Intensive Care Med (1999) 25: 977–982


ORIGINAL

**P. A. Patel**
**B. J. B. Grant**

# Application of mortality prediction systems to individual intensive care units

Received: 21 December 1998
Final revision received: 26 May 1999
Accepted: 1 June 1999

P. A. Patel · B. J. B. Grant (✉)
Division of Pulmonary and Critical Care Medicine, Veterans Affairs Health Care System of Western New York,
State University of New York at Buffalo, New York, USA

*Mailing address:*
VAMC (111-S), 3495 Bailey Avenue,
Buffalo, NY 14215, USA
Fax: + 1 (716) 862 4729
email: grant@buffalo.edu

**Abstract** *Objective:* To evaluate the predictive accuracy of the severity of illness scoring systems in a single institution.
*Design:* A prospective study conducted by collecting data on consecutive patients admitted to the medical intensive care unit over 20 months. Surgical and coronary care admissions were excluded.
*Setting:* Veterans Affairs Medical Center at Buffalo, New York.
*Patients and participants:* Data collected on 302 unique, consecutive patients admitted to the medical intensive care unit.
*Interventions:* None.
*Measurements and results:* Data required to calculate the patients' predicted mortality by the Mortality Probability Model (MPM) II, Acute Physiology and Chronic Health Evaluation (APACHE) II and Simplified Acute Physiology Score (SAPS) II scoring systems were collected. The probability of mortality for the cohort of patients was analyzed using confidence interval analyses, receiver operator characteristic (ROC) curves, two by two contingency tables and the Lemeshow–Hosmer chi-square statistic. Predicted mortality for all three scoring systems lay within the 95 % confidence interval for actual mortality. For the MPM II, SAPS II and APACHE II, the c-index (equivalent to the area under the ROC curve) was 0.695 ± 0.0307 SE, 0.702 ± 0.063 SE and 0.672 ± 0.0306 SE, respectively, which were not statistically different from each other but were lower than values obtained in previous studies.
*Conclusion:* Although the overall mortality was consistent with the predicted mortality, the poor fit of the data to the model impairs the validity of the result. The observed outcoume could be due to erratic quality of care, or differences between the study population and the patient population in the original studies. The data cannot be used to distinguish between these possibilities. To increase predictive accuracy when studying individual intensive care units and enhance quality of care assessments it may be necessary to adapt the model to the patient population.

**Key words** Predictive scoring systems · Mortality rates · Intensive care units · Quality assurance · Quality improvement · Veterans

## Introduction

The Acute Physiology and Chronic Health Evaluation (APACHE II), Mortality Probability Model (MPM II) and the Simplified Acute Physiology Score (SAPS II) scoring systems were developed based on data acquired from a large number of hospitals with diverse patient populations. The APACHE II model was developed

using a theoretical approach: physiologic variables were selected and weights assigned to these variables based on clinical judgment and documented physiologic relations [1]. The MPM II model was developed using an empirical approach that involved collecting information on groups of patients and then contrasting the physiologic patterns of survivors and non-survivors [2]. Developers of the SAPS II model used multiple regression analysis to assist in the selection of variables that would constitute the SAPS II scoring system, to identify appropriate groupings and point assignments, and convert the scores to a probability of hospital mortality [3].

These scoring systems have been evaluated in several large Canadian and European studies which have confirmed their predictive accuracy in those settings [4–6]. The area under the receiver operator characteristic (ROC) curve for the three models, in these studies, ranged from 0.74 to 0.86 [4–6]. In general, these studies have involved over 1500 patients each and have been multicenter studies, which has ensured a heterogenous patient population. There have been only a few studies limited to single hospitals and involving smaller patient populations to evaluate predictive accuracy in these situations, and most of these have been outside the United States. A review of the literature revealed only one published study evaluating the APACHE II scoring system in the Veterans Affairs (VA) patient population. Mortality in that study was higher than predicted by the APACHE II model, but no measure of predictive accuracy was given [7]. To our knowledge no published studies have been undertaken yet to validate SAPS II or MPM II in the VA population.

