CHEST Postgraduate Education Corner

CONTEMPORARY REVIEWS IN CRITICAL CARE MEDICINE

Severity Scoring in the Critically Ill

Part 1—Interpretation and Accuracy of Outcome Prediction Scoring Systems

Michael J. Breslow, MD; and Omar Badawi, PharmD

This review examines the use of scoring systems to assess ICU performance. APACHE (Acute Physiology and Chronic Health Evaluation), MPM (mortality probability model), and SAPS (simplified acute physiology score) are the three major ICU scoring systems in use today. Central to all three is the use of physiologic data for severity adjustment. Differences in the size, nature, and time horizon of the data set translate into minor differences in accuracy and difficulty of data abstraction. APACHE IV provides ICU and hospital predictions for mortality and length of stay, whereas MPM and SAPS only provide hospital mortality predictions (although new algorithms generated from MPM data elements may predict ICU length of stay adequately). The primary use of scoring systems is for assessing ICU performance, with the ratio of actual-to-predicted outcomes in the study cohort providing performance comparisons to the reference ICUs. The reliability of scoring system predictions depends on the completeness and accuracy of the abstracted data; accordingly, ICUs must implement robust data quality control processes. CIs of the ratios are inversely related to sample size, and care must be taken to avoid overinterpreting changes in outcomes. ICU structural and process issues also can affect scoring system performance measures. Despite good discrimination and calibration, scoring systems are used in only 10% to 15% of US ICUs. Without ICU performance data, there is little hope of improving quality and reducing costs. Current demands for transparency and computerization of documentation are likely to drive future use of ICU scoring systems. ***CHEST 2012; 141(1):245–252***

Abbreviations: APACHE = Acute Physiology and Chronic Health Evaluation; LOS = length of stay; MPM = mortality prediction model; ROC = receiver operating characteristic; SAPS = simplified acute physiology score; SMR = standardized mortality ratio

ICU scoring systems were introduced almost 30 years ago with the goal of using physiologic data available at ICU admission to predict individual patient outcomes. Although these predictions have little utility for managing individual patients, they provide a mechanism to assess ICU performance by comparing actual outcomes in a given population to the outcomes observed in the reference population used to develop the prediction algorithms. Two recent review articles provide useful basic information on ICU scoring systems.[1,2] The current review, which is presented in two parts, focuses on the use of ICU scoring systems for measuring ICU performance. Part 1 focuses on current usage patterns of the three major scoring systems—APACHE (Acute Physiology and Chronic Health Evaluation) (Cerner Corporation), mortality probability model (MPM), and simplified acute physiology score (SAPS)—and examines how they differ. Considerable attention is devoted to potential sources of error and strategies for optimizing the accuracy of outcome predictions. Part 2, to be published in an upcoming issue of *CHEST*, will focus on maximizing the value derived from scoring system data and considers the use of ICU scoring system data for quality benchmarking.

Affiliations: From the Department of Research and Product Marketing, Philips Healthcare (Drs Breslow and Badawi), and Department of Pharmacy Practice and Science (Dr Badawi), University of Maryland School of Pharmacy, Baltimore, MD.
Correspondence to: Michael J. Breslow, MD, Department of Research and Product Marketing, Philips Healthcare, Ste 1900, 217 E Redwood St, Baltimore, MD 21202; e-mail: michael.breslow@philips.com

DOI: 10.1378/chest.11-0330

Comparison of ICU Scoring Systems

The three most commonly used ICU scoring systems are APACHE, MPM_0 (where the 0 indicates from time of admission), and SAPS.[3-7] Although the initial versions were introduced many years ago, each has undergone multiple revisions over the ensuing years.[8-10] In addition to the regularly updated models introduced by the developers, other investigators have created alternate prognostic models using APACHE, SAPS, and MPM data elements that are customized to fit a specific patient population.[11-15] These include the Veterans Administration hospital system in the United States (APACHE), the California ICU Outcomes Study/CalHospitalCompare project (MPM_0-III predominantly and APACHE and SAPS), The Netherlands (APACHE II, MPM_0-III, and SAPS II), Great Britain (APACHE II), and others.[11-13,15-17] Although some groups have used the same variable weights that were used in the original model and simply changed the regression equations to better fit their patient population (first-level customization), others have derived new weights and created new regression equations (second-level customization). It is important to recognize that these new models are not superior to the originals because they have not included any new data elements or refinements. They simply have modified the model to better fit their own population. Whether this is the most desirable strategy will be considered later.

Despite many similarities, there are important differences among the three systems. First, APACHE IV is a proprietary tool, although Cerner Corporation has recently made hospital mortality and ICU LOS algorithms freely available (https://apachefoundations.cernerworks.com/apachefoundations/login/auth).

MPM_0-III and SAPS 3 have focused on simplifying data collection, using fewer elements and only data from the first hour in the ICU, as opposed to APACHE, which uses data from the entire first "APACHE day." APACHE and MPM were developed predominantly from patients cared for in US ICUs, whereas SAPS 3 included patients from 35 countries.[3-7] APACHE IV and MPM_0-III used large data sets for their development (131,618 and 124,885, respectively), whereas SAPS 3 used data from 22,791 patients. Despite the relatively large number of patients in the data sets used to develop these scoring systems, they still represent a small subset of nonrandomly selected ICUs that may have greater focus on quality than other nonparticipating hospitals. APACHE is also the only scoring system that has demonstrated good discrimination and calibration in predicting ICU and hospital length of stay (LOS) for US ICUs. (MPM recently has been shown to predict ICU LOS adequately for California hospitals.[13]) Given the current and future focus on reducing health-care costs, evaluating LOS performance is of considerable importance.

Despite an increased focus on health-care safety and the push for greater transparency, ICU scoring systems are not used in many US ICUs. Although some countries have converged on a single methodology and mandated use for all ICUs, there is no similar coordinated effort in the United States to require universal ICU quality reporting.[11,12] Current estimates of use in the United States are shown in Table 1.[18]

As Table 1 demonstrates, APACHE currently is used more commonly than other scoring systems in the United States. The data in Table 1 are from validation publications and large organizations that serve many hospitals (eg, Project Impact, CalHospitalCompare). The table does not include unreported use by individual ICUs. There also may be some ICUs using multiple scoring systems that are double counted in Table 1. However, even with these caveats, it seems likely that ICU scoring systems currently are used in ~10% to 15% of US ICU patients. The endorsement

Table 1—*Estimated Use of Scoring Systems in US ICUs*

ICU Scoring System	No. of ICUs	No. of Patients/Year	% of Total ICU Population[a]
APACHE IV[b]	390	300,000	6.7
APACHE IV[c]	104	65,809	1.5
MPM_0-III[d]	103	55,000	1.2
Modified MPM_0-III (California)[e]	212	84,800[f]	1.9
Veterans Administration-Adjusted APACHE[g]	62	33,000	0.7
Total	871	538,609	12.0

APACHE = Acute Physiology and Chronic Health Evaluation; MPM_0 = mortality prediction model, where 0 indicates from time of admission.
[a]Based on an estimated 4,400,000 annual US ICU admissions.[18]
[b]2010 data (approximate) for Philips Healthcare remote ICU programs using APACHE IV.
[c]2002-2003 data.[3]
[d]2004-2005 data.[7]
[e]2010 CalHospitalCompare data (R. Adams Dudley, MD, MPH, personal communication, December 2010).
[f]Minimum no. based on each hospital reporting data on a minimum of 100 patients per quarter.
[g]2001-2004 data.[16]

of ICU scoring system reporting by the California Healthcare Foundation (CalHospitalCompare) may spur other states to implement ICU reporting. The National Quality Forum also endorsed the use of MPM in late 2010 (www.qualityforum.org).

MPM was selected based on having a smaller required data set. However, manual data collection burden may become less relevant as we move into an era of electronic charting. Table 2 provides an overview of the data elements required for APACHE IV, MPM_0-III, and SAPS 3.

Evaluating Scoring Systems

Prior to discussing the performance of the different systems, several concepts bear brief review. First, it is important to understand how risk scoring systems are developed and evaluated. To generate a predictive model, most researchers divide their data set into two pools. The first is used to develop the model, generally through multivariable regression methods, whereas the second is reserved for validation after the final model is developed. Ideally, the performance in the development and validation sets is similar, and the model can be applied to populations beyond that used during development.

