Prediction of hospital mortality among critically ill patients in a single centre in Asia: comparison of artificial neural networks and logistic regression–based model

Swan Lau *, HP Shum, Carol CY Chan, MY Man, KB Tang, Kenny KC Chan, Anne KH Leung, WW Yan

ABSTRACT

Introduction: This study compared the performance of the artificial neural network (ANN) model with the Acute Physiologic and Chronic Health Evaluation (APACHE) II and IV models for predicting hospital mortality among critically ill patients in Hong Kong.

Methods: This retrospective analysis included all patients admitted to the intensive care unit of Pamela Youde Nethersole Eastern Hospital from January 2010 to December 2019. The ANN model was constructed using parameters identical to the APACHE IV model. Discrimination performance was assessed using area under the receiver operating characteristic curve (AUROC); calibration performance was evaluated using the Brier score and Hosmer–Lemeshow statistic.

Results: In total, 14 503 patients were included, with 10% in the validation set and 90% in the ANN model development set. The ANN model (AUROC=0.88, 95% confidence interval [CI]=0.86-0.90, Brier score=0.10; P in Hosmer–Lemeshow test=0.37) outperformed the APACHE II model (AUROC=0.85, 95% CI=0.80-0.85, Brier score=0.14; P<0.001 for both comparisons of AUROCs and Brier scores) but showed performance similar to the APACHE IV model (AUROC=0.87, 95% CI=0.85-0.89, Brier score=0.11; P=0.34 for comparison of AUROCs, and P=0.05 for comparison of Brier scores). The ANN model demonstrated better calibration than the APACHE II and APACHE IV models.

Conclusion: Our ANN model outperformed the APACHE II model but was similar to the APACHE IV model in terms of predicting hospital mortality in Hong Kong. Artificial neural networks are valuable tools that can enhance real-time prognostic prediction.

Hong Kong Med J 2024;30:130–8
https://doi.org/10.12809/hkmj2210235

[1] **S Lau *,** BSc, MB, BS
[2] **HP Shum,** MD, FRCP
[2] **CCY Chan,** FHKCA, FHKAM (Anaesthesiology)
[2] **MY Man,** MRCP (UK), FHKAM (Medicine)
[2] **KB Tang,** FHKCA, FHKAM (Anaesthesiology)
[3] **KKC Chan,** MStat, FHKAM (Anaesthesiology)
[4] **AKH Leung,** FHKCA (IC), FCICM
[2] **WW Yan,** FRCP, FHKAM (Medicine)

[1] *Department of Anaesthesia, Pain and Perioperative Medicine, Queen Mary Hospital, Hong Kong SAR, China*
[2] *Department of Intensive Care, Pamela Youde Nethersole Eastern Hospital, Hong Kong SAR, China*
[3] *Department of Anaesthesia and Intensive Care, Tuen Mun Hospital, Hong Kong SAR, China*
[4] *Department of Intensive Care, Queen Elizabeth Hospital, Hong Kong SAR, China*

* Corresponding author: ls037@ha.org.hk

This article was published on 28 Mar 2024 at www.hkmj.org.

New knowledge added by this study

- An artificial neural network model outperformed the Acute Physiologic and Chronic Health Evaluation (APACHE) II model but was similar to the APACHE IV model in terms of predicting hospital mortality.
- The three most important predictor variables were the highest sodium level, highest bilirubin level, and lowest white cell count within 24 hours of intensive care unit admission.
- External validation studies using data from other hospitals are recommended to confirm these findings.

Implications for clinical practice or policy

- Prediction of mortality among critically patients is challenging.
- Artificial neural networks, along with other machine learning techniques, are valuable tools that can enhance real-time prognostic prediction.