Since these scoring systems were developed using a large patient population, evaluation of predictive accuracy in a single hospital becomes especially important if the hospital has an unusual patient population. Previous studies have shown that the MPM II model and the APACHE II model do not adjust for unusual patient populations [8, 9]. Murphy-Filkins and colleagues demonstrated that increasing the frequency of patients with a particular condition causes the discrimination and calibration of the MPM II model to deteriorate [8]. Goldhill and Withington [9] showed that mortality ratios (number of observed deaths divided by number of predicted deaths) varied widely in different subgroups of patients whose predicted mortality was calculated using the APACHE II system. The general impression of the patient base in the VA system is that of elderly male patients with multisystem pathology and hence less physiologic reserve (Table 1). Therefore these scoring systems need to be evaluated in terms of predictive accuracy within a single institution before applying them to make quality of care assessments. This study illustrates the caveats of interpreting the results of severity of illness scoring systems that can occur when they are applied to a single institution.

## Materials and methods

This study was conducted at the Veterans Affairs Medical Center at Buffalo. Consecutive patients admitted to the medical intensive care unit (ICU) of 11 beds between 1 January 1996 and 31 August 1997 were studied prospectively. Patients were followed until discharge from hospital. Patients still in the hospital as of October 1997 were dropped from the study. Surgical and coronary care patients were excluded in this study. Data collection was done by four registered nurses. For each patient, variables required to calculate the APACHE II, MPM II and SAPS II scores were collected and entered into a computer program designed to provide an estimate of mortality for intensive care patients based on the three severity of illness scoring systems [1–3].

The APACHE II mortality estimate was obtained by measuring 12 physiologic variables (heart rate, systolic blood pressure, temperature, oxygenation, respiratory rate, arterial pH, serum sodium, potassium and creatinine, hematocrit, white blood cell count and Glasgow Coma Score) as well as the patient's age and chronic health status [1]. The MPM II (24-h) was estimated from 13 variables obtained at the 24-h time point: age, malignancy, intracranial mass effect, cirrhosis, coma or deep stupor at 24 h, mechanical ventilation, intravenous vasoactive drug therapy, hospital admission not for elective surgery, confirmed infection, urine ouput less than 150 ml in an 8-h period, serum creatinine concentration greater than 2 mg/dl, prothrombin time greater than 3 s above standard, partial pressure of oxygen less than 60 mmHg [2]. The SAPS II probability of mortality was derived from 11 physiologic measurements: heart rate, systolic blood pressure, oxygenation, temperature, white blood cell count, serum sodium, potassium and bilirubin, blood urea nitrogen level, urinary output and Glasgow Coma Score. The patient's age, presence of malignancy and/or acquired immunodeficiency syndrome and type of hospital admission were included in addition to the physiologic measurements [3].

The predictive ability of the three scoring systems was evaluated by four methods: (i) confidence interval (CI) analysis, (ii) the c-index, which is equivalent to the area under the ROC curve, (iii) two by two decision tables and (iv) goodness-of-fit assessed by the Lemeshow–Hosmer chi-square statistic.

Overall predictive accuracy for the entire cohort of patients was evaluated by comparing the predicted mortality by each scoring system to the observed mortality using CI analysis. A ROC curve was constructed for each scoring system from the patients' predicted outcome and observed outcomes. A plot of the sensitivity against the false-positive rate at several decision thresholds (cutoffs) yielded the ROC curve. The sensitivity is the proportion of patients who died that was predicted correctly by the model. The false-positive rate is the proportion of patients who survived who were predicted incorrectly to die. Diagnostic accuracy was assessed by the c-index, which is equivalent to the area under the ROC curve. The c-index and its standard error (SE) were calculated by the bootstrap method. The c-index is a measure of the overall discriminatory power of the prognostic model in distinguishing those who died from those who lived [10]. A value of 0.5 indicates random chance, while a value of 1.0 indicates perfect prediction. The actual and predicted outcomes were compared using two by two decision matrices at several cutoffs. The sensitivity, specificity and accuracy or correct classification rate (the sum of true positives and true negatives divided by the total number of patients) were calculated. For example, with a decision criterion of 0.5, predicted risk of death greater than 50 % is considered to predict hospital death, while a predicted risk of death less than 50 % is considered to predict survival. These predicted outcomes were then compared with observed outcomes to calculate the sensitivity, specificity and accuracy. Goodness-of-fit of the models was determined by

