Intensive Care Med
DOI 10.1007/s00134-008-1187-4

BRIEF REPORT

Didier Ledoux
Jean-Luc Canivet
Jean-Charles Preiser
Joëlle Lefrancq
Pierre Damas

SAPS 3 admission score: an external validation in a general intensive care population

Received: 18 December 2007
Accepted: 28 May 2008

© Springer-Verlag 2008

Electronic supplementary material The online version of this article (doi:10.1007/s00134-008-1187-4) contains supplementary material, which is available to authorized users.

D. Ledoux (✉) · J.-L. Canivet ·
J.-C. Preiser · J. Lefrancq · P. Damas
Soins Intensifs Généraux,
Centre Hospitalier Universitaire de Liège,
Domaine Universitaire de Sart Tilman
Bat B35, 4000 Liège, Belgium
e-mail: dledoux@chu.ulg.ac.be
Tel.: +32-4-3667495
Fax: +32-4-3668898

Abstract *Objectives:* To validate the SAPS 3 admission score in an independent general intensive care case mix and to compare its performances with the APACHE II and the SAPS II scores. *Design:* Cohort observational study. *Setting:* A 26-bed general ICU from a Tertiary University Hospital. *Patients and participants:* Eight hundred and fifty-one consecutive patients admitted to the ICU over an 8-month period. Of these patients, 49 were readmissions, leaving 802 patients for further analysis. *Intervention:* None. *Measurements and results:* APACHE II, SAPS II and SAPS 3 variables were prospectively collected; scores and their derived probability of death were calculated according to their original manuscript description. The discriminative power was assessed using the area under the ROC curve (AUROC) and calibration was verified with the Hosmer–Lemeshow goodness-of-fit test. The AUROC of the APACHE II model (AUROC = 0.823) was significantly lower than those of the SAPS II (AUROC = 0.850) and SAPS 3 models (AUROC = 0.854) ($P = 0.038$). The calibration of the APACHE II model ($P = 0.037$) and of the SAPS 3 global model ($P = 0.035$) appeared unsatisfactory. On the contrary, both SAPS II model and SAPS 3 model customised for Central and Western Europe had a good calibration. However, in our study case mix, SAPS II model tended to overestimate the probability of death. *Conclusion:* In this study, the SAPS 3 admission score and its prediction model customised for Central and Western Europe was more discriminative and better calibrated than APACHE II, but it was not significantly better than the SAPS II.

Keywords Severity of illness index · SAPS 3 · APACHE · Outcome assessment · Critical care

Introduction

The first scoring systems dedicated to the assessment of severity of illness of ICU patients were launched more than 25 years ago. Among these severity of illness scoring systems, the second version of the Acute Physiology and Chronic Health Evaluation score (APACHE II) [1] became used worldwide. Although more recent severity scores versions were developed in the nineties [2–4], the APACHE II remains, to date, the most widely used scoring system for ICUs assessment and for clinical trials conducted in the field of critical care medicine. Nevertheless, several studies showed a deterioration of both APACHE II and SAPS II scores performances [5–8]. The recently published SAPS 3 admission score [9] is a model built to predict hospital mortality from admission data (recorded within ±1 h). This model is based on a large cohort of patients (16,784 patients) consecutively admitted to 303 intensive care units from 35 countries around the world [10]. From this admission score are derived not only a global equation for hospital mortality prediction based on the whole case mix, but also equations customised for different geographic regions. Although the SAPS 3 admission model is a promising and elegant tool, there is a need for its external validation to verify its performances on an independent population sample.

The first aim of the present study was to assess SAPS 3 admission score in a patients' cohort from a mixed medico-surgical ICU located in a Western Europe country. A secondary end point of the study was to compare the SAPS 3 score performances with those of the older APACHE II and SAPS II scores.

