Prospective validation of the intensive care unit admission Mortality Probability Model (MPM_0-III)*

Thomas L. Higgins, MD, MBA; Andrew A. Kramer, PhD; Brian H. Nathanson, PhD; Wayne Copes, PhD; Maureen Stark; Daniel Teres, MD

Objective: **To validate performance characteristics of the intensive care unit (ICU) admission mortality probability model, version III (MPM_0-III) on Project IMPACT data submitted in 2004 and 2005. This data set was external from the MPM_0-III developmental and internal validation data collected between 2001 and 2004.**

Design: **Retrospective analysis of clinical data collected concurrently with care.**

Setting: **One hundred three (103) adult ICUs in North America that voluntarily collect and submit data to Project IMPACT.**

Subjects: **A total of 55,459 patients who were eligible for MPM scoring (age ≥18; first ICU admission for hospitalization, excludes burns, coronary care, and cardiac surgical patients).**

Interventions: **None.**

Measurements: **Prevalence of MPM risk factors and their relationship to hospital mortality; calibration and discrimination of MPM_0-III model applied to new data.**

Main Results: **Seventy-eight ICUs contributed data to both this study and the original development set. Fifty-six ICUs from the original MPM_0-III study were replaced by 25 new ICUs in this external validation set. Patient characteristics (type of patient, risk factors, and resuscitation status) were similar to the original 2001–2004 cohort, except for slightly more patients on mechanical ventilation at admission (32% vs. 27%, $p < 0.01$) and the percentage of patients having no MPM_0-III risk factors except age (11% vs. 14%, $p < 0.01$). Observed deaths were 7331 (13.2%) vs. 7456 predicted, yielding a standardized mortality ratio of 0.983, 95% CI (0.963–1.001).**

Conclusions: **MPM_0-III calibrates on a new population of 55,459 North American patients who include many patients from new ICUs, which helps confirm that the model is robust and was not overfitted to the development sample. Although Project IMPACT participants change over time, 2004–2005 patient risk factors and their relationship to hospital mortality have not significantly changed. The increase in mechanically ventilated patients and reduction in admissions with no risk factors are trends worth following. (Crit Care Med 2009; 37:1619–1623)**

KEY WORDS: transparency; severity of illness; mortality probability model; intensive care; outcomes; benchmarking

Transparency in health care is facilitated by the release of standardized performance metrics that are easily accessible to any interested observer. Historically, this type of data has been used by physicians and other healthcare providers to improve performance by benchmarking their results against others. With the public release of data, insurers may choose to reward quality and efficiency, and patients may make better decisions when choosing health care. The Society of Critical Care Medicine presciently recognized the need for standardized data collection in the 1990s with the creation of Project IMPACT (1). From the outset, Project IMPACT used the Mortality Probability Model at intensive care unit (ICU) admission, version 2 (MPM_0-II) (2) as the primary tool for adjusting outcome results by a patient's presenting severity of illness. MPM_0-II was developed on 12,610 patients undergoing critical care in 1989–1990 in the United States and Europe. However, by 2002 it was becoming clear that MPM_0-II was poorly calibrated to current clinical practice. In the Project IMPACT group of North American ICUs, observed mortality rates were consistently lower than those predicted by the aging model (3). As a result, the MPM was updated, using Project IMPACT data from 124,885 patients in 135 ICUs at 98 hospitals participating between October 2001 and March 2004 (4). The data were randomly split into development (60%) and internal validation (40%) sets. Analysis demonstrated the continuing value of the 15 independent variables used in MPM_0-II, albeit weighted somewhat differently than in the original model. Age-interaction terms were added to MPM_0-III to reflect surprisingly favorable results in elderly patients. Variables were also added to reflect the favorable prognosis observed in elective surgical patients with no other risk factors other than age, and for the unfavorable prognosis when "Do Not Resuscitate" orders were present at the time of ICU admission. With these changes, MPM_0-III demonstrated excellent calibration (Hosmer-Lemeshow goodness of fit [HL-GOF] statistic 11.62; $p = 0.31$ where nonsignificant is desirable) and discrimination (area under the receiver operating characteristic curve = 0.823) on its internal, split-sample validation set.

