Journal of Critical Care (2012) **27**, 423.e1–423.e7

ELSEVIER

Journal of Critical Care

Caution when using prognostic models: A prospective comparison of 3 recent prognostic models☆,☆☆

Antonio Paulo Nassar Jr*, Amilcar Oshiro Mocelin, André Luiz Baptiston Nunes, Fabio Poianas Giannini, Leonardo Brauer, Fabio Moreira Andrade, Carlos Augusto Dias

Adult Intensive Care Units—Hospital e Maternidade São Camilo—São Paulo, Brazil

Keywords:
Prognostic;
Model;
Score;
Critical care;
Benchmarking

Abstract

Purpose: Prognostic models have been developed to estimate mortality and to compare outcomes in different intensive care units. However, these models need to be validated before their use in different populations. In this study, we assessed the performance of 3 recently developed general prognostic models (Acute Physiologic and Chronic Health Evaluation [APACHE] IV, Simplified Acute Physiology Score [SAPS] 3 and Mortality Probability Model III [MPM_0-III]) in a population admitted at 3 medical-surgical Brazilian intensive care units.

Materials and Methods: All patients admitted from July 2008 to December 2009 were evaluated for inclusion in the study. Standardized mortality ratios were calculated for all models. Calibration was assessed by the Hosmer-Lemeshow goodness-of-fit test. Discrimination was evaluated using the area under the receiver operator curve.

Results: A total of 5780 patients were included. Inhospital mortality was 9.1%. Discrimination was very good for all models (area under the receiver operator curve for APACHE IV, SAPS 3 and MPM_0-III was 0.883, 0.855 and 0.840, respectively). APACHE IV showed better discrimination than SAPS 3 and MPM_0-III ($P < .001$ for both comparisons). All models calibrated poorly and overestimated hospital mortality (Hosmer–Lemeshow statistic was 53.7, 134.2, 226.6 for APACHE IV, MPM_0-III, and SAPS 3, respectively; $P < .001$ for all).

Conclusions: In this study, all models showed poor calibration, while discrimination was very good for all of them. As this has been a common finding in validation studies, caution is warranted when using prognostic models for benchmarking.

1. Introduction

Prognosis in critically ill patients varies according to diagnosis, previous health status, physiologic alterations, and care provided. Over the past 3 decades, prognostic scores have been developed to predict hospital mortality taking these variables into consideration [1]. The 3 most widely used models are Acute Physiologic and Chronic Health

☆ The authors do not have conflict of interest to declare.

☆☆ This study was presented as an abstract at the 23rd European Society of Intensive Care Medicine Congress in Barcelona, Spain, October 9-13, 2010.

* Corresponding author. Rua Tavares Bastos, 715, Zip Code 05012-020, São Paulo-SP, Brazil.

E-mail address: paulo_nassar@yahoo.com.br (A.P. Nassar).

0883-9441/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcrc.2011.08.016

Evaluation (APACHE) [2], Simplified Acute Physiology Score (SAPS) [3] and Mortality Probability Model (MPM) [4]. Their first versions were developed in the 1980s. Some years later, they were updated based on larger populations and case mix.

Theoretically, the comparison between observed and predicted mortality rates could serve as an indicator of intensive care unit (ICU) performance, facilitate resource allocation and lead to overall improvement in healthcare services. However, ICU profiles vary greatly worldwide, depending on the proportion of medical and surgical patients, admission and discharge policies, availability of intermediate care units, and staffing with intensive care specialists. Therefore, any statement about a given ICU performance based on severity scores requires validation of the score at this new ICU, a process called external validation [5]. Assessment of prognostic models is accomplished evaluating calibration and discrimination of the models. Calibration assesses the degree of correspondence between the estimated probability of hospital mortality and that actually observed. Discrimination assesses the ability of a model to distinguish patients who died from those who survived [1].