Material and methods

The study was conducted in a 26-bed general intensive care unit at the Liege University Hospital, Belgium. Data were analysed on all consecutive admissions over an 8-month period. For patients admitted more than once to the ICU during their hospital stay, only data recorded during the first ICU admission were analysed. As for the APACHE II and SAPS 3 scores [10, 11], patients under 16 years of age were excluded from the analysis. Burns were also excluded from the study, since in our institution, these patients are treated in a specific burns unit. Finally, we decided to include patients admitted after heart surgery in the case mix, since those patients are taken into account in the SAPS 3 admission score. In addition, previous studies showed that performance of the APACHE II and SAPS II is adequate in case mix of patients admitted to the ICU after heart surgery [12]. Additional information on exclusion criteria may be found in the electronic supplementary material (ESM). The institutional human research ethics committee waived the need for informed consent, since no patient intervention was required and patient anonymity was preserved.

Data collection, computer entry and processing

Data were collected prospectively by a research nurse with a previous experience in data collection for the APACHE II and SAPS II scores. That nurse was trained for SAPS 3 variables collection and she had access to the variables definitions published in the ESM from the original SAPS 3 paper [10]. Data were collected on paper form complying with the original publications methodology [1, 3, 9] and then encoded on the ICU database. Data were imported into a spreadsheet (Microsoft® Excel 2003, Microsoft® Corporation) for the calculation of the scores and their derived probabilities of death using the published equations and coefficients.

Statistical analysis

The study was powered to assess a difference between areas under the ROC curves (AUROC) of the APACHE II model and SAPS 3 model customized for Central and Western Europe [13, 14]. The AUROC used for power calculation were obtained from historical observations made in our ICU for APACHE II model (AUROC = 0.816 on a cohort of 1,221 consecutive patients admitted to the ICU during the year 2004) and from the AUROC published in the original SAPS 3 paper (AUROC = 0.861 for Central and Western Europe) [9]. The decision to use APACHE II AUROC rather than SAPS II AUROC to define sample size was based on two considerations: first, we wanted to challenge the APACHE II; second, although this was not statistically significant, in our historical case mix, APACHE II AUROC (0.816, 95% CI 0.787–0.846) was higher than SAPS II AUROC (0.803, 95% CI 0.768–0.839); sample size calculation was therefore also adequate to assess SAPS II. A sample of 780 patients was required in the study to reach an 80% power. Continuous variables are reported as the median and interquartile range (IQR) or as mean ± standard deviation. Categorical variables are reported as the count and percentage. For each score, the discriminative power was assessed using the AUROC. The calibration was evaluated with the Hosmer–Lemeshow goodness-of-fit $\hat{C}$ test. A P value above 0.05 at the Hosmer–Lemeshow goodness-of-fit test indicated a good calibration. The areas under the ROC curves were compared using the method described by DeLong et al. [15]. The standard mortality ratios (SMRs) were calculated as the ratio between observed mortality and expected mortality based on the severity score models. The 95% confidence interval of the SMRs were computed using the method described by Rapoport et al. [16]. Statistical analyses were performed using SAS (version 9.1.3 Service Pack 4, SAS Institute Inc., Cary, NC, USA) and STATA software (version 8.0, Stata Corporation, Texas, USA). A two-tailed P value below 0.05 was considered statistically significant.

Results

Patients' case mix

Over the 8-month period, from December 2005 to July 2006, 865 patients were admitted to the ICU. Forty-nine of these patients (5.7%) were readmitted during the same hospital stay and 14 patients (1.6%) were younger than 16 years of age. Those patients were not included in the study, leaving 802 (92.3%) patients for analysis. The data used to derive APACHE II, SAPS2 and SAPS 3 scores and probabilities of death were collected in all these patients. Patient's characteristics are presented in Table 1. Apart from basic and observational admission ($n = 105/802$, 13%), the main reasons for ICU admission were as follows: cardiovascular, respiratory and neurological. These reasons encountered for 70% of the ICU admissions. Additional details on patients' characteristics may be found in the ESM (Tables E2 and E3, ESM). During the study period, the overall hospital mortality was 140 (17.5%) patients.

