BJA

CRITICAL CARE

Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit

Y. Sakr[1], C. Krauss[1], A. C. K. B. Amaral[2], A. Réa-Neto[2], M. Specht[1],
K. Reinhart[1] and G. Marx[1]*

[1]*Department of Anaesthesiology and Intensive Care, Friedrich-Schiller-University Hospital, Erlanger Allee 103, 07743 Jena, Germany.* [2]*Department of Critical Care, Hospital Brasilia, Rede ESHO, Brasilia, Brazil*
Corresponding author. E-mail: gernot.marx@med.uni-jena.de

Background. The Simplified Acute Physiology Score (SAPS) 3 has recently been developed, but not yet validated in surgical intensive care unit (ICU) patients. We compared the performance of SAPS 3 with SAPS II and the Acute Physiology and Chronic Health Evaluation (APACHE) II score in surgical ICU patients.

Methods. Prospectively collected data from all patients admitted to a German university hospital postoperative ICU between August 2004 and December 2005 were analysed. The probability of ICU mortality was calculated for SAPS II, APACHE II, adjusted APACHE II (adj-APACHE II), SAPS 3, and SAPS 3 customized for Europe [C-SAPS3 (Eu)] using standard formulas. To improve calibration of the prognostic models, a first-level customization was performed, using logistic regression on the original scores, and the corresponding probability of ICU death was calculated for the customized scores (C-SAPS II, C-SAPS 3, and C-APACHE II).

Results. The study included 1851 patients. Hospital mortality was 9%. Hosmer and Lemeshow statistics showed poor calibration for SAPS II, APACHE II, adj-APACHE II, SAPS 3, and C-SAPS 3 (Eu), but good calibration for C-SAPS II, C-APACHE II, and C-SAPS 3. Discrimination was generally good for all models [area under the receiver operating characteristic curve ranged from 0.78 (C-APACHE II) to 0.89 (C-SAPS 3)]. The C-SAPS 3 score appeared to have the best calibration curve on visual inspection.

Conclusions. In this group of surgical ICU patients, the performance of SAPS 3 was similar to that of APACHE II and SAPS II. Customization improved the calibration of all prognostic models.

Severity scoring systems integrate clinical data to estimate the probability of mortality, which can be used to facilitate resource utilization or continuing quality improvement and to stratify patients for clinical research.[1] Several criteria should be taken into consideration when judging the value of any scoring system in clinical practice. Reliability and validity are important issues that allow confident use of a scoring system in intensive care unit (ICU) patients with different case-mixes and baseline characteristics.

In critically ill patients, several scoring systems have been developed over the last three decades. The Acute Physiology and Chronic Health Evaluation (APACHE)[2 3]

and the Simplified Acute Physiology (SAPS)[4] scores are the most widely used scoring systems in the ICU. The SAPS 3 score was developed recently in a worldwide cohort.[5 6] SAPS 3 is based on 20 different variables (Appendix E1 in the online data supplement) that are easily measured at patient admission (within the first hour) and allows early appraisal of risk, dissociating patient status from the quality of care in the ICU. After extensive use of cross-validation techniques, the SAPS 3 score showed very good internal validity.[5] Nevertheless, prospective validation in separate populations and in more defined ICU patients improves the generalizability, or

applicability, of the model to these settings. Recently, Soares and Salluh[7] validated SAPS 3 in a cohort of cancer patients. The performance of SAPS 3 has not previously been validated in surgical ICU patients.

The aim of this study was, therefore, to validate the SAPS 3 score and compare its performance with those of the commonly used APACHE II and SAPS II scores in a large cohort of surgical ICU patients.

Methods

The study was approved by the Institutional Review Board of Friedrich Schiller University Hospital. All consecutive patients admitted to our surgical ICU between August 2004 and December 2005 were screened for eligibility. For patients who were readmitted to the ICU during the study period, only the first admission was considered. Patients who had missing components of SAPS II, APACHE II, or SAPS 3 score were excluded from the analysis. Patients with ICU length of stay (LOS) <24 h were also excluded from the analysis as SAPS II and APACHE II cannot be calculated in these patients.