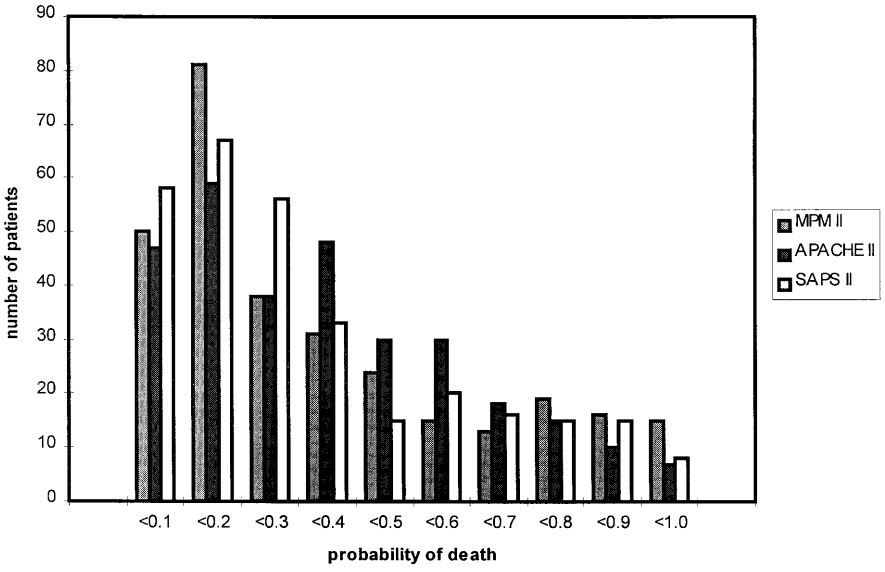


**Fig. 1** Distribution of patient's predicted risk of hospital death estimated by Mortality Probability Model II *MPM II,* Acute Physiology and Chronic Health Evaluation *APACHE II* and Simplified Acute Physiology Score *SAPS II*

calculating the Lemeshow–Hosmer chi-square statistic, which is used to assess calibration of the models and the strength of the association between the predicted and observed outcome over the entire range of probabilities [11]. Calibration defines the ability of the model to describe the mortality pattern in the data and indicates the accuracy of risk prediction by the model. The observed death rates for each scoring system were plotted against the predicted death rates stratified by 10% risk ranges to obtain the calibration curve. A $p < 0.05$ was considered statistically significant.

## Results

The study was conducted prospectively by collecting data on 302 consecutive patients admitted to the medical ICU between 1 January 1996 and 31 August 1997. Altogether, 107 (35.4%) patients died and 195 (64.6%) survived. The distribution of predicted probability of hospital death was skewed toward the lower probabilities of death for all three models (Fig. 1). The median age of patients in the study was 70.5 years (range 30–99 years), with a preponderance of males (97%) (Table 1).

For the entire cohort of patients, the mean MPM II (0.347), APACHE II (0.344) and SAPS II (0.3219) predicted mortality were within the 95% CI (0.298 to 0.406) for the actual overall mortality of 0.354. Both the MPM II and APACHE II predicted mortality were almost identical with the observed mortality, while the SAPS II predicted mortality was lower than the observed mortality.

ROC curves were drawn for the three scoring systems to assess predictive accuracy (Fig. 2). The c-index was found to be $0.702 \pm 0.063$ (SEM) for APACHE II, $0.695 \pm 0.0307$ (SEM) for MPM II and $0.672 \pm 0.0306$ (SEM) for SAPS II. The differences between them were not statistically significant.

**Table 1** Baseline characteristics of sample population

| | |
|---|---|
| Male: female | 293 : 9 |
| Median age (years) | 70.5 (range 30–99) |
| Disease category | Frequency ($n$ = 302) |
| Cardiovascular | 69 |
| Gastrointestinal | 61 |
| Infectious | 23 |
| Metabolic | 29 |
| Neoplastic | 5 |
| Neurologic | 20 |
| Respiratory | 83 |
| Surgical | 2 |
| Toxic | 10 |

**Table 2** Comparing the sensitivity *Se,* specificity *Sp* and accuracy *Ac* of the MPM II, APACHE II and SAPS II models at decision thresholds of 0.3, 0.4, 0.5 and 0.6, values are percent

| Cutoff | MPM II | | | APACHE II | | | SAPS II | | |
|---|---|---|---|---|---|---|---|---|---|
| | Se | Sp | Ac | Se | Sp | Ac | Se | Sp | Ac |
| 0.3 | 61 | 65 | 64 | 70 | 57 | 62 | 53 | 67 | 62 |
| 0.4 | 48 | 74 | 65 | 55 | 73 | 67 | 45 | 80 | 67 |
| 0.5 | 36 | 79 | 64 | 40 | 81 | 67 | 39 | 84 | 68 |
| 0.6 | 29 | 84 | 64 | 21 | 86 | 63 | 27 | 87 | 66 |

