

APACHE II and SAPS II are Poorly Calibrated in a Hong Kong Intensive Care Unit

I K S Tan,**MRCP (UK), FFICANZCA, FANZCA*

Abstract

This study seeks to determine if the Acute Physiology and Chronic Health Evaluation II (APACHE II) and the new Simplified Acute Physiology Score (SAPS II) model predictions are well calibrated in our adult Intensive Care Unit (ICU) patient population. 1064 successive ICU discharges were enrolled with 222 deaths at hospital discharge. APACHE II predicted 287.44 deaths, thus giving an APACHE II standardised mortality ratio (SMR) of 0.77 (95% confidence interval ± 0.07). SAPS II predicted 269.59 deaths, giving a SAPS II SMR of 0.82 (95% confidence interval ± 0.07). The Hosmer-Lemeshow goodness-of-fit χ^2 statistic was 43.96 for APACHE II and 49.06 for SAPS II ($P < 0.001$ for both, inferring significant departures from the null hypotheses of good calibration). The area under the Receiver Operating Characteristic plot for SAPS II risk of death was 0.87 (95% confidence interval ± 0.028) while that for APACHE II risk of death was 0.88 (95% confidence interval ± 0.026). Although the APACHE II and SAPS II models provide good discriminatory performance, this study finds the APACHE II and SAPS II models to be poorly calibrated in that they over-predict mortality in our ICU population.

Ann Acad Med Singapore 1998; 27:318-22

Key words: Critical care, Mortality prediction, Severity of illness index, Validation

Introduction

The Acute Physiology and Chronic Health Evaluation II (APACHE II)¹ and the new Simplified Acute Physiology Score (SAPS II)² scoring systems are widely used prognostic tools in Intensive Care Units (ICUs). These scoring systems calculate a “risk of death” (ROD) for ICU patients while adjusting for case-mix with severity of illness, chronic health, age and, in the case of the APACHE II model, diagnosis. For meaningful comparison of performance between ICUs, or over time in the same ICU, these factors must be taken into account. It remains unclear if the models developed in European and North American ICUs are applicable in Asian ICUs in the 1990s. Application of these models requires an assessment of how well the models are calibrated, i.e. the degree of agreement between predicted and observed mortality rates. This study seeks to determine if APACHE II and SAPS II model predictions are well calibrated in our patient population.

Materials and Methods

The Pamela Youde Eastern Hospital is an acute general hospital with 1600 beds, serving a population of approximately 0.7 million. The ICU admits approximately 700 patients annually. It is a “general adult” ICU that admits patients aged above two years from all

specialities of the hospital. The ICU does not receive patients post-cardiac surgery, or transplantation. Administratively the ICU is a “closed” unit in which the ICU medical staff have full and primary responsibility for patient care, and decide all admissions and discharges.

All patients admitted to the ICU from 1 October 1995 to 30 June 1997 were entered into the study. APACHE II and SAPS II data were collected by four medical officers rostered solely to the ICU, and verified by the author. Approval from the Ethics Committee or informed consent was not required as the data were collected as part of a routine quality assurance. Data definitions follow those of the Australian and New Zealand Intensive Care Society Patient Database.³ SAPS II calculates for each patient a risk of hospital death (ROD) based on 12 physiological variables, age, type of admission and underlying disease.² APACHE II calculates ROD based on 12 physiological variables, age, type of admission, underlying chronic health and diagnosis.¹ Data were entered and scored using a spreadsheet software (Microsoft Excel). The worst physiological values within the first 24 hours of ICU admission were used. Missing physiological variables were assumed to be normal, attracting no scores. Patients were followed up in hospital until 1 September 1997. Patients transferred to

* Director

Intensive Care Unit

Pamela Youde Eastern Hospital, Hong Kong SAR, People's Republic of China

Address for Reprints: Dr I K S Tan, Intensive Care Unit, Pamela Youde Nethersole Eastern Hospital, Chai Wan, Hong Kong SAR, People's Republic of China. e-mail: iantan@cuhk.edu.hk

another hospital, or still remaining in hospital at this date, were excluded. Data from patients less than 12 years of age were excluded. Data from patients re-admitted to ICU within the same hospitalisation were excluded (only the first admission was analysed), to prevent over-counting of outcome. Status on hospital discharge (dead or alive) was used as outcome. The predicted death rate for the group is obtained by summing the individual ROD and dividing by the number of patients.¹