To assess performance, researchers generally focus on discrimination and calibration. An ICU risk scoring system that has high discrimination is able to accurately identify the patients at highest risk for mortality. It is standard practice when describing discrimination to report the area under the receiver operating characteristic (ROC) curve, which is a graphical representation of the sensitivity against the false-positive rate.[19] For ICU models predicting mortality, this represents the probability that a randomly selected patient who dies has a higher predicted risk than a randomly selected patient who survives. An ROC of 0.5 is no better than chance, whereas values >0.7, 0.8, and 0.9 are considered acceptable, excellent, and outstanding, respectively.[20]

To measure calibration, which examines how well actual outcomes match their predicted incidence, the Hosmer-Lemeshow C statistic often is used.[20] The most common way to do this is to divide the sample into deciles of risk and evaluate the actual and expected number of events in each group. The test is essentially a global χ^2 test that examines whether there are significant differences between actual and predicted outcomes across groups; an ideal model performs equally well across all risk strata. Not surprisingly, given the influence of sample size on CIs, significant differences are common when large sample sizes are examined.[21] Calibration plots may be provided in conjunction with the Hosmer-Lemeshow test results so that readers can visually inspect the variation in performance over risk strata. Calibration is only useful in the context of discrimination; predictive models that perform no better than chance will perform consistently across all risk strata and exhibit excellent calibration. Calibration also has relevance across different ICU types, admission diagnoses, and geographic regions. A scoring system that performs well in a large heterogeneous data set but poorly in specific patient populations within that data set will yield unreliable results in ICUs with large numbers of such patients. Concerns about regional calibration led to the creation of scoring system models by several European governments using only data from their own patients. This approach was largely the result of poor calibration of the APACHE II algorithms when used in Great Britain ICUs.[22] Whether the observed problem with calibration reflected regional differences in care, differences in acuity mix or the use of outdated algorithms that reflected performance of older populations of patients was

Table 2—*Data Elements Required for APACHE IV, MPM_0-III, and SAPS 3*

Scoring System	No. of Physiologic Data Elements	Additional Data	Data Timing
APACHE IV	17	Age, chronic health variables (6 variables), ICU admission diagnosis, ICU admission source, LOS prior to ICU admission, emergency surgery, thrombolytic therapy, FIO_2, mechanical ventilation[a]	First ICU day[b]
MPM_0-III	3	Age, chronic health variables (3 variables), acute diagnoses (5 variables), admission type (eg, medical-surgical) and emergency surgery, CPR within 1 h of ICU admission, mechanical ventilation, code status	Prior to and within 1 h of ICU admission
SAPS 3	10	Age, chronic health variables (6 variables), ICU admission diagnosis, ICU admission source, LOS prior to ICU admission, emergency surgery, infection on admission, type of surgery (4 variables)	Prior to and within 1 h of ICU admission

LOS = length of stay; SAPS = simplified acute physiology score. See Table 1 for expansion of other abbreviations.

[a]Additional data are required for patients admitted after coronary artery bypass graft surgery.

[b]First ICU day duration varies based on admission time (range, 16-32 h)

never determined. However, the decision to adopt regional models precludes comparison of outcomes across national boundaries and has the potential to hinder identification of superior care processes because there is no universal comparison tool.

When evaluating ICU performance, the most common approach is to calculate the ratio of number of deaths observed to the number of deaths predicted by the reference scoring system. This is known as the standardized mortality ratio (SMR). Similarly, actual-to-predicted ratios are used to assess LOS performance. Although most users focus exclusively on the reported actual-to-predicted ratios, it is important to also consider the CIs for these ratios. It is unlikely that the point estimate SMR exactly represents the true value because there is always some degree of random error; CIs provide insight into the degree of precision of the observed ratio.[23]

There is extensive literature evaluating the accuracy of the three primary ICU scoring systems and numerous studies comparing one to another.[1,2,13,14] All demonstrate good discrimination, with APACHE IV usually having a slightly higher area under the ROC curves (Table 3). Most calibration studies focus exclusively on consistency across the severity spectrum. Some studies suggest that calibration at the extremes of severity may be less accurate.[3,24] APACHE appears to calibrate reasonably well across diagnostic categories[3] perhaps because diagnosis is part of the model.

Data Rules and Data Quality

Data completeness rules and data quality can have major effects on outcome predictions and deserve major focus by all users of ICU scoring systems. Most ICU scoring systems generate predictions based on the data at hand, although APACHE will not generate predictions in the absence of certain data elements. Each system treats missing data as normal data. Afessa et al[25] evaluated the impact of missing data on APACHE predictions and found worse outcomes in patients with missing data, suggesting that undocumented abnormalities can result in inaccurate mortality predictions. Ideally, all data elements should be collected to ensure maximum accuracy.

All of the major ICU scoring systems require data that are not routinely contained in administrative databases. In the past, most ICUs using scoring systems employed dedicated personnel to manually abstract scoring system data from paper charts. Simpler data collection has been cited as a rationale for selecting MPM over APACHE, despite the slightly superior accuracy of the APACHE methodology.[13] Not only is manual data collection time consuming but it also requires skilled data abstractors who must exercise clinical judgment and follow very specific data collection rules. Pilot studies evaluating the feasibility of mandatory ICU scoring noted high error rates. These concerns become somewhat less important as hospitals adopt use of ICU clinical information systems. Although electronic data abstraction and automated algorithm execution eliminate computational problems, they do not address data completeness or data accuracy. Moreover, some data items used in scoring systems (eg, emergency surgery, CPR within 1 h of ICU admission) are not routinely documented in many ICU clinical information systems, and processes to capture these data must be developed.

Regardless of how data are abstracted, ICUs must implement oversight procedures to ensure data accuracy. For example, the APACHE methodology is very specific in defining how the admission diagnosis should be selected. The diagnosis must be documented within the first ICU day; should reflect the primary reason for ICU admission; and when multiple diagnoses are relevant, should be the diagnosis with the worst prognosis (eg, sepsis rather than hyperglycemia). When scoring system data come directly from clinical documentation, which may be done without full consideration of the importance of the diagnosis in prognostic scoring, lower acuity diagnoses may be used, which can underestimate mortality risk. For example, in the *e*ICU Program 2008 database, the "Respiratory—medical, not otherwise categorized" selection was the second most frequently entered admission diagnosis (Table 4).[26] This diagnosis carries a lower weight than almost all of the more specific respiratory system diagnoses and, thus, may have resulted in underestimation of mortality risk.[3]

Table 3—*Comparison of Common Scoring Systems at Predicting Hospital Mortality*

	No. of Patients[a]	ROC[b]	Hosmer-Lemeshow C Statistic[b]	*P* Value[c]
APACHE IV	66,270	0.88	16.9	.08
MPM_0-III	74,578	0.82	11.6	.31
SAPS 3	13,427	0.85	14.3	.16

ROC = receiver operating characteristic. See Tables 1 and 2 for expansion of other abbreviations.

[a]No. patients in the development set.

[b]Measured in validation set.

[c]$P > .05$ considered adequate calibration.

Clinicians often make this more generic selection because none of the more precise selections appear to be absolutely correct to them.

Missing data, omission of chronic health conditions, and selection of lower consequence ICU admission diagnoses all result in lower mortality predictions. In contrast, charting low Glasgow coma scores in sedated patients will incorrectly increase severity of illness scores and predicted mortality. Although APACHE specifies an approach in this situation, low Glasgow coma scores can be a problem in ICUs that admit large numbers of patients after major surgery who arrive with residual anesthesia. With automated data extraction from computerized medical records, incorrect documentation is not altered as it might be during manual data abstraction, and artifacts can be immortalized (eg, erroneous vital signs that are inadvertently validated). As a result, there is a greater need to educate bedside clinicians about potential sources of error. There also can be errors from incorrectly implemented interfaces. Focused education programs and real-time (and retrospective) audit processes can help to avoid this problem.

The Impact of ICU Structure and Care Processes on Score Interpretation

Variations in duration of care provided prior to admission to the ICU can introduce lead-time bias, which has been shown to affect severity-adjusted mortality rates.[2,27-29] Some EDs may define their primary mission as accurate triage and prompt transfer to the appropriate care locale, whereas others focus on ensuring optimal early treatment. The former approach results in patients possibly coming to the ICU with less treatment and, thus, more abnormal vital signs (eg, higher heart rates) than the latter. More abnormal vital signs may generate higher mortality predictions (and thus a lower actual-to-predicted ratio), even though the only difference is where treatment was delivered. Similarly, preoperative and surgical care prior to ICU admission may be reflected in ICU admission acuity scores and outcomes. There has been little discussion in the ICU literature on this source of lead-time bias, and it may only be a significant factor in trauma units or ICUs where the majority of patients are admitted after elective surgical procedures.

When interpreting scoring system data, it is important to consider whether your ICU processes may differ from the reference population ICUs. Several investigators have suggested that ICU and hospital discharge practices can affect ICU performance scores.[30] Hospitals that are able to easily transfer patients to other hospitals or alternative care sites, such as long-term acute care units, will have shorter hospital stays and potentially decreased mortality rates than facilities that must continue to care for these patients. Vasilevskis et al[31] found a correlation between ICU transfers to other hospitals and observed SMR (each 1% increase in transfers was associated with a 0.02 decrease in SMR). They also noted a correlation between early postdischarge mortality and SMR and proposed that 30-day postdischarge mortality might be a more suitable outcome than hospital mortality. Patients transferred to the ICU of another hospital will appear to have both shorter stays and fewer deaths than if they completed the stay in their original ICU. This occurs most often in community hospitals with limited critical care resources and could potentially result in substantial reductions in actual-to-predicted mortality ratios. Examining discharge location data can help quality personnel to evaluate whether ICU transfers are artificially lowering mortality ratios. Currently, tracking posthospital outcomes is a manual process beyond the resources of most ICU personnel. Recent attention to hospital readmissions and payment denial for patients readmitted within 30 days of hospital discharge has increased focus on postdischarge care.[32] Until such time as hospitals routinely collect postdischarge outcome data, it seems reasonable to incorporate patient discharge location into ICU quality reporting.

