Bangladesh Crit Care J March 2013; 1: 27-32

Original Article

A Comparison of Severity Systems APACHE II and SAPS II in Critically ill Patients

Mohammad Omar Faruq[1], Mohammad Rashed Mahmud[2], Tanjima Begum[3], ASM Areef Ahsan[4], Kaniz Fatema[5], Fatema Ahmed[6], Md Rezaul Karim[7]

Abstract

Objective: *To assess the performance of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) in Bangladeshi critically ill patients.*

Material and Method: *Prospective observational cohort study conducted between January 1, 2008 and December 31, 2008 in the Intensive Care Unit (ICU) of BIRDEM Hospital, an 600-beds tertiary referral Postgraduate hospital and October to December 2008 in ICU, Ibn Sina Hospital Dhaka.*

Results: *One hundred ninety four patients were enrolled. There were 58 deaths (42.65%) at ICU discharge. APACHE II and SAPS II predicted hospital mortality 35.32 ± 21.81and 37.11 ± 27.34 respectively. Both models showed excellent discrimination. The overall discriminatory capability, as measured by the aROC, was generally good for two models and ranged from 0.78 to 0.89. APACHE II is slightly better compared to SAPS II score but not significantly better than SAPS II. Both systems exhibited good calibration (= 8.304, p = 0.40 for APACHE II, = 9.040, p = 0.34 for SAPS II). Hosmer-Lemeshow goodness-of-fit test revealed a good performance for APACHE II scores.*

Conclusion: *APACHE II provided better performance than SAPS II in predicting mortality in our ICU patients but SAPS II also performed well. Our observed mortality was similar with the predicted mortality from APACHE II and SAPS II scores, which suggests that the result of this study reveals good intensive care quality.*

Key Word: *Severity of illness, Intensive care, Mortality prediction, Acute Physiology and Chronic Health Evaluation (APACHE II), Simplified Acute Physiology Score (SAPS II)*

Introduction:

In recent decades the emphasis on developing systems to measure the severity of illness in the intensive care units (ICUs) has increased. Several models have been made for mortality prediction in critically ill patients[1,2]. By using these indices, in addition to making decisions about the cost effectiveness of these services[3,4] and assess the performance of different ICUs [5], evaluation of the results of new treatments and technologies is also possible.

The main reasons that augmented the importance of these scoring systems are: 1- the scoring systems are used in clinical trials for matching, 2- these systems are used to quantify the severity of illness for the administrative decisions such as resource allocation, 3- the scoring systems assess the ICU performance, and compare the quality of care; and 4- they are used to appraise the prognosis of individual patients [6].

The APACHE II and SAPS II systems can be used to calculate the individual risk of hospital death by converting the score into probability of death using logistic regression. Acute Physiology and Chronic Health Evaluation (APACHE) II and Simplified Acute Physiology Score (SAPS) II measure severity of illness by a numeric score based on physiologic variables selected because of their impact on mortality: the sicker the patient, the more deranged the values and the higher the score. The numeric scores are then converted into predicted mortality by using a logistic regression formula developed and validated on populations of ICU patients.

One interesting characteristics of APACHE II and SAPS II is that they propose to create homogeneous patients'

1. Dr. Mohammad Omar Faruq, MD, FACP, FACEP, FCPS (Med), Professor of Critical Care Medicine, BIRDEM General Hospital, Dhaka, Bangladesh
2. Dr. Mohammad Rashed Mahmud, MBBS, MPH, ICU Incharge, Ibn Sina Hospital Sylhet Ltd
3. Tanjima Begum, Statistitian, BIRDEM General Hospital
4. ASM Areef Ahsan, MBBS, FCPS, MD (Chest), MD (CCM), Asso. Professor, Department of Critical Care Medicine, BIRDEM General Hospital & Consultant, Intensive Care Unit, Ibn Sina Hospital, Dhanmondi, Dhaka.
5. Dr. Kaniz Fatema, MBBS, FCPS, Assistant Professor, Department of Critical Care Medicine, BIRDEM General Hospital, Dhaka, Bangladesh
6. Dr. Fatema Ahmed, MBBS, FCPS, Junior Consultant, Department of Critical Care Medicine, BIRDEM General Hospital & Consultant, Intensive Care Unit, Renaissance Hospital & Research Institute Ltd., Dhanmondi, Dhaka.
7. Dr. Md Rezaul Karim, ICU Incharge, Ibn Sina Hospital, Dhaka