A severity-adjustment model can truly be considered robust only when its performance is proven on data separate from

*See also p. 1803.

From the Critical Care Division (TLH), Departments of Medicine, Surgery and Anesthesia, Springfield, MA; Tufts University School of Medicine (TLH, DT), Boston, MA; Cerner Corporation (AAK, WC, MS), Kansas City, MO; and OptiStatim (BHN), LLC, Longmeadow, MA.

Presented in abstract form at the 20th European Society of Intensive Care Medicine Annual Congress, Berlin, Germany, October 2007.

Dr. Higgins is a consultant to Cerner Corporation, which owns the rights to the MPM_0-III predictive model. Dr. Kramer and Ms. Stark are employees of Cerner Corporation and own stock in that company. Dr. Nathanson is an employee of OptiStatim, which has a consulting contract with Cerner Corporation. Dr. Copes is a former employee of Cerner Corporation.

For information regarding this article, E-mail: thomas.higgins@bhs.org

DOI: 10.1097/CCM.0b013e31819ded31

that on which it has been developed. Although internal validation, using split-sample or bootstrapping techniques, provides some confidence that the model is not overfitted to its development set, the most rigorous form of validation is testing the model on a new, external population on data gathered at a different point in time. Unfortunately, this is infrequently accomplished because of the time and cost of data acquisition. The ongoing data collection with Project IMPACT provided an opportunity to efficiently analyze temporal performance of MPM_0-III. Our hypothesis is that MPM_0-III is robust when applied to new data, as judged by discrimination and calibration measures.

MATERIALS AND METHODS

The initial MPM-III project was reviewed by the institutional review board at Baystate Medical Center, and was determined to be exempt from the need for institutional review board approval because it met all requirements for anonymized data. We subsequently confirmed with the institutional review board that the additional analysis to complete the validation was also considered exempt.

Data Sources. Deidentified patient data were provided by Cerner/Project IMPACT. Participating units submit data at least quarterly to a central repository for all admissions, or data on a random sample of 50% or 75% of all admissions. Data collectors undergo web-based technical (software) and clinical training, are provided thorough documentation, including detailed operational definitions for each data element, and must pass a challenging certification examination before actual data collection and entry can begin. Technical, customer, and clinical support for participant questions are available each business day. User software automatically identifies ICU admissions to be randomized into the unit's sample and performs extensive checks for data accuracy, quality, and completeness that must be passed before record submission for comparative reporting. Additional data checks are performed at the central site and dialog with participants occurs when questionable data are identified. Health Insurance Portability and Accountability Act requirements are fully met. These processes are identical to what was performed for the data used in the MPM_0-III model's development (4).

Data were obtained retrospectively on 87,375 ICU admissions by 103 adult ICUs at 77 participating hospitals in the United States and Canada between July 1, 2004 and June 30, 2005. Of these admissions, 28,194 (32%) were not included because of random sampling. For the remaining 59,181 admissions, the following data were collected: information on all MPM_0-III variables (see Table 4; a complete description is given in our earlier work [4]), patient-specific characteristics (age, race, gender, reason for ICU admission, life support status at admission, and operative status), site-specific variables (bed size, region, teaching status, and ICU type), admission and discharge data (date and time of ICU admission and discharge, date of hospital admission and discharge), location before ICU admission, discharge (destination from the ICU and hospital, respectively), and hospital mortality. There were 3,724 records that did not meet MPM_0-III applicability criteria (i.e., cardiac surgery, acute myocardial infarction, burns, patients under the age of 18, and subsequent ICU readmission during a hospitalization), and these were excluded from analysis. A total of 55,459 admissions were eligible for analysis.

Analyses Performed. MPM_0-III was applied to the eligible ICU admissions to generate a mortality probability for each patient. The standardized mortality ratio was calculated by dividing the observed hospital mortality rate by the sum of the predicted hospital mortality rates, and 95% confidence intervals were calculated. Model discrimination was assessed by calculating the area under the receiver operating characteristic curve (5). This measure ranges from 0.50 (no discrimination) to 1.0 (perfect discrimination); values above 0.80 are considered good. It is interpreted as the proportion of randomly chosen pairs of patients, one of whom dies and the other survives, where the former has a larger predicted mortality than the latter. Model calibration was assessed by various methods: graphic representation of observed vs. expected mortality outcomes by deciles of risk, the HL-GOF test (6), and Brier score. The HL-GOF test summarizes differences between predicted and observed mortality and survival rates within deciles of predicted mortality. The Brier score is similar to the mean square error in regression, and values <0.10 are considered very good (7). All of these results were compared with those obtained during internal model validation of the original MPM_0-III model. SAS version 9.1 was used for all analyses.