Two by two decision matrices were constructed for the three scoring systems at cutoff points for predicted probability of death of 0.3, 0.4, 0.5 and 0.6 (Table 2). Sensitivity, specificity (the proportion of true negatives) and accuracy were calculated (Table 2). MPM II was most accurate (65%) at a cutoff of 0.4, sensitivity and

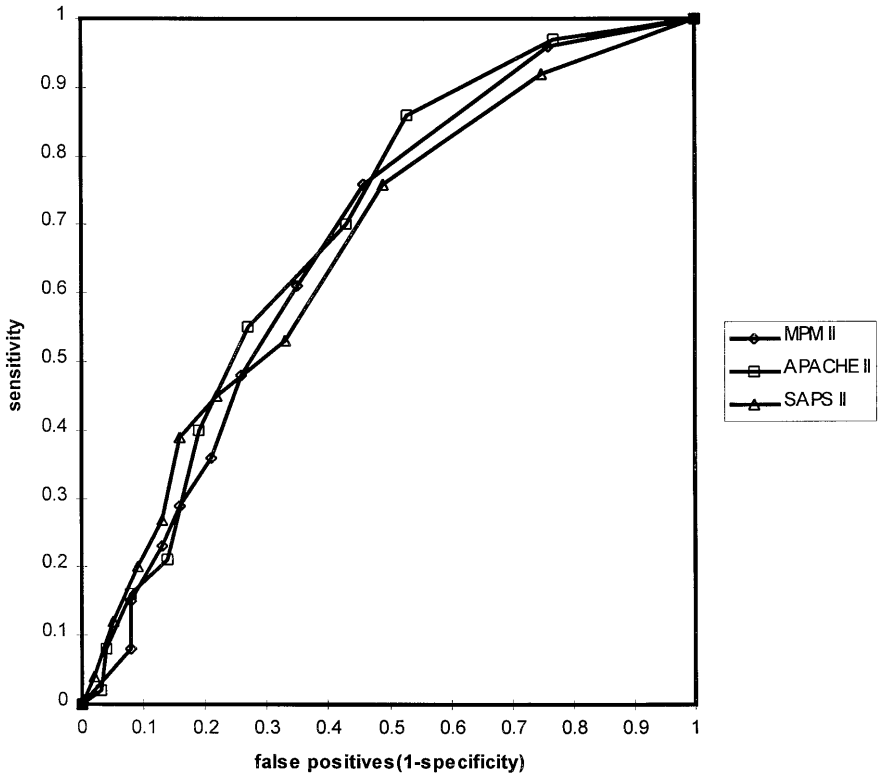


**Fig. 2** Comparison of the areas under the receiver operator characteristic curve demonstrating predictive ability of the MPM II, APACHE II and SAPS II. The areas under the curves are, respectively, 0.6953, 0.7016 and 0.6721. For any decision criterion, the sensitivity is the percentage of patients correctly predicted to die among those who actually died. The false-positive rate is the percentage of patients predicted to die who actually survived

specificity at this cutoff were 48 and 74%, respectively. The APACHE II system achieved a maximal accuracy of 67%, and this was obtained at thresholds of both 0.4 and 0.5. The sensitivity and specificity at 0.4 was 55 and 73%, respectively, and at 0.5 was 40 and 81%, respectively. For the SAPS II system, the best accuracy obtained was 68% at a threshold of 0.5. At this threshold the sensitivity was 39% and the specificity was 84%.

Calibration appeared only modest (Fig. 3). In all three models as the predicted risk of death increased, the proportion of patients who died increased. However, for all three models, the observed mortality was higher than predicted when the predicted probability of death was less than 0.4. The observed mortality was lower than predicted when the predicted probability of death was greater than 0.6 (Fig. 3). The calculated Lemeshow–Hosmer chi-square statistics were: 14.33 ($p = 0.073$) for APACHE II, 20.70 ($p < 0.05$) for MPM II and 22.58 ($p < 0.05$) for SAPS II. Based on the Lemeshow–Hosmer chi-square statistic, the APACHE II model demonstrates marginally better calibration to the data than the MPM II and SAPS II systems.