Corresponding Author : Dr. Mohammad Rashed Mahmud, ICU Incharge, Ibn Sina Hospital, Sylhet Ltd. Subhanighat Point, Sylhet, Bangladesh, Phone: Office- 0821-2832735-43, ext-1806, Cell: +88 01712 614026, +88 01612 614026, Email: rashedsb28@gmail.com, rashedsb28@yahoo.com

categories from an inhomogeneous patients' case mix. These models try to avoid patient selection bias by including consecutive admission to the ICU in the development database. However one cannot exclude that some specific patient diagnoses have more weight than other and hence influence the outcome prediction. In addition, these models exclude some subgroups from analysis.

A variety of statistical methods are used to compare predictions with actual outcomes. Two principal considerations, namely, discrimination and calibration, should be taken into account during the validation process of a scoring system. Discrimination defines how well the model discriminates between patients who are likely to either die or not die; calibration refers to the correlation between the predicted and the actual outcome for the entire range of risk. The discriminating ability of the model can often be expressed by the area under the ROC curve. As this area approaches 1.0, the model becomes more 'perfect'; as the performance of the model becomes more random, the area under the curve trends towards 0.5. Calibration of the scoring system is often assessed by the use of Lemeshow and Hosmer's goodness-of-fit statistics.

Both systems have been evaluated in many population samples individually, in comparison to each other, or in comparison to other scoring systems. Their accuracy and predictive ability have also been tested in subgroups of critically ill patients, such as patients experiencing surgery, head trauma, and myocardial infarction. Using the severity of illness scoring systems has not been common in the Bangladeshi ICUs. In this study we evaluated the performance of two scoring systems: APACHE II and SAPS II on a sample of Bangladeshi patients in the ICUs of a postgraduate and referral hospitals in Dhaka.

Materials and Methods:

The research design of this study was non-experimental, a descriptive comparative prospective cohort study. The main purpose was to describe the performance of APACHE II and SAPS II to predict the probability of mortality in a well defined ICU patient cohort in Bangladesh.

We identified all individuals over 16 years old that had been admitted to the ICU at BIRDEM Hospital in January to December 2008 and at Ibn Sina Hospital ICU in October to December 2008. A total of 264 ICU admissions were identified in BIRDEM hospital and 64 admissions in Ibn Sina Hospital; however, we included only first-time ICU admissions, and did not include coronary care patients and cardiac surgery or other patients admitted for planned post-operative observations for less than 24 hours as defined in the original models. In addition, we excluded patients who had been readmitted to the ICU, during the same hospital stay. Every recurrent admission with a defined outcome of last hospitalization (*e.g*., discharge) was recorded as a separate ICU entry. Thus, total 194 patients were analyzed further. Because some of the clinical data were obtained from medical records, we confirmed that all the selected ICU patients had been in the ICU during the study period.

The authors used the APACHE II and SAPS II variables specified in their original publications to prepare the formal research instrument. All data regarding the variables were collected manually by the author himself. Enrolled patients under study were screened with respect to their demographic profile (age and sex), presence of chronic disease, past history of hospitalization and ICU admission, surgical status (elective or emergency surgery), major reason for ICU admission (*i.e*., predominant diagnostic category) and severity of illness (acute physiologic state). Initial and worst values were taken during the patient's first 24 h of ICU admission in respect of 12 variables constituting the acute physiology score (A). However, points were allocated to the worst values as per protocol of APACHE II scoring system. Age (B) and chronic disease (C) were also assigned points in similar manner. Sum of A, B and C constituted APACHE II score for a patient, derivation of which facilitated the subsequent calculation of predicted risk of mortality. In sedated patients, the 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 normal. ICU stay, hospital length of stay (LOS) and lead time (the interval from hospital admission to ICU admission) were recorded. Patients were followed up until ICU and hospital discharge in order to registrar their survival status.

Statistical Analysis

Predicted mortality for APACHE II and SAPS II was calculated for each individual using the calculative software available free in the internet. SPSS v12 was used for statistical analysis.

In this study, categorical data were presented as n (%) and continuous data as mean±SD. Chi-square statistics were used to test for the statistical significance of category variables. All statistical tests were two-sided, and a significance level of 0.05 was used.