Data were collected from the monitoring equipment (heart rate, ventilatory frequency, and arterial pressure monitoring) and the ventilators, and automatically recorded by a clinical information system (CIS) introduced in our ICU in 1998. The CIS, manufactured by Copra System GmbH Sasbachwalden, provides a complete electronic documentation, order entry (e.g. medication), and direct access to laboratory and microbiology results.

The APACHE II and SAPS II scores were calculated from the CIS within 24 h of admission by the attending physician who was in charge of the patient. Data recorded on admission also included age, gender, referring facility, primary and secondary admission diagnoses, and surgical procedures preceding admission. Admission categories, the presence of infection, and the use of vasopressors on admission to the ICU were extracted retrospectively by a trained physician (C.K.) and were validated periodically by a senior intensivist (Y.S.). Inconsistency between the two raters was resolved by consensus. These data were available from the CIS in text format in a special section and the consistency between the two raters was, therefore, high (100% for admission categories, 97% for vasopressor use, and 98% for the presence of infection). Other data required for the calculation of SAPS 3 were extracted electronically and were subjected to a plausibility check by the attending physician in charge of each patient before being used in the calculations. Infection was defined as the presence of clinical or microbiological evidence of infection necessitating the administration of antibiotics. The presence of infection was documented daily in a special section of the CIS. Hospital mortality data were available for all patients.

The probability of death for all models was calculated according to standard formulas (Appendix 1). The probabilities of in-hospital death were calculated for SAPS 3 according to the general equation provided by Moreno and colleagues,[5] in addition to the customized probability equation for west and central Europe [C-SAPS 3 (Eu)]. The adjusted probability of death according to the diagnostic category of the APACHE II score (adj-APACHE II)[2] [3] was also calculated.

Data were analysed using SPSS 13.0 for windows (SPSS Inc., Chicago, IL, USA). Calibration of the prognostic models was assessed using Hosmer–Lemeshow Ĉ and Ĥ statistics and by calibration curves. Lower Hosmer–Lemeshow χ^2 values and higher P-values (>0.05) indicate good fit. Model discrimination, defined as the ability of the model to discriminate in-hospital non-survivors from survivors, was assessed using the receiver operating characteristic (ROC) area under the curve (aROC) and 95% confidence intervals (CI) were computed. Comparison of ROC curves was performed using the method of DeLong and colleagues.[8] The three original scores were also customized to our specific patient population, creating three new models (C-SAPS 3, C-SAPS II, and C-APACHE II). The data set was randomly split into development (1051 patients) and validation (800 patients) samples. To improve calibration of the original models, customization was performed using the development sample, using logistic regression with mortality as the dependent variable and the original probability of death in each score as the independent variable. Probability of death was then calculated for each patient in the validation sample based on the output of the former procedure. Calibration and discrimination were assessed in these models as previously described for the original models.

Categorical data are presented as n (%) and continuous data as mean (SD) unless otherwise indicated.

Results

Among 2168 patients admitted consecutively to our surgical ICU between December 2004 and August 2005, 241 readmissions and 76 patients with missing data (mostly due to ICU LOS<24 h) were excluded. The study group, therefore, comprised 1851 patients, 1173 males (63.4%) and 678 females (36.6%), mean age 62 yr. The characteristics of the study group are shown in Table 1. Elective surgery was performed before admission to the ICU in 61.5% of the patients and emergency surgery in 24.3%. Patients were most commonly admitted after cardiac surgery (n=488, 26.4%). Two hundred and sixty postoperative patients were referred from other facilities and did not undergo any surgical procedure in the 48 h preceding ICU admission. The median ICU LOS was 1 day (IQR 1–4 days), and the overall ICU and hospital mortality rates were 6.4% and 9% (n=118 and 167), respectively.