All 3 models have been recently updated to their newest versions (APACHE IV [6], SAPS 3 [7,8] and MPM_0-III [9]), and there are some important differences between them. While SAPS 3 and MPM_0-III only use data collected in the first hour after admission, APACHE IV uses data obtained within 24 hours of admission. APACHE IV was developed based on data from hospitals in the United States. The variables used in MPM_0-III were mostly derived from hospitals in the United States but were also obtained in Canada and Brazil. SAPS 3 was developed in a heterogeneous population from various countries in Europe, North, South and Central Americas and Australasia, which may have contributed to improve its external validity. Some studies have validated these new versions in other populations [10-19], but to the best of our knowledge, they have never been compared with each other before. The objective of this study is to assess external validity of the 3 scores and to compare their performance in a Brazilian population.

2. Methods

2.1. Design and setting

This prospective cohort study was conducted between July 1, 2008, and December 31, 2009, in 3 Brazilian ICUs. All ICUs are medical-surgical and staffed with full-time intensive care specialists, nurses, and physical therapists. During the study period, 2 ICUs had 31 beds and 1 had 24. None of them has an explicit ICU admission policy, that is, patients may come from the emergency department, the operating room, the wards, or the hemodynamic laboratory or transferred from other hospitals at the discretion of the attending physician. Decisions to discharge patients are taken by the intensive care specialist and the physician in charge of the patient. During the study period, none of the hospitals had intermediate care units, and all patients were discharged to wards.

2.2. Selection of participants

All consecutively admitted patients 18 years or older were included. Patients who were readmitted, transferred to another hospital (during ICU stay or after ICU discharge if still hospitalized), or admitted for acute coronary syndromes were excluded. We decided to exclude patients admitted for acute coronary syndromes because these patients were not included in the MPM_0-III original cohort [9]. Patients with incomplete data which prevented an adequate calculation of one or more of the scores were also excluded. These missing data could be: pre-ICU length of stay, reason for ICU admission, chronic health variables and mechanical ventilation on first day. Missing physiologic variables, namely bilirubin, acid-base abnormalities, PaO_2 or PaO_2/FiO_2 ratio, were considered as normal for purpose of calculations. The study was approved by the local ethics committees and informed consent was waived because no intervention was required and no individual data were expected to be disclosed.

2.3. Data collection

Data were manually collected according to the general rules and definitions for the 3 models [6,8,9] using a specific form to be completed by the intensive care specialist on duty at the moment of the patient's admission. It included demographics (age, sex), type of admission (medical or surgical), patient origin (emergency room, surgery room, ward or other [transferred from another hospital or hemodynamic laboratory]), reason for admission (defined as sepsis, cardiovascular, neurological, respiratory, gastrointestinal, genitourinary, trauma, orthopedic, metabolic, hematologic and/or other), and variables for the 3 models. As SAPS 3 and MPM_0-III variables were collected in the first hour after admission, they were entered on the form by the intensive care specialist who admitted the patient. The APACHE IV diagnosis was determined and completed at admission. All other APACHE IV variables were collected within 24 hours of admission from medical records by a nursing student trained in severity scores or by the local ICU medical coordinator. All data were entered on a Microsoft Excel spreadsheet used to estimate mortality risks. For SAPS 3, the global equation was utilized.

2.4. Statistical analysis

Statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS) version 10.0 and the Medical Calculator (MedCalc) version 9.0. Continuous