Validation of the systems was tested by assessing calibration and discrimination. Calibration (the ability to provide risk estimate corresponding to the observed mortality) was assessed by calibration curves and the Lemeshow–Hosmer goodness of fit C-statistic. Calibration curves were drawn by plotting predicted against actual

mortality for groups of the patient population stratified by 10% increments of predicted mortality. To calculate the C-statistic, the study population was stratified into ten deciles with approximately equal numbers of patients. The predicted and actual number of survivors and non-survivors were compared statistically with the use of formal goodness-of-fit testing to determine whether or not the discrepancy was statistically insignificant (P > 0.05). Lower Hosmer–Lemeshow x2 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 interval.

Results:

During the study period there were 194 admissions in ICU. Male and female was 55.7% and 44.3% respectively. The characteristics of ICU patients are shown in Table 1. Mean age was 61.06 ± 15.42.The mean ICU length of stay was 8.15 ± 6.97. 58.2% was diabetic and 41.8% was non diabetic patients.

In comparison with survivors, non survivors had higher APACHE II and SAPS II scores. APACHE II score in survivor and non-survivor was (16.88 v 23.91, p=.000) and SAPS II scores was (40.58 v 55.14, p=0.000). There was no significant difference in age and length of stay in ICU between survivors and non survivors subjects.

Table I

Characteristics of ICU patients (n = 194)

Variables	Mean±SD
Age(years)	61.06 ± 15.42
LOS in ICU	8.15 ± 6.97
ICU mortality	42.65
APACHE II Score	18.98 ± 7.86
SAPS II Score	44.93 ± 17.38
Predicted Mortality APACHE II	35.32 ± 21.81
Predicted Mortality SAPS II	37.11 ± 27.34

Table II

Comparison of survivors and non survivors in ICU patients

Patients demographics				
Variables	Total	Non survivors	Survivors	P value
No. of females	86	16 (18.6)	70 (81.4)	.002
Age	194	58 (59.93 ± 16.57)	136 (63.69 ± 12.05)	.121 (NS)
ICU LOS in days	191	58 (7.51 ± 8.42)	133 (8.42 ± 6.51)	.411 (NS)
APACHE II score	194	58(23.91 ± 7.45)	136 (16.88 ± 7.07)	.000
SAPS II score	194	58 (55.14 ± 18.70)	136 (40.58 ± 14.82)	.000

Figure 1 shows the admission type in APACHE II. Non operative, emergency and elective postoperative type in APACHE II was 89.2, 4.6 and 6.2% respectively.

Scheduled, unscheduled surgery and medical type in SAPS II was 5.2, 4.6 and 89.2% respectively as in Figure 2. There is no significant difference between survivor and non-survivor ICU patients in admission type in APACHE II and SAPS II score.

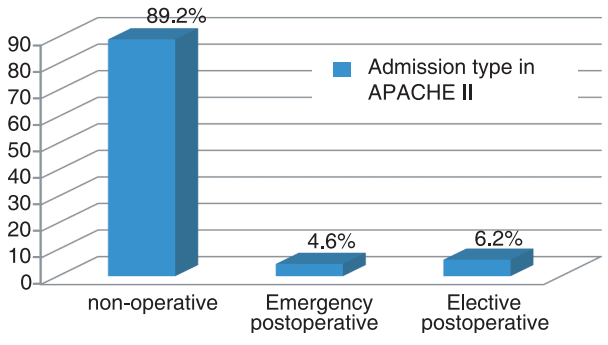

Fig.-1: *Admission type in APACHE II*

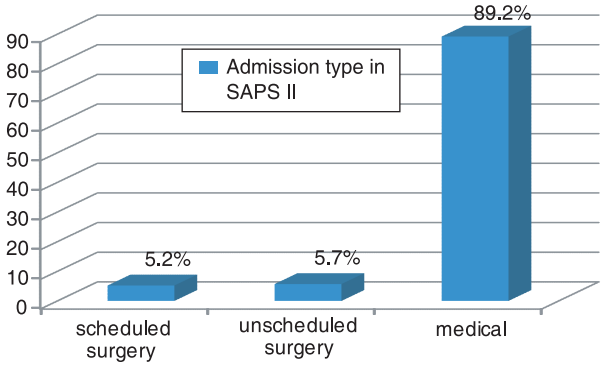

Fig.-2: *Admission type in SAPS II*

Table-III

Hosmer and lemeshow test and Area under the receiver operating characteristics curves