The mean (SD) SAPS 3 score on admission to the ICU was 48.6 (14.4); the mean APACHE II and SAPS II scores, calculated within 24 h of admission to the ICU, were 22 (8.3) and 34.4 (18), respectively. The distributions of SAPS II, SAPS 3, and APACHE II scores are presented in

Table 1 Characteristics of the study group on admission to the ICU. ENT, ear, nose, throat; OR, operating room

n	1851
Age (yr), mean (range)	61.6 (16–97)
Gender (male/female)	1173/678
Source of admission [*n* (%)]	
OR/recovery	1468 (79.3)
Emergency room	121 (6.5)
Hospital floor	198 (10.7)
Other ICU	64 (3.5)
Type of admission [*n* (%)]	
No surgery	260 (14.0)
Elective surgery	1139 (61.5)
Emergency surgery	449 (24.3)
Type of surgery [*n* (%)]	
Cardiac surgery	488 (26.4)
Gastrointestinal surgery	346 (18.7)
Neurosurgery	150 (8.1)
Vascular surgery	105 (5.7)
Trauma surgery	140 (7.5)
ENT surgery	56 (3.0)
Thoracic surgery	52 (2.8)
Liver transplantation	26 (1.4)
Others	228 (12.3)
APACHE II score [mean (SD)]	22.0 (8.3)
SAPS II score [mean (SD)]	34.4 (18.0)
SAPS 3 score [mean (SD)]	48.6 (14.4)
ICU mortality [*n* (%)]	118 (6.4%)
Hospital mortality [*n* (%)]	167 (9.0%)
ICU length of stay (days), median (IQR)	1 (1–4)
Hospital length of stay (days), median (IQR)	11 (8–17)

Table 2 Performance of the different prognostic models. SAPS, Simplified Acute Physiology Score; C-SAPS II, SAPS II customized to local study population; C-SAPS 3 (Eu), SAPS 3 customized for central and western Europe; C-SAPS 3, SAPS 3 customized for local study population; APACHE, Acute Physiology and Chronic Health Evaluation; Adj-APACHE II, adjusted probability of death according to the diagnostic category of the APACHE II score; C-APACHE II, APACHE II customized to local study population

	H–L Ĉ-statistics		H–L Ĥ-statistics		aROC (95% CI)
	χ^2	*P*-value	χ^2	*P*-value	
SAPS II	452.03	<0.0001	458.89	<0.0001	0.83 (0.80–0.86)
C-SAPS II	8.59	0.37	7.85	0.09	0.83 (0.79–0.88)
SAPS 3	208.49	<0.0001	211.84	<0.0001	0.84 (0.81–0.88)
C-SAPS 3 (Eu)	126.79	<0.0001	177.37	<0.0001	0.84 (0.81–0.88)
C-SAPS 3	7.9	0.44	11.06	0.08	0.89 (0.86–0.93)
APACHE II	1417.04	<0.0001	1188.81	<0.0001	0.80 (0.76–0.83)
Adj-APACHE II	714.45	<0.0001	690.72	<0.0001	0.83 (0.80–0.86)
C-APACHE II	11.11	0.19	4.78	0.18	0.78 (0.73–0.84)