Table 1 Baseline patient characteristics and main outcomes

Characteristics	
Age (years; median [IQR])	66 (47-79)
Female (n [%])	3046 (52.7)
Location before ICU admission (n [%])	
Emergency room	3562 (63.2)
Operating room	1206 (20.9)
Ward	636 (11.0)
Other	286 (4.9)
Reason for admission (n [%])	
Cardiovascular	1170 (20.2)
Neurologic	1126 (19.6)
Sepsis	920 (15.9)
Respiratory	690 (12.0)
Gastrointestinal	626 (10.8)
Orthopedic	363 (6.3)
Metabolic	319 (5.5)
Genitourinary	299 (5.1)
Trauma	185 (3.2)
Other	82 (1.4)
ICU LOS (days; median [IQR])	2.51 (1.55-4.26)
In-hospital mortality (n [%])	528 (9.1)

variables are presented as median and interquartile ranges (IQR). Categorical variables are presented as absolute values and percentages. The prognostic performance of the different scores was evaluated in terms of calibration and discrimination. Calibration was assessed with the Hosmer-Lemeshow goodness-of-fit test C-statistic, which evaluates the agreement between the observed and expected numbers of survivors and nonsurvivors across all of the strata of probabilities of death. A high P value ($P > .05$) indicates a good fit for the model [20]. Calibration curves were also constructed by plotting predicted mortality rates stratified by 10% intervals of mortality risk (x-axis) against observed mortality rates (y-axis). The score discrimination was assessed by calculating the area under the receiver operating characteristic curve (AUROC) and its 95% confidence interval (CI). Discrimination was considered excellent, very good, good, moderate and poor with AUROC values of 0.9-0.99, 0.8-0.89, 0.7-0.79, 0.6-0.69 and <0.6, respectively. Pairwise comparisons of the AUCs were performed with the De-Long method [21]. Standardized mortality ratios (SMRs) with their respective 95% CIs were calculated by dividing observed by predicted rates. SMRs and AUROCs were calculated in the overall population and per subgroup as follows: medical and surgical patients; reason for admission (cardiovascular, neurological, sepsis, and respiratory).

3. Results

During the study period, 8085 patients were admitted. A total of 2061 patients were excluded from the analysis due to age <18 years (n = 95), readmission during the same hospital stay (n = 651), transfer to another hospital (n = 86), and admission for acute coronary syndrome (n = 1229). We excluded 244 patients because a mortality prediction for one or more of the scores could not be calculated due to incomplete data. Patient characteristics are displayed in Table 1. Most admissions were for medical reasons and most patients originated from the emergency department. Inhospital mortality was 9.1%.

All models showed poor calibration. APACHE IV exhibited the most appropriate calibration and SAPS 3, the worst. All models overestimated inhospital mortality. Discrimination, however, was very good with AUROCs >0.85 for all scores (Table 2). The comparison between discriminations showed that APACHE IV performed better than SAPS 3 and MPM_0-III ($P < .001$ for both comparisons), and SAPS 3 was superior to MPM_0-III ($P = .043$).

Calibration curves (Fig. 1) indicated that in low-risk patients (<10% predicted mortality), all scores overestimated mortality. However, as this overestimation persisted for SAPS 3 in all strata of probabilities of death, it seems that APACHE IV underestimated death probability in higher risk strata (20%-70%). MPM_0-III had a variable performance across strata but seemed to work better in strata with mortality risk between 20% and 60%.

When cases were stratified by type of admission (medical or surgical) mortality was still overestimated, and again, APACHE IV performed better. When data were analyzed by medical or surgical admission, the APACHE IV performance seemed even more adequate with an SMR including 1 in the 95% CI. When the performance of the models was evaluated according to the reason for admission, again there was a trend toward overestimation of mortality with all scores. Some SMR scores included 1, but their CIs were too broad. Discrimination was very good in all subgroups analyzed (Table 3).