Introduction

Intensive care treatments are primarily intended to improve patient outcomes. Considering the high operating costs of intensive care units (ICUs), a reliable, decision-supporting, risk stratification system is needed to predict patient outcomes and facilitate cost-effective use of ICU beds. Several disease severity scoring systems, such as the Acute Physiology and Chronic Health Evaluation (APACHE) system and the Simplified Acute Physiology Score system, are currently used to objectively assess outcomes and recovery potential in this complex and diverse group of patients.[1,2]

The APACHE system, one of the most commonly used benchmark severity scoring systems worldwide, can measure disease severity and predict hospital mortality among ICU patients. In the 40 years since its initial development, the APACHE system has undergone multiple revisions to improve statistical power and discrimination performance by modifying the numbers and weights of included variables.[3-6] The underlying statistical principle is multivariable logistic regression based on data from an American population. The results are easy to interpret and allow robust outcome prediction for individuals with characteristics similar to the original population. However, the APACHE system has limited capacity to manage non-linear relationships between predictor and outcome variables, interactions between variables, and missing data. Although the value of the APACHE system for mortality prediction has been established, especially in Western countries, its discrimination performance and calibration are inconsistent when applied outside of the US.[7-10] Since 2008, the Hospital Authority in Hong Kong has utilised the APACHE IV model to assess outcomes in critically ill patients. Nevertheless, the APACHE II model remains the most extensively validated version; it is widely used for research and reference purposes.[11]

In the early 1990s, artificial neural networks (ANNs), a type of machine learning algorithm, were proposed as alternative statistical techniques to logistic regression–based method. Similar to the organisation and data processing configurations in human brains, these networks consist of input and output layers with at least one or more intermediate (hidden) layers for pattern recognition. Each layer contains several ‘artificial neurons’, known as nodes, for data extraction; these nodes are connected with each other through variable ‘weights’.[12] Artificial neural networks identify representative patterns from input data and observed output data within a training set, then fine-tune the variable weights; thus, they can predict outcomes when provided novel information. This method has considerable advantages in terms of managing non-linear relationships and multivariable interactions.[13]

A review of 28 studies comparing ANN and regression-based models showed that ANN outperformed regression-based models in 10 studies (36%), was outperformed by regression-based models in four studies (14%), and had similar performance in the remaining 14 studies (50%).[14] Multiple recent studies also demonstrated that the integration of machine learning with electronic health records provided more accurate and reliable predictive performance compared with conventional prognostic models.[15,16]

亞洲單一中心重症患者住院死亡率預測：人工神經網絡與邏輯迴歸模型的比較

劉詩韻、沈海平、陳卓茵、文敏儀、鄧建邦、陳勁松、
梁結雄、殷榮華

引言：本研究比較人工神經網絡模型與急性生理學和慢性健康評估（APACHE）II和IV模型在預測香港重症患者住院死亡率的效能。

方法：這項回顧性研究納入2010年1月至2019年12月期間所有入住東區尤德夫人那打素醫院深切治療部的患者。本研究的人工神經網絡模型構建參數與APACHE IV模型相同。我們分別利用受試者工作特徵曲線下面積（AUROC），以及布賴爾分數和Hosmer–Lemeshow檢驗來比較人工神經網絡和APACHE的區分及校準性能。

結果：研究共包括14 503名深切治療部患者，當中10%用於驗證人工神經網絡模型，90%用於模型開發。結果顯示，人工神經網絡模型的表現優於APACHE II模型（人工神經網絡模型：AUROC=0.88，95%置信區間=0.86-0.90，布賴爾分數=0.10，Hosmer–Lemeshow檢驗的P值=0.37；APACHE II模型：AUROC=0.85，95%置信區間=0.80-0.85，布賴爾分數=0.14；AUROC比較和布賴爾分數比較的P值均為<0.001）。然而，人工神經網絡模型的預測性能與APACHE IV模型相若（AUROC=0.87，95%置信區間=0.85-0.89，布賴爾分數=0.11；AUROC比較的P值=0.34，布賴爾分數比較的P值=0.05）。此外，人工神經網絡模型在校準性方面優於APACHE II和APACHE IV模型。

結論：在預測香港危重患者住院死亡率方面，人工神經網絡模型優於APACHE II模型，但與APACHE IV模型相若。人工神經網絡是能夠增強實時預測預後的重要工具。

This study was conducted to compare ANN performance with the performances of extensively validated and benchmark scoring systems—APACHE II and APACHE IV—in terms of predicting hospital mortality among critically ill patients in Hong Kong.