Prediction models	Score (Mean ± SD)	Predicted mortality (Mean ± SD)	Hosmer and lemeshow goodness of fit test		ROC curvea ROC ± SE (95% CI)
			C statistics	P value	
APACHE II	18.98 ± 7.86	35.32 ± 21.81	8.304	0.40	0.75±0.04 (0.67-0.82)
SAPS II	44.93 ± 17.38	37.11 ± 27.34	9.040	0.34	0.74±0.04 (0.66-0.81)

The APACHE II and SAPS II model exhibited good calibration (= 8.304, p = 0.40 for APACHE II, = 9.040, p = 0.34 for SAPS II). Hosmer-Lemeshow goodness-of-fit test revealed a good performance for APACHE II and SAPS II scores. There was no significant difference between observed and expected values of the predicted mortality which suggests that the results of this study reveals good intensive care quality.

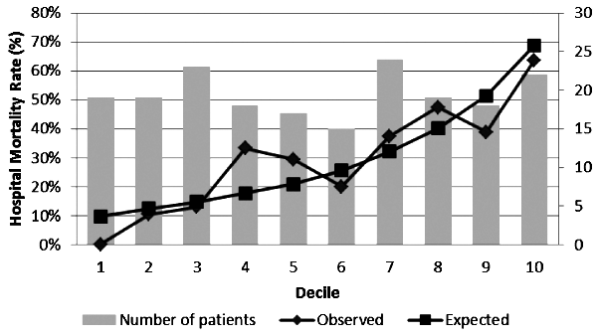

Fig.-3: *Calibration curve for APACHE II*

Although APACHE II had a greater aROC, suggesting APACHE II had slightly better discriminative power than the SAPS II, both models had aROC values less than 0.8. [aROC 0.75,CI (0.67-0.82) for APACHE and 0.74, CI (0.66-0.81) for SAPS II]

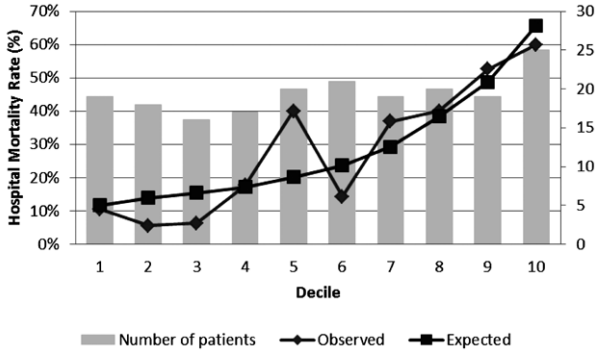

Fig.-4: *Calibration curve for SAPS II*

The overall discriminatory capability, as measured by the aROC, was generally good for two models and ranged from 0.78 to 0.89. APACHE II is slightly better compared to SAPS II score but not significantly better than SAPS II.

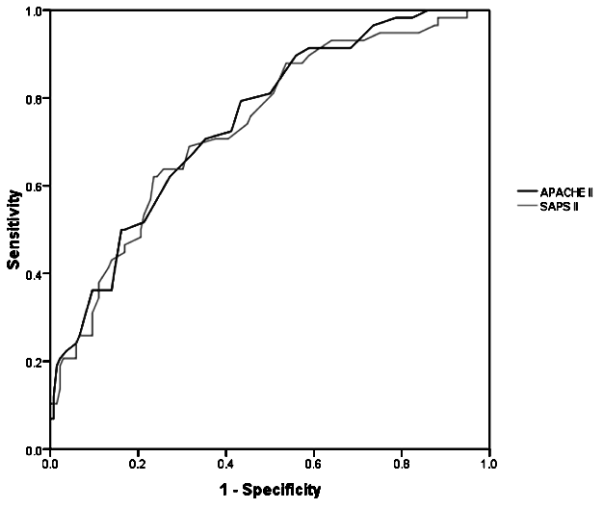

Fig.-5: *ROC curves for APACHE II and SAPS II scores on prediction of hospital mortality.*

Discussion

In the present study, we evaluated the ability and validity of APACHE II and SAPS II systems to accurately predict hospital mortality in a Bangladeshi adult mixed-case ICU. Both models showed good discrimination and calibration. Intermodel comparison showed that both the models perform equally, although a slightly better performance was found with APACHE II.