4. Discussion

In our study, the 3 modern prognostic models, namely APACHE IV, SAPS 3, and MPM_0-III, overestimated inhospital mortality, therefore showing poor calibration. However, their discrimination was nearly excellent. APACHE IV

Table 2 Performance of the models

Score	H-L statistics	P	SMR (95% CI)	AUROC (95% CI)
APACHE IV	53.7	<.001	0.79 (0.60-0.98)	0.883 (0.874-0.891)
MPM_0-III	134.2	<0.001	0.61 (0.53-0.70)	0.840 (0.830-0.849)
SAPS 3	226.6	<0.001	0.46 (0.37-0.54)	0.855 (0.846-0.864)

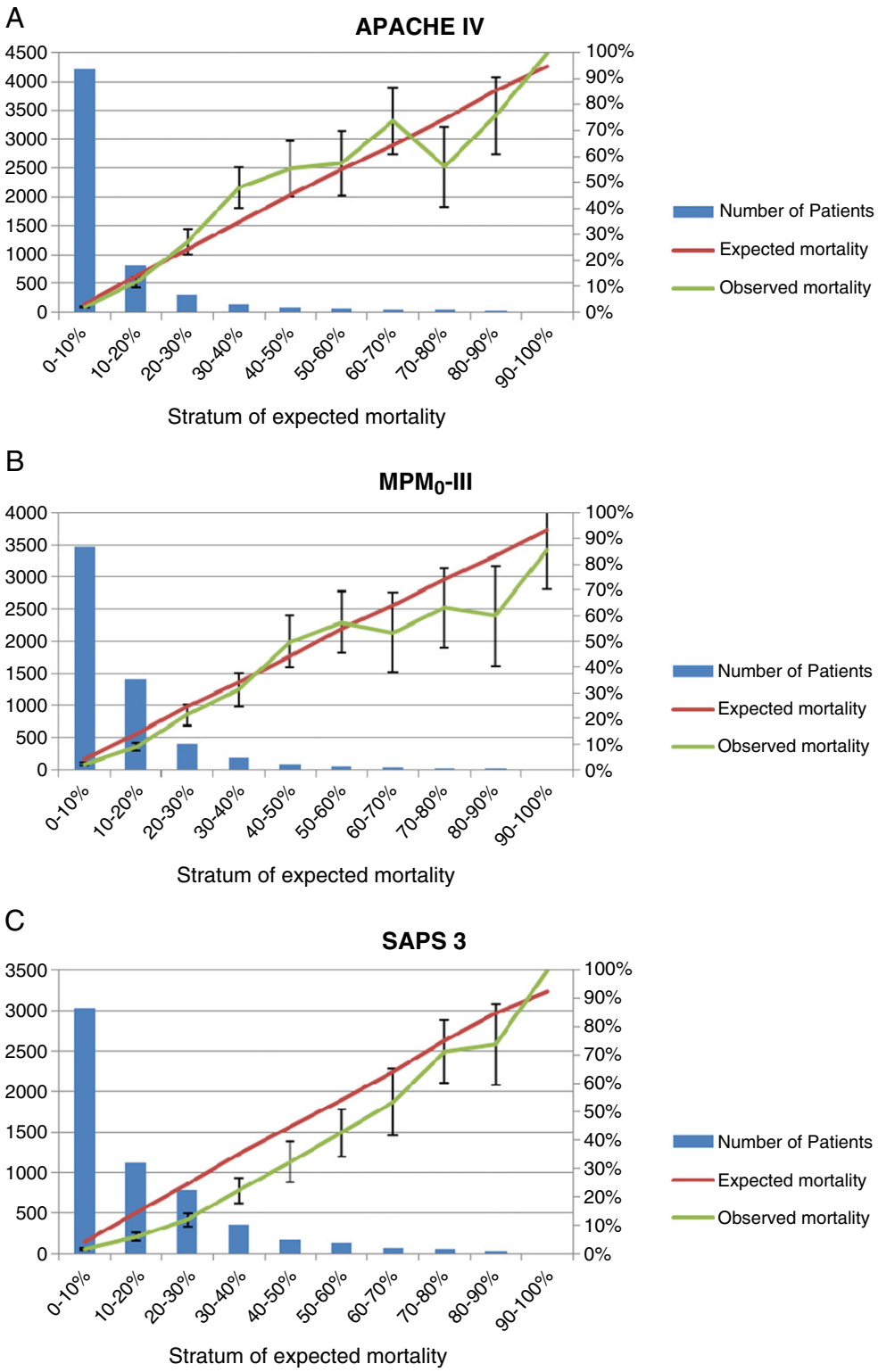

Fig. 1 Calibration curves. A, APACHE IV score. B, MPM_0-III score. C, SAPS 3 score. Observed mortalities are presented as means and 95% CIs.

