VALIDATION OF THE LOD SCORE COMPARED WITH APACHE II SCORE IN PREDICTION OF THE HOSPITAL OUTCOME IN CRITICALLY ILL PATIENTS

Bodin Khwannimit

Division of Critical Care, Department of Internal Medicine, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand

Abstract. The Logistic Organ Dysfunction score (LOD) is an organ dysfunction score that can predict hospital mortality. The aim of this study was to validate the performance of the LOD score compared with the Acute Physiology and Chronic Health Evaluation II (APACHE II) score in a mixed intensive care unit (ICU) at a tertiary referral university hospital in Thailand. The data were collected prospectively on consecutive ICU admissions over a 24 month period from July1, 2004 until June 30, 2006. Discrimination was evaluated by the area under the receiver operating characteristic curve (AUROC). The calibration was assessed by the Hosmer-Lemeshow goodness-of-fit H statistic. The overall fit of the model was evaluated by the Brier's score. Overall, 1,429 patients were enrolled during the study period. The mortality in the ICU was 20.9% and in the hospital was 27.9%. The median ICU and hospital lengths of stay were 3 and 18 days, respectively, for all patients. Both models showed excellent discrimination. The AUROC for the LOD and APACHE II were 0.860 [95% confidence interval (CI) =0.838-0.882] and 0.898 (95% CI=0.879-0.917), respectively. The LOD score had perfect calibration with the Hosmer-Lemeshow goodness-of-fit H chi-2=10 (p=0.44). However, the APACHE II had poor calibration with the Hosmer-Lemeshow goodness-of-fit H chi-2=75.69 (p<0.001). Brier's score showed the overall fit for both models were 0.123 (95%CI=0.107-0.141) and 0.114 (0.098-0.132) for the LOD and APACHE II, respectively. Thus, the LOD score was found to be accurate for predicting hospital mortality for general critically ill patients in Thailand.

INTRODUCTION

In critical care, there are several severity scoring systems that have been developed to calculate scores and hospital outcome. The Acute Physiology and Chronic Health Evaluation II (APACHE II) score is a popular severity scoring system (Knaus *et al*, 1985). Multiple organ dysfunction (MOD) is one of the leading causes of death in intensive care unit (ICU) patients (Tran *et al*, 1993; Zimmerman *et al*, 1996). Therefore, scores which describe MOD should be able to assess and describe morbidity as well as mortality. There are several organ dysfunction scores, such as the Multiple Organ Dysfunction score (MOD) (Marshall *et al*, 1995) and Sequential Organ Failure Assessment (SOFA) (Vincent *et al*, 1996), that describe organ dysfunction and predict morbidity in critically ill patients. The logistic organ dysfunction score (LOD) differs from other organ failure scores in that it also allows calculation of predicted hospital mortality based on organ dysfunction on the day of ICU admission. Severity scores and organ dysfunction scores are used in many areas with critically ill patients, such as clinical research, to demonstrate equivalency between studied and

Correspondence: Bodin Khwannimit, Division of Critical Care, Department of Internal Medicine, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
Tel: +66 (074) 451452; Fax: +66 (074) 429385
E-mail: kbordin@medicine.psu.ac.th

control patients, clinical decision making, and resource allocation (Teres, 2004; Le Gall, 2005). Good discrimination using APACHE II has been found in several studies, however, most of these studies reported poor calibration (Rowan *et al*, 1993; Moreno and Morais, 1997; Tan, 1998; Katsaragakis *et al*, 2000; Livingston *et al*, 2000; Arabi *et al*, 2002; Harrison *et al*, 2006). The LOD score is easy to use and calculate for predicting hospital mortality and previous studies have reported good discrimination (Le Gall *et al*, 1996; Metnitz *et al*, 2001; Pettila *et al*, 2002; Timsit *et al*, 2002). It was not clear, however, to what extent these findings could be extrapolated to ICU patients in different ICUs or in different countries (Le Gall, 2005). Before applying severity scoring systems in a specific country or different type of ICU, their prognostic performance must be validated (Le Gall, 2005).