Performance of the scoring systems

Performances of the three models are summarized in Table 2. The discriminative power, assessed using the AUROC, was significantly lower for the APACHE II model (AUROC 0.823 ± 0.020) as compared with SAPS II (AUROC 0.850 ± 0.019) and SAPS 3 (AUROC 0.854 ± 0.019) model ($P = 0.037$) (Fig. E1, ESM). The Hosmer–Lemeshow goodness-of-fit test ($\hat{C}$) revealed a poor calibration for the APACHE II models ($\hat{C} = 16.38$, $P = 0.037$) and for the SAPS 3 global model ($\hat{C} = 16.59$, $P = 0.035$). On the contrary, the calibration of SAPS II model ($\hat{C} = 5.78$, $P = 0.671$) and SAPS 3 customized for Central and Western Europe ($\hat{C} = 8.30$, $P = 0.405$) was appropriate (Fig. 1). The analysis of the standardised mortality ratios revealed that the best predictive results were achieved with the SAPS 3 model customized for Central and Western Europe. The global SAPS 3 model significantly overestimated hospital mortality; the 95% confidence interval did not indeed contain 1 (SMR = 0.82; 95% CI 0.70–0.93). While APACHE II tended to underestimate mortality, the SAPS II model, on the contrary, tended to overestimate mortality (Fig. 2E, ESM).

Table 1 Patients' demographic characteristics

Patients' characteristics	
Age (years), median (IQR)	66 (53–75)
Male, *n* (%)	486 (60.6)
No surgery, *n* (%)	232 (28.9)
Scheduled surgery, *n* (%)	397 (49.5)
Unscheduled surgery, *n* (%)	173 (21.6)
Origin	
Home	109 (13.6)
Same hospital	551 (68.7)
Chronic care facility	1 (0.1)
Public place	11 (1.4)
Other hospital	130 (16.2)
Co-morbidities	
Alcoholism	69 (8.6)
Arterial hypertension	444 (55.6)
Chemotherapy	10 (1.3)
Chronic heart failure	355 (44.4)
Chronic pulmonary failure	18 (2.3)
COPD	127 (15.9)
Chronic renal failure	39 (4.9)
Cirrhosis	25 (3.1)
EV drug addict	6 (0.8)
Haematological cancer	17 (2.1)
HIV positive	3 (0.4)
Immunosuppression, other	15 (1.9)
Diabetes	191 (23.9)
Cancer	69 (8.6)
Radiotherapy	7 (0.9)
Steroid treatment	13 (1.6)
Ventilated on admission, *n* (%)	594 (74.1)
Length of stay in ICU (days), median (IQR)	3 (2–7)
Length of stay in hospital (days), median (IQR)	14 (10–26)
ICU mortality, *n* (%)	106 (13.2)
Hospital mortality, *n* (%)	140 (17.5)

Definition for co-morbidities can be found in the electronic supplementary material of the original SAPS 3 paper

Discussion

Both the global model and the model customised for Central and Western Europe of the SAPS 3 admission score had a very good discriminative power as shown by an AUROC very close to the one published in the original SAPS 3 paper [9]. However, the fit of the global SAPS 3 mortality prediction model was inadequate in our patients' sample from a Western Europe ICU. The global SAPS 3 model significantly overestimated hospital mortality in our studied patients' cohort. These findings are not surprising, since, in the original SAPS 3 hospital outcome cohort, Moreno et al. already reported that the SAPS 3 global mortality prediction model fit was poor for Central and Western Europe ICUs [9]. On the contrary, the SAPS 3 model customized for Central and Western Europe region was adequate. The discriminative power was very good, close to the one published in the original publication, and the calibration was appropriate. Moreover, this model produced the best predictive results as shown by a standardised mortality ratio close to one.