performed better than SAPS 3 and MPM_0-III, but it also requires more effort and time for data collection. These results were consistent in many subgroups analyzed. Of note, SAPS 3, showed the worst calibration level, despite being developed based on a very large sample from different countries, although most were European [7].

Table 3 Performance of the scores by type of admission (medical/surgical) and reason for admission

Subgroup	No. of patients (%)	APACHE IV		MPM_0-III		SAPS 3	
		SMR (95% CI)	AUROC (95% CI)	SMR (95% CI)	AUROC (95% CI)	SMR (95% CI)	AUROC (95% CI)
Medical	4574 (79.1)	0.83 (0.67-1.00)	0.874 (0.857-0.891)	0.67 (0.57-0.77)	0.829 (0.809-0.848)	0.47 (0.36-0.58)	0.844 (0.826-0.862)
Surgical	1206 (20.9)	0.75 (0.45-1.05)	0.890 (0.830-0.949)	0.40 (0.22-0.59)	0.843 (0.783-0.904)	0.42 (0.23-0.61)	0.873 (0.827-0.919)
Cardiovascular	1170 (20.2)	1.12 (0.28-1.96)	0.869 (0.823-0.916)	0.50 (0.34-0.66)	0.813 (0.761-0.866)	0.52 (0.13-0.92)	0.854 (0.809-0.899)
Neurological	1126 (19.6)	0.63 (0.36-0.90)	0.862 (0.818-0.906)	0.34 (0.24-0.44)	0.879 (0.843-0.916)	0.35 (0.22-0.48)	0.846 (0.803-0.888)
Sepsis	920 (15.9)	0.82 (0.59-1.05)	0.877 (0.847-0.907)	1.07 (0.77-1.33)	0.797 (0.755-0.839)	0.55 (0.43-0.66)	0.812 (0.774-0.851)
Respiratory	690 (12.0)	0.95 (0.60-1.29)	0.837 (0.794-0.879)	0.84 (0.57-1.11)	0.799 (0.754-0.845)	0.50 (0.36-0.64)	0.818 (0.773-0.862)

These results are not totally unexpected, since many studies have shown the poor calibration and good discrimination of former generation models [22,23]. Although we could not find the reasons to explain these findings, there are some possibilities that deserve further discussion.

Firstly, and probably most important, there are differences in case mix. Usually, the population intended for validation of the model is different from the population where it was generated. Our population is quite different from those in which the 3 prognostic models were developed. For example, our ICU patients have mostly medical conditions. We had only 20.9% surgical patients. In all the original cohorts, this proportion was higher (30.9% for APACHE IV, 38.5% for SAPS 3 and 30.4% for MPM_0-III). Even small differences in case mix have been shown to influence the score's calibration [24]. Another important difference is that we studied a predominantly low-risk population, a fact that may be related to the absence of intermediate care units in the 3 hospitals [25]. In our cohort, 73% of the patients had a predicted risk of death of less than 10% for APACHE IV, 52% for SAPS 3, and 60% for MPM_0-III. This proportion was higher than that found in APACHE IV and SAPS 3 original cohorts but similar to that found in MPM_0-III. It has been shown that even small differences in mortality predictions impact calibration [26].

Another possibility is related to regional variability of end-of-life decisions. It has been demonstrated that the way such decisions are conducted impacts on mortality [27]. In our country, there is little limitation to the use of life-sustaining therapies [28], which could offer an explanation to the lower in-hospital mortality found in our study. This subject was not specifically addressed in our study.