Model fit (calibration) was analysed using the Hosmer-Lemeshow statistics H and C.⁴ The Hosmer-Lemeshow H and C statistics are calculated by summing the χ^2 values for deaths and survivors in each decile of predicted mortality. The statistic tests the null hypothesis that mortality is accurately predicted across the range of risk (goodness-of-fit). The standardised mortality ratio (SMR) is the number of observed hospital deaths divided by the number of predicted hospital deaths, and is crude summary measure of model calibration. Calibration is poor when H and C statistics depart significantly ($P < 0.01$) from the null hypotheses of good fit, and when the 95% confidence intervals for the SMR do not include 1. Calibration curves plot the observed to predicted deaths over the deciles of risk. Model discrimination was examined using Receiver Operating Characteristic (ROC) plots (Kurt Rossmann Laboratories, University of Chicago).⁵ The ROC plot is a graphic representation of true and false positive fractions associated with varying the decision criterion for prediction from 0% to 100%. The area under the ROC plot is a summary measure of the discriminative ability of the model,⁶ ranging from 0.5 (representing chance performance) to 1.0 (representing perfect prediction), and is considered "good" if > 0.8 .

Results

A total of 1064 consecutive ICU discharges were enrolled. There were 59 re-admissions, 9 patients were less than 12 years of age, 130 patients had been transferred to rehabilitation hospitals on discharge and 7 patients remained in hospital at the time of analysis. All these patients were excluded leaving 859 hospital discharges for analysis. The mean age was 62 years and 57% of patients were male. The median (mean) APACHE II score and SAPS II score were 17 (19) and 35 (39), respectively. The median length of stay was 2 days. Sixty-seven per cent of the patients received mechanical ventilation, 58% of the patients were postoperative, and 26% were elective (planned) postoperative admissions. The diagnostic categories of the patients on ICU admission are shown in Table I. Compared to patients who died, survivors had lower APACHE II ROD (0.22 ± 0.22 vs 0.66 ± 0.26 , $P < 0.001$), and lower SAPS II ROD (0.20 ± 0.24 vs 0.65 ± 0.30 , $P < 0.001$).

There were 222 deaths (26%) at hospital discharge.

APACHE II predicted 287.44 deaths, thus giving an APACHE II SMR of 0.77 (95% confidence interval ± 0.073). SAPS II predicted 269.59 deaths, giving a SAPS II SMR of 0.82 (95% confidence interval ± 0.071). The calibration curves of APACHE II and SAPS II are shown in Figure 1. The Hosmer-Lemeshow goodness-of-fit H and C statistics departed significantly ($P < 0.001$) from the null hypotheses of good fit for both APACHE II (C statistic = 43.96, H statistic = 53.31) and SAPS II (C statistic = 49.06, H statistic = 38.63). Tables II and III are goodness-of-fit tables for the Hosmer-Lemeshow C statistic for APACHE II and SAPS II, respectively. Figures 2 and 3 depict the ROC plot for the APACHE II and SAPS II ROD respectively in their prediction of hospital mortality. The area under the ROC plot for SAPS II ROD was 0.87 (95% confidence interval ± 0.028) while that for APACHE II ROD was 0.88 (95% confidence interval ± 0.026). These areas were not significantly different from one another ($P = 0.477$).

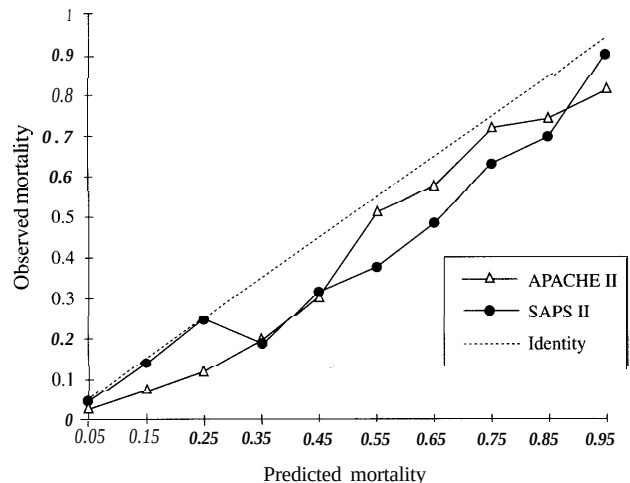


Fig. 1. Calibration curves for APACHE II and SAPS II.