Internal post-ICU care also can alter hospital outcomes in ICU patients. Slightly more than one-half of

Table 4—*Most Common APACHE Admission Diagnoses in the 2008 eICU Program Database*

Diagnosis	No. Patients (%)[a]
Acute myocardial infarction	9,065 (5.1)
Respiratory—medical, not otherwise categorized	7,773 (4.4)
CABG alone	6,665 (3.8)
CHF	5,930 (3.4)
Unstable angina	4,650 (2.6)
Rhythm disturbance (atrial, supraventricular)	4,594 (2.6)
CVA/stroke	4,630 (2.6)
Cardiovascular—medical, not otherwise categorized	4,233 (2.4)
Diabetic ketoacidosis	3,478 (2.0)
Chest pain, unknown origin	3,453 (2.0)
Cardiac arrest (with or without respiratory arrest)	3,014 (1.7)
Bleeding, GI-location undefined	3,074 (1.7)
Pneumonia, bacterial	3,036 (1.7)
Pneumonia, other	2,973 (1.7)
Sepsis, pulmonary	2,520 (1.4)
Total	69,088 (39.2)

Data from Reference 26. CABG = coronary artery bypass graft; CHF = congestive heart failure; CVA = cerebrovascular accident. See Table 1 for expansion of other abbreviation.

[a]Among 176,302 cases for which APACHE IV data were available.

all deaths in ICU patients occur in the ICU. It seems likely that post-ICU outcomes are affected by the quality of medical-surgical floor care. This impact of floor care on hospital mortality has been used to justify the use of ICU mortality ratios rather than the use of hospital mortality ratios. There are several counter arguments to focusing on ICU mortality ratios. First, premature patient discharge by the ICU team can result in postdischarge complications and deaths. Second, both recognized and unrecognized ICU-acquired problems can continue to play out after ICU discharge. Central to both is the concept that ICU care affects post-ICU outcome. An even more compelling argument is that we need to develop methodologies to assess the quality of services provided by the entire hospital system, not simply the ICU. Although ICU scoring systems were developed to measure the quality of care delivered by individual critical care units, the current focus on accountable care organizations has highlighted the need to look beyond limited epochs of care and instead evaluate care across multiple times, disciplines, and locales. This perspective argues that next-generation critical illness scoring systems should evaluate care from the minute the patient enters the acute care system until a defined time beyond hospital discharge.

Table 5 summarizes the most common causes of distortion of actual-to-predicted ratios. ICU and quality leaders should consider these distortions when selecting and implementing an ICU scoring system and periodically evaluate their relevance as part of the review of performance data.

Barriers to Implementation

Currently, only 10% to 15% of US ICUs (by patient volume) use ICU scoring systems, which is clearly at odds with standard recommendations by quality experts in other fields for frequent measurement and close scrutiny of quality data.[33] Commonly articulated obstacles to use include cost and concerns about accuracy and applicability to the patient population. Although annual licensing fees for APACHE dissuaded adoption by some hospitals in the past, open access to hospital mortality and ICU LOS algorithms have removed this obstacle. Hospitals, however, continue to balk at the cost of data collection. As mentioned previously, costs associated with manual data abstraction will largely disappear with electronic charting and abstraction. One health system estimated savings in excess of $300,000 annually after conversion from manual to automated data abstraction for APACHE data.[34]

Another frequently cited concern relates to the accuracy of ICU scoring systems. Despite good discrimination and calibration, many ICU thought leaders cite the existence of potential confounders (described in this article) as a reason for not using scoring systems to measure ICU performance. Others raise concerns about exclusion of potentially important data elements in the prognostic models. Although these are legitimate concerns, the majority of potential confounders can be mitigated through effective training and implementation of quality controls for internal use. Moreover, when used for internal quality improvement, many of the potential confounders discussed in this article become less relevant because major changes to ICU structures and processes are uncommon. Although all models degrade over time, the developers of MPM and APACHE have invested considerable energy to regularly update their calibration through periodic revisions based on more current data. So, although not perfect, ICU scoring systems are clearly the best outcome-focused ICU quality metric available today. No studies indicate that performance information derived from data available in administrative databases has adequate discrimination or calibration for use in ICU patients. As a result, we believe that failure to implement an ICU scoring system is equivalent to not having any meaningful ICU outcome data. As a result, we strongly support initiatives to increase use of ICU scoring systems. Table 6 lists recommendations for ICU leaders seeking to implement an ICU scoring system and garner maximum value from performance data.

Table 5—*Common Causes of Distortion of ICU Performance Data*

Use of outdated prognostic models or models that perform poorly in select patient populations
Missing data
Inaccurate or incorrect data
Use of scoring systems that inadequately correct for care prior to ICU admission
Discharge of patients requiring ongoing critical care to other hospitals or long-term-care facilities
Abnormal acuity mix (will be discussed in part 2 of this review)

Conclusions

APACHE, MPM, and SAPS have evolved over the past 25 years in an effort to improve predictive accuracy and keep pace with evolving critical care practices. Each provides accurate hospital mortality predictions; APACHE (and potentially MPM) also generate LOS performance data. Given the dominant role of LOS in explaining ICU cost,[35] hospital executives and ICU leaders will require data on this key financial and operational metric. Manual data collection burden is lower with MPM and SAPS. However, governmental incentives are accelerating adoption of electronic

Table 6—*Recommendations To Maximize Scoring System Validity and Utility*

	Recommendations
Validity	
	Training to ensure accurate documentation
	Systems for real-time review/correction of subjective data and robust retrospective quality reviews
	Use of a scoring system calibrated against data from many ICUs that care for a wide variety of patient conditions
	Use of a regularly recalibrated scoring system
Utility	
	Use of a scoring system that provides mortality and LOS performance data
	Regular review of performance data with ICU staff and hospital leadership
	Performance of subgroup analyses for low-, medium-, and high-acuity populations (will be discussed in part 2 of this review)
	Analysis of hospital discharge location data to monitor for "leakage" of adverse outcomes
	Focused chart reviews of adverse outcomes in low-risk patients

See Table 2 legend for expansion of abbreviation.

charting, and this will enable automation of data collection and largely obviate this problem. Data completeness and data accuracy are essential for reliable mortality and LOS predictions. These will remain key issues even with automation, and ICUs must implement robust quality control processes. Currently, only 10% to 15% of US ICUs (by patient volume) use scoring systems. Although cost and accuracy concerns have been advanced as arguments against the use of ICU scoring systems, several European countries have decided that these are insufficient excuses for not measuring the quality of ICU care and have mandated their use. We believe that societal demands for transparency and the need for improvements in quality and reductions in cost will increase the use of ICU scoring systems in US ICUs.[36,37] Part 2 of this review focuses on getting maximal value from ICU scoring system data both for providers and quality personnel who aspire to improve ICU performance and for societal consumers of benchmarking data who seek to compare performance across facilities.

Acknowledgments

Financial/nonfinancial disclosures: The authors have reported to *CHEST* the following conflicts of interest: Drs Breslow and Badawi are employees of Philips Healthcare Inc, where this work was performed.
Other contributions: We thank David Stone, MD, for his assistance in preparing this manuscript.

References

1. Vincent J-L, Moreno RP. Clinical review: scoring systems in the critically ill. *Crit Care*. 2010;14(2):311.
2. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Crit Care Med*. 2011; 39(1):163-169.
3. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34(5):1297-1310.
4. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med*. 2006;34(10):2517-2529.
5. Metnitz PGH, Moreno RP, Almeida E, et al; SAPS 3 Investigators. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. *Intensive Care Med*. 2005;31(10):1336-1344.
6. 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(10):1345-1355.
7. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med*. 2007;35(3):827-835.
8. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*. 1981;9(8):591-597.
9. Le Gall JR, Loirat P, Alperovitch A, et al. A simplified acute physiology score for ICU patients. *Crit Care Med*. 1984; 12(11):975-977.
10. Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Crit Care Med*. 1985; 13(7):519-525.
11. Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med*. 2007;35(4):1091-1098.
12. de Lange DW, Dusseljee J, Brinkman S, et al. Severity of illness and outcome in ICU patients in the Netherlands: results from the NICE registry 2006–2007. *Neth J Crit Care*. 2009;13(1):16-22.
13. Vasilevskis EE, Kuzniewicz MW, Cason BA, et al. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. *Chest*. 2009;136(1):89-101.
14. 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(6):1319-1327.
15. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland—II: outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *BMJ*. 1993;307(6910):977-981.
16. Render ML, Deddens J, Freyberg R, et al. Veterans Affairs intensive care unit risk adjustment model: validation, updating, recalibration. *Crit Care Med*. 2008;36(4):1031-1042.
17. Metnitz PG, Vesely H, Valentin A, et al. Evaluation of an interdisciplinary data set for national intensive care unit assessment. *Crit Care Med*. 1999;27(8):1486-1491.
18. Young MP, Birkmeyer JD. Potential reduction in mortality rates using an intensivist model to manage intensive care units. *Eff Clin Pract*. 2000;3(6):284-289.
19. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

