PROCEEDINGS OF SINGAPORE HEALTHCARE

Original Article

Proceedings of Singapore Healthcare
2019, Vol. 28(3) 147–152

DOI: 10.1177/2010105818812896
journals.sagepub.com/home/psh
SAGE

Performance of the Acute Physiology and Chronic Health Evaluation II (APACHE II) in the prediction of hospital mortality in a mixed ICU in Singapore

Charles Chin Han Lew[1,2], Gabriel Jun Yung Wong[2], Chee Keat Tan[3] and Michelle Miller[1]

Abstract

Background: The Acute Physiology and Chronic Health Evaluation II (APACHE II) is used to quantify disease severity and hospital mortality risk in critically ill patients. It is widely used in intensive care units (ICUs) in Singapore, but its prognostic validity remains questionable as it has not been thoroughly assessed by established statistical methods.

Objectives: This study aimed to: (a) evaluate the discrimination and calibration accuracy of the APACHE II in the prediction of hospital mortality in a mixed ICU, and (b) customise the APACHE II in an effort to maximise its prognostic performance.

Methods: A prospective cohort study was conducted and all adult patients with >24 h of ICU admission in a tertiary care institution in Singapore were included. The outcome measure was hospital mortality, and all patients were followed-up until hospital discharge or death for up to one year after ICU admission.

Results: There were 503 patients, and their mean (SD) age and APACHE II score were 61.2 (15.8) years and 24.5 (8.2), respectively. Hospital mortality was 31%, and no patients were lost to follow-up. The APACHE II has good discrimination (receiver operating characteristic: 0.76) but poor calibration (Hosmer–Lemeshow C test: <0.001). Customisation did not significantly improve calibration accuracy.

Conclusions: The APACHE II and its customised version should not be used in the local setting as they both have poor calibration. There is an urgent need for larger studies to perform second-level customisation or to develop a new prognostic model tailored to the Singapore critical care setting.

Keywords
APACHE II, critical care, mortality, prognosis

Introduction

The Acute Physiology and Chronic Health Evaluation II (APACHE II) was developed in 1985 to objectively quantify disease severity and predict hospital mortality risk.[1] Despite newer versions such as the APACHE III and IV,[2,3] the APACHE II continues to be widely used in research and clinical practice. This is in part due to the ease of calculation, and the possibility of comparative consistency by reason of its long history of usage.

The hospital mortality rates of critically ill patients are used to assess the performance of intensive care units (ICUs) because it reflects important characteristics that are associated with good clinical practices (e.g. accurate diagnosis and timely therapies).[4] Since some hospitals will inherently admit patients with higher disease severity and thus have higher mortality rates than others, the APACHE II plays an essential role in the adjustment of mortality risk. That is, the predicted mortality rate derived from the APACHE II can be compared

[1]Nutrition and Dietetics Department, Flinders University, Adelaide, Australia
[2]Dietetics and Nutrition Department, Ng Teng Fong General Hospital, Singapore
[3]Department of Intensive Care Medicine, Ng Teng Fong General Hospital, Singapore

Corresponding author:
Charles Chin Han Lew, Dietetics and Nutrition Department, Ng Teng Fong General Hospital, 1 Jurong East Street 21, 609606, Singapore.
Email: Charles_lew@nuhs.edu.sg

with the observed mortality rate; this is termed as the standardised mortality ratio (SMR).[4] Of note, the accuracy of the SMR in assessing ICU performance is underpinned by the accuracy of the predicted hospital mortality risk since under- or overestimation of such risk will, respectively, inflate or understate the actual performance of the ICU.

All the ICUs of the public hospitals in Singapore use the APACHE II for quality audit. Despite the pervasive use of the APACHE II, its prognostic accuracy in Singapore remains questionable since it has never been validated locally with established statistical methods. Therefore, this study primarily aims to determine the performance of the APACHE II in the prediction of hospital mortality in a mixed ICU. The secondary aim is to customise the APACHE II and evaluate the performance of this new model.

Methods

Patients and setting

This was a prospective cohort study conducted in a 35-bed mixed ICU at Ng Teng Fong General Hospital. The ICU functions as a closed unit in which board-certified intensivists and residents provide care for both medical and surgical patients. Between August 2015 and October 2016, all adult ICU patients ⩾21 years old who had ⩾24 h length-of-stay were enrolled. For patients who were readmitted to the ICU during the same hospitalisation, only the data on the first admission was included. The Domain Specific Review Board approved this study (NHG DSRB Ref: 2014/00878), and informed consent was not required as this study was deemed as a clinical audit.