## Discussion

### Comparison of our results to others

Three models for predicting outcome in ICU patients have been evaluated in this study. All three models were developed from large heterogeneous cohorts of medical and surgical patients and it was important to evaluate their predictive accuracy in a smaller setting with a different disease spectrum before applying them to make quality of care assessments.

Overall predicted mortality by all three models was similar to the observed mortality. The MPM II and APACHE II models displayed almost identical mean predicted mortality and were closer to the observed mortality than the SAPS II model. However, these results were different from those obtained in the study conducted at the University of California at San Francisco VA Medical Center [7]. In that study, the actual mortality in a population of mechanically ventilated patients was higher than that predicted by the APACHE II model [7]. Perhaps this discrepancy can be explained by the fact that all the cases in that study were on mechanical ventilation and the APACHE II system does not weigh mechanical ventilation as an adverse outcome predictor.

There is general agreement that these models cannot be utilized to make recommendations for specific pa-

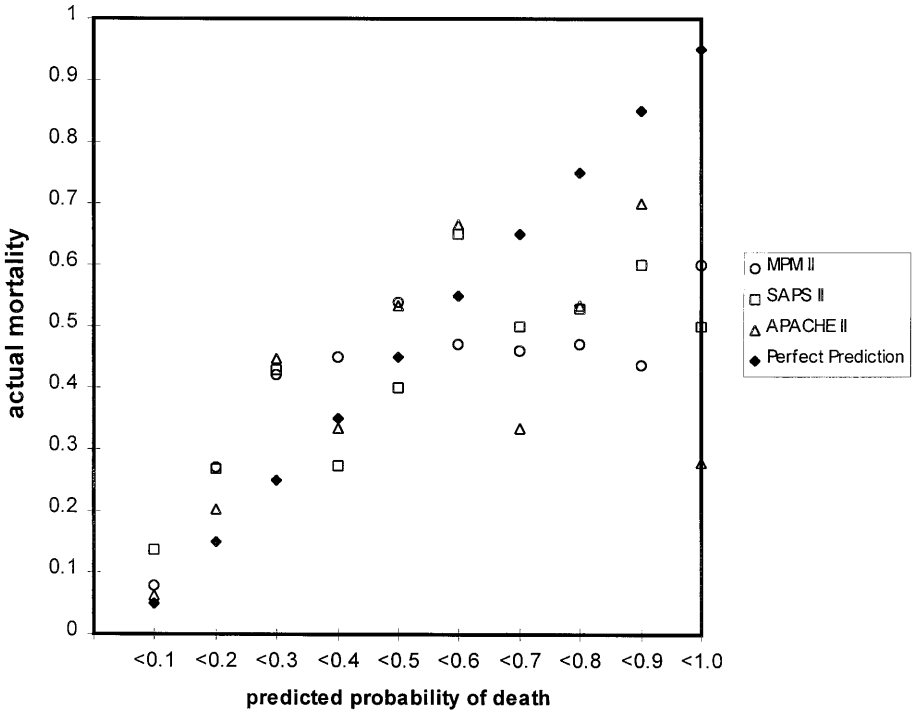


**Fig. 3** Comparison of the calibration curves for MPM II, APACHE II and SAPS II. The line of perfect prediction is where the number of actual and predicted deaths is equal

tients because of their relatively high false-positive rate. In our study, the SAPS II model proved the most accurate, correctly classifying 68 % of the patients at a cutoff of 0.5. Even with this model, the false-positive rate at a cutoff of 0.6 was 13 %, which indicates that it lacks specificity in predicting individual patient outcome.

Calibration as assessed by the Lemeshow–Hosmer chi-square statistic was only modest for the MPM II and SAPS II systems and marginally better with the APACHE II model. Nevertheless, at lower predicted probabilities of death, the observed mortality was higher than predicted mortality, while at higher predicted probabilities of death observed mortality was lower than predicted mortality. These results are consistent with those in large studies conducted in the United Kingdom and Japan [6, 12].