Methods

This retrospective analysis included all patients aged ≥18 years with first-time admissions to the ICU of Pamela Youde Nethersole Eastern Hospital between 1 January 2010 and 31 December 2019. The hospital is a 2000-bed tertiary care regional hospital that provides comprehensive services except for cardiothoracic surgery, transplant surgery, and burn management. The ICU is a 24-bed, closed, mixed medical-surgical unit with an average of 1600 patients admitted annually.

Demographic characteristics and hospital mortality data were retrospectively recorded. The worst value of each physiological parameter during the first 24 hours after ICU admission was used to generate an APACHE score. The predicted mortality risk was calculated based on published methods.[3,5] Included parameters were age, sex, systolic and diastolic blood pressures, temperature, heart rate, respiratory rate, glucose level, blood urea nitrogen level, serum sodium level, creatinine level, haematocrit level, white cell count, albumin level, bilirubin level, pH, fraction of inspired oxygen, partial pressures of carbon dioxide and oxygen, bicarbonate, and urine output during the first 24 hours after ICU admission. For patients who had multiple ICU admissions during a single hospital stay, only the first admission was included. Patients were excluded if they died or were discharged from the ICU within 4 hours after admission.

Instances of incomplete data were resolved by multiple imputation using the Markov chain Monte Carlo algorithm (ie, fully conditional specification). This method fits a univariate (single dependent variable) model using all other available variables in the model as predictors, then imputes missing values for the dependent variable. The method continues until the maximum number of iterations is reached; the resulting imputed values are saved to the imputed dataset.

Neural network models were constructed with SPSS software (Windows version 25.0; IBM Corp, Armonk [NY], US) using the same parameters as in the APACHE IV model (online supplementary Fig); SPSS software was also used to examine model precision. The multilayer perceptron procedure, a class of feed-forward learning model, consists of ≥3 layers of nodes: input, hidden, and output.[17] Automatic architecture building, which computes the best number of units in a hidden layer, was performed with SPSS software. Each hidden unit is an activation function of the weighted sum of the inputs; the values of the weights are determined by an estimation algorithm. In this study, the hidden layer consisted of 12 units (nodes). A hyperbolic tangent activation function was also employed for the hidden layers. Softmax activation and cross-entropy error functions were used for the output layer. The multilayer perceptron procedure utilised a backpropagation technique for supervised training. Learning occurred in the recognition phase for each piece of data via changes to connection weights based on the amount of error in the output compared with the expected result (gradient descent method).[18]

The training process was terminated when no further decreases in calculated error were observed. Subsequently, network weights were identified and used to compute test values. The importance of an independent variable was regarded as a measure of the extent to which network model–predicted values differed from observed values. Normalised importance, expressed as a percentage, constituted the ratio between the importance of each predictor variable and the largest importance value. Model stability was assessed by tenfold cross-validation. Oversampling of minority classes was performed via duplication to manage imbalances in outcome data.

Categorical and continuous variables were expressed as numbers (percentages) and medians (interquartile ranges). The Chi squared test or Fisher's exact test was used for comparisons of categorical data; the Mann-Whitney *U* test was used for comparisons of continuous data. The performances of ANN, APACHE II, and APACHE IV models were evaluated in terms of discrimination and calibration power. Discrimination, which constitutes the ability of a predictive model to separate data into classes (eg, death or survival), was evaluated using the area under the receiver operating characteristic curve (AUROC). The AUROCs of the models were compared using the DeLong test. Calibration, which represents the closeness of model probability to the underlying probability of the study population, was evaluated using the Brier score, Hosmer–Lemeshow statistic, and calibration curves.[19] All P values were two-sided, and values <0.05 were considered statistically significant. All analyses were performed with SPSS software and MedCalc statistical software (version 19.6.1).

