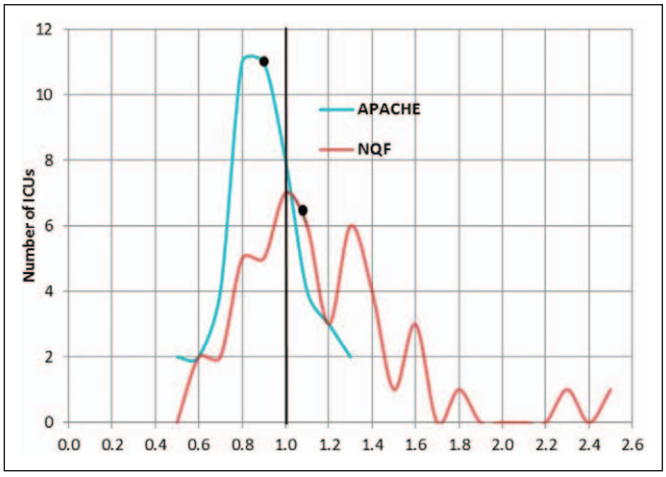


**Figure 1.** Smoothed distribution of standardized mortality ratios for 47 ICUs based on Acute Physiology and Chronic Health Evaluation (APACHE) (*blue*) and National Quality Forum (NQF) (*red*) models.

a median of 0.956. **eFigure 7** (Supplemental Digital Content 1, http://links.lww.com/CCM/B106) demonstrates that recalibration of both models did not substantially change the disparity in ICU performance rankings compared to those shown in Figure 2: there were now eight ICUs with nonoverlapping CIs as opposed to 11 such ICUs before recalibration.

## DISCUSSION

Clinical data from 47 ICUs with a large number of admissions revealed wide discrepancies in APACHE and NQF-based performance assessments. Only 21 (44.7%) ICUs had concordant performance assessment, and at four ICUs, the SMR based on APACHE suggested superior ICU performance, but inferior performance using NQF. Of particular concern were two ICUs with a NQF predicted mortality rate that was less than 50% of what was observed, effectively stigmatizing them, but with expected performance using APACHE. These results clearly demonstrate that ICU benchmarking is heavily impacted by the model used to predict mortality.

The question then becomes: Which model should be used? Traditionally, this has been addressed by comparing model accuracy at the patient level. We recently reported a detailed patient level comparison of APACHE IVa, $MPM_0$-III, and NQF mortality predictions (18). Our patient-level results were similar in the current study: APACHE underpredicted hospital mortality by 1.3% while NQF overpredicted mortality by 1.0%, and APACHE had superior discrimination (AU-ROC = 0.883) compared to NQF (AU-ROC = 0.808). Further, accuracy as reflected by the adjusted Brier score was superior for APACHE (30.9%) compared to NQF (18.0). We believe the issue of data collection burden has been minimized (17, 18) with the proliferation of electronic medical record systems. For this reason, model simplicity should not trump model accuracy.

Results at the patient level, however, can be misleading if a small number of ICUs have a disproportionally large number of patients. In this database, the five ICUs with the largest number of admissions accounted for 32.6% of all patients; their outcomes heavily impact results at the patient level. The APACHE model resulted in SMRs with a Gaussian-shaped distribution, whereas the NQF model produced SMRs that were multimodal and scattered in their distribution. For ICU benchmarking purposes, the former is better suited. Recalibrating the models using the study database did not change these results.

The main issue raised by our results is why did the two models differ so substantially in assessing unit-level performance? One explanation could be differences in how ICU admission diagnosis is included as a predictor: APACHE has 116 diagnostic