20. Hosmer DW, Lemeshow S. Assessing the fit of the model. In: Hosmer DW, Lemeshow S, eds. *Applied Logistic Regression*. 2nd ed. New York, NY: Q Wiley-Interscience Publication; 2000:143-202.
21. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35(9):2052-2056.
22. Wunsch H, Rowan KM, Angus DC. International comparisons in critical care: a necessity and challenge. *Curr Opin Crit Care*. 2007;13(6):725-731.
23. Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott-Williams-Wilkins; 2008: 148-167.
24. Beck DH, Smith GB, Taylor BL. The impact of low-risk intensive care unit admissions on mortality probabilities by SAPS II, APACHE II and APACHE III. *Anaesthesia*. 2002;57(1):21-26.
25. Afessa B, Keegan MT, Gajic O, Hubmayr RD, Peters SG. The influence of missing components of the Acute Physiology Score of APACHE III on the measurement of ICU performance. *Intensive Care Med*. 2005;31(11):1537-1543.
26. Lilly CM, Zuckerman IH, Badawi O, Riker RR. Benchmark data from more than 240,000 adults that reflect the current practice of critical care in the United States. *Chest*. 2011;140(5):1232–1242.
27. Tunnell RD, Millar BW, Smith GB. The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care—a pilot study. *Anaesthesia*. 1998;53(11):1045-1053.
28. Goldhill DR, McNarry AF, Hadjianastassiou VG, Tekkis PP. The longer patients are in hospital before intensive care admission the higher their mortality. *Intensive Care Med*. 2004;30(10):1908-1913.
29. Rosenberg AL, Hofer TP, Strachan C, Watts CM, Hayward RA. Accepting critically ill transfer patients: adverse effect on a referral center's outcome and benchmark measures. *Ann Intern Med*. 2003;138(11):882-890.
30. Kahn JM, Kramer AA, Rubenfeld GD. Transferring critically ill patients out of hospital improves the standardized mortality ratio: a simulation study. *Chest*. 2007;131(1): 68-75.
31. 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(7):803-812.
32. Jweinat JJ. Hospital readmissions under the spotlight. *J Healthc Manag*. 2010;55(4):252-264.
33. Chassin MR, Loeb JM, Schmaltz SP, Wachter RM. Accountability measures—using measurement to promote quality improvement. *N Engl J Med*. 2010;363(7):683-688.
34. Rincon T, Welcher B, Srikanth D, Seiver A. Economic implications of data collection from a remote center utilizing technological tools [abstract]. *Crit Care Med*. 2007; 35(12):A161.
35. Rapoport J, Teres D, Zhao Y, Lemeshow S. Length of stay data as a guide to hospital economic performance for ICU patients. *Med Care*. 2003;41(3):386-397.
36. Murphy JG, Dunn W. Transparency in health care: an issue throughout US history. *Chest*. 2008;133(1):9-10.
37. Mongan JJ, Ferris TG, Lee TH. Options for slowing the growth of health care costs. *N Engl J Med*. 2008;358(14):1509-1514.

CHEST Postgraduate Education Corner
CONTEMPORARY REVIEWS IN CRITICAL CARE MEDICINE

Severity Scoring in the Critically Ill

Part 2: Maximizing Value From Outcome Prediction Scoring Systems

Michael J. Breslow, MD; and Omar Badawi, PharmD

Part 2 of this review of ICU scoring systems examines how scoring system data should be used to assess ICU performance. There often are two different consumers of these data: lCU clinicians and quality leaders who seek to identify opportunities to improve quality of care and operational efficiency, and regulators, payors, and consumers who want to compare performance across facilities. The former need to know how to garner maximal insight into their care practices; this includes understanding how length of stay (LOS) relates to quality, analyzing the behavior of different subpopulations, and following trends over time. Segregating patients into low-, medium-, and high-risk populations is especially helpful, because care issues and outcomes may differ across this severity continuum. Also, LOS behaves paradoxically in high-risk patients (survivors often have longer LOS than nonsurvivors); failure to examine this subgroup separately can penalize ICUs with superior outcomes. Consumers of benchmarking data often focus on a single score, the standardized mortality ratio (SMR). However, simple SMRs are disproportionately affected by outcomes in high-risk patients, and differences in population composition, even when performance is otherwise identical, can result in different SMRs. Future benchmarking must incorporate strategies to adjust for differences in population composition and report performance separately for low-, medium- and high-acuity patients. Moreover, because many ICUs lack the resources to care for high-acuity patients (predicted mortality > 50%), decisions about where patients should receive care must consider both ICU performance scores and their capacity to care for different types of patients. ***CHEST 2012; 141(2):518–527***

Abbreviations: APACHE = Acute Physiology and Chronic Health Evaluation; eRI = eICU Research Institute; LOS = length of stay; SMR = standardized mortality ratio

As described in part 1[1] of this review, ICU scoring systems evolved to meet the desire of clinical and administrative leaders to assess the quality of care provided by their ICUs. Measuring ICU performance and using this information to guide quality improvement activities remains an important rationale for their use today. Yet ICU scoring systems differ from other quality metrics in several important ways. Unlike best practice compliance, in which 100% compliance is a logical goal, optimal ICU care will never result in all patients surviving their ICU stay and ICU length of stay (LOS) will never equal zero days. Many users simply aspire to a standardized mortality ratio (SMR) value < 1.0 without really considering how they can derive additional value from ICU scoring system data. So how should users obtain maximal value from a relative performance metric like SMRs or actual-to-predicted LOS ratios? Another unusual characteristic of ICU scoring systems is their use across a wide variety of diagnoses and patient acuities. Although there is value in having a single tool that generates a single score for the entire ICU population, this aggregation can obscure important variability in performance. The goal of this review is to help

Manuscript received February 9, 2011; revision accepted August 29, 2011.
Affiliations: From the Department of Research and Product Marketing (Drs Breslow and Badawi), Philips Healthcare; and the Department of Pharmacy Practice and Science (Dr Badawi), University of Maryland School of Pharmacy, Baltimore, MD.
Correspondence to: Michael J. Breslow, MD, Research and Product Marketing, Philips Healthcare, Ste 1900, 217 E Redwood St, Baltimore, MD 21202; e-mail: michael.breslow@philips.com

DOI: 10.1378/chest.11-0331

ICU clinicians and quality personnel maximize the value derived from ICU scoring system data, better understand the performance of their ICU, and use this information to identify areas for focused quality improvement efforts. We also address several key issues in the use of ICU scoring system data for benchmarking. This latter issue deserves particular attention in light of recent efforts by several countries (and the State of California) to use ICU scoring systems to benchmark ICUs. Throughout the review we use data from the eICU Research Institute (eRI) to illustrate key points. The eRI contains aggregate deidentified data from > 500 ICUs with remote ICU care programs[2] and constitutes an important resource for understanding scoring system characteristics. As employees of Philips VISICU, and cognizant of our potential conflict of interest, we have tried in this review to focus on concepts and data of general applicability to all users of ICU scoring system data.

Use of Scoring Systems for Internal Quality Improvement

Hospitals generally use ICU scoring systems in order to better understand how well their ICUs are performing. As discussed in part 1, only APACHE (Acute Physiology and Chronic Health Evaluation) (Cerner Corp) provides both mortality and LOS predictions,[3-5] although Vasilevskis et al[5] recently developed LOS prediction algorithms from Mortality Probability Model data elements. Given the current focus on health-care costs and the preeminent role of LOS in determining ICU costs, LOS is becoming an increasingly essential ICU performance metric.[6] APACHE is also the only scoring system that provides separate ICU and hospital predictions. Although most would agree that hospital mortality is the key mortality metric, inconsistencies between the two can identify quality gaps (eg, premature discharge, poor floor care). Similarly, threefold-higher ICU costs make this the key LOS metric.[7] Here also, discordance between ICU and hospital LOS can indicate the presence of systemic problems (eg, lack of floor beds).