The aim of this study was to evaluate the ability of the LOD score, compared with APACHE II score, to predict hospital mortality in a mixed ICU at a tertiary referral university hospital in southern Thailand.

MATERIALS AND METHODS

This study was carried out in Songklanagarind Hospital, an 800-bed tertiary referral university teaching hospital at Prince of Songkla University, Hat Yai, Songkhla, Thailand. In our hospital, there are two units in the adult ICU: a ten-bed surgical ICU and a ten-bed mixed medical and coronary care unit. The surgical ICU serves all postoperative and trauma patients.

Data collection took place over a 24 month period from July1, 2004 until June 30, 2006. All the data were collected concurrently from consecutive ICU admissions. Patients who were excluded from this study included those who: were younger than 15 years of age, had coronary artery disease and cardiac surgery cases, suffered burn injuries, had not received attempted cardiac resuscitation, died within four hours of ICU admission or who stayed in the ICU less than 24 hours. If patients had been admitted more than once to the ICU during the study period, only the first admission was included. Approval for the project was obtained from the faculty's Ethics Committee.

The following data were collected according to Knaus *et al* (1985): basic demographic data including sex, age, the presence of any comorbidities and the principal diagnostic categories leading to ICU admission. In sedated patients, a Glasgow Coma Score (GCS) was determined either from medical records before sedation or through interviewing the physician who ordered the sedation. However, if a variable could not be measured the GCS was assumed to be normal (Le Gall *et al*, 1996). The predicted hospital mortality was calculated using the original formulas for the APACHE II and LOD scores (Knaus *et al*, 1985; Le Gall *et al*, 1996). Patients were followed up until hospital discharge in order to register their survival status.

Data are presented as mean±SD, when indicated. Student's *t*-test and Wilcoxon's rank sum test were used to compare normally distributed continuous variables and nonparametric data, respectively. The chi-squared statistic was used to test for the statistical significance of categorical variables. A p-value of less than 0.05 was considered statistically significant.

The ability and accuracy of the models to predict the hospital mortality were determined by examining their discrimination (the ability of the model to distinguish survivors from non-survivors), calibration (the accuracy of the estimated probability of survival) and overall fit. The discrimination was tested through the area under the receiver operating characteristic (AUROC) curve that was computed by a modification of the Wilcoxon statistic as described by McNeil and Hanley

(1984) and also a 2x2 classification table. An AUROC of one was perfect discrimination and an AUROC of 0.5 was random chance. The model has good discrimination when AUROC >0.8. The Hosmer-Lemeshow goodness-of-fit H statistic was used to evaluate calibration (Hosmer *et al*, 1997). Patients were rank-ordered into ten groups according to their probability of death to calculate the H statistic. A good fit was defined as p>0.05. A calibration curve was constructed by plotting the predicted mortality rate stratified by 10% intervals of mortality against the observed mortality rates. The overall fit of the model was assessed by Brier's score (Harrison *et al*, 2006). Brier's score was developed in relation to metrological forecasting, as an overall measure of accuracy. It is the mean square error between outcome and prediction. For perfect prediction, the Brier's score will be 0; for constant predictions of 0.5, each individual Brier's score will be 0.25. Statistical analysis was performed using Stata 7 software (Stata Corporation, College Station, Tx, USA).

RESULTS

A total of 1,429 patients were including during the study period. Overall, 299 patients (20.9%) died in the ICU and 399 patients (27.9%) died in the hospital. The patients' demographic characteristics, diagnostic categories and APACHE comorbidities for the patients are shown in Table 1. In comparison to patients who survived, the patients who died were nonoperative, post-cardiac arrest, had sepsis, gastrointestinal disease or patients with comorbidities. The survivor group had significantly more respiratory problems, were postoperative and no comorbidities. Age, gender and patients with neurological disease were not significantly different between the survivor and non-survivor groups. Severity of patient illness was assessed by the LOD score, APACHE II score and LOS for both the ICU and the hospital are shown in Table 2.

Table 1
Demographic and clinical characteristics of patients in the study.