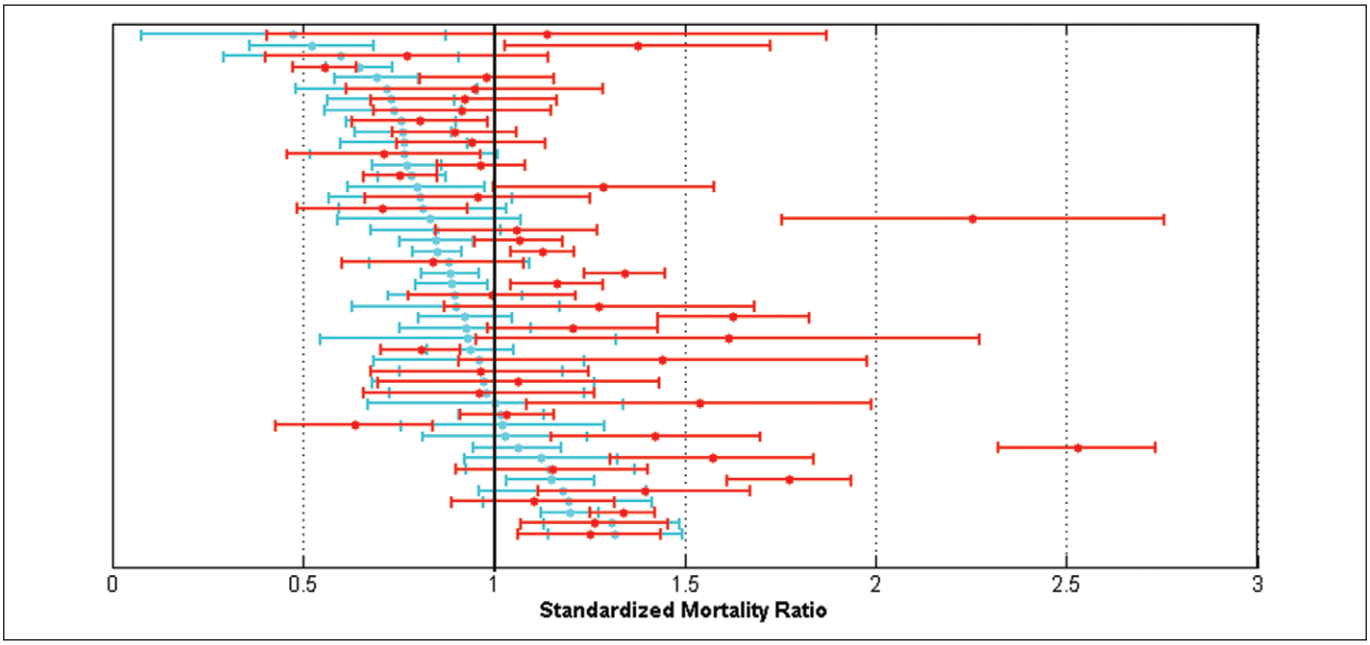


**Figure 2.** Mean standardized mortality ratio and 99% CIs for 47 ICUs based on Acute Physiology and Chronic Health Evaluation (*blue lines*) and National Quality Forum (*red lines*).

**TABLE 3. Performance of 47 ICUs Using Acute Physiology and Chronic Health Evaluation IVa and National Quality Forum Standardized Mortality Ratios Significantly Above or Below 1.0 ($p < 0.01$)**

| | National Quality Forum | | |
|---|---|---|---|
| Acute Physiology and Chronic Health Evaluation IV | SMR < 1.00 | SMR = 1.00 | SMR > 1.00 |
| SMR < 1.00 | **3** | 11 | 4 |
| SMR = 1.00 | 4 | **14** | 7 |
| SMR > 1.00 | 0 | 0 | **4** |

SMR = standardized mortality ratio.
$\kappa = 0.126$ ($p = 0.22$).
Bowker test of symmetry, asymptotically a chi-square with 3 $df = 14.27$ ($p = 0.003$).
Bold values indicate agreement between Acute Physiology and Chronic Health Evaluation IV and National Quality Forum.

groups versus three for NQF. A second explanation might be differences in accounting for physiological abnormalities, which are more important relative contributors in APACHE (65%) than $MPM_0$-III (10%) (3, 4). Although NQF enhances $MPM_0$-III by including 23 interaction terms, half of these do not include physiology.

Major teaching (COTH) hospitals generally care for patients with more complex diagnoses and greater severity of illness (39–41). Differences in model adjustment for diagnoses and physiological abnormalities may account for the significantly greater proportion of ICU SMRs with nonoverlapping 99% CIs in COTH hospitals. By contrast, there was no meaningful difference in SMRs among ICUs that treat patients with a mean age above versus below the median for all units. This might be explained by the similarity in how APACHE (use of splines) and NQF (use of splines plus 16 interaction terms) adjust for patient age.