These models have been broadly used in European and North American countries with a good reported performance. However, some care must be exercised in the interpretation of this calibration test as it is sensitive to the sample size. Significant Hosmer-Lemeshow goodness-of-fit chi-square statistics could be achieved only with increasing sample size of the population despite an excellent fit. However, calibration statistics were less accurate (Higher H and C values) in our study compared with those reported in the European/north American studies, because of the fact that larger cohorts of ICU patients were included in the western databases.

Various reasons might be considered to explain the problem of calibration of the tested models in our ICU patients.

These reasons include actual differences in the quality of care between Bangladeshi and western ICUs, or the consequences of other factors such as differences in the disease leading to ICU admission, lead time bias, or availability of resources.

Both models showed excellent discrimination, although the authors found that discrimination was better for APACHE II than for SAPS II. Good discrimination of both models has been reported in previous studies [7,8,9,10-12]. The area under the ROC of both systems was similar with the original reports. Previously reported area under the ROC curve of APACHE II and SAPS II included 0.839 and 0.870 in Greece, 0.787 and 0.817 in Portugal[9], 0.83 and 0.79 in Saudi Arabia[7], 0.819 and 0.840 in Tunisia[8] and 0.88 and 0.87 in Hong Kong[13], respectively and 0.88 in the original SAPS II[17]. In Thailand, area under the ROC curve of APACHE II include 0.723[16], 0.788[15] and 0.838[14]. Lertsithichi et al[14] found the area under the ROC curve of SAPS II was 0.818 in thai surgical patients.

The reliability of the data collected is important because poor data can influence the predictions of mortality. Holt et al[19] showed that the main causes of data error in scoring APACHE II are inconsistent choice between highest and lowest value of acute physiologic score and GCS. The variability of GCS determination in sedated patients may affect the predicted death in both models. In the present study, the authors used the pre-sedation GCS in sedated patients as in previous studies[8], an approach which has been shown to be associated with better performance of APACHE II than the approach that is normal GCS for sedated patients[20]. Systematic differences in medical definitions and inclusion criterias in the databases can also lead to calibration problems. However, all the definitions used in our study were in agreement with their original publications. Coronary care and post-cardiac surgical patients were not included in the present study.

The potential role of difference in case mix between the presented database and the development database may have had a negative impact of calibration assessment. In general, medical patients have a higher mortality risk than postoperative surgical patients, and in the present study population, medical patients constituted a larger proportion (89.2%) than in the original SAPS II database (48%)[17]. These differences in case mix could contribute to predicted death for the APACHE II and SAPS II model in the presented patients.

Lead time bias is another factor that could adversely affect the accuracy of risk prediction. Tunnell et al[18] revealed that lead time bias increased the APACHE II and SAPS II scores by 14 and 23 points, respectively, leading to the APACHE II and SAPS II prediction of hospital mortality being increased as much as 42.7% and 33.4%, respectively. Lead time bias occurs when patients are partially treated before ICU admission. Doing so would underestimate the severity of underlying disease. This factor is difficult to quantify in our study, but we can assume that most of our patients admitted to the Emergency department were transferred to the ICU without significant vital support since intensive care facilities are limited there. The lead time was no different between survivors and non-survivors. Thus, the authors believed that the influence of lead time bias on calibrations of both models is minimal in the present study.

The likelihood of some management deficiencies in our ICUs with the consequent higher observed over expected mortality is another issue to consider. It could argued that Bangladeshi ICUs do not have the same quality of care compared with western ICUs. Both models give very similar predicted probabilities of death (35% to 37%), while the actual mortality is 42%, suggests that there are probably some shortcomings in the quality of care in our ICUs. In a study comparing a French ICU with a Tunisian ICU, it was found that death rates were similar in both ICUs for diseases with extremes of predicted risks, while in midrange severity diseases, Tunisian patients had a higher death rate. Technology availability and the level of therapy, as assessed by omega scores, did not seem to explain these differences.