Historically, some ICUs performed one-time analyses of ICU performance as a form of "spot check." Currently, most ICUs use scoring systems to track performance over time, with the goal of continuously improving quality and rapidly detecting problems that might reflect gaps in care.[8] However, consumers of scoring system data must avoid too frequent measurement of actual-to-predicted ratios, because results from small numbers of patients (eg, monthly in most ICUs and quarterly in low-census ICUs) can result in less reliable data and wide CIs. Use of larger population samples (longer observation periods) and persistence of changes over time increase the likelihood that changes observed are real. Other tools such as run charts, which show results over multiple time periods, can help visualize longitudinal performance and reduce reliance on larger sample sizes.[9]

Another issue in the longitudinal tracking of performance is compensating for general trends in care. Several studies have reported on global reductions in hospital mortality over time.[10,11] Although there are no ICU-specific data, the recalibration of the APACHE algorithms provides some helpful insights.[3,4] The APACHE III algorithms, when applied to the APACHE IV reference population, generated a hospital SMR of 0.93, suggesting that mortality decreased by slightly less than 1% a year over the 10 years between the two calibrations.[3] Our group performed a similar comparison for LOS performance using the eRI data set and observed negligible difference in the actual-to-predicted ratios, suggesting little change in LOS performance over the same interval. Although these data suggest that time-related changes in ICU performance have been small over the past decade, larger changes in aggregate performance may be seen in the future. Specifically, the current focus on quality improvement and cost reduction, the introduction of new therapeutic approaches to several high-impact diseases (eg, severe sepsis), and the implementation of new ICU care models may accelerate future improvements.[12-14] In order to compensate for temporal trends, scoring systems need periodic recalibration; the CalHospitalCompare (California Intensive Care Outcomes) project is considering yearly updates of their algorithms (R. Adams Dudley, MD, oral communication, December 2010). Although regular recalibration is necessary to ensure that ICU outcomes are not compared with an old reference population that used therapies and practices that are different from those in current use, it is equally important to track ICU performance trends by referencing newly calibrated systems to their predecessors, and making this information available to consumers of public health information.

Measuring Mortality Performance

Mortality is a key ICU quality metric and reflects many aspects of ICU care, including use of best practices, accurate diagnosis, and effective and timely therapies. ICU scoring systems provide mortality predictions based on ICU admission status/severity of illness. Although these predictions provide little value in managing individual patients, aggregate predictions correlate very well with observed outcomes in the reference population. Because of this population-level accuracy, ICU scoring systems measure mortality performance by comparing the actual number of deaths in an ICU population with the sum of the individual

mortality predictions of the group. This method is known as the indirect method of standardization, and the resulting measure is the SMR. The benefits of indirect standardization are its ease of calculation and its stability in small sample sizes. However, despite its widespread acceptance, the SMR has some intrinsic limitations. High-risk patients contribute disproportionately to the SMR, because more of these patients die. An ICU with better than average outcomes in its low-risk population can have an aggregate SMR >1.0 if outcomes are below average in the high-risk group, even though the high-risk group represents a smaller percentage of the total ICU population. (Tables 1 and 2 demonstrate this effect).

Unlike the case with most quality metrics, in which hospitals set goals for how they rank relative to other institutions (eg, top 25%, top 10%), there are insufficient data in the public domain for similar calibration of ICU scoring system results. As a result, many ICUs simply target SMR/actual-to-predicted ratios <1.0. Kuzniewicz et al[15] and Vasilevskis et al[5] provided individual mortality and LOS scores for the 29 hospitals in their original California Intensive Care Outcomes publications; the 25th to 75th percentile scores were approximately 0.8 and 1.20 for mortality and 0.85 and 1.15 for ICU LOS. Whether these data reflect actual variance across ICUs in the United States is unknown. It is also worth noting that both APACHE IV and Mortality Probability Model III (from time of admission) were calibrated to patients cared for in ICUs that chose to measure ICU performance and invested resources for this purpose.[3,4,16] The performance of these self-selected hospitals may not reflect that of the average US ICU.

Measuring LOS Performance

ICU LOS is another important quality and financial metric (ICU LOS is the primary determinant of ICU cost).[6] LOS is also an important measure of operating efficiency, because occupancy rates are high in many ICUs. Capacity constraints affect ED throughput, ambulance diversion status, elective surgical schedules, and acceptance of intrahospital transfers. LOS is affected by many factors, including quality of ICU care, end-of-life policies, discharge planning, and downstream bed availability. However, unlike mortality, where increasing severity of illness is linearly related to predicted mortality rate, LOS demonstrates a more complex relationship to admission acuity. As Figure 1 illustrates, predicted LOS increases with increasing acuity, and then decreases at the highest levels of severity. To better understand this behavior, we examined year 2006 data from 62,000 patients in the eRI database (153 ICUs), and analyzed survivors and nonsurvivors separately. Details of the ICUs contributing data to the eRI database are described elsewhere.[2,17]

Figure 2 shows LOS data for survivors and nonsurvivors as a function of acuity/mortality risk. For surviving patients, LOS increased linearly as predicted mortality rose. In contrast, LOS was largely unrelated to acuity for nonsurvivors. It thus appears that the relationship between severity of illness and LOS in Figure 1 reflects the behavior of two distinct populations, survivors and nonsurvivors. It seems reasonable to attribute the linear relationship between severity and LOS in survivors to sicker patients requiring longer times to recover from their illness and be stable enough for ICU discharge. The explanation for the lack of relationship between LOS and acuity in nonsurvivors is unknown. We speculate that some patients who have low mortality risk at ICU admission (and short predicted LOS) develop complications and eventually succumb to these new problems, whereas some high-mortality-risk patients, who would have a long LOS if they lived, die within 1 or 2 days of admission despite maximal therapy. In some ICUs, care limitations may also contribute to this behavior.

Table 1—*Comparison of Aggregate SMRs in Two Hypothetical ICUs With Similar Performance Within Risk Groups But Different Severity Distribution*

Example ICUs	Low-Risk Patients	Medium-Risk Patients	High-Risk Patients	All Patients
ICU A				
Patients, No.	600	300	100	1,000
Predicted mortality rate, %	3	15	60	12.3
Actual deaths, No.	9	45	90	144
Predicted deaths, No.	18	45	60	123
SMR	0.50	1.0	1.50	1.17
ICU B				
Patients, No.	400	300	300	1,000
Predicted mortality rate, %	3	15	60	23.7
Actual deaths, No.	6	45	270	321
Predicted deaths, No.	12	45	180	237
SMR	0.50	1.0	1.50	1.35

SMR = standardized mortality ratio.

Table 2—*Subgroup SMR Data Showing How Aggregate SMR Data Can Obscure Poor Low-Risk Population Performance*

ICU A	Low-Risk Patients	Medium-Risk Patients	High-Risk Patients	All Patients
Patients, No.	400	300	200	900
Actual deaths, No.	24	42	110	176
Predicted mortality rate, %	3.0	15.0	60.0	19.7
Actual deaths, %	6.0	14.0	55.0	19.6
Predicted deaths, No.	12	45	120	177
SMR	2.00	0.93	0.92	0.99

See Table 1 legend for expansion of abbreviations.

The divergent LOS behavior between survivors and nonsurvivors has potentially important consequences. The APACHE-predicted LOS for high-risk patients is a blend of the average survivor and nonsurvivor LOS (eg, if 8 days and 4 days, respectively, for patients with 75% predicted mortality, the predicted LOS would be 5 days). ICUs that have higher than predicted survival for these patients have more