Differences in SMRs based on whether a unit's number of patients receiving mechanical ventilation was above or below median values may also reflect more extensive adjustment for diagnosis and physiology by APACHE. This may account for the greater proportion of ICUs (34.8%) with significantly different SMRs at units with above median frequency of mechanical ventilation compared to a smaller proportion of ICUs (12.5%) with below median frequency. These results are supported by previous analyses of model accuracy across patients with an increasing risk of hospital mortality (18, 20).

The median severity of illness at each ICU, measured using the APS, also had a significant impact on SMR differences. A significantly different SMR occurred in a greater proportion

**TABLE 4. Agreement of Acute Physiology and Chronic Health Evaluation and National Quality Forum Standardized Mortality Ratios Across 47 ICUs, Stratified by Potential Confounders**

| Strata | No. (%) of ICUs With Overlapping 99% CIs for SMRs | No. (%) of ICUs With Nonoverlapping 99% CIs for SMRs |
|---|---|---|
| COTH member | 17 (70.8) | 7 (29.2) |
| Non-COTH teaching and nonteaching hospitals | 19 (82.6) | 4 (17.4) |
| Mixed medical-surgical ICU | 22 (78.6) | 6 (21.4) |
| Specialized ICUs[a] | 14 (73.7) | 5 (26.3) |
| ICU < age median | 18 (75.0) | 6 (25.0) |
| ICU > age median | 18 (78.3) | 5 (21.7) |
| ICU < median severity of illness | 20 (83.3) | 4 (16.7) |
| ICU > median severity of illness | 16 (69.6) | 7 (30.4) |
| ICU < median % of mechanically ventilated patients | 21 (87.5) | 3 (12.5) |
| ICU > median % of mechanically ventilated patients | 15 (65.2) | 8 (34.8) |

SMR = standardized mortality ratio, COTH = Council of Teaching Hospitals.
[a]Specialized ICUs included 13 medical, four surgical, one coronary, and one neurological ICU.

of ICUs (30.4%) with above median severity compared to the smaller proportion of ICUs (16.7%) with below median severity. Similar to other investigators (20, 42), we attribute these differences in SMR to the use of 17 physiological abnormalities and the use of splines to avoid linearity assumptions in APACHE compared to the use of three physiological abnormalities plus 10 interaction terms in NQF.

Another reason for discordant assessment of ICU performance is differing model exclusion criteria. In our study, patient exclusion criteria resulted in elimination of 10.6% of eligible admissions using APACHE and 27.9% using NQF. A study by Wunsch et al (43) showed that model exclusion criteria altered crude hospital mortality for individual ICUs by as much as 15%. To assess the quality of ICU care, a prognostic model should not only accurately predict mortality (20, 29, 44, 45) but also exclude as few patients as possible (43, 45). Although highly accurate models are available to benchmark outcomes for cardiac surgery (46) and trauma (47) patients, they require that hospitals use additional models.

Our analysis is subject to several limitations. First, our results cannot be interpreted as representative of all ICUs in the United States or other countries because our data collection was confined to U.S. units that elected to measure their performance using $MPM_0$-III and APACHE IV. ICU performance assessment may be affected by worldwide differences in patient characteristics (patient type, severity, and frequency of ventilation), as well as variations in patient selection based on structural, financial, and cultural considerations, and admission/discharge criteria. Second, in addition, studying only 47 ICUs limited our ability to assess reasons for differences in ICU SMRs using the two models. Third, we were unable to compare the impact of interhospital transfer on ICU performance rankings. This is because NQF excludes these patients (31) but APACHE adjusts for their adverse prognostic impact (3). Fourth, SMR-based rankings do not account for the impact of differing discharge destinations among the study ICUs (17, 25). SMRs using both models, however, were similarly affected by this limitation. Finally, as prognostic scoring systems age, they tend to overpredict mortality (3–5). Our recent patient-level analysis of current models, however, showed similar deterioration in the accuracy of APACHE and NQF between 2008 and 2012 (18). However, after recalibrating both models, substantial differences in ICU assessment still remained.