In third place, there is the temporal bias. This was part of the rationale for updating the prognostic scores. SAPS3 was generated from 2002 data [7,8], APACHE IV from 2002-2003 data [6] and MPM_0-III from 2001-2004 data [9]. There was a 5- to 7-year interval between the development of these models and our study. This could partly explain our results, as demonstrated by other authors in external validations of SAPS 3 and APACHE IV [14,15,17]. However, if this is the case, then how often should a prognostic model be updated? A recent review points toward 4 years as an adequate interval [1], but not all data support this suggestion. SAPS3 was published in 2005, and Italian data from 2007 showed it was not calibrated to that population [15]. APACHE IV was published in 2006, and Dutch data from 2006-2009 showed it was not calibrated to that population [17]. We arrived basically at the same conclusion and also showed poor calibration for MPM_0-III.

In addition, there is a limitation of the statistical methods used to evaluate performance. The Hosmer-Lemeshow statistics is very sensitive to sample size and tends to show better calibrations in small series [29]. Interestingly, this trend was seen in SAPS 3 validation studies, with calibrations found to be better in smaller samples [10-13,16] and worse in larger ones [14,15], as what happened in our

cohort. The Dutch study [17] also had a larger sample than the North American study that showed a good calibration for APACHE IV [18].

Considering all these aspects, validating a score in an external population may be problematic. If the score is not calibrated in a new population, it means that this population is not comparable to the one in which the score was originally developed. Since benchmarking is "a process of comparing an ICU with a reference population" [9], how can we do it with different populations? Whenever the calibration is not good, the usual approach is to customize the tool, but although this could improve the score's performance in another population, the question remains if the results of this second population are comparable to those of the first one? How can we compare the results of this ICU with the others? Should every ICU perform a new logistic regression model in a prognostic model before adopting it? Therefore, benchmarking by comparing observed and predicted mortality ratios must be performed with caution, taking into consideration potential confounding factors.

More recently, a new issue emerged and should be taken into consideration when comparing different ICUs: a single whole sample SMR and its 95% CI does not reveal if a score performs well in all strata of severity [30]. Our data also demonstrate heterogeneity in performance among strata in our cohort. It seems that our overall low mortality was basically derived from a better performance in low-risk patients.

Despite all that, all scores had very good discrimination and even if they did not show a perfect fit, higher predicted mortality meant higher observed mortality. Therefore, these recent prognostic models may still be considered clinically useful, although not showing an excellent statistical performance in our cohort. Probably, there are some important variables that are absent from the models and could lead to quite different performances in our ICUs. Similar findings have been presented in many studies [14,15,17]. Since it is not feasible for each country or region to develop its own prognostic models based on its singular case-mix, admission and discharge policies, we believe that the available prognostic models are clinically useful and may be used for comparing ICUs with similar profiles or even the same ICU over time since its features remain similar.

Our study has potential limitations. Although designed as a multicenter study, all ICUs were from the same city in Brazil and provide similar standards of care. Therefore, our cohort may not be representative of a region or of the Brazilian healthcare system. Secondly, our subgroup analysis was merely exploratory, since we analyzed many subgroups and did not control for multiple comparisons. In addition, customization might have provided a better calibration, but we decided not to take this approach because our objective was to evaluate the original scores as they are intended to be used for benchmarking in non-teaching ICUs.

In conclusion, our study showed that APACHE IV, SAPS 3, and MPM_0-III had poor calibration, overestimating mortality in this population. Discrimination was very good for all scores. APACHE IV had the best performance, but also requires more effort and time for data collection. As poor calibration is a common finding in studies intended to validate severity scores worldwide, benchmarking by comparing observed and predicted mortality ratios must be performed with caution, taking into consideration potential confounding factors which interfere in ICU profiles.

Acknowledgments

We would like to thank Lia Delphino Salles for helping in data acquisition and Lais Andrade for the critical review of the manuscript.