Fig 2 ROC curves for APACHE II, adjusted APACHE II, SAPS II, and SAPS 3 scores on prediction of hospital mortality.
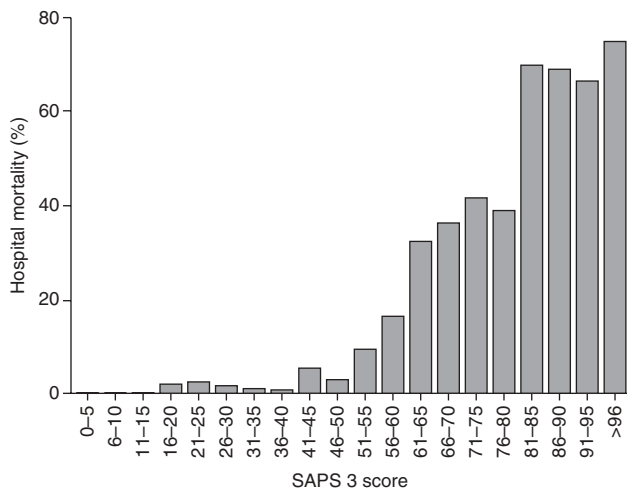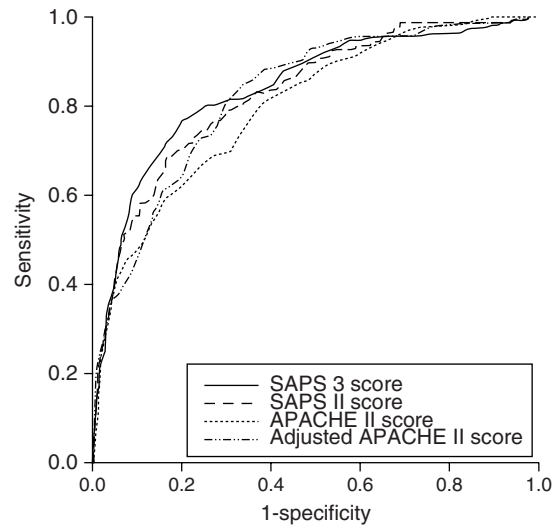

Fig 1 Bar chart representing hospital mortality rates (%) according to SAPS 3 score.

Figure E1 in the online data supplement. Hospital mortality was substantially greater in patients with higher SAPS 3 scores (Fig. 1). A hospital mortality of <3% was observed in patients with SAPS 3 scores ≤40, increasing to around 10% in patients with SAPS 3 scores between 40 and 60. The highest hospital mortality rate (around 70%) was observed in patients with a SAPS 3 score greater than 80.

H–L statistics showed poor calibration for SAPS II, APACHE II, adj-APACHE II, SAPS 3, and C-SAPS 3 (Eu) (H–L Ĉ and Ĥ-statistics: *P*<0.05), whereas the scores customized to our study population, C-SAPS II, C-APACHE II, and C-SAPS 3, showed good calibration in the validation sample (Table 2). This was confirmed by the calibration curves of the corresponding scores (Fig. E2 in the online data supplement). Visual inspection of the calibration curve for C-SAPS 3 suggests that it might have a better calibration along the full spectrum of severity of disease than the other models.

The overall discriminatory capability, as measured by the aROC, was generally good for all models (Table 2 and Fig. 2) and ranged from 0.78 (C-APACHE II score) to 0.89 (C-SAPS 3). Customization of the models did not change the discriminatory ability of the original scores. Discrimination was identical for the SAPS 3 and C-SAPS 3 (Eu) scores. The APACHE II and C-APACHE II scores had significantly lower aROC compared with the other scores (Table 3).

Table 3 Comparison of aROC for the prognostic models (*P*-values). SAPS, Simplified Acute Physiology Score; C-SAPS II, SAPS II customized to local study population; C-SAPS 3 (Eu), SAPS 3 customized for central and western Europe; C-SAPS 3, SAPS 3 customized for local study population; APACHE, Acute Physiology and Chronic Health Evaluation; Adj-APACHE II, adjusted probability of death according to the diagnostic category of the APACHE II score; C-APACHE II, APACHE II customized to local study population

	SAPS II	C-SAPS II	SAPS 3	C-SAPS 3 (Eu)	C-SAPS 3	APACHE II	Adj-APACHE II
C-SAPS II	1.0						
SAPS 3	0.367	0.363					
C-SAPS 3 (Eu)	0.365	0.362	1.0				
C-SAPS 3	0.363	0.359	1.0	1.0			
APACHE II	0.016	0.016	0.028	0.027	0.027		
Adj-APACHE II	0.773	0.773	0.539	0.538	0.537	<0.001	
C-APACHE II	0.018	0.017	0.028	0.028	0.028	1.0	<0.001

Discussion

The performance of prognostic models encompasses two objective measures: calibration and discrimination.[9] Calibration refers to how closely the estimated probabilities of mortality correlate with the observed mortality over the entire range of probabilities. Discrimination refers to how well the model discriminates between individuals who will live and those who will die. It is important to understand that it is impossible for any model to have perfect calibration and perfect discrimination at the same time.[10] From the individual patient's point of view, it would be interesting to have perfect discrimination, however, for clinical trials or comparison of care between ICUs better calibration is needed.