Conclusion:

Based on our findings, we are unable to show a significant superiority of one scoring system over another as both models performed equally in our ICU patients. However, we believe that even without proper calibration, these models could still be used in Bangladeshi ICUs. These models would be accurate enough for a general description of our ICU patients; they also could be used for comparison of Bangladeshi ICUs to each other and to stratify patients by level of severity for national therapeutic trials. However, to support clinical decisions, these models require an appropriate adjustment to reflect more precisely the mortality in our own ICU patients. It would be possible to customize actual models as it was performed with SAPS II in specific groups of patients with sepsis. There is no doubt, however, that even when well calibrated, the resulting new versions or "customized" models must be used as a supplement, rather than a substitute for good clinical judgment.

Although generalizing our results to all Bangladeshi ICUs would be hazardous, we believe that our study population represents a reliable reflection of our specific conditions. However, we need further prospective validation studies on a larger Bangladeshi ICU population before a firm conclusion.

Acknowledgements:

Department of critical care medicine, BIRDEM Hospital , Dhaka

Intensive Care Unit, Ibn Sina Hospital, Dhaka

References:

1. Lewandowski K, Lewandowski M. Scoring systems in the intensive care unit. *Anaesthesist* 2003; 52 (10): 965-87; quiz 988- 9. (Grmec and Gasparovic 2001)
2. Ohno-Machado L, Resnic FS, Matheny ME. Prognosis in critical care. *Annu Rev Biomed Eng* 2006; 8: 567-99.
3. Hsiun Tang ,Chao, Che Ming Yang, Chi Yuang Chuang, Ming Lee Chang, Yu Chwen Huang, and Chin Feng Huang. A comparative study of clinical severity scoring systems in ICUs in Taiwan. *TZU Chi Med J* 2005: 239-45.
4. Glance LG, Osler T, Shinozaki T. Intensive care unit prognostic scoring systems to predict death: a costeffectiveness analysis. *Crit Care Med* 1998; 26 (11): 1842- 9.
5. Glance LG, Osler TM, Dick A. Rating the quality of intensive care units: is it a function of the intensive care unit scoring system? *Crit Care Med* 2002; 30 (9): 1976- 82.
6. Gregoire, G, Russell JA. Assessment of severity of illness. In Principle of Critical Care, eds JB Hall, GA Schmidt, LDH Wood, McGraw Hill , New York, 1998
7. 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 (2): 166- 74.
8. Nouira, S., M. Belghith, et al. Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units. *Critical care medicine* 1998; 26(5): 852.
9. Moreno R. and Morais P. et al. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. *Intensive Care Medicine* 1997; 23(2): 177-186.
10. Castella X, Artigas A, Bion J, Kari A. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European/ North American Severity Study Group. *Crit Care Med* 1995; 23:1327-35.
11. Capuzzo M, Valpondi V, Sgarbi A, Bortolazzi S, Pavoni V, Gilli G, et al. Validation of severity scoring systems SAPS II and APACHE II in a single center population. *Intensive Care Med* 2000; 26: 1779-85.
12. Katsaragakis S, Papadimitropoulos K, Antonakis P, Strergiopoulos S, Konstadoulakis MM, Androulakis G. Comparison of acute physiology and chronichealth 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.
13. 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.
14. Lertsithichai P, Euanorasetr C. Preliminary evaluation of APACHE II, APACHE III, MPM OII, MPM 24II and SAPS II predictive systems in a surgical ICU. *Ramathibodi Med* 1997; 20: 32-41.
15. Ratanarat R, Thanakittiwirun M, Vilaichone W, Thongyoo S, Permpikul C. Prediction of mortality by using the standard scoring systems in a medical intensive care unit in Thailand. *J Med Assoc Thai* 2005; 88: 949-55.
16. Wilairatana P, Noan NS, Chinprasatsak S, Prodeengam K, Kityaporn D, Looareesuwan S. Scoring systems for predicting outcomes of critically ill patients in northeastern Thailand. *Southeast Asian J Trop Med Public Health* 1995; 26: 66-72.
17. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association* 1993; 270: 2957–63.
18. Tunnell RD, Millar BW, Smith GB. The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care—a pilot study. *Anaesthesia* 1998; 53: 1045–53
19. 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.
20. Livingston BM, Mackenzie SJ, MacKirdy FN, Howie JC. Should the pre-sedation Glasgow Coma Scale value be used when calculating acute physiology and chronic Health Evaluation scores for sedated patients? Scottish Intensive Care Society Audit Group. *Crit Care Med* 2000; 28: 389-94.