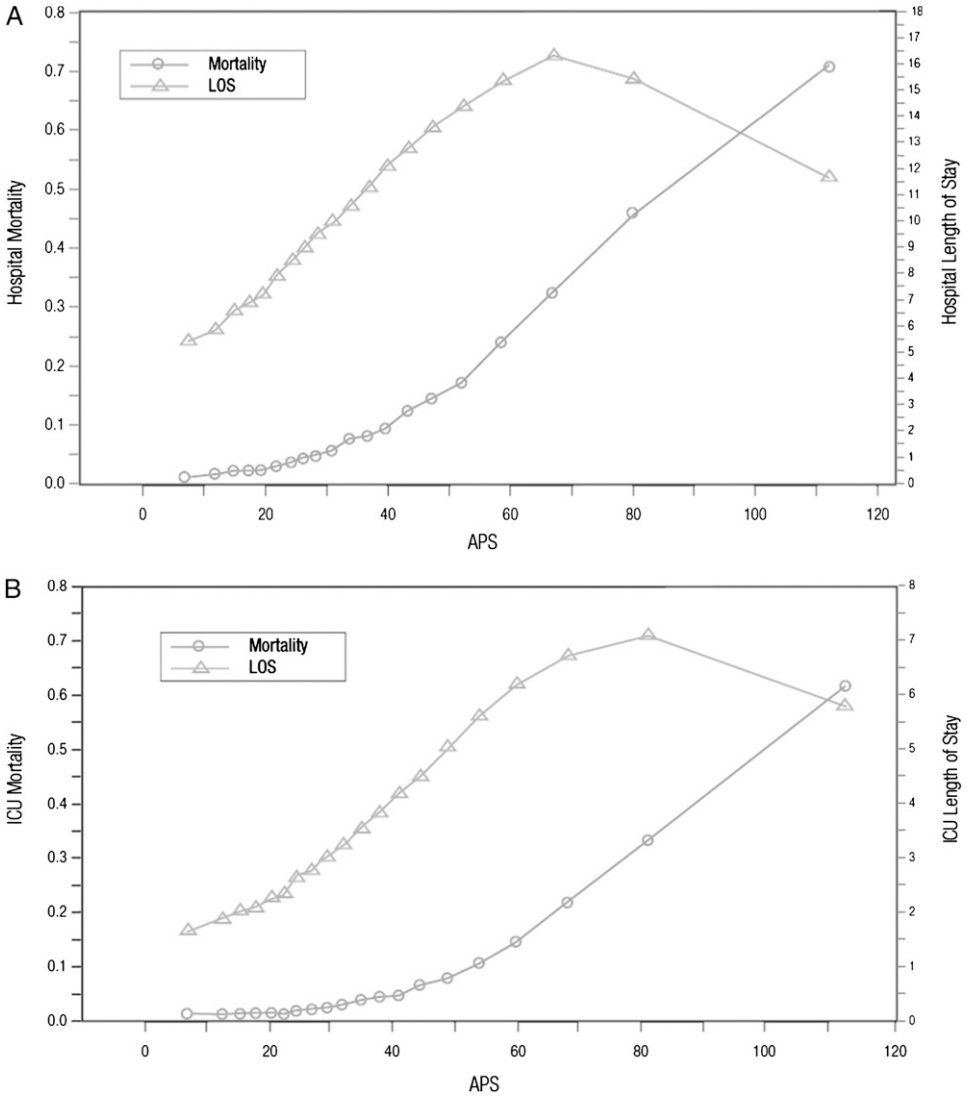

FIGURE 1. A, ICU mortality and LOS, non-coronary artery bypass graft (CABG) patients. B, Hospital mortality and LOS, non-CABG patients. Actual mortality and LOS data from the APACHE (Acute Physiology and Chronic Health Evaluation) III validation data set, shown as a function of the APACHE first ICU day acute physiology score (APS). Data are displayed by fifth percentiles. LOS = length of stay. (Reproduced with permission from the Cerner Corporation, Kansas City, MO).

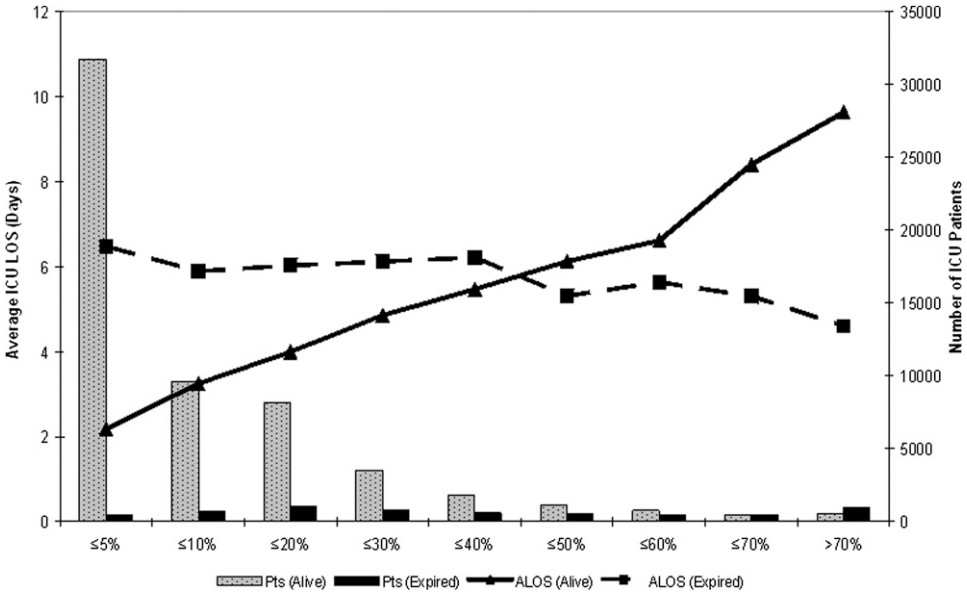

FIGURE 2. Average ICU LOS by predicted mortality. eICU Research Institute data for all patients (62,397) discharged from eICU Program ICUs in 2006 showing average ICU LOS data for surviving and nonsurviving patients as a function of APACHE-III-predicted hospital mortality. Patients are aggregated into deciles of predicted mortality. Mortality predictions were generated using the APACHE III first ICU day mortality prediction algorithm. ALOS = average length of stay; Pts = patients. See Figure 1 legend for expansion of other abbreviations.

survivors than the reference population, and thus the APACHE-blended LOS prediction underestimates what their LOS should be. Using the example here, 50% mortality in the 75% predicted mortality group would translate into an expected LOS of 6 days. This phenomenon increases this ICU's actual-to-predicted LOS ratio, and penalizes high-performing ICUs that have better-than-predicted high-risk patient mortality rates. The converse is also true: poor-performing ICUs that have excess mortality in this group benefit from the assumption of a fixed high-acuity population mortality rate. For this reason, we report actual-to-predicted LOS performance for low-, medium-, and high-risk patients separately, and provide aggregate actual-to-predicted LOS data both with and without patients with mortality predictions $>50\%$.

Using ICU Scoring Systems for Benchmarking

The move toward benchmarking ICUs is driven primarily by regulator and consumer interest in identifying high- and low-performing institutions. As discussed in Part 1 of this review, several countries have mandatory ICU reporting; the State of California appears to be moving in this direction as well. In all of these regions, standard ICU scoring systems have been recalibrated against their patient population. Although calibration does not affect the ability to compare ICUs within the region, local calibration precludes comparisons with ICUs that use scoring systems calibrated against other populations. The adoption of regionally calibrated scoring systems makes it more difficult to determine whether certain countries have developed care models that achieve superior outcomes. Publication of regression models would enable such comparisons and would be desirable.

Despite the limited adoption of ICU scoring systems in the United States, other countries have mandated their use in all ICUs.[18-20] There is a broad audience for quality data, and we can anticipate increased use of scoring systems for benchmarking ICU performance. Consumers of this information, however, must be cautious in how they use these data, because differences in the numbers of high- and low-risk patients can affect calculated SMR. Two ICUs with identical mortality rates for their low- and high-risk patients can have different aggregate SMRs if they have different numbers of low- and high-risk patients. This behavior reflects the disproportionate impact of high-risk patients on SMR, and ICUs with fewer such patients will have this effect diluted. Epidemiologists have long recognized this phenomenon and generally advocate against the practice of comparing simple SMRs.[21] Table 1 shows how population mix can affect the overall SMR ranking for two hypothetical ICUs.

This bias in SMR can be addressed through another simple technique referred to as the method of direct standardization.[21,22] This can be done by assuming a distribution of high-, medium-, and low-risk patients

equal to that in the reference population (or any population deemed appropriate for standardization). Unlike the SMR, this generates an adjusted mortality rate that adequately addresses the confounding introduced by different distributions of patient acuity between populations. Unfortunately, the adjusted mortality rate has no real meaning; we, therefore, recommend transforming the adjusted mortality rate of sample populations into a standardized rate ratio.[22] This can be done by dividing the adjusted mortality rate by the mortality rate of the reference population, which generates a "population-adjusted SMR" that is more intuitive to most consumers of benchmarking data. Direct standardization and subsequent calculation of the population-adjusted SMR requires large samples with at least 30 total events (deaths) and multiple events within each stratum.[22] Therefore, this may only be feasible over fairly long time horizons, especially for smaller ICUs (eg, yearly), or perhaps more frequently if ICU data are aggregated at the hospital level.

Population adjustment requires stratifying the population into multiple risk groups; more groups equate to more accurate adjustment but this also creates mathematical instability when events within strata are rare. We have used this technique on data from ICUs in the eRI and confirmed that population variability is present, but actual population variability is generally insufficient to induce large degrees of bias into the SMR. Figure 3 shows 105 eRI ICUs that cared for at least 1,000 patients in 2010, and displays the correlation between the standard SMR and the "population-adjusted SMR" generated using six risk groups. The SMR and "population-adjusted SMR" are highly correlated in this group. These data suggest that, although population-adjusted SMRs can measure performance more accurately when there are major differences in population distributions, the adjustment will not significantly impact most ICUs.