The present study demonstrates that in a large sample of critically ill surgical patients, the original scores had poor calibration, which improved after first-level customization. The prognostic performance of SAPS 3 was similar to that of SAPS II. Discrimination was generally good for all models, except for the original and customized APACHE II scores. It is not unexpected that calibration improved after first-order customization due to the very specific case-mix in our sample (surgical patients). Although this methodology is simple, we have demonstrated its effectiveness in improving calibration. However, as no new terms were added, it is also obvious that discrimination would not change.

The SAPS 3 score was developed using a worldwide database of 303 ICUs and 16 784 patients.[5 6] However, the SAPS 3 database was not collected to be representative of global case-mix, especially of specific regional areas or patient types, such as individual diseases. External validation is, therefore, important before applying this score to other case-mixes. In 952 cancer patients admitted to the ICU, Soares and Salluh[7] found that the SAPS II and SAPS 3 prognostic models had excellent discrimination. The calibration of SAPS II was poor. However, the calibration of SAPS 3 and its customized equation for Central and South American countries was appropriate. The poor calibration of the original SAPS 3 model in our study may be due to the difference in case-mix between our study and that of Soares and Salluh.[7] Disparity between the case-mix of test and reference databases is one of the main sources

for the decay in the predictive accuracy of prognostic models when applied to populations other than those for which they were developed.[9 11] In addition, the use of an automatic data management system in our study may be expected to overestimate the mortality due to the high sampling rate, as reported in previous studies.[12 13]

Our results are in agreement with other reports on the performance of the APACHE scoring system in the UK.[14−16] The same pattern was observed in the external validation of the SAPS II, APACHE II, and APACHE III models in Scottish intensive care patients.[17] One study reported good calibration for the APACHE II model, but again imperfect calibration for the two other scores tested.[18] The SAPS II model also failed to adjust adequately for differences in the case-mix profiles of ICU patients from various European countries.[19−22] Recently, Beck and colleagues[23] validated the SAPS II and APACHE II and III prognostic models in 16 646 adult intensive care patients in Southern UK. The external validation showed a similar pattern for all three models tested: good discrimination, but imperfect calibration.

The admission SAPS 3 prediction model is based exclusively on data available within 1 h of ICU admission.[5 6] Interestingly, about half of the predictive power of the original SAPS 3 score is derived from information available before ICU admission.[5] Prognostic scoring systems that are meant to include measurements over the first 24 h period in the ICU are not valid for use in ICU triage. Moreover, values obtained over a 24 h period frequently capture the standard of care more than the real clinical status of the patient. This major advantage of the SAPS 3 score, in the absence of evidence of any superiority of the other scoring systems calculated within 24 h of admission to the ICU, merits its use in the postoperative ICU setting after proper customization. External validation is required to assess the performance of this score in other ICU populations.

Our study provides an external validation of the SAPS 3 score in a large cohort of surgical ICU patients; however, several limitations need to be addressed. First, the case-mix in our study may differ from that in other ICUs as a great proportion of our patients were admitted after cardiac surgery, limiting the extrapolation of our results to

other populations. Secondly, postoperative patients are not homogeneous; however, the small sample size in the various subgroups in our study hinders exploration of the uniformity of fit among subgroups. Thirdly, our study may be limited by the retrospective calculation of the SAPS 3 score; however, data collection was mostly prospective. Finally, second rather than first-level customization may have provided a better fit of the prognostic models; however, the ease of first-level customization, using only original models rather than their components, favours this method and provides a practical approach to improving calibration in various ICUs.

We conclude that, in this group of surgical ICU patients, the performance of the SAPS 3 score was similar to that of SAPS II and both had better discrimination than the APACHE II score. Customization improved calibration of all prognostic models. Since SAPS 3 can be calculated within the first hour of admission to the ICU, it could be helpful in triage.