The present study shows that older severity of illness scoring systems performances may not be satisfactory anymore. In our patients' case mix, the APACHE II score suffered from both a lower discriminative power, as compared with the other assessed severity scores, and from a significant lack of calibration. These findings were previously described by several authors [5, 17]. Nevertheless, other authors found acceptable calibration of the APACHE II score even in recent case mix population sample [18, 19]. It appears however that the APACHE II score is nowadays outdated. Interestingly, Knaus, the APACHE II original developer, advised that researchers should discontinue the use of the APACHE II for outcome assessment [20].

In our patients' sample, the SAPS II score performed well, its discriminative power was very good and its

Table 2 Area under the receiver-operating characteristic curve, Hosmer–Lemeshow goodness-of-fit test and standardised mortality ratios for the APACHE II, SAPS II and SAPS 3 (global and customised for Central and Western Europe) prognostics models

Prediction models	Score (mean ± SD)	Predicted mortality (mean ± SD)	Area under ROC curve		Goodness-of-fit $\hat{C}$ test		SMR
			AUC (95% CI))	*P* value*	$\hat{C}$	*P* value	(95% CI)
APACHE II equation	13.3 ± 6.5	15.9 ± 19.1	0.82 (0.78–0.86)	0.037	16.38	0.037	1.10 (0.97–1.24)
SAPS II equation	33.1 ± 14.5	19.7 ± 22.0	0.85 (0.81–0.89)		5.78	0.671	0.89 (0.77–1.01)
SAPS 3 global equation	48.9 ± 15.2	21.4 ± 21.9	0.85 (0.82–0.89)		16.59	0.035	0.82 (0.70–0.93)
SAPS 3 Central, Western Europe equation		18.1 ± 21.0	0.85 (0.82–0.89)		8.30	0.405	0.96 (0.84–1.08)

ROC curve receiver-operating characteristic curve, *AUC* area under the curve, *SD* standard deviation, *95% CI* 95% confidence interval, *SMR* standardised mortality ratio, *APACHE* acute physiology and chronic health evaluation, *SAPS* simplified acute physiology score

*Comparison of APACHE II, SAPS II, SAPS 3 and customized SAPS 3 using DeLong methods

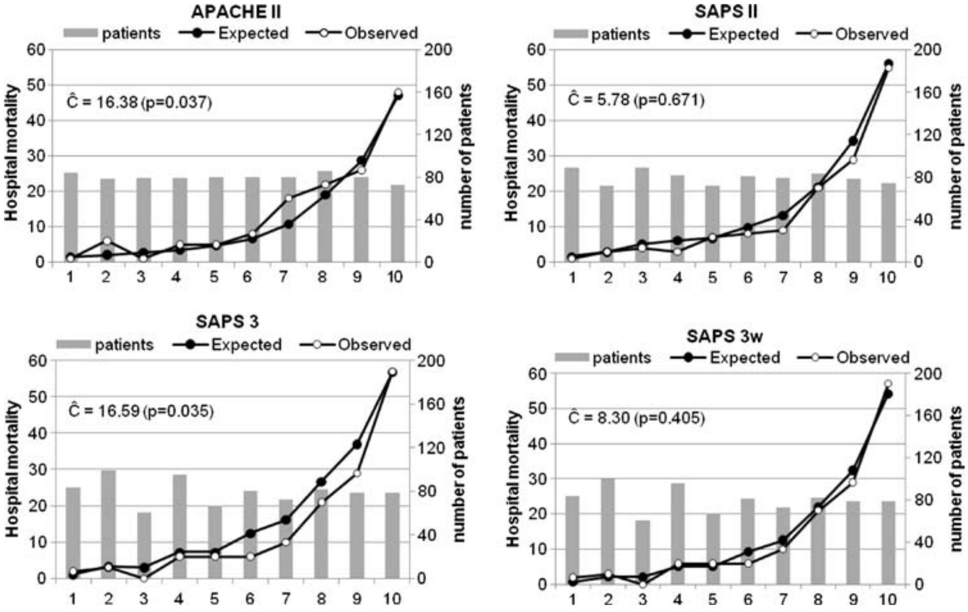

Fig. 1 Hosmer–Lemeshow $\hat{C}$ goodness-of-fit test; calibration curves for the APACHE II, SAPS II, global SAPS 3 (SAPS 3) and SAPS 3 customised for Central and Western Europe (SAPS 3w) models