Use of population-adjusted SMRs addresses the confounding introduced by case mix, but it does not provide insight into heterogeneity in performance across risk groups. We believe that heterogeneous performance is a major quality concern, and have provided separate outcome data for low-, medium-, and high-acuity patients as part of the routine performance data set provided to all ICUs with electronic ICU care programs for the past 6 years. We initially segregated

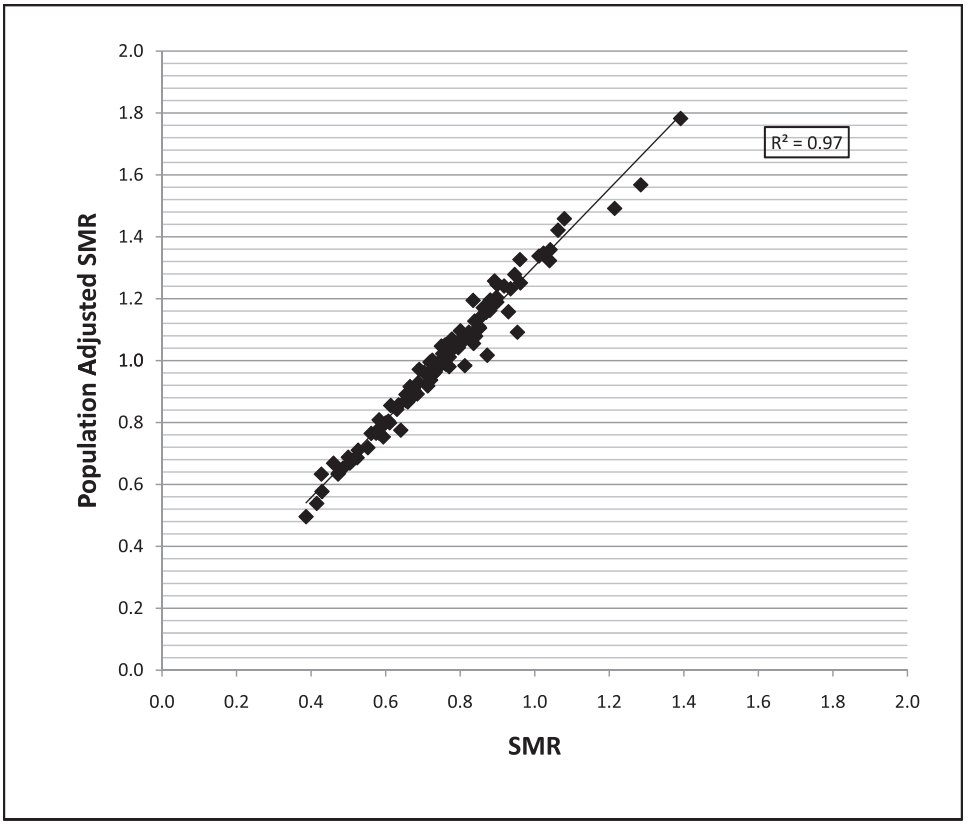

FIGURE 3. Correlation between direct and indirect standardization. Simple SMR data compared with population-adjusted (direct standardization) mortality ratio data from 105 ICUs in the eICU Research Institute data set. Only ICUs with at least 1,000 patients with APACHE IV hospital mortality predictions were included. Data are from patients discharged from the hospital in 2010. SMR = standardized mortality ratio. See Figure 1 legend for expansion of other abbreviations.

patients by acuity because we speculated that processes important for preventing complications in low-risk patients might be different from those that result in improved survival of high-acuity patients.[23] To examine this hypothesis, we used 10% and 50% APACHE III-predicted hospital mortality cutoffs to differentiate low-, medium-, and high-risk populations, respectively. Although these cutoffs result in dissimilar-sized groups, the number of deaths in each group is similar, and this maximizes the reliability of the SMR calculations. This definition of low risk is also used by APACHE for their "low risk monitor" subset of patients.[24] Across the electronic ICU program install base, approximately two-thirds of ICU patients are low risk on admission, which is similar to the APACHE IV validation cohort.[1,2] Larger tertiary care hospital ICUs tend to have slightly fewer low-risk patients (although frequently >50%) and more patients with very high mortality risk. Figure 4 shows the lack of correlation between mortality performance in low- and high-acuity patients in the same 105 ICUs from Figure 3. These data demonstrate wide disparity within ICUs in their performance in the different acuity groups, and highlight the rationale for examining low-, medium-, and high-acuity groups separately.

Using Scoring Systems to Study Subpopulations

As stated previously, because of heterogeneity in performance across risk groups, we advocate independent assessment of outcomes in low-, medium-, and high-acuity patients for both ICU quality assessment and for benchmarking. Cutoffs of 10% and 50% predicted hospital mortality to define low-, medium-, and high-risk groups are easily understood, have some historical precedence, and improve the stability of the calculations. Of particular concern, aggregate SMRs can mask quality problems in the low-risk cohort of patients. The patients in this group, which represents more than one-half of all admissions in most ICUs, are often admitted because they are at risk of complications. Table 2 illustrates this problem in a hypothetical tertiary care ICU.

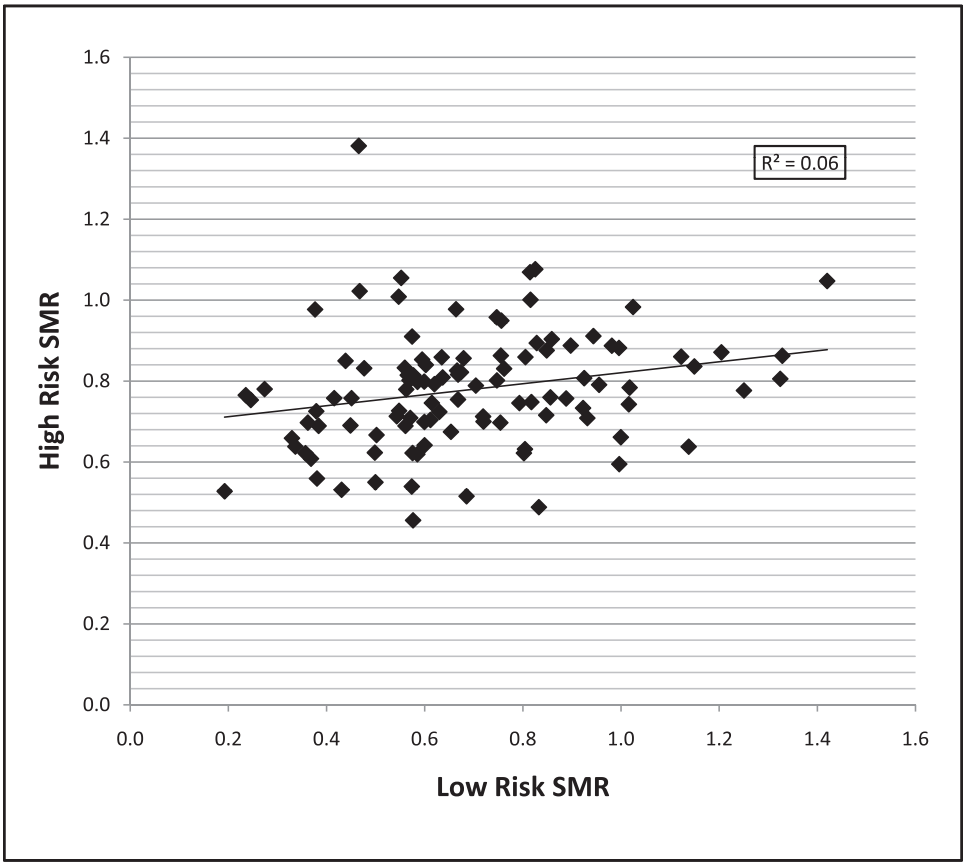

Figure 4. Lack of correlation between high-risk and low-risk SMRs. SMR data for low- and high-mortality-risk patients from 105 ICUs in the eICU Research Institute data set. Only ICUs with at least 1,000 patients with APACHE IV hospital mortality predictions were included. Data are from patients discharged from the hospital in 2010. Low- and high-risk populations had APACHE-IV-predicted hospital mortality below 10% and above 50%, respectively. There was little intra-ICU correlation between performance in the two populations. See Figure 1 and 3 legends for expansion of abbreviations.

Despite having an SMR of 2.0 in the low-risk population, which represents 45% of the total patients, the aggregate SMR for this ICU is slightly below 1.0. This hypothetical example resembles actual observations that we presented to ICU leaders at an academic medical center. Although the cause of discrepant performances may vary, it is easy to see how excessive focus on high-acuity patients, who can require considerable attention when acutely unstable, can divert attention from more "stable" patients. Regardless of the basis for this problem, outcomes in this subpopulation in these ICUs improved markedly once the problem was recognized.

Subgroup analysis by acuity is also important for LOS. In a recent analysis of July to December 2007 data from four ICUs within a single large health system, 16% of patients had ICU stays in excess of 6 days. LOS outliers came from all three acuity groups (low, medium, and high), with the lowest incidence in the low-mortality-risk group and the highest incidence in the high-risk population. However, because two-thirds of the patients were in the low-risk group, this population made up 50% of the total LOS outliers. These low-risk LOS outliers accounted for 25% of the total ICU days. They also had fivefold higher mortality (actual to predicted) than low-risk patients with shorter LOS, suggesting that complications accounted for both longer stays and lower survival. LOS outliers are important because they have substantially higher costs, and ICU leaders should determine whether their long stays are attributable to high acuity on arrival to the ICU (mostly unavoidable LOS) or to potentially avoidable complications in low-risk patients. Examining low- and high-risk populations enables ready differentiation of these two different causes for long LOS, and has relevance to both quality and financial personnel. We have reported low-risk LOS outlier data as a component of our ICU quality metrics for the past 6 years. Over this time period, there was a 30% decrease in the number of low-risk LOS outliers (and a decrease in aggregate actual-to-predicted LOS ratios). We speculate that providing these data prompted organizational efforts to reduce avoidable complications in this population.