Data collection

All data required to calculate the APACHE II score and predicted mortality risk (i.e. demography, physiological parameters, admission diagnoses and comorbidities) were prospectively recorded in the electronic medical records. Calculation of the APACHE II was carried out by methods described by Knaus et al.[1] However, several established modifications were also carried out. In most cases, the lowest Glasgow Coma Score (GCS) during the first 24 h of ICU admission was used to calculate the APACHE II. However, in patients who were anaesthetised before ICU admission, the GCS recorded before anaesthesia was used.[5] The diagnosis of acute kidney injury (AKI) was in accordance with the latest definition, that is, increase in serum creatinine by ⩾26.5 μmol/L within 48 h or by ⩾1.5 times baseline, or urine volume <0.5 mL/kg/h for 6 h.[6] For missing data, parameters not measured in the first 24 h of ICU admission were considered normal.[2]

The predicted hospital mortality was calculated using a formula that comprised of a constant, the APACHE II score multiplied by a coefficient, exposure status for emergency surgery multiplied by a coefficient as well as the admission diagnostic coefficient outlined in Knaus et al.,[1] for example, $\ln(R/1 - R) = -3.517 + (\text{APACHE II score} \times 0.146) +$ admission diagnostic coefficient + 0.603 if exposed to emergency surgery, where ln = natural logarithm and R = risk of hospital mortality.[1] In the circumstance of multiple admission diagnoses, the condition with the worst prognosis (e.g. haemorrhagic shock rather than sepsis) would be taken.[4] For observed hospital mortality, patients were followed until hospital discharge or death for up to one year after ICU admission.

Statistical analysis

Performance of the APACHE II. Performance was assessed by its discriminative ability and calibration accuracy. Discrimination refers to the ability of the APACHE II in distinguishing discrete outcomes (e.g. died/survived). This was measured by the area under the receiver operating characteristic (ROC) curve, in which perfect, excellent, very good, good, moderate and poor discrimination are defined as ROC of 1.00, 0.90–0.99, 0.80–0.89, 0.70–0.79, 0.60–0.69 and <0.60, respectively.[7] In contrast to discrimination, calibration accuracy refers to the ability of the APACHE II in quantifying risk across the continuum of mortality risk. Calibration was measured using two methods. First, by the Hosmer–Lemeshow C test, in which accurate calibration is defined as p-value > 0.05, indicating no significant difference between the observed and predicted mortality.[8] Second, by plotting a calibration curve, the observed and predicted mortality across all risk ranges was presented in a graphical plot.

The SMR (a ratio of the observed versus predicted hospital mortality (estimated by APACHE II)) and its 95% confidence interval (CI) was also calculated for the purpose of future comparisons. The 95% CI was derived by dividing the 95% CI of the observed mortality by the predicted mortality.[9] An SMR with 1.0 within the 95% CI indicates an overall good ICU performance.

Customisation and validation of the customised APACHE II. The study population was randomly split into equal training and validation groups. The training group was used to customise the APACHE II in which new coefficients for the APACHE II score and exposure to emergency surgery as well as a new constant were computed from logistic regression with hospital mortality as the dependent variable. Thereafter, in the validation group, the discriminative ability and calibration accuracy of the customised model were determined by methods described above.

Patient characteristics of the training and validation groups were reported as mean and standard deviation, medians and inter-quartile range or counts and percentages; and the Student's t-test, Mann–Whitney U-test or Chi-square test were used appropriately to compare patient characteristics. All statistical analyses were performed using STATA 14.2 (Stata Corp, College Station, TX, USA) and significance assumed at $p < 0.05$.