## CONCLUSIONS

Mortality benchmarking should use the most accurate model at both the patient and ICU levels. Model-based variations in ICU performance assessment are due to differences in the variables that are included in the models, their weighting, and how many patients are excluded from mortality predictions.

## REFERENCES


1. Sirio CA, Shepardson LB, Rotondi AJ, et al: Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure: Implications for critical care delivery from Cleveland Health Quality Choice. *Chest* 1999; 115:793–801
2. Grover FL, Shroyer AL, Hammermeister K, et al: A decade's experience with quality improvement in cardiac surgery using the Veterans Affairs and Society of Thoracic Surgeons national databases. *Ann Surg* 2001; 234:464–472
3. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
4. Higgins TL, Teres D, Copes WS, et al: Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med* 2007; 35:827–835
5. Moreno RP, Metnitz PG, Almeida E, et al; SAPS 3 Investigators: SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; 31:1345–1355
6. Flaatten H: The present use of quality indicators in the intensive care unit. *Acta Anaesthesiol Scand* 2012; 56:1078–1083
7. Render ML, Freyberg RW, Hasselbeck R, et al: Infrastructure for quality transformation: Measurement and reporting in veterans administration intensive care units. *BMJ Qual Saf* 2011; 20:498–507
8. Higgins TL: Quantifying risk and benchmarking performance in the adult intensive care unit. *J Intensive Care Med* 2007; 22:141–156
9. Zimmerman JE, Alzola C, Von Rueden KT: The use of benchmarking to identify top performing critical care units: A preliminary assessment of their policies and practices. *J Crit Care* 2003; 18:76–86
10. Lilly CM, Zuckerman IH, Badawi O, et al: Benchmark data from more than 240,000 adults that reflect the current practice of critical care in the United States. *Chest* 2011; 140:1232–1242
11. Keegan MT, Gajic O, Afessa B: Severity of illness scoring systems in the intensive care unit. *Crit Care Med* 2011; 39:163–169
12. Breslow MJ, Badawi O: Severity scoring in the critically ill: Part 2: Maximizing value from outcome prediction scoring systems. *Chest* 2012; 141:518–527
13. Render ML, Deddens J, Freyberg R, et al: Veterans Affairs intensive care unit risk adjustment model: Validation, updating, recalibration. *Crit Care Med* 2008; 36:1031–1042
14. Afessa B, Keegan MT, Hubmayr RD, et al: Evaluating the performance of an institution using an intensive care unit benchmark. *Mayo Clin Proc* 2005; 80:174–180
15. Afessa B, Gajic O, Morales IJ, et al: Association between ICU admission during morning rounds and mortality. *Chest* 2009; 136:1489–1495
16. Kramer AA, Higgins TL, Zimmerman JE: The association between ICU readmission rate and patient outcomes. *Crit Care Med* 2013; 41:24–33
17. Kuzniewicz MW, Vasilevskis EE, Lane R, et al: Variation in ICU risk-adjusted mortality: Impact of methods of assessment and potential confounders. *Chest* 2008; 133:1319–1327
18. Kramer AA, Higgins TL, Zimmerman JE: Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: Implications for national benchmarking. *Crit Care Med* 2014; 42:544–553
19. Iezzoni LI: The risks of risk adjustment. *JAMA* 1997; 278:1600–1607
20. Brinkman S, Abu-Hanna A, van der Veen A, et al: A comparison of the performance of a model based on administrative data and a model based on clinical data: Effect of severity of illness on standardized mortality ratios of intensive care units. *Crit Care Med* 2012; 40:373–378
21. Harrison DA, Brady AR, Parry GJ, et al: Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med* 2006; 34:1378–1388
22. Park RE, Brook RH, Kosecoff J, et al: Explaining variations in hospital death rates. Randomness, severity of illness, quality of care. *JAMA* 1990; 264:484–490
23. Peek N, Arts DG, Bosman RJ, et al: External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol* 2007; 60:491–501
24. Rosenberg AL, Hofer TP, Strachan C, et al: Accepting critically ill transfer patients: Adverse effect on a referral center's outcome and benchmark measures. *Ann Intern Med* 2003; 138:882–890