Another potential use for ICU scoring systems is to better understand the behavior of select populations of patients (eg, patients cared for by individual providers or patients with specific diagnoses). However, care must be taken to ensure adequate sample size when using scoring systems for this purpose. Simplified Acute Physiology Score (SAPS) researchers have presented subgroup performance by geographic region,[25] and APACHE researchers have produced sepsis and coronary artery bypass graft-specific models to facilitate evaluating outcomes in these subgroups more effectively.[26,27] Recent programs aimed at improving outcomes in patients with severe sepsis have created interest in using scoring system data to assess the impact of these efforts.[12] Similar cautions regarding adequate sample size apply to this use case as well.

Other Uses

Clinical investigators use severity of illness scores and/or mortality and LOS predictions to compare outcomes across different experimental groups.[28,29] Two different methodologies are generally used. In one, actual-to-predicted ratios of the experimental groups are compared directly with the control group in order to assess whether outcomes differ. Drawbacks to this approach include the added variability in patient predictions, the potential for bias, and the assumption that the prediction model, which was derived in a different population, optimally explains outcome differences in the study population. Certain scoring systems (eg, APACHE) have been customized for particular populations such as cardiac surgery by including predictors specific to that population.[2] Although this may be useful in unique situations, in general, the usefulness of the SMR remains limited as an outcome for clinical research. Whenever possible, a customized model that adjusts for confounders specifically related to the exposure and outcome of interest should be used, which may not be adequately captured by the risk scoring system alone. For this reason, many investigators create customized regression models for their study population that include a severity of illness score as a single covariate.[14]

ICU scoring systems have been proposed to have value in ICU discharge and admission decisions. APACHE provides data on patients who were admitted to the ICU with a low predicted mortality and required no major intervention during the first ICU day, referred to as "low-risk monitor" patients. While intended to provide retrospective information about the appropriateness of ICU admission practices (because patients cannot be classified as low-risk monitor until after the first ICU day), Zimmerman and Kramer[24] recently proposed that APACHE could help identify patients not requiring ICU interventions. It may also be reasonable to assume that severity of illness might correlate with risk of deterioration after ICU discharge. Investigators have developed predictive algorithms using patient and physiologic variables to differentiate patients who did and did not develop postdischarge problems (readmission or death).[30,31] Although these models showed moderate discrimination, there are as yet no data establishing the usefulness of such algorithms in discharge decision making.

Conclusions

ICU scoring systems have the potential to provide useful information for both ICU quality personnel and the general public. Garnering maximal value from scoring system data requires in-depth knowledge of how these scoring systems behave in different populations, and how care changes over time. Sophisticated users who evaluate low-, medium-, and high-acuity subgroup performance independently can use these data to target quality issues and improve quality of care. The increasing use of ICU scoring for benchmarking has the potential to provide helpful comparative performance data. However, adoption of simple SMR reporting for this purpose is problematic, because it fails to adjust for differences in population composition and lacks information about performance of different acuity subpopulations. Because the audience for benchmarking data often knows little about these tools, quality leaders need to advocate for refinements in current reporting strategies. One potential value of benchmarking is the identification of superior systems of care. The current practice of calibrating scoring systems in each region (eg, Great Britain, the Netherlands) precludes comparison of international outcomes; it is hoped that broader adoption of scoring systems across multiple countries will lead to the use of internationally generated algorithms.

Acknowledgments

Financial/nonfinancial disclosures: The authors have reported to *CHEST* the following conflicts of interest: Drs Breslow and Badawi are employees of Philips Healthcare, which sells a tele-ICU solution.
Other contributions: We thank David Stone, MD, for his assistance in preparing this manuscript.

References

1. Breslow MJ, Badaw O. Severity scoring in the critically ill: part 1—interpretation and accuracy of outcome prediction scoring systems. *Chest*. 2011;141(1):245-252.
2. McShea M, Holl R, Badawi O, Riker RR, Silfen E. The eICU research institute - a collaboration between industry, healthcare providers, and academia. *IEEE Eng Med Biol Mag*. 2010;29(2):18-25.
3. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34(5):1297-1310.
4. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med*. 2006;34(10):2517-2529.
5. Vasilevskis EE, Kuzniewicz MW, Cason BA, et al. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. *Chest*. 2009;136(1):89-101.
6. Rapoport J, Teres D, Zhao Y, Lemeshow S. Length of stay data as a guide to hospital economic performance for ICU patients. *Med Care*. 2003;41(3):386-397.
7. Halpern NA, Pastores SM, Greenstein RJ. Critical care medicine in the United States 1985-2000: an analysis of bed numbers, use, and costs. *Crit Care Med*. 2004;32(6):1254-1259.
8. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Crit Care Med*. 2011;39(1):163-169.
9. Perla RJ, Provost LP, Murray SK. The run chart: a simple analytical tool for learning from variation in healthcare processes. *BMJ Qual Saf*. 2011;20(1):46-51.
10. Bueno H, Ross JS, Wang Y, et al. Trends in length of stay and short-term outcomes among Medicare patients hospitalized for heart failure, 1993-2006. *JAMA*. 2010;303(21):2141-2147.
11. Capelastegui A, España PP, Quintana JM, et al. Declining length of hospital stay for pneumonia and postdischarge outcomes. *Am J Med*. 2008;121(10):845-852.
12. Levy MM, Dellinger RP, Townsend SR, et al; Surviving Sepsis Campaign. The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Crit Care Med*. 2010;38(2):367-374.
13. Lilly CM, Thomas EJ. Tele-ICU: experience to date. *J Intensive Care Med*. 2010;25(1):16-22.
14. Lilly CM, Cody S, Zhao H, et al; University of Massachusetts Memorial Critical Care Operations Group. Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. *JAMA*. 2011;305(21):2175-2183.
15. 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(6):1319-1327.
16. Higgins TL, Kramer AA, Nathanson BH, Copes W, Stark M, Teres D. Prospective validation of the intensive care unit admission Mortality Probability Model (MPM_0-III). *Crit Care Med*. 2009;37(5):1619-1623.
17. Lilly CM, Zuckerman IH, Badawi O, Riker RR. Benchmark data from more than 240,000 adults that reflect the current practice of critical care in the United States. *Chest*. 2011;140(5):1232-1242.
18. Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med*. 2007;35(4):1091-1098.
19. de Lange DW, Dusseljee J, Brinkman S, et al. Severity of illness and outcome in ICU patients in the Netherlands: results from the NICE registry 2006–2007. *Neth J Crit Care*. 2009;13:16-22.
20. Metnitz PG, Vesely H, Valentin A, et al. Evaluation of an interdisciplinary data set for national intensive care unit assessment. *Crit Care Med*. 1999;27(8):1486-1491.
21. Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd. Philadelphia, PA: Lippincott-Williams-Wilkins; 2008:148-167.
22. Schoenbach VJ. Standardization of rate and ratios. In: Schoenbach VJ, Rosamond WD, eds. *Understanding the Fundamentals of Epidemiology: An Evolving Text*. Chapel Hill, NC: University of North Carolina at Chapel Hill; 2000:129-151.
23. Moreno RP, Bauer P, Metnitz PG. Characterizing performance profiles of ICUs. *Curr Opin Crit Care*. 2010;16(5):477-481.
24. Zimmerman JE, Kramer AA. A model for identifying patients who may not need intensive care unit admission. *J Crit Care*. 2010;25(2):205-213.
25. Moreno RP, Metnitz PGH, 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(10):1345-1355.
26. Knaus WA, Harrell FE, Fisher CJ Jr., et al. The clinical evaluation of new drugs for sepsis. A prospective study design based on survival analysis. *JAMA*. 1993;270(10):1233-1241.
27. Becker RB, Zimmerman JE, Knaus WA, et al. The use of APACHE III to evaluate ICU length of stay, resource use, and mortality after coronary artery by-pass surgery. *J Cardiovasc Surg (Torino)*. 1995;36(1):1-11.
28. Sorbello A, Komo S, Valappil T, Nambiar S. Registration trials of antibacterial drugs for the treatment of nosocomial pneumonia. *Clin Infect Dis*. 2010;51(suppl 1):S36-S41.
29. Gursel G, Demirtas S. Value of APACHE II, SOFA and CPIS scores in predicting prognosis in patients with ventilator-associated pneumonia. *Respiration*. 2006;73(4):503-508.
30. Fernandez R, Serrano JM, Umaran I, et al; Sabadell Score Study Group. Ward mortality after ICU discharge: a multicenter validation of the Sabadell score. *Intensive Care Med*. 2010;36(7):1196-1201.
31. Gajic O, Malinchoc M, Comfere TB, et al. The Stability and Workload Index for Transfer score predicts unplanned intensive care unit patient readmission: initial development and validation. *Crit Care Med*. 2008;36(3):676-682.