Table 1 Characteristics of hospitals and intensive care units in the external validation and the original MPM_0-III study, respectively

	External Validation Study	MPM_0-III Study	*p*
Hospitals	n = 77	n = 102	
Region			0.762
North East	21.6%	19.6%	
South	29.7%	25.5%	
Midwest	35.1%	43.1%	
West	13.5%	11.8%	
Bed size (median, IQR)	400 (295, 580)	410 (320, 580)	0.761
ICUs	n = 103	n = 135	
Type			0.993
Medical/surgical combined	23.0%	23.6%	
Medical/surgical/coronary care combined	18.0%	22.9%	
Surgical/trauma ICU combined	12.0%	9.3%	
Other mixed medical/surgical/specialty Combined	12.0%	10.7%	
Medical ICU only	10.0%	10.0%	
Surgical ICU only	8.0%	7.9%	
Medical ICU/coronary care combined	5.0%	5.7%	
Coronary care/medical cardiac care only	4.0%	3.6%	
Other	8.0%	6.4%	
Number of beds (median, IQR)	16 (10, 20)	15 (12, 20)	0.902

MPM_0, Mortality Probability Model; ICU, intensive care unit; IQR, interquartile range.

RESULTS

Table 1 displays the characteristics of the study hospitals. The hospitals are diverse in terms of geographic distribution, organization, and size. There are no significant differences between the present cohort and the hospitals that participated in the development of MPM_0-III. Only 20% of the ICUs in the Project IMPACT group are "closed," meaning critical care consultation or management is mandated, although critical care physicians were available for discretionary consultation or management at 98% of participating units.

A total of 55,459 ICU admissions met the entry criteria. Data for these patients were supplied by 103 ICUs, 78 of which were in the development set, and 25 were new participants. Thus, 24.3% of the current study's ICUs did not participate in the MPM_0-III model's development or internal validation. The demographic characteristics of this population are given in Table 2. Mean age, hospital mortality, and the distribution of elective and emergency surgical patients were clinically similar in both time periods (Table 2),

Table 2 Comparison of demographic data for external validation group vs. the cohort used in the development of MPM_0-III

Variable	External Validation Study n = 55,459 %	MPM_0-III Study n = 124,855 %	p
Gender			<0.001
Male	55.6	54.6	
Female	44.4	45.4	
Race			<0.001
White	76.6	78.1	
Black	12.8	12.6	
Hispanic	5.0	3.5	
Asian	1.1	1.0	
Other, unknown	4.9	4.5	
Reason for ICU admission			<0.001
Treatment	57.0	55.5	
Postoperative observation	10.7	10.9	
Safety monitoring	7.9	8.6	
Cardiovascular monitoring	7.8	7.1	
Respiratory monitoring	4.6	4.8	
CNS monitoring	2.1	2.1	
Bleeding monitoring	0.8	0.7	
Other, unknown	9.1	10.2	
Life support			<0.001
Full code	95.0	95.1	
DNR	3.3	3.7	
Other, unknown	1.7	1.2	
Operative status			<0.001
Medical (nonoperative)	64.9	65.8	
Emergency surgery	12.6	10.9	
Elective surgery	22.6	23.4	
Hospital mortality			0.08
Yes	13.2	13.5	
No	86.8	86.5	
ICU mortality			0.18
Yes	8.3	8.5	
No	91.8	91.6	
% Ventilated at admission	32.1	29.7	<0.001
% Zero factors at admission[a]	10.9	11.7	<0.001
Median (IQR)			
Age	63 (48–75)	64 (49–76)	<0.001
Hospital LOS	7.0 (3.0–13.0)	8.0 (4.0–13.0)	<0.001
ICU LOS	1.9 (1.0–3.8)	1.8 (1.0–3.6)	<0.001

MPM, Mortality Probability Model; ICU, intensive care unit; CNS, central nervous system; DNR, do not resuscitate; LOS, length of stay.