25. Kahn JM, Kramer AA, Rubenfeld GD: Transferring critically ill patients out of hospital improves the standardized mortality ratio: A simulation study. *Chest* 2007; 131:68–75
26. Vasilevskis EE, Kuzniewicz MW, Dean ML, et al: Relationship between discharge practices and intensive care unit in-hospital mortality performance: Evidence of a discharge bias. *Med Care* 2009; 47:803–812
27. Bakhshi-Raiez F, Peek N, Bosman RJ, et al: The impact of different prognostic models and their customization on institutional comparison of intensive care units. *Crit Care Med* 2007; 35:2553–2560
28. Rothen HU, Stricker K, Einfalt J, et al: Variability in outcome and resource use in intensive care units. *Intensive Care Med* 2007; 33:1329–1336
29. Glance LG, Osler TM, Dick A: Rating the quality of intensive care units: Is it a function of the intensive care unit scoring system? *Crit Care Med* 2002; 30:1976–1982
30. Lombardozi K, Bible S, Eckman J, et al: Evaluation of efficiency and accuracy of a streamlined data entry process into an outcomes database. *Abstr Crit Care Med* 2009; 37:758
31. Philip R: Lee Institute for Health Policy Studies: ICU Outcomes (Mortality and Length of Stay) Methods, Data Collection Tool, and Data. Available at: http://healthpolicy.ucsf.edu/content/icu-outcomes. Accessed March 12, 2014
32. Moreno R, Apolone G: Impact of different customization strategies in the performance of a general severity score. *Crit Care Med* 1997; 25:2001–2008
33. Hanley M, McNiel B: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29–36
34. Delong ER, Delong D, Clarke-Peterson DL: Comparing the area under two or more correlated receiver operating curves: A nonparametric approach. *Biometrics* 1988; 4:387–845
35. Brier G: Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950; 75:1–3
36. Kramer AA, Zimmerman JE: Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med* 2007; 35:2052–2056
37. Rapoport J, Teres D, Lemeshow S, et al: A method for assessing the clinical performance and cost-effectiveness of intensive care units: A multicenter inception cohort study. *Crit Care Med* 1994; 22:1385–1391
38. Wolfe RA: The standardized mortality ratio revisited: Improvements, innovations, and limitations. *Am J Kidney Dis* 1994; 24:290–297
39. Zimmerman JE, Shortell SM, Knaus WA, et al: Value and cost of teaching hospitals: A prospective, multicenter, inception cohort study. *Crit Care Med* 1993; 21:1432–1442
40. Block BM, Sirio CA, Cooper GS, et al: Use of intensive care-specific interventions in major teaching and other hospitals: A regional comparison. *Crit Care Med* 2000; 28:1204–1207
41. Lott JP, Iwashyna TJ, Christie JD, et al: Critical illness outcomes in specialty versus general intensive care units. *Am J Respir Crit Care Med* 2009; 179:676–683
42. Keegan MT, Gajic O, Afessa B: Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. *Chest* 2012; 142:851–858
43. Wunsch H, Brady AR, Rowan K: Impact of exclusion criteria on case mix, outcome, and length of stay for the severity of disease scoring methods in common use in critical care. *J Crit Care* 2004; 19:67–74
44. Breslow MJ, Badawi O: Severity scoring in the critically ill: Part 1–Interpretation and accuracy of outcome prediction scoring systems. *Chest* 2012; 141:245–252
45. Harrison DA, Parry GJ, Carpenter JR, et al: A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med* 2007; 35:1091–1098
46. Shahian DM, Edwards FH, Ferraris VA, et al; Society of Thoracic Surgeons Quality Measurement Task Force: Quality measurement in adult cardiac surgery: Part 1–Conceptual framework and measure selection. *Ann Thorac Surg* 2007; 83(4 Suppl):S3–S12
47. Nathens AB, Jurkovich GJ, Maier RV, et al: Relationship between trauma center volume and outcomes. *JAMA* 2001; 285:1164–1171