TABLE I: DIAGNOSTIC CATEGORIES ON ADMISSION (RE-ADMISSIONS AND PAEDIATRIC PATIENTS EXCLUDED)

Non-operative admissions		Postoperative admissions	
	Number (%)		Number (%)
Cardiovascular	47 (4.7%)	Cardiovascular	17 (1.7%)
Respiratory	91 (9.1%)	Respiratory	38 (3.8%)
Gastrointestinal	79 (7.9%)	Gastrointestinal	224 (22.5%)
Neurological	62 (6.2%)	Neurological	185 (18.6%)
Sepsis	57 (5.7%)	Trauma	45 (4.5%)
Trauma	38 (3.8%)	Renal	23 (2.3%)
Metabolic	18 (1.8%)	Gynaecological	14 (1.4%)
Haematological	(0.1%)	Orthopaedic	29 (2.9%)
Renal	17 (1.7%)		
Other	11 (1.1%)		
Total	421 (42%)	Total	575 (58%)

TABLE II: GOODNESS-OF-FIT TABLE FOR APACHE II

Deciles of risk (%)	Number	Observed deaths	Expected deaths	Observed survivors	Expected survivors
0.000 - 0.031	86	2	1.50	84	84.50
0.032 - 0.062	86	3	3.88	83	82.12
0.063 - 0.103	86	2	6.89	84	79.11
0.104 - 0.150	86	6	10.93	80	75.07
0.151 - 0.223	86		15.78	77	70.22
0.224 - 0.322	86	12	23.39	74	62.61
0.323 - 0.452	86	18	33.27	68	52.73
0.453 - 0.657	86	42	48.47	44	37.53
0.658 - 0.846	86	57	65.04	29	20.97
0.847 - 0.991	85	71	78.29	14	6.71
Total	859	222	287.44	637	571.57

Hosmer-Lemeshow $C = 43.96$, 8 df ($P < 0.001$)

TABLE III: GOODNESS-OF-FIT TABLE FOR SAPS II

Deciles of risk (%)	Number	Observed deaths	Expected deaths	Observed survivors	Expected survivors
0.000 - 0.015	86	2	0.66	84	85.34
0.016 - 0.029	86	0	1.82	86	84.18
0.030 - 0.052	86		3.54	79	82.46
0.053 - 0.079	86	5	5.58	81	80.43
0.080 - 0.167	86	8	10.19	78	75.81
0.168 - 0.285	86	19	19.68	67	66.33
0.286 - 0.484	86	20	32.92	66	53.09
0.485 - 0.661	86	37	49.40	49	36.61
0.662 - 0.861	86	51	65.78	35	20.22
0.862 - 0.996	85	73	80.05	12	4.95
Total	859	222	269.62	637	589.42

Hosmer-Lemeshow $C = 49.06$, 8 df ($P < 0.001$)

Discussion

Intensive care units are expensive and, despite the outlay in resources, mortality of patients admitted to ICUs remains high. It is therefore necessary to quantify its effectiveness and mortality prediction systems are widely used quality assurance tools in ICUs. It is important that these systems are independently validated before they are introduced into each ICU or region. Validation of the predictive model requires analysis of two aspects: how close the observed mortality rate is to the predicted mortality rate over different mortality strata (calibration), and the accuracy with which patients can be classified into the outcome groups based on whether the predicted mortality exceeds a given threshold (discrimination). For application in a new setting and in quality of care assessments, calibration rather than discrimination assumes primary importance.

The performance of the APACHE II and SAPS II systems have been reported from a number of ICUs around the world. Calibration of APACHE II was found to be better than APACHE III (an updated version of

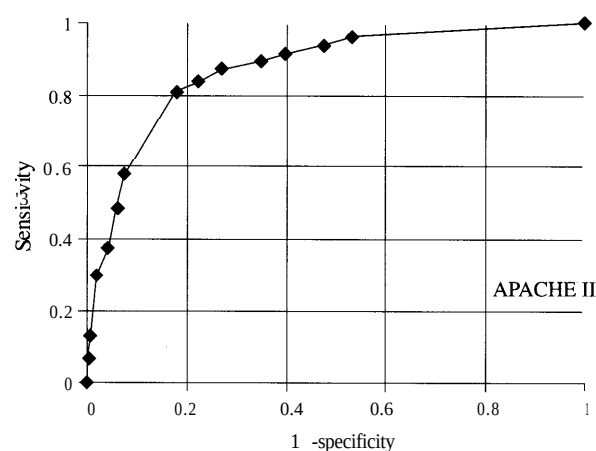


Fig. 2. Receiver Operating Characteristic plot for APACHE II risk of death in its prediction of hospital outcome.