Results

In total, 14 503 patients were included. The demographic characteristics and hospital mortality data of the study cohort were shown in Table 1, while the physiological and laboratory parameters required to generate an APACHE score were presented in Table 2. Among the recruited patients, 4.93% had at least one missing data point, and the overall rate of missing data was 0.48%. Furthermore, 1400 (9.7%) of the recruited patients were randomly assigned to the validation set; the remaining patients (n=13 103, 90.3%) were assigned to the model development set. With respect to the ANN model, 70% and 30% of the development set were used for training and testing purposes, respectively. The median age was 67 years (interquartile range [IQR]=54-78), median APACHE II score was 18 (IQR=13-25), and median APACHE IV score was 66 (IQR=46-91). The overall hospital and ICU mortality rates were 19.3% (n=2799) and 9.6% (n=1392), respectively.

The baseline co-morbidities, source of admission, disease category, APACHE II score, and APACHE IV score were similar in the test and validation sets (Table 1). More patients in the validation set received continuous renal replacement therapy (18.3% vs 16.1%; P=0.04). Concerning the worst physiological and laboratory parameters within the first 24 hours (Table 2), there were almost no significant differences between the development and validation sets; notably, the haemoglobin level was lower in the validation set (11.3 g/dL vs 11.5 g/dL; P=0.02).

TABLE 1. Patient characteristics and outcome parameters*

	Total (n=14 503)	Development set (n=13 103)	Validation set (n=1400)	P value
Age, y	67 (54-78)	67 (54-78)	66 (53-78)	0.36
Male sex	8576 (59.1%)	7745 (59.1%)	831 (59.4%)	0.86
Source of admission				0.58
General ward	6330 (43.6%)	5718 (43.6%)	612 (43.7%)	
OT/recovery	5480 (37.8%)	4954 (37.8%)	526 (37.6%)	
AED	2351 (16.2%)	2118 (16.2%)	233 (16.6%)	
Others	342 (2.4%)	313 (2.4%)	29 (2.1%)	
Specialty				0.99
Medical	6300 (43.4%)	5678 (43.3%)	622 (44.4%)	
Surgical	4104 (28.3%)	3716 (28.4%)	388 (27.7%)	
NS	2451 (16.9%)	2224 (17.0%)	227 (16.2%)	
ORT	783 (5.4%)	699 (5.3%)	84 (6.0%)	
ENT	454 (3.1%)	410 (3.1%)	44 (3.1%)	