[a]Elective surgical patients with no other MPM risk factors other than age.

although the extremely large size of the database causes these small differences to be statistically significant. Clinically important differences were seen in the percentage of patients on mechanical ventilation at ICU admission (more in the validation group) and in the number of patients admitted to ICU with no risk factors other than age (lower in the validation group). The prevalence of MPM risk factors (Table 3) differs slightly between groups, but not in any particular pattern.

Table 4 summarizes the results of applying MPM_0-III to new data. Observed hospital deaths (13.2%) are slightly lower than the predicted rate (13.4%), yielding a standardized mortality ratio of 0.983 (95% confidence interval 0.963–1.001). Discrimination of the model on the new dataset as measured by the area under the receiver operating characteristic curve is 0.830, comparable to the value previously obtained on the internal validation set (4). While the HL-GOF is 50.6 ($p < 0.001$), the largest difference between predicted and observed mortality rate within deciles was 1.7%. Figure 1 shows a graphic representation of model performance vs. "perfect" prediction at each decile of predicted risk. There is very close agreement between deciles of observed and predicted mortality.

A perfect forecasting model would have a Brier score of zero, whereas a "perfectly" incorrect forecast would have a score of 1 (i.e., the model would assign a mortality probability of 1 to each survivor and a score of 0 to each survivor). A model assigning a probability of 0.5 to each patient would produce a Brier score of 0.25. The Brier score for this dataset was 0.088. The Brier score on the original MPM_0-III validation set was 0.092, indicating that the predictive ability of the model is very good, and stable in this new dataset.

DISCUSSION

MPM_0-III successfully validates on patients in a succeeding cohort from the Project IMPACT database. The close association of observed and predicted mortality occurs despite some changes in the composition of Project IMPACT participants. Of note, the percentage of elective surgical patients without risk factors (what we call "zero factor" patients) has decreased from 14% to fewer than 11%. Recent clinical experience suggests that this may reflect the ongoing increase in demand for critical care beds (8), and consequent decisions to manage some low-risk elective surgical patients in stepdown or medical surgical wards rather than in a critical care unit. This supposition is further supported by increases in the percentage of patients receiving active treatment (vs. monitoring/observation) from 51.5% to 57.0%. Patients receiving mechanical ventilation at or within an hour of ICU admission were 23.8% in the original MPM_0-III study, and now comprise 32.1% of the population. MPM_0-III predicted probability of survival was 86.3% during 2001–2004, and is now 86.6%. Actual survival increased from 86.2% to 86.8%. The ratio of observed hospital mortality to mortality predicted by MPM_0-III remains within the 95% confidence limits of 1.0, indicating adequate calibration of the model to new data.

Another method of assessing calibration is the HL-GOF test, which should be nonsignificant. In our analysis, the HL-GOF was highly significant ($p < 0.001$), which suggests a calibration problem. However, caution must be exercised in interpreting the HL-GOF when very large sample sizes (>10,000 patients) are analyzed (9). This is due to the sensitivity of the HL-GOF inference test being affected by increasing sample sizes. The calibration graph shown here and the small within-decile differences between observed and predicted mortality indicate that the model still calibrates well. Fur-

Table 3 MPM_0-III risk factors in external validation group vs. the cohort used in the development of MPM_0-III

MPM_0-III Risk Factors (%)	External Validation Study n = 55,459	MPM_0-III Study n = 124,855	*p*
Intracranial mass	4.39	4.45	0.57
Cerebrovascular event	3.96	4.67	<0.001
Gastrointestinal bleed	4.79	5.29	<0.001
Cardiac dysrhythmia	5.92	6.44	<0.001
Acute renal failure	5.95	5.54	<0.001
Coma or deep stupor (GCS 3 or 4)	5.28	6.13	<0.001
Tachycardia (HR >150)	2.26	2.39	0.09
Hypotension (SBP <90)	18.09	15.53	<0.001
CPR before admission	3.03	3.26	0.01
Cirrhosis	2.35	3.07	<0.001
Metastatic Neoplasm	3.47	4.74	<0.001
MV within 60 min of admission	32.10	26.64	<0.001
Chronic renal insufficiency	6.01	6.84	<0.001
Patient type	77.42	76.57	<0.001
Zero factors	10.91	11.73	<0.001
Full code status	95.01	95.14	0.25

MPM, Mortality Probability Model; GCS, Glasgow Coma Scale; HR, heart rate; CPR, cardiopulmonary resuscitation; MV, mechanical ventilation.