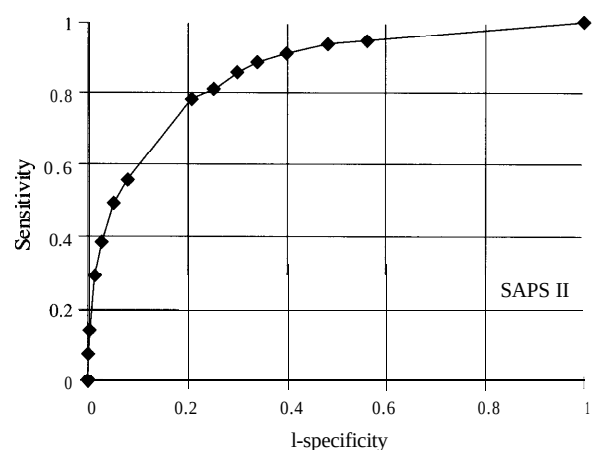


Fig. 3. Receiver Operating Characteristic plot for SAPS II risk of death in its prediction of hospital outcome.

APACHE II) in a cohort of 1144 British patients.⁷ The GiViTI study of 2202 patients admitted to 99 Italian ICUs found the calibration of SAPS II to be poor.⁸ An Australian study of 645 patients found calibration of APACHE II and SAPS II to be unsatisfactory.⁹ A large Hong Kong APACHE II series (1573 patients) found good discriminating ability but did not analyse calibration.¹⁰ There has not been a large Asian study analysing calibration of ICU outcome predictive models. The present study finds the discriminating ability of the model to be good, but the calibration of APACHE II and SAPS II to be poor. On inspection of the calibration curves, APACHE II over-predicted mortality across the range of risk, while SAPS II over-predicted mortality in the mid-to-higher ranges of mortality risk.

Fundamentally, what it means to be an “ICU”, which patients are admitted into this location, and what sort of hospital the ICU finds itself in, varies from country to country and within the same country. It was found that removing patients with given characteristics did not

influence the discrimination or calibration of an ICU predictive model, while increasing the frequency of patients with particular characteristics will cause the discrimination and calibration of the model to deteriorate." That is, the model prediction will be inaccurate in "speciality" ICUs that admit primarily patients with certain pathologies or characteristics. A preponderance of patients admitted for cardiovascular pathologies is commonly acknowledged to lower the SMR, and patients post-coronary artery bypass grafts are specifically excluded from analyses of this kind. Our study involves a "general" ICU in an acute "general" hospital, and the diagnostic categories of patients admitted to the ICU do not appear to show a great preponderance of any particular patient group.

ICU admission and discharge policies will affect the model's calibration.¹² Admitting patients with extremes of illness severity, whether "too well" or "too sick" to benefit from ICU care, will tend to cause the SMR to regress to 1, and is a waste of resources. An ICU admission policy designed to admit patients for whom care in the ICU will make a difference in outcome would be a factor affecting model calibration. Our ICU has a defined admission and discharge policy. Although it is likely that this will bias the SMR, its contribution to the total error in the calibration of APACHE II and SAPS II is unclear.

Poor data collection may lead to poor calibration. A previous study found that up to 20% of patients may have an error in data collection.¹³ In general, random rather than systematic error results. Applying a specific set of data definitions can lead to systematic error if the definitions deviate from that used in the original studies. The data definitions of the Australasian Intensive Care Database Project are derived from original studies. Applying a specific diagnosis (reason for ICU treatment) is required for APACHE II scoring. However, the similarity of the APACHE II results with the SAPS II results, which does not require applying a diagnosis, implies that this is not a significant source of calibration error.