Parameters	No. (%)
Male	794 (55.6)
Operative status	
Nonoperation	836 (58.5)
Elective operation	306 (21.4)
Emergency operation	287 (20.1)
Categories of diseases	
Non-operative	
Respiratory disease	110 (7.7)
Cardiovascular disease	218 (15.3)
Post-cardiac arrest	57 (4.0)
Sepsis	322 (22.5)
Neurological disease	37 (2.6)
Gastrointestinal disease	50 (3.5)
Other	42 (2.9)
Post-operative disease	
Brain and spinal cord	193 (13.5)
Gastrointestinal	124 (8.7)
Other	276 (19.3)
APACHE comorbidities	
Liver cirrhosis	30 (2.1)
Severe COPD	15 (1.1)
Chronic renal failure	28 (1.9)
Heart failure class IV	4 (0.3)
Hematologic malignancy	76 (5.3)
Metastasis carcinoma	48 (3.4)
Immunocompromised	46 (3.2)
AIDS	25 (1.8)
None of the above	1,157 (80.9)

COPD = Chronic obstructive pulmonary disease; AIDS = Acquired immune deficiency syndrome

The receiver operating characteristic (ROC) curves for both systems are shown in Fig 1. The AUROC for the LOD score was 0.860 (95% CI=0.838-0.882), and was 0.898 (95% CI=0.879-0.917) for the APACHE II score. The AUROC for the APACHE II score showed a significantly higher prediction rate than the LOD score did (p<0.001). The results of the 2x2 classification table for the LOD and APACHE II score are shown in Table 3.

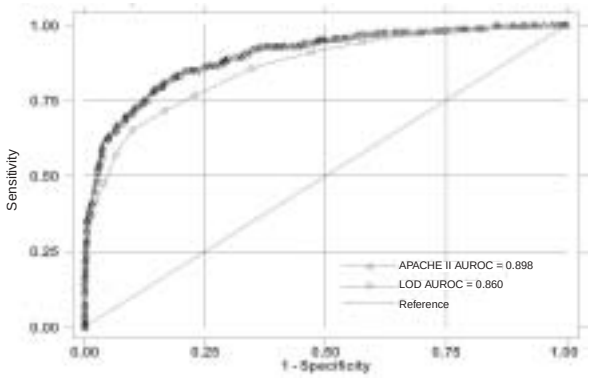

Fig 1–Receiver Operating Characteristic Curves (ROC) for LOD and APACHE II systems.

The LOD score showed a good fit, the Hosmer-Lemeshow goodness-of-fit H chi-2=10 (p=0.44). In contrast, the APACHE II score had a Hosmer-Lemeshow goodness-of-fit H chi-2= 75.69 (p<0.001). These findings indicated a significant lack of fit for the APACHE II score. Calibration curves for the LOD and APACHE II scores are shown in Fig 2. Overall, the calibration curve for observed mortality for the LOD score predicted mortality similar to the APACHE II score. The LOD score curve accurately predicted mortality in six groups of patients and overestimated mortality in other strata. The APACHE II score overestimated predicted mortality in all strata except at predicted hospital deaths more than 80%. The Brier's score showed an overall fit for both models of 0.123 (95%CI=0.107-0.141) and 0.114 (0.098-0.132) for the LOD and APACHE II scores, respectively.

DISCUSSION

This study, evaluated the validity of the LOD score compared to the APACHE II score to accurately predict hospital mortality in a Thai adult mixed-case ICU. The results show that both models had excellent discrimination and overall fit. However, only the LOD score had perfect calibration in predicting hospital deaths.

The discrimination of the APACHE II score was slightly better than the LOD score. The ability for both models to correctly predict group prognosis was also assessed by means of a 2x2 decision table. The results show the highest correct classification was obtained with a decision criterion of 50% for both models. The APACHE II score had a slightly higher

Table 2
LOD and APACHE II scores, predicted risk of hospital death and LOS of patients in this study.