Table 4 External validation results

Statistic	Results
Observed deaths	7331 (13.2%)
Predicted deaths	7456 (13.4%)
Standardized mortality rate	0.983 ($p = 0.06$) 95% CI (0.963, 1.001)
Discrimination	AUROC = 0.830
Calibration	HL-GOF statistic = 50.6 ($p < 0.001$); biggest decile difference is 1.7%
Brier score	0.088

AUROC, area under the receiver operating characteristic curve; HL-GOF, Hosmer-Lemeshow goodness of fit statistic; CI, confidence interval.

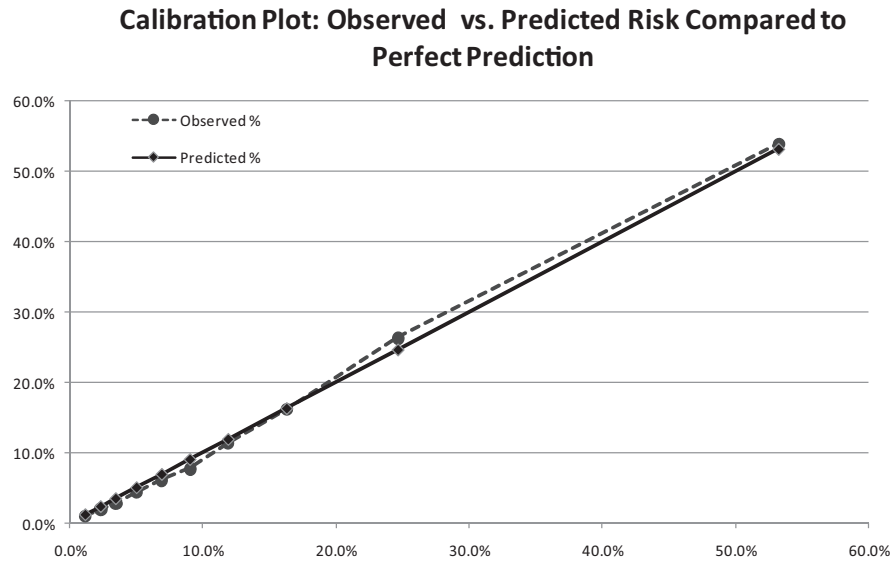

Figure 1. Calibration plot of Mortality Probability $Model_0$-III on the external validation sample.

thermore, the Brier scores, both in this study and in the internal, split sample validation set of the MPM_0-III study, are virtually the same and close to zero. This shows that the model does have good predictive ability and its performance has not degraded as the population in Project IMPACT changed. Also of importance is that the model's discrimination remained high.

Based on results from this analysis, MPM_0-III performs well on a follow-up cohort of patients, suggesting that its original success was not simply because of the idiosyncrasies of its development data set. Indeed, almost a quarter (24.3%) of all ICUs present in this external validation study are "new."Although the original validation cohort and the one studied here contain many of the same hospitals, some patient characteristics have changed over time. This is most clearly manifested by the decrease in mortality, despite the increase in percentage of patients on a mechanical ventilator and a decrease in the percentage of patients lacking risk factors besides age.

It remains to be determined if the model performs well on ICUs not in the Project IMPACT database and this is a limitation of the study. Project IMPACT is composed primarily of US hospitals and inner city/urban hospitals as well as academic centers and transplant/quaternary ICUs are underrepresented. Some studies have shown that model accuracy tends to erode when a model is applied to external populations (10–13), although this has largely meant applying US models to European ICUs. Thus, validation of the MPM_0-III model in other populations is warranted.