calibration was appropriate. These results are divergent from most published results. Several authors indeed showed that if the SAPS II model has a good discriminative power, its calibration is poor when applied to an independent case mix [8, 17]. However, although it seemed to perform adequately in our patients' sample, we found, like other authors, that the SAPS II predictive model tended to overestimate the hospital mortality [8, 17, 18].

The present study has potential limitations. One could criticize the relatively small sample size of the study; however, it was designed to be adequately powered to detect differences between APACHE II or SAPS II and SAPS 3 AUROC. Another potential limitation is related to the fact that the present work is a single-centre study with a different patients' case mix as compared to the original SAPS 3 hospital outcome cohort, and these results may not be generalisable to other ICUs. Finally, one could criticize the fact that in the present study there was no assessment of data collector reliability. Although this is an important issue, we are quite confident that, in this work, bias related to inadequate data collection was limited, since collection was done by one trained research nurse. Previous studies showed indeed that, in such condition, interobserver variability was reduced [21].

In conclusion, in the present study, we found that the SAPS 3 admission score was superior to the APACHE II model. However, in our case mix, it was not significantly better than the SAPS II score, both having a good discriminative power and calibration.

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. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A et al (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100:1619–1636
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. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270:2478–2486
5. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP (1993) 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 307:977–981
6. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, Melotti RM (1996) The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. Intensive Care Med 22:1368–1378
7. Moreno R, Miranda DR, Fidler V, Van Schilfgaarde R (1998) Evaluation of two outcome prediction models on an independent database. Crit Care Med 26:50–61
8. Metnitz PG, Valentin A, Vesely H, Alberti C, Lang T, Lenz K, Steltzer H, Hiesmayr M (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. Intensive Care Med 25:192–197
9. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR (2005) 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 31:1345–1355
10. Metnitz PG, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR (2005) SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. Intensive Care Med 31:1336–1344
11. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1986) An evaluation of outcome from intensive care in major medical centers. Ann Intern Med 104:410–418
12. Martinez-Alario J, Tuesta ID, Plasencia E, Santana M, Mora ML (1999) Mortality prediction in cardiac surgery patients: comparative performance of Parsonnet and general severity systems. Circulation 99:2378–2382
13. Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843
14. Mandrekar JN, Mandrekar SJ, (2005) Statistical methods in diagnostic medicine using SAS® software. In: Proceedings of the 30th SAS Users Group International Conference (SUGI), Philadelphia, 10–13 April 2005
15. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
16. Rapoport J, Teres D, Lemeshow S, Gehlbach S (1994) A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 22:1385–1391
17. Moreno R, Morais P (1997) Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 23:177–186
18. Capuzzo M, Valpondi V, Sgarbi A, Bortolazzi S, Pavoni V, Gilli G, Candini G, Gritti G, Alvisi R (2000) Validation of severity scoring systems SAPS II and APACHE II in a single-center population. Intensive Care Med 26:1779–1785
19. Ho KM, Lee KY, Williams T, Finn J, Knuiman M, Webb SA (2007) Comparison of Acute Physiology and Chronic Health Evaluation (APACHE) II score with organ failure scores to predict hospital mortality. Anaesthesia 62:466–473
20. Knaus W (2005) APACHE II. http://www.cerner.com/public/FileDownload.asp?LibraryID=24648&iphl=apachede:apaches:apache:apach:iye:II:ii:iy:. Retrieved 12 Dec 2007
21. Polderman KH, Jorna EM, Girbes AR (2001) Inter-observer variability in APACHE II scoring: effect of strict guidelines and training. Intensive Care Med 27:1365–1369