	All (n = 1429)	Survivors (n = 1030)	Non-survivors (n = 399)	p-value
Age (years)	54.7±18.9	54.6±18.9	55.2±19.1	0.585
LOD score	5.3±4.3	3.7±2.8	9.5±4.5	<0.001
Acute physiologic score	15.9±9.4	12.4±6.3	25.4±9.5	<0.001
APACHE II score	19.7±9.9	15.7±6.9	29.9±9.5	<0.001
LOD prediction of hospital death (%)	29.4±28.7	18.3±18.0	58.1±31.2	<0.001
APACHE II prediction of hospital death (%)	35.5±29.3	23.1±19.8	67.4±25.3	<0.001
ICU LOS (day)[a]	3 (1-5)	3 (2-5)	3 (1-6)	0.719
Hospital LOS (day)[a]	18 (9-34)	21 (12-36)	8 (2-23)	<0.001

[a]Median and interquartile range; LOD = Logistic Organ Dysfunction; APACHE II = Acute Physiology and Chronic Health Evaluation II; LOS = length of stay

Table 3
Classification table of the LOD and APACHE II systems.

	LOD		APACHE II	
	Predicted to live (n)	Predicted to die (n)	Predicted to live (n)	Predicted to die (n)
Decision criterion 10%				
Observed survivors	436	594	349	681
Observed non-survivors	23	376	10	389
Sensitivity	94.24 (91.48-96.31)		97.49 (95.44-98.79)	
Specificity	42.33 (39.29-45.41)		33.88 (30.99-36.87)	
Positive predictive value	38.76 (35.68-41.91)		36.36 (33.47-39.32)	
Negative predictive value	94.99 (92.58-96.71)		97.71 (95.55-99.01)	
Overall correct classification	56.82 (54.21-59.41)		51.64 (49.02-54.26)	
Decision criterion 50%				
Observed survivors	932	98	901	129
Observed non-survivors	141	258	99	300
Sensitivity	64.68 (59.75-69.35)		75.19 (70.65-79.35)	
Specificity	90.44 (86.53-92.21)		87.48 (85.30-89.44)	
Positive predictive value	72.47 (67.52-77.05)		69.93 (65.35-74.24)	
Negative predictive value	86.86 (84.69-88.82)		90.10 (88.08-91.88)	
Overall correct classification	83.24 (81.16-85.11)		84.04 (82.04-85.91)	
Decision criterion 90%				
Observed survivors	1,025	5	1,028	2
Observed non-survivors	288	111	306	93
Sensitivity	27.73 (23.48-32.50)		23.31 (19.25-27.78)	
Specificity	99.52 (98.87-99.84)		99.81 (99.30-99.98)	
Positive predictive value	95.69 (90.23-98.59)		97.87 (92.60-99.74)	
Negative predictive value	78.07 (75.73-80.28)		77.06 (74.71-79.29)	
Overall correct classification	79.50 (77.31-81.56)		78.17 (75.93-80.28)	

In parenthesis are 95% CI; LOD = Logistic Organ Dysfunction; APACHE II = Acute Physiology and Chronic Health Evaluation II

overall percentage of correct classification at the decision criterion of 50% than the LOD score. However, the LOD score had a higher overall correct classification at a decision criterion of both 10% and 90% than did the APACHE II score. Other reports showed a lower correct classification of the APACHE II system, ranging from 77-85.5% (Rowan *et al*, 1993; Moreno and Morais, 1997; Tan, 1998; Katsaragakis *et al*, 2000; Arabi *et al*, 2002). There are no previous reports of the overall correct classification for the LOD score. The AUROC for both the scores in our study was higher than that found in previous reports. Previous reports showed an AUROC for the APACHE II score as 0.839 (standard error 0.02) in Greece (Katsaragakis *et al*, 2000), 0.787 (standard error 0.015) in Portugal (Moreno *et al*, 1997), 0.88 in Hong Kong (Tan, 1998), 0.83 in Saudi Arabi (Arabi *et al*, 2002), 0.805 in Scotland (Livingston *et al*, 2000), 0.83 in England and Ireland (Rowan *et al*, 1993) and 0.804 (95%CI=0.802-0.806) in England (Harrison *et al*, 2006). The AUROC for the ini-