Results

There were 844 admissions, of which 503 patients were enrolled (Figure 1). A majority of them were from the emergency department and admitted to the ICU for medical reasons. Other characteristics of the enrolled patients in

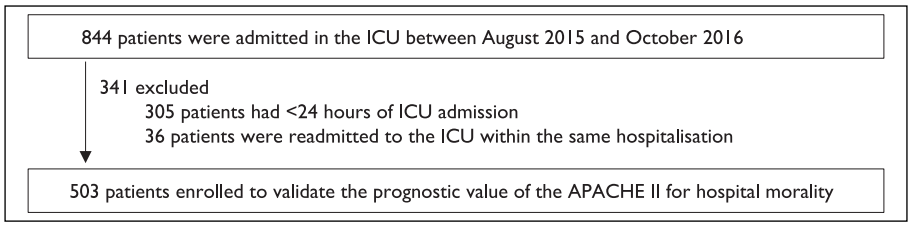

Figure 1. Patient enrolment.

Table 1. Characteristics of the enrolled patients in the overall, training and validation groups.

	All patients (n = 503)	Training group (252)	Validation group (251)	p-value
Age (years)	61.2 (15.8)	60.9 (16.2)	61.5 (15.4)	0.674
Male	302 [60.0]	153 [60.7]	149 [59.4]	0.757
Location before admission				0.886
Emergency Department	219 [43.5]	111 [44.0]	109 [43.4]	
High dependency	82 [16.3]	38 [15.1]	44 [17.5]	
Operation theatre	114 [22.7]	57 [22.6]	56 [22.3]	
Wards	88 [17.5]	46 [18.3]	42 [16.7]	
Type of admission				0.445
Medical	333 [66.2]	164 [65.1]	169 [67.3]	
Elective surgery	18 [3.6]	7 [2.8]	11 [4.4]	
Emergency surgery	152 [30.2]	81 [32.1]	71 [28.3]	
APACHE II	24.5 (8.2)	24.7 (8.6)	24.3 (7.7)	0.561
Lead time (days)	1.0 (0.0, 1.0)	1.0 (0.0, 1.0)	1.0 (0.0, 2.0)	0.447
Admission reasons				0.373
Cardiovascular	102 [20.7]	56 [22.2]	48 [19.1]	
Respiratory	88 [17.5]	47 [18.7]	41 [16.3]	
Sepsis	113 [22.5]	50 [19.8]	63 [25.1]	
Trauma	13 [2.6]	8 [3.2]	5 [2.0]	
Metabolic/renal	11 [2.2]	5 [2.0]	6 [2.4]	
Gastrointestinal	48 [9.5]	20 [7.9]	28 [11.2]	
Post operation	17 [3.4]	10 [4.0]	7 [2.8]	
Orthopaedics	7 [1.4]	6 [2.4]	1 [0.4]	
Neurological	102 [20.3]	50 [19.8]	52 [20.7]	
ICU LOS (days)	2.0 (2.0, 5.0)	2.0 (1.0, 5.0)	2.0 (2.0, 4.0)	0.841
Hospital LOS (days)	13.0 (6.0, 24.0)	13.0 (6.0, 24.0)	13.0 (6.0, 25.0)	0.985
ICU mortality	93 [18.5]	46 [18.3]	47 [18.7]	0.892
Hospital mortality	156 [31.0]	83 [32.9]	73 [29.1]	0.350

Values are mean (SD), median (q1, q3) or count [percentage]. APACHE II: Acute Physiology and Chronic Health Evaluation II, LOS: length of stay, ICU: intensive care unit.

the overall, training and validation groups are summarised in Table 1. Hospital mortality was 31% in the overall group, and no patients were lost to follow-up since the longest hospital length of stay was 255 days. A small number of patients had missing data. That is, 6.6% had missing haematocrit and white blood count, and 6.0% had missing serum sodium and potassium.

Overall sample

Discrimination was good as evidenced by the ROC, but calibration accuracy measured by the Hosmer–Lemeshow C test was poor (Table 2). This was consistent with the calibration curve which showed an overestimation of predicted hospital mortality risk in nearly all deciles (Figure 2).

Customisation

The new customised equation to quantify predicted hospital mortality risk was as follows: logit = −4.587 + (APACHE II score × 0.143) + existing diagnostic weight outlined in Knaus et al.[1]

Exposure to emergency surgery was not significantly associated with hospital mortality (p-value: 0.324) and hence was omitted in this new model.

Discrimination was good in the validation group and very good in the training group (Table 2 and Figure 3). Although customisation of the APACHE II considerably improved the accuracy of the predicted hospital mortality risk in all deciles (Figure 2: overall versus validation group), there was still significant inaccuracies in which predicted hospital mortality

Table 2. Discriminative ability and calibration accuracy of the APACHE II in all patients, training and validation groups.