Others	411 (2.8%)	376 (2.9%)	35 (2.5%)	
Postoperative cases	5480 (37.8%)	4954 (37.8%)	526 (37.6%)	0.86
Emergency admission	12 539 (86.5%)	11 331 (86.5%)	1208 (86.3%)	0.84
GCS score	14 (9-15)	14 (9-15)	14 (9-15)	0.54
Disease category				0.95
Sepsis	2703 (18.6%)	2444 (18.7%)	259 (18.5%)	
Gastrointestinal/hepatobiliary	2691 (18.6%)	2436 (18.6%)	255 (18.2%)	
Cardiovascular	1486 (10.2%)	1343 (10.2%)	143 (10.2%)	
Neurological	3031 (20.9%)	2745 (20.9%)	286 (20.4%)	
Respiratory	1630 (11.2%)	1469 (11.2%)	161 (11.5%)	
Metabolic	1038 (7.2%)	927 (7.1%)	111 (7.9%)	
Renal/genitourinary	754 (5.2%)	674 (5.1%)	80 (5.7%)	
Trauma	748 (5.2%)	683 (5.2%)	65 (4.6%)	
Others	422 (2.9%)	382 (2.9%)	40 (2.9%)	
Co-morbidities				
HT	8488 (58.5%)	7675 (58.6%)	813 (58.1%)	0.72
DM	3584 (24.7%)	3256 (24.8%)	328 (23.4%)	0.24
NYHA class IV heart failure	23 (0.2%)	22 (0.2%)	1 (0.1%)	0.72
Chronic respiratory insufficiency with or without PH	200 (1.4%)	175 (1.3%)	25 (1.8%)	0.17
Receipt of chronic renal dialysis	634 (4.4%)	562 (4.3%)	72 (5.1%)	0.14
Hepatic failure	155 (1.1%)	143 (1.1%)	12 (0.9%)	0.42
Cirrhosis with documented portal HT	212 (1.5%)	198 (1.5%)	14 (1.0%)	0.13
AIDS	24 (0.2%)	23 (0.2%)	1 (0.1%)	0.72
Lymphoma	147 (1.0%)	131 (1.0%)	16 (1.1%)	0.61
Metastatic cancer	527 (3.6%)	467 (3.6%)	60 (4.3%)	0.17
Leukaemia/myeloma	160 (1.1%)	142 (1.1%)	18 (1.3%)	0.49
Use of immunosuppressive agents	726 (5.0%)	647 (4.9%)	79 (5.6%)	0.25
Length of stay, d				
ICU	1.8 (1.0-4.0)	1.8 (1.0-4.0)	1.9 (1.0-3.9)	0.45
Hospital	13.1 (6.9-27.1)	13.1 (6.9-27.1)	13.3 (7.2-27.8)	0.26
APACHE IV				
Score	66 (46-91)	66 (46-91)	66 (46-91)	0.92
Predicted risk of death	0.16 (0.05-0.40)	0.16 (0.05-0.40)	0.16 (0.05-0.40)	0.64
APACHE II				
Score	18 (13-25)	18 (13-25)	18 (13-26)	0.94
Predicted risk of death	0.26 (0.10-0.52)	0.26 (0.10-0.52)	0.26 (0.10-0.53)	0.49
Treatment received				
CRRT or HD	2370 (16.3%)	2114 (16.1%)	256 (18.3%)	0.04
Mechanical ventilation or NIV	7128 (49.1%)	6440 (49.1%)	688 (49.1%)	0.10
Vasopressors/inotropes	3561 (24.6%)	3220 (24.6%)	341 (24.4%)	0.86
Mortality				
ICU	1392 (9.6%)	1253 (9.6%)	139 (9.9%)	0.66
Hospital	2799 (19.3%)	2530 (19.3%)	269 (19.2%)	0.93