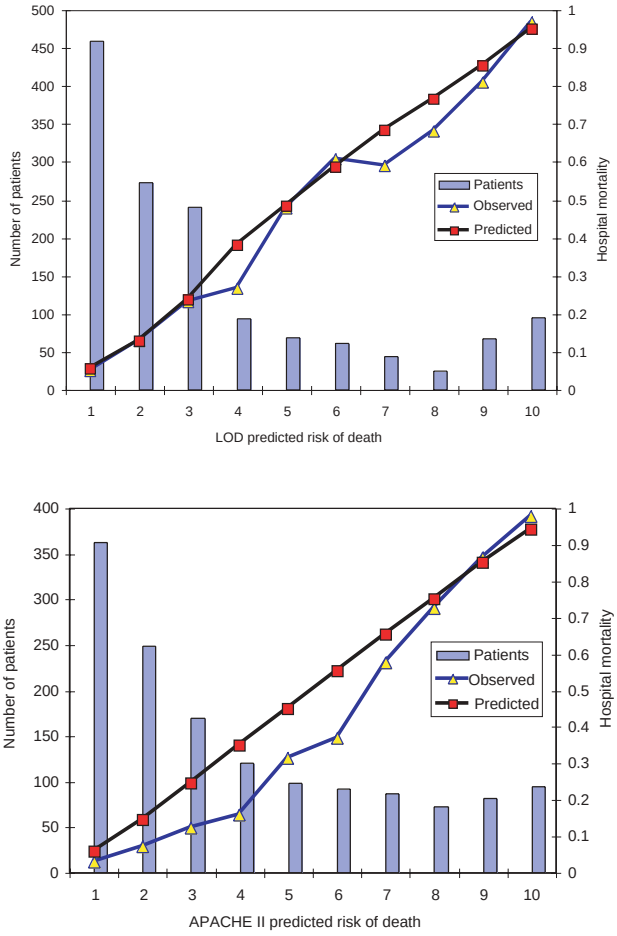

Fig 2–Calibration curves for LOD and APACHE II systems.

tial LOD score in other reports was 0.726-0.805 (Pettila *et al*, 2002; Timsit *et al*, 2002) and 0.843 in the original LOD system (Le Gall *et al*, 1996). Thus, both the LOD and APACHE II scores are able to discriminate outcomes in Thai ICU patients.

In this study only the LOD score accurately predicted hospital mortality. Overall, the poor calibration of the APACHE II score was the same as other previous studies (Moreno *et al*, 1997; Tan, 1998; Katsaragakis *et al*, 2000; Arabi *et al*, 2002). Nevertheless, the sample size had a major influence on the measured calibration when using the Hosmer-Lemeshow goodness-of-fit test (den Boer *et al*, 2005): small samples result in an apparently good fit and large sample result in a poor fit (Rowan *et al*, 1994; Zhu *et al*, 1996).

Potential reasons for the poor calibration may include: data collection and definitions, differences in the case-mix compared with other studies and the quality and policies of our ICU. The reliability of the data collected is important because poor data may influence predictions of mortality. Holt *et al* (1992) showed the main cause of data error in the APACHE II score is the inconsistent choice between the highest and lowest values for the acute physiologic score and the GCS. The variability of GCS determination in sedated patients may affect the predicted death in both models. Similar to other studies of the LOD score, we used the pre-sedation GCS in patients (Le Gall *et al*, 1996). The number and type of missing physiological variables might affect the prediction of mortality (Afessa *et al*, 2005). In this study, missing physiological variables were found for only 5.5% of APACHE II score compared to 13% of the cases in the original APACHE II model (Knaus *et al*, 1985), and none for the LOD score. All data collection was performed by a single research assistant and rechecked by the author. Therefore, the influence of data collection and definitions probably had minimal effect on the calibration for both models. The potential difference in case-mix between our database and the development database may have negative impact on calibration assessment.

This study had some limitations. First, as a single center study there may be bias concerning the case-mix, quality of ICU care and ICU policy. Secondly, the relatively small sample size was a relevant limiting factor in performing stratified analysis of calibration for both models. A multicenter study would have given fewer concerns over the case-mix and a better sample size. Finally, a single assessment of severity score and organ dysfunction

score within the first 24 hours of ICU admission is not accurate in patients with a long ICU stay. The severity score showed an acceptable accuracy only in patients with a brief ICU stay (Lemeshow *et al*, 1994; Sicignano *et al*, 1996).