Patient groups	ROC (95% CI)	HL-C (*p*-value)	SMR (95% CI)
All patients	0.756 (0.715–0.792)	146.54 (<0.001)	0.609 (0.532–0.692)
Medical (*n* = 333)	0.762 (0.713–0.807)	86.97 (<0.001)	0.648 (0.557–0.745)
Surgical (*n* = 170)	0.728 (0.656–0.795)	75.30 (<0.001)	0.513 (0.383–0.672)
Training group	0.804 (0.744–0.865)	9.95 (0.445)	1.015 (0.830–1.223)
Medical (*n* = 169)	0.794 (0.720–0.867)	9.24 (0.510)	1.024 (0.818–1.252)
Surgical (*n* = 82)	0.806 (0.691–0.921)	5.87 (0.826)	0.986 (0.616–1.504)
Validation group	0.722 (0.654–0.790)	31.47 (0.001)	1.131 (0.940–1.339)
Medical (*n* = 164)	0.734 (0.654–0.815)	11.58 (0.314)	1.061 (0.858–1.281)
Surgical (*n* = 88)	0.681 (0.553–0.809)	29.51 (0.001)	1.382 (0.936–1.950)

CI: confidence interval, HL-C: Hosmer–Lemeshow C test, ROC: receiver operating characteristic, SMR: standardised mortality ratio.

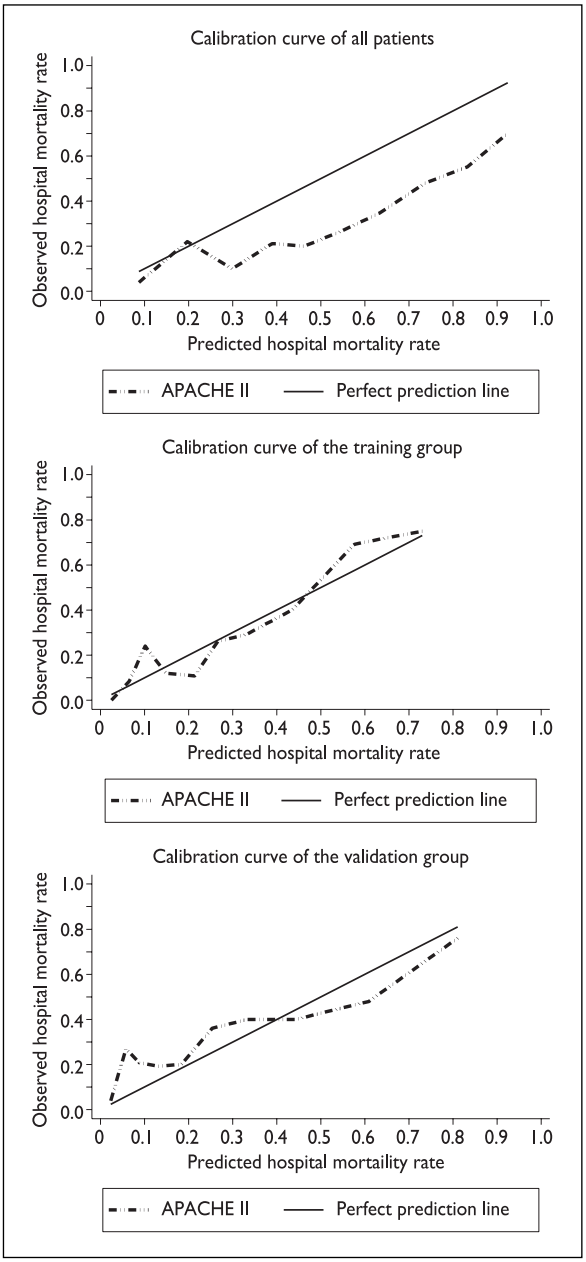

Figure 2. Calibration curves for all patients, the training and the validation group.