References

[1] Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. Crit Care Med 2011;39:163-9.
[2] Zimmerman JE, Kramer AA. Outcome prediction in critical care: the Acute Physiology and Chronic Health Evaluation models. Curr Opin Crit Care;14:491-7.
[3] Capuzzo M, Moreno RP, Le Gall JR. Outcome prediction in critical care: the Simplified Acute Physiology Score models. Curr Opin Crit Care 2008;14:485-90.
[4] Higgins TL, Teres D, Nathanson B. Outcome prediction in critical care: the Mortality Probability Models. Curr Opin Crit Care 2008;14: 498-505.
[5] Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338: b605.
[6] Zimmerman JE, Kramer AA, McNair DS, et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006;34: 1297-310.
[7] Metnitz PG, Moreno RP, Almeida E, et al. 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;31:1336-44.
[8] Moreno RP, Metnitz PG, Almeida E, et al. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 2005;31:1345-55.
[9] 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;35:827-35.
[10] Soares M, Salluh JI. Validation of the SAPS 3 admission prognostic model in patients with cancer in need of intensive care. Intensive Care Med 2006;32:1839-44.
[11] Soares M, Silva UV, Teles JM, et al. Validation of four prognostic scores in patients with cancer admitted to Brazilian intensive care units: results from a prospective multicenter study. Intensive Care Med 2010;36:1188-95.
[12] Khwannimit B, Bhurayanontachai R. The performance and customization of SAPS 3 admission score in a Thai medical intensive care unit. Intensive Care Med 2010;36:342-6.
[13] Ledoux D, Canivet JL, Preiser JC, et al. SAPS 3 admission score: an external validation in a general intensive care population. Intensive Care Med 2008;34:1873-7.
[14] Metnitz B, Schaden E, Moreno R, et al. Austrian validation and customization of the SAPS 3 Admission Score. Intensive Care Med 2009;35:616-22.

[15] Poole D, Rossi C, Anghileri A, et al. External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. Intensive Care Med 2009;35:1916-24.
[16] Silva Junior JM, Malbouisson LM, Nuevo HL, et al. Applicability of the simplified acute physiology score (SAPS 3) in Brazilian hospitals. Rev Bras Anestesiol 2010;60:20-31.
[17] Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, et al. External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. J Crit Care. 2011;26:105 e11-8.
[18] Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 2008;133:1319-27.
[19] Higgins TL, Kramer AA, Nathanson BH, et al. Prospective validation of the intensive care unit admission Mortality Probability Model (MPM_0-III). Crit Care Med 2009;37:1619-23.
[20] Lemeshow S, Hosmer Jr DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92-106.
[21] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
[22] Beck DH, Smith GB, Pappachan JV, et al. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 2003;29: 249-56.
[23] Pappachan JV, Millar B, Bennett ED, et al. Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. Chest 1999;115: 802-10.
[24] Glance LG, Osler T, Shinozaki T. Effect of varying the case mix on the standardized mortality ratio and W statistic: a simulation study. Chest 2000;117:1112-7.
[25] Eachempati SR, Hydo LJ, Barie PS. The effect of an intermediate care unit on the demographics and outcomes of a surgical intensive care unit population. Arch Surg 2004;139:315-9.
[26] Glance LG, Osler TM, Papadakos P. Effect of mortality rate on the performance of the Acute Physiology and Chronic Health Evaluation II: a simulation study. Crit Care Med 2000;28:3424-8.
[27] Azoulay E, Pochard F, Garrouste-Orgeas M, et al. Decisions to forgo life-sustaining therapy in ICU patients independently predict hospital death. Intensive Care Med 2003;29:1895-901.
[28] Soares M, Terzi RG, Piva JP. End-of-life care in Brazil. Intensive Care Med 2007;33:1014-7.
[29] Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med 2007;35:2052-6.
[30] Moreno RP, Hochrieser H, Metnitz B, et al. Characterizing the risk profiles of intensive care units. Intensive Care Med 2010;36: 1207-12.