Abbreviations: AED = Accident and Emergency Department; AIDS = acquired immunodeficiency syndrome; APACHE = Acute Physiology and Chronic Health Evaluation; CRRT = continuous renal replacement therapy; DM = diabetes mellitus; ENT = ear, nose, and throat; GCS = Glasgow Coma Scale; HD = haemodialysis; HT = hypertension; ICU = intensive care unit; NIV = non-invasive ventilation; NS = neurosurgical; NYHA = New York Heart Association; ORT = orthopaedics; OT = operating theatre; PH = pulmonary hypertension

* Data are shown as median (interquartile range) or No. (%), unless otherwise specified

TABLE 2. Physiological and laboratory parameters during the first 24 hours after admission to the intensive care unit*

	Total (n=14 503)	Development set (n=13 103)	Validation set (n=1400)	P value
Physiological parameters				
Core temp (high), °C	37.8 (37.4-38.4)	37.8 (37.4-38.4)	37.9 (37.4-38.5)	0.12
Core temp (low), °C	36.5 (36.0-37.0)	36.5 (36.0-37.0)	36.5 (36.0-37.0)	0.44
Heart rate (high), beats/min	108 (93-126)	108 (93-126)	109 (94-126)	0.38
Heart rate (low), beats/min	73 (62-86)	73 (62-86)	74 (63-87)	0.06
Respiratory rate (high), breaths/min	26 (22-31)	26 (22-31)	26 (22-31)	0.66
Respiratory rate (low), breaths/min	13 (11-15)	13 (11-15)	13 (11-15)	0.65
Mean blood pressure (high), mm Hg	102 (91-115)	102 (91-115)	102 (91-114)	0.68
Mean blood pressure (low), mm Hg	63 (55-73)	63 (55-73)	64 (55-73)	0.61
Urine output in 24 hours, mL	1388 (856-2120)	1388 (860-2110)	1388 (831-2150)	0.68
Laboratory parameters				
Sodium (high), mmol/L	139 (137-142)	139 (137-142)	139 (137-142)	0.64
Sodium (low), mmol/L	137 (134-139)	137 (134-139)	137 (134-139)	0.98
Potassium (high), mmol/L	4.2 (3.8-4.7)	4.2 (3.8-4.7)	4.2 (3.8-4.7)	0.92
Potassium (low), mmol/L	3.6 (3.3-3.9)	3.6 (3.3-3.9)	3.6 (3.3-3.9)	0.10
Urea (high), mmol/L	7.5 (5.0-13.3)	7.5 (5.0-13.3)	7.6 (5.0-13.4)	0.97
Creatinine (high), μmol/L	92 (67-174)	92 (67-174)	94 (68-176)	0.23
Creatinine (low), μmol/L	76 (60-129)	76 (60-129)	78 (61-131)	0.19
Albumin (high), g/L	31.4 (25.8-36.6)	31.4 (25.8-36.7)	31.5 (25.9-36.6)	0.90
Albumin (low), g/L	29.0 (23.4-34.2)	29.0 (23.4-34.2)	28.8 (23.0-33.9)	0.48
Bilirubin (high), μmol/L	13.5 (8.8-22.8)	13.5 (8.8-22.7)	14.0 (8.7-23.1)	0.90
White cell count (high), × 10^9/L	13.6 (9.9-18.4)	13.6 (10.0-18.4)	13.7 (9.6-17.9)	0.33
White cell count (low), × 10^9/L	10.3 (7.4-13.9)	10.3 (7.4-13.9)	10.2 (8.6-12.0)	0.55
Haemoglobin (high), g/dL	11.5 (9.8-13.2)	11.5 (9.9-13.2)	11.3 (9.7-13.1)	0.02
Haemoglobin (low), g/dL	10.4 (8.6-12.1)	10.4 (8.6-12.1)	10.2 (8.6-12.0)	0.14
Haematocrit (high)	0.34 (0.90-0.39)	0.34 (0.30-0.39)	0.34 (0.29-0.39)	0.02
Haematocrit (low)	0.31 (0.26-0.36)	0.31 (0.26-0.36)	0.31 (0.26-0.35)	0.17
Platelet (high), × 10^9/L	210 (153-276)	210 (154-276)	211 (153-274)	0.49
Platelet (low), × 10^9/L	181 (126-240)	181 (126-240)	180 (124-240)	0.70
Glucose (high), mmol/L	9.6 (7.8-12.6)	9.6 (7.8-12.6)	9.6 (7.7-12.4)	0.44
Glucose (low), mmol/L	5.8 (4.8-7.2)	5.8 (4.8-7.2)	5.8 (4.8-7.3)	0.95
1st pH	7.38 (7.30-7.43)	7.38 (7.30-7.43)	7.38 (7.31-7.43)	0.70
1st $PaCO_2$ (mm Hg)	36.8 (30.8-43.2)	36.8 (30.8-43.2)	36.6 (30.9-43.5)	0.97
1st PaO_2 (mm Hg)	112.3 (86.3-150.8)	111.7 (86.3-150.0)	113.3 (87.0-151.5)	0.54
FiO_2 use during 1st blood gas test	0.30 (0.25-0.40)	0.30 (0.25-0.40)	0.30 (0.25-0.41)	0.86

Abbreviations: FiO_2 = fraction of inspired oxygen; $PaCO_2$ = partial pressure of carbon dioxide; PaO_2 = partial pressure of oxygen

* Data are shown as median (interquartile range), unless otherwise specified

In the development set, the ANN model (AUROC=0.89, 95% confidence interval [CI]=0.88-0.92, Brier score=0.10; P in Hosmer–Lemeshow test=0.34) outperformed the APACHE II model (AUROC=