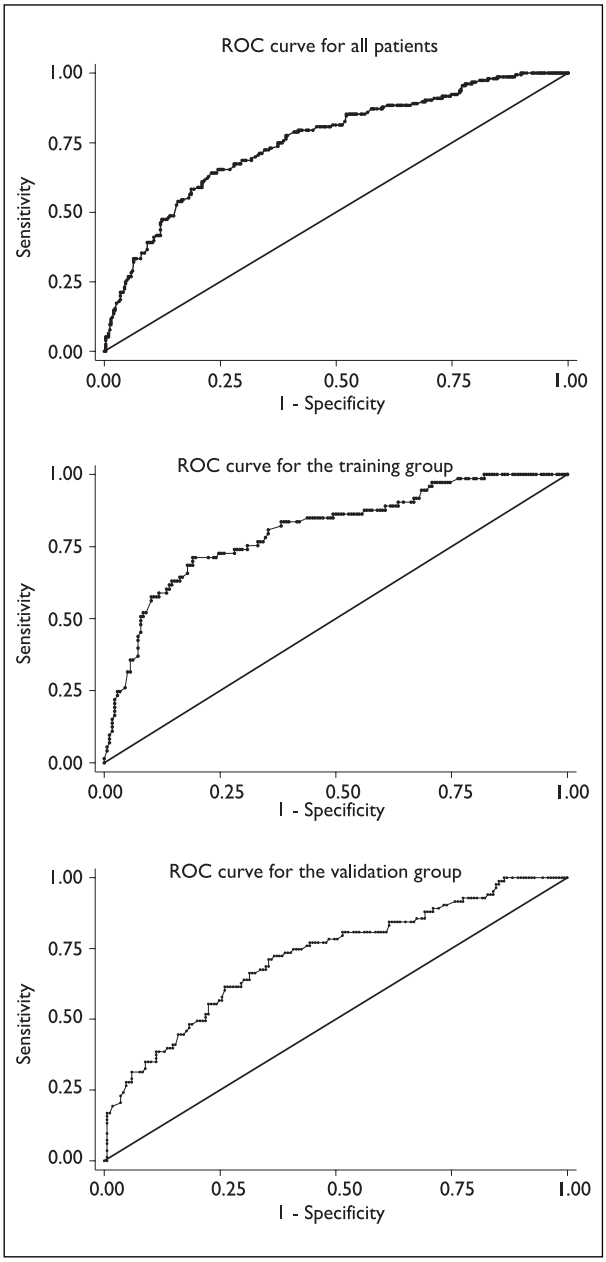

Figure 3. Receiver operating characteristic curves for all patients, the training and the validation group.

risks were under-estimated in patients with ⩽40%, and over-estimated in patients with >40% observed hospital mortality risk. Of note, calibration accuracy was good for medical patients in both the training and validation group whereas it was poor in surgical patients in the validation group. Similarly, the SMR in medical patients, as opposed to surgical patients, appears to be more reliable as evidenced by the tighter CI.

Discussion

To our knowledge, this is the largest study conducted in Singapore to evaluate the validity of the APACHE II in predicting hospital mortality. The APACHE II has demonstrated good discrimination but poor calibration accuracy for hospital mortality, and customisation of the APACHE II did not significantly improve its calibration accuracy in the local setting.

In 1985, Knaus et al. used data (i.e. 12 physiological parameters, comorbidities and emergency surgery, age and admission diagnosis) from a reference population of 5815 patients from 13 hospitals in the USA to develop APACHE II,[1] whereby it quantifies the predicted hospital mortality risk of critically ill patients via an equation. Therefore, all subsequent evaluations of ICU performance using the APACHE II are in effect weighing against the reference population. Given the advances in ICU treatment modalities since 1985, it is crucial to validate the APACHE II before using it in local settings.

This is the fifth validation study performed in Singapore, but results of our study cannot be compared with three previous studies as they did not report the discrimination and calibration accuracy of APACHE II.[10–12] Nevertheless, we were able to estimate the SMR and calibration accuracy from the crude results reported by Lee et al.[10] These authors prospectively calculated the APACHE II scores of 131 patients in the medical ICU, and the SMR was estimated to be 0.89 and there was good correlation ($r = 0.95$, p-value: 0.001) between observed and predicted mortality, suggestive of good calibration. Similar results were demonstrated in the surgical ICU in which there were very good discrimination and likely good calibration (correlation between observed and predicted mortality was 0.97 (p-value unreported)).[13] The good prognostic performance of the APACHE II in these studies was likely due to the close proximity of APACHE II development (i.e. 1985) and the validation period (1991) in which treatment modalities were likely similar.[10,13] Evidently, reduction in observed hospital mortality with time due to advances in treatment modalities gradually reduces the discrimination and calibration accuracy of the APACHE II.[14,15] This may account for the discordance of results between our study and those of Lee et al. and Chen et al.[10,13]