In conclusion, this study demonstrates the LOD and APACHE II scores show excellent discrimination and overall fit; however, only the LOD score had a perfect calibration. Thus, the LOD score is suitable for predicting hospital mortality in general critically ill patients in a Thai ICU. Periodic reassessment is beneficial to ensure calibration is maintained.

ACKNOWLEDGEMENTS

This study was supported through a Faculty of Medicine research grant, Prince of Songkla University.

REFERENCES

Afessa B, Keegan MT, Gajic O, Hubmayr RD, Peters SG. The influence of missing components of the Acute Physiology Score of APACHE III on the measurement of ICU performance. *Intensive Care Med* 2005; 31: 1537-43.

Arabi Y, Haddad S, Goraj R, Al-Shimemeri A, Al-Malik S. Assessment of performance of four mortality prediction systems in a Saudi Arabian intensive care unit. *Crit Care* 2002; 6: 166-74.

Den Boer S, de Keizer NF, de Jonge E. Performance of prognostic models in critically ill cancer patients - a review. *Crit Care* 2005; 9: R458-63.

Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med* 2006; 34: 1378-88.

Holt AW, Bury LK, Bersten AD, Skowronski GA, Vedig AE. Prospective evaluation of residents and nurses as severity score data collectors. *Crit Care Med* 1992; 20: 1688-91.

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997; 16: 965-80.

Katsaragakis S, Papadimitropoulos K, Antonakis P, Strergiopoulos S, Konstadoulakis MM, Androulakis G. Comparison of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) scoring systems in a single Greek intensive care unit. *Crit Care Med* 2000; 28: 426-32.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818-29.

Le Gall JR. The use of severity scores in the intensive care unit. *Intensive Care Med* 2005; 31: 1618-23.

Le Gall JR, Klar J, Lemeshow S, *et al*. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA* 1996; 276: 802-10.

Lemeshow S, Klar J, Teres D, *et al*. Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. *Crit Care Med* 1994; 22: 1351-8.

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.

Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. *Crit Care Med* 1995; 23: 1638-52.

McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making* 1984; 4: 137-50.

Metnitz PG, Lang T, Valentin A, Steltzer H, Krenn CG, Le Gall JR. Evaluation of the logistic organ dysfunction system for the assessment of organ dysfunction and mortality in critically ill patients. *Intensive Care Med* 2001; 27: 992-8.

Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. *Intensive Care Med* 1997; 23: 177-86.

Pettila V, Pettila M, Sarna S, Voutilainen P, Takkunen O. Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill. *Crit Care Med* 2002; 30: 1705-11.

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. *BMJ* 1993; 307: 977-81.

Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. *Crit Care Med* 1994; 22: 1392-401.

Sicignano A, Carozzi C, Giudici D, Merli G, Arlati S, Pulici M. The influence of length of stay in the ICU on power of discrimination of a multipurpose severity score (SAPS). ARCHIDIA. *Intensive Care Med* 1996; 22: 1048-51.

Tan IK. APACHE II and SAPS II are poorly calibrated in a Hong Kong intensive care unit. *Ann Acad Med Singapore* 1998; 27: 318-22.

Teres D. The value and limits of severity adjusted mortality for ICU patients. *J Crit Care* 2004; 19: 257-63.

Timsit JF, Fosse JP, Troche G, *et al*. Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting hospital mortality in critically ill patients. *Crit Care Med* 2002; 30: 2003-13.

Tran DD, Cuesta MA, van Leeuwen PA, Nauta JJ, Wesdorp RI. Risk factors for multiple organ system failure and death in critically injured patients. *Surgery* 1993; 114: 21-30.

Vincent JL, Moreno R, Takala J, *et al*. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; 22: 707-10.

Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. *Crit Care Med* 1996; 24: 57-63.

Zimmerman JE, Knaus WA, Wagner DP, Sun X, Hakim RB, Nystrom PO. A comparison of risks and outcomes for patients with organ system failure: 1982-1990. *Crit Care Med* 1996; 24: 1633-41.