Supplementary material

Supplementary material is available at *British Journal of Anaesthesia* online.

Appendix 1

Logistic regression 'logit' equations [probability of death$=e^{logit}/(1+e^{logit})$] for calculating the probability of hospital death. SAPS, Simplified Acute Physiology Score; C-SAPS II, SAPS II customized to local study population; C-SAPS 3 (Eu), SAPS 3 customized for central and western Europe; C-SAPS 3, SAPS 3 customized for local study population; APACHE, Acute Physiology and Chronic Health Evaluation; Adj-APACHE II, adjusted probability of death according to the diagnostic category of the APACHE II score; C-APACHE II, APACHE II customized to local study population

Score	Logit equation
SAPS II	$-7.7631+0.0737\times$(SAPS II)$+0.9971\times$ln(SAPS II$+1$)
C-SAPS II	$-4.101+0.046\times$(SAPS II probability of death)
APACHE II	$-3.517+$(APACHE II)$\times0.146$
Adj-APACHE II	$-3.517+$(APACHE II)$\times0146+$diagnostic category weight
C-APACHE II	$-5259+0053\times$(APACHE II probability of death)
SAPS 3	$-32.6659+$ln(SAPS 3 score$+20.5958)\times7.3068$
C-SAPS 3 (Eu)	$-36.0877+$ln(SAPS 3 score$+22.2655)\times7.9867$
C-SAPS 3	$-4.069+0.057\times$(SAPS 3 probability of death)

References

1 Moreno R, Matos R. New issues in severity scoring: interfacing the ICU and evaluating it. *Curr Opin Crit Care* 2001; **7**: 469–74
2 Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med* 1981; **9**: 591–7
3 Vassar MJ, Lewis FR, Jr, Chambers JA, *et al.* Prediction of outcome in intensive care unit trauma patients: a multicenter study of Acute Physiology and Chronic Health Evaluation (APACHE), Trauma and Injury Severity Score (TRISS), and a 24-hour intensive care unit (ICU) point system. *J Trauma* 1999; **47**: 324–9
4 Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *J Am Med Assoc* 1993; **270**: 2957–63
5 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
6 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
7 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
8 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
9 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; **19**: 453–73
10 Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol* 1992; **45**: 85–9
11 Harrell FE, Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; **3**: 143–52
12 Bosman RJ, Oudemane van Straaten HM, Zandstra DF. The use of intensive care information systems alters outcome prediction. *Intensive Care Med* 1998; **24**: 953–8
13 Suistomaa M, Kari A, Ruokonen E, Takala J. Sampling rate causes bias in APACHE II and SAPS II scores. *Intensive Care Med* 2000; **26**: 1773–8
14 Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland—II: outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *Br Med J* 1993; **307**: 977–81
15 Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland—I: variations in case mix of adult admissions to general intensive care units and impact on outcome. *Br Med J* 1993; **307**: 972–7
16 Pappachan JV, Millar B, Bennett ED, Smith GB. Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. *Chest* 1999; **115**: 802–10
17 Livingston BM, MacKirdy FN, Howie JC, Jones R, Norrie JD. Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med* 2000; **28**: 1820–7
18 Markgraf R, Deutschinoff G, Pientka L, Scholten T. Comparison of acute physiology and chronic health evaluations 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
19 Apolone G, Bertolini G, D'Amico R, *et al.* 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* 1996; **22**: 1368–78

20 Moreno R, Miranda DR, Fidler V, Van Schilfgaarde R. Evaluation of two outcome prediction models on an independent database. *Crit Care Med* 1998; **26**: 50–61

21 Moreno R, Apolone G, Miranda DR. Evaluation of the uniformity of fit of general outcome prediction models. *Intensive Care Med* 1998; **24**: 40–7

22 Metnitz PG, Vesely H, Valentin A, *et al.* Evaluation of an interdisciplinary data set for national intensive care unit assessment. *Crit Care Med* 1999; **27**: 1486–91

23 Beck DH, Smith GB, Pappachan JV, Millar B. 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