Compared to recent studies conducted in other countries, patients in our study had higher disease severity, as evidenced by the higher mean APACHE II score (24.5 versus 17–21).[14,16–21] Similar to recent studies, the APACHE II in our study also had good discriminative validity, that is, 0.756 versus 0.729 to 0.805,[14,16–21] and poor calibration accuracy.[14,16–19] It is established that the latter will have a negative impact on the statistical risk adjustment in research studies and the SMR used in clinical audits.[4] Therefore, customisation of the APACHE II is often carried out in the literature in an effort to improve calibration accuracy.

There are two levels of customisation. First-level customisation refers to computing a new constant and new coefficients for the APACHE II score and exposure to emergency surgery. Second-level customisation involves all steps described above and additionally computes new coefficients for the admission diagnoses.[22] In our study, second-level customisation was not performed because it should only be performed in studies with a large sample size as this will reduce the risk of overfitting.[23] Although first-level customisation considerably improved calibration, it remained insufficient to significantly improve calibration accuracy. This is similar to the results of Brinkman et al. and Mann et al.,[15,24] in which customisation also did not improve calibration accuracy.

This study has some strengths. Selection, attrition and treatment biases were respectively minimised by the use of consecutive recruitment, complete follow-up and blinding of the treatment team to the objectives of the study. However, there are some limitations. This was a single centre study and hence lacks generalisability. Although our study was the largest in the local setting, the sample size did not allow robust subgroup analyses.

Future direction

Poor calibration after customisation is indicative of the dire need to either conduct a local multicentre study to perform robust second-level customisation or develop a new prognostic model with the addition of strong prognostic variables such as exposure to cardiopulmonary resuscitation before ICU admission and baseline nutritional status.[25,26] This will allow comparison of ICU performance among the local hospitals as well as internal quality assessment and benchmarking based on historical baseline performance data (e.g. baseline SMR).

The performance of an ICU is commonly measured by the SMR. For future studies, it is best practice to stratify the SMR by low-, medium- and high-risk patients to better understand the performance of the APACHE II in different risk groups. This is because the proportion of high-risk patients within a cohort will disproportionately affect the aggregated SMR since most high-risk patient will die.[4] In our study, we did not stratify the SMR by risk groups because the customised APACHE II did not demonstrate good calibration.

Conclusion

The APACHE II demonstrated good discrimination but poor calibration accuracy in the prediction of hospital mortality in a mixed ICU in Singapore. Customisation was attempted to improve calibration accuracy, but such effort proved to be futile. Therefore, there is an urgent need for future studies to recruit a larger sample of patients from multiple hospitals to perform second-level customisation or develop a new prognostic model that will better predict hospital mortality.

Acknowledgements

We are grateful for the statistical support provided by Wong Chiew Meng Johnny, Biostatistician, Clinical Research Unit, Ng Teng Fong General Hospital.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data

The datasets generated and/or analysed during the current study are not publicly available due to the data confidentiality requirements of the ethics committee, but are available from the corresponding author on reasonable request and approval from the ethics committee.

Authors' contributions

CCH Lew, GJY Wong, CK Tan and M Miller equally contributed to the conception and design of the research; CCH Lew, and GJY Wong contributed to the acquisition of the data; and CCH Lew contributed to the analysis and interpretation of the data, CCH Lew drafted the manuscript. All authors critically revised the manuscript, agree to be fully accountable for ensuring the integrity and accuracy of the work, and read and approved the final manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

Ethical approval

Ethical approval was obtained from the Domain Specific Review Board (NHG DSRB Ref: 2014/00878).

Informed consent

Informed consent was not sought for the present study because this was an observational study where no attempt was made to change the standard of care.