Discrimination in risk stratifying patients as assessed by the ROC curve was only moderate. The differences between the c-indices for the three models were not statistically significant. Our findings showed that the c-index was significantly lower for all three systems than the equivalent measures of predictive accuracy obtained in larger studies in Canada, Spain, United Kingdom, Italy and Portugal [4–6, 13, 14]. An explanation for this discrepancy may be that those studies were performed in a large (over 1500 cases) population with a wide spectrum of disease. It has been demonstrated previously that the discrimination and calibration of a mortality model will deteriorate when applied to a subgroup of a cohort of patients on which the model had shown good calibration and discrimination [9]. Similarly, an increase in the frequency of patients with a particular condition beyond a critical percentage will cause a decrease in the predictive accuracy of the model [8].

In smaller settings with a homogeneous or unusual patient population, the importance of certain physiologic variables in predicting mortality may diminish and other factors may become important predictors of mortality. As a result, these factors may be weighted inadequately in a scoring system developed on a large patient population. In a single ICU, with a unique spectrum of disease or increased prevalence of a particular condition, it may become necessary to customize the variables to be utilized in calculating the predicted mortality. It appears that a scoring system based on a testing and validation set from one population when transferred to another population without modification will often lose predictive accuracy.

There are problems with using scoring systems for evaluation of a particular ICU. The predictive accuracy of mortality models is generally assessed by determining the area under the ROC curve or by calculating the Lemeshow–Hosmer chi-square statistic. The calculation of both these statistics depends on comparing the expected to observed mortality. In a given ICU, if the model appears to have poor discrimination and calibration there are two possibilities: (i) the quality of care is better or worse than expected and more patients than expected survived/died, or (ii) the model's applicability to a given patient population is poor due to the unusual

nature of the patient population. In the first scenario, the scoring system would lose its predictive accuracy as better care would assist more patients than excepted to survive, and worse care would allow patients who should have survived to die. In the second scenario, the accuracy of the prediction instrument would be reduced because of limited applicability to the given population. From our results it was not possible to determine which of the two factors played a role in decreasing the predictive accuracy of the severity of illness scoring systems.

Moreover, even if initially the model discriminates well, it is possible that following an improvement in quality of care (or a deterioration in quality of care for that matter) calibration and discrimination would deteriorate, reducing applicability of the severity of illness scoring system to the situation. These limitations may be overcome by recalibrating the model frequently to take into account changes in quality of care and improved survival.

In conclusion, there was no significant difference between the three models in predicting group and individual outcome. Based on our results and the results of others, all three models appear to lack the specificity and the discrimination required to predict survival in individual patient outcome in our setting. To improve the predictive accuracy of these models in an individual ICU such as ours, it may be necessary to customize the models, or perhaps to utilize scoring systems specific for particular disease conditions to estimate mortality. Because discrimination of the models is dependent on both the nature of the population being evaluated and the quality of care being rendered and since it is not possible to distinguish between these two factors, using the models to assess or compare quality of care may be limited.

**Acknowledgements** We thank Sandra Brucato, Ray Carter, Susan Schaffer and Sally Shimmell for their assistance with data collection.

## References

1. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13: 818–829
2. Lemeshow S, Teres D, Klar J et al (1993) Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270: 2478–2486
3. Le Gall JR, Lemeshow S, Saulnier F (1993) A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 270: 2957–2963
4. Wong DT, Crofts SL, Gomez M et al (1995) Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. Crit Care Med 23: 1177–1183
5. Castella X, Artigas A, Bion J, Karei A (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. Crit Care Med 23: 1327–1335
6. Rowan KM, Kerr JH, Major E et al (1994) 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 for adult intensive care patients. Crit Care Med 22: 1392–1401
7. Papadakis MA, Browner WS (1987) Prognosis of noncardiac medical patients receiving mechanical ventilation in a veterans hospital. Am J Med 83: 687–692
8. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer DW (1996) Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. Crit Care Med 24: 1968–1973
9. Goldhill DR, Withington PS (1996) The effect of casemix adjustment on mortality as predicted by APACHE II. Intensive Care Med 22: 415–419
10. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver-operating characteristic (ROC) curve. Radiology 143: 29–36
11. Lemeshow S, Hosmer DW (1982) A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 115: 92–106
12. Sirio CA, Tajimi K, Knaus WA et al (1992) An initial comparison of intensive care in Japan and the United States. Crit Care Med 20: 1207–1215
13. Apolone G, Bertolinin G, D'Amico R et al (1996) The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Intensive Care Med 22: 1368–1378
14. Moreno R, Morias P (1997) Outcome prediction in intensive care: results of a prospective, multicenter, Portuguese study. Intensive Care Med 23: 177–186