There are other limitations to this study that are identical to those in the original MPM_0-III study: MPM excludes certain patient subsets (e.g., cardiac surgery, acute myocardial infarction, and ICU readmissions), which reduces its usefulness to some ICU. MPM may underestimate severity of illness in patients whose condition is rapidly changing as they are admitted, because physiologic abnormalities, such as tachycardia and hypotension, ideally should be corrected as soon as they are identified, regardless of location.

The nonsignificant trend toward lower hospital mortality in the current data set vs. the original data used to develop MPM_0-III will need to be followed up in a new cohort of patients. At this stage, random chance alone could produce these

results, but our observations based on applying the MPM-II model to current patient populations do suggest improvements in medical care as one possible explanation of the failure for old models to perform well on new patient populations. Other explanations for changes in model calibration over time include increasing use of rapid response teams, early goal-directed therapy, rapid transfer to long-term acute care hospitals (where patients may expire without contributing to an acute care hospital's mortality rate), and population differences between the original multinational MPM-II study group (European and North American) and the current Project IMPACT (primarily North American) participants. However, model deterioration over time has also been observed for the Simplified Acute Physiology Score (14) and the Acute Physiology and Chronic Health Evaluation (15), and all three models have required recalibration. Our data cannot easily answer the question whether capacity constraints in ICU beds has reduced the ability of ICUs to admit lower-risk patients, not requiring ICU-level interventions, such as mechanical ventilation.

CONCLUSIONS

MPM_0-III calibrates and discriminates well on a new population of 55,459 patients, suggesting that the relationships between 2004 and 2005 clinical risk factors and hospital mortality have not changed from that seen with 2001–2004 Project IMPACT data. The standardized mortality ratio for 2004–2005 patients is 0.983; not significantly different from 1.0. Graphic calibration suggests the model is robust and not overfitted to the developmental sample. A low Brier score suggests that predictive ability of the model continues to be very good. MPM_0-III can, thus, provide an accurate assessment of mortality risk for large groups of patients, using just 17 variables obtained at or within 1 hour of ICU admission, and without needing to specify an admission diagnosis. Future studies should determine whether the standardized mortality ratio continues to trend downward. This would suggest continuing improvement in outcomes if the trend became significant in a consistent group of hospitals. Finally, this study identified a trend toward Project IMPACT units admitting fewer low-risk patients and more patients requiring mechanical ventilation, but additional studies are required to see if this trend remains valid in larger, more contemporary databases.

REFERENCES

1. Project IMPACT. Available at: www.cerner.com/piccm. Accessed April 3, 2008
2. Lemeshow S, Teres D, Klar J, et al: Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; 270:2478–2486
3. 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
4. Higgins TL, Teres D, Copes WS, et al: Assessing contemporary intensive care unit outcome: An updated mortality probability admission model (MPM_0-III). *Crit Care Med* 2007; 35:827–835
5. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36
6. Hosmer DW, Lemeshow S: Applied Logistic Regression. Second Edition. New York, John Wiley & Sons, 2000
7. Brier GW: Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950; 78:1–3
8. Halpern NA, Pastores SM, Thaler HT, et al: Changes in critical care beds and occupancy in the United States 1985–2000: Differences attributable to hospital size. *Crit Care Med* 2006; 34:2105–2112
9. 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
10. Moreno R, Morais P: Outcome prediction in intensive care: Results of a prospective, multicenter Portuguese study. *Intensive Care Med* 1997; 23:177–186
11. Rowan KM, Kerr JH, Major E, et al: Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: A prospective, multicenter, cohort study comparing two methods for predicting outcome in adult intensive care patients. *Crit Care Med* 1994; 22:1392–1401
12. Livingston B, MacKirdy FN, Howie JC, et al: Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med* 2000; 28: 1820–1827
13. Markgraf R, Deutschinoff G, Pientka L, et al: Comparison of Acute physiology and Chronic Health Evaluation II and III and Simplified Acute Physiology Score II: A prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. *Crit Care Med* 2000; 28: 26–33
14. Metnitz PG, Moreno RP, Almeida E, et al: SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part I: Objectives, methods and cohort description. *Intensive Care Med* 2005; 31:1336–1344
15. 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