ORCID iD

Charles Chin Han Lew https://orcid.org/0000-0001-6410-3859

References

1. Knaus WA, Draper EA, Wagner DP, et al. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818–829.
2. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100: 1619–1636.
3. Zimmerman JE, Kramer AA, McNair DS, et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34: 1297–1310.
4. Breslow MJ and Badawi O. Severity scoring in the critically ill. Part 2. Maximizing value from outcome prediction scoring systems. *Chest* 2012; 141: 518–527.
5. Manganaro L and Stark M. *APACHE foundations user guide*. Kansas City, MO: Cerner Corporation, 2010.
6. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl* 2012; 2: 1–138.
7. Afessa B, Gajic O and Keegan MT. Severity of illness and organ failure assessment in adult intensive care units. *Crit Care Clin* 2007; 23: 639–658.
8. Hosmer DW and Lemesbow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat Theory Methods* 1980; 9: 1043–1069.
9. Morris JA and Gardner MJ. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed)* 1988; 296: 1313–1316.
10. Lee KH, Hui KP, Lim TK, et al. Acute Physiology And Chronic Health Evaluation (APACHE II) scoring in the medical intensive care unit, National University Hospital, Singapore. *Singapore Med J* 1993; 34: 41–44.
11. Leong IY and Tai DY. Is increasing age associated with mortality in the critically ill elderly. *Singapore Med J* 2002; 43: 33–36.
12. Lim SC, Fok AC and Ong YY. Patient outcome and intensive care resource allocation using APACHE II. *Singapore Med J* 1996; 37: 488–491.
13. Chen FG, Koh KF and Goh MH. Validation of APACHE II score in a surgical intensive care unit. *Singapore Med J* 1993; 34: 322–324.
14. Harrison DA, Lone NI, Haddow C, et al. External validation of the intensive care national audit & research centre (ICNARC) risk prediction model in critical care units in Scotland. *BMC Anesthesiol* 2014; 14: 116.
15. Mann SL, Marshal MR, Woodford BJ, et al. Predictive performance of acute physiological and chronic health evaluation releases II to IV: a single New Zealand centre experience. *Anaesth Intensive Care* 2012; 40: 479–489.
16. Bilgili B, Dikmen Y, Demirkiran O, et al. Comparison of the performance of four intensive care scoring systems. *Haseki Tip Bülten* 2013; 51: 45–50.
17. Ho KM, Williams TA, Harahsheh Y, et al. Using patient admission characteristics alone to predict mortality of critically ill patients: a comparison of 3 prognostic scores. *J Crit Care* 2016; 31: 21–25.
18. Ilker I, Mehmet K, Mehmet A, et al. Study of effectiveness of the SAPS II-III, APACHE II-IV and MPM II scores in the determination of prognosis of the patients in reanimation intensive care unit. *Acta Med Mediterr* 2015; 31: 127–131.
19. Kim JY, Lim SY, Jeon K, et al. External validation of the acute physiology and chronic health evaluation II in Korean intensive care units. *Yonsei Med J* 2013; 54: 425–431.
20. Parajuli BD, Shrestha GS, Pradhan B, et al. Comparison of acute physiology and chronic health evaluation II and acute physiology and chronic health evaluation IV to predict intensive care unit mortality. *Indian J Crit Care Med* 2015; 19: 87–91.
21. Serpa Neto A, Assuncao MS, Pardini A, et al. Feasibility of transitioning from APACHE II to SAPS III as prognostic model in a Brazilian general intensive care unit. A retrospective study. *Rev Paul Med* 2015; 133: 199–205.
22. Khwannimit B and Bhurayanontachai R. The performance of customised APACHE II and SAPS II in predicting mortality of mixed critically ill patients in a Thai medical intensive care unit. *Anaesth Intensive Care* 2009; 37: 784–790.
23. Royston P, Moons KG, Altman DG, et al. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009; 338: b604.
24. Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, et al. External validation of acute physiology and chronic health evaluation IV in Dutch intensive care units and comparison with acute physiology and chronic health evaluation II and simplified acute physiology score II. *J Crit Care* 2011; 26: 105, e11-e18.
25. Lew CCH, Cheung KP, Chong MFF, et al. Combining 2 commonly adopted nutrition instruments in the critical care setting is superior to administering either one alone. *J Parenter Enteral Nutr* 2018; 42: 872–876.
26. Lew CCH, Wong GJY, Cheung KP, et al. Association between malnutrition and 28-day mortality and intensive care length-of-stay in the critically ill: a prospective cohort study. *Nutrients* 2017; 10: 10.