It may be argued that a poor calibration can be obtained by increasing sample size despite excellent goodness-of-fit, since the power for detecting departures from the null hypothesis increases. However, the validation database for SAPS II involved 4628 patients and obtained a *P* value of 0.104 for the goodness-of-fit statistic. Lead-time bias may also be a source of calibration error. However, APACHE III, which includes a component contributed by patient location prior to ICU admission, found this factor to contribute only 1% as a direct explanatory factor.¹⁴

Finally, our poor calibration may reflect modern ICU performance rather than poor model performance. Using APACHE III, SMRs varied from 0.67 to 1.25 across

the validation database of 42 ICUs in the United States. It was stated that the number of hospitals with SMRs significantly different from 1 is far higher than would have occurred randomly by chance.¹⁵ It is believed that ICUs with low (good) SMRs have superior organisation and technology. In particular, the transition from an "open" ICU to a "closed" ICU, where full-time intensivists control admission, discharge, and clinical management of all patients in the ICU, has been repeatedly shown to improve ICU outcomes.¹⁷ Since APACHE II is derived from a 1979 to 1982 North American database, it is likely that current levels of management and technology should allow better ICU performance than predicted by APACHE II. Some suggestion that this has occurred was found when prognoses from multi-organ failure were re-examined.¹⁸ This may explain the over-prediction of mortality for patients in the mid-to-higher risks of death in our study.

Although the APACHE II and SAPS II provide good discriminatory performance, this study finds the APACHE II and SAPS II models to be poorly calibrated in that they over-predict mortality in our ICU population. It is important for each region with similar patient case-mixes and similar ICU resources to derive its own customised probability model from its own regional database. Furthermore, these predictive models require regular calibration with time as patients' profile, disease spectra and treatment evolved. This will allow the mortality prediction system to better analyse quality of care among ICUs in the target population.

REFERENCES

1. Knaus W A, Draper E A, Zimmerman J E. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13:818-29.
2. LeGall J R, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European / North American multicenter study. *JAMA* 1993; 270:2957-63.
3. Australasian Intensive Care Database Project. Data Dictionary. Australian and New Zealand Intensive Care Society Patient Database 1995.
4. Lemeshow S, Hosmer D W. A review of goodness-of-fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 1982; 115:92-106.
5. Metz C E. Basic principles of ROC analysis. *Semin Nucl Med* 1978; 8:283-98.
6. Zwieg M H, Campbell G. Receiver Operating Characteristic (ROC) plots: a fundamental evaluating tool in clinical medicine. *Clin Chem* 1993; 39:561-7.
7. Beck D H, Taylor B L, Millar B, Smith G B. Prediction of outcome from intensive care: a prospective cohort study comparing Acute Physiology and Chronic Health Evaluation II and III prognostic systems in a United Kingdom Intensive Care Unit. *Crit Care Med* 1997; 25:9-15.
8. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, et al. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GIVI. *Intensive Care Med* 1996; 22:1368-78.
9. Finnis M, Moran J, Leppard I, Herman B, Grealy B, Peisach A. A comparison of outcome prediction algorithms in an adult intensive care unit. *Anaesth Intensive Care* 1997; 25:555-6.

10. Oh T E, Hutchison R, Short S, Buckley T, Lin E, Leung D. Verification of the Acute Physiology and Chronic Health Evaluation scoring system in a Hong Kong intensive care unit. *Crit Care Med* 1993; 21:698-706.
 11. Murphy-Filkins R L, Teres D, Lemeshow S, Hosmer D W. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: How to distinguish a general from a speciality intensive care unit. *Crit Care Med* 1996; 24:1968-73.
 12. Angus D C, Sirio C A, Clermont G, Bion J. International comparisons of critical care outcome and resource consumption. *Crit Care Clin* 1997; 13:389-407.
 13. Holt A W, Bury L K, Bersten A D, Skowronski G A, Vedig A K. Prospective evaluation of residents and nurses as severity score collectors. *Crit Care Med* 1992; 20:1688-91.
 14. Knaus W A, Wagner D P, Draper E A, Bergner M, Bastos P G, Sirio C A, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100:1619-36.
 15. Knaus W A, Wagner D P, Zimmerman J E, Draper E A. Variations in mortality and length of stay in intensive care units. *Ann Intern Med* 1993; 118:753-81.
 16. Shortell S M, Zimmerman J E, Rousseau D M, Gillies R R, Wagner D P, Draper E A, et al. The performance of intensive care units: Does good management make a difference? *Med Care* 1994; 32:508-17.
 17. Carson S S, Stocking C, Podsadecki T, Christenson J, Pohlman A, MacRae S, et al. Effects of organizational change in the medical intensive care unit of a teaching hospital: a comparison of 'open' and 'closed' formats. *JAMA* 1996; 276:322-8.
 18. Zimmerman J E, Knaus W A, Wagner D P, Sun X, Hakim R B, Nystrom P. A comparison of risks and outcomes for patients with organ system failure: 1982-1990. *Crit Care Med* 1996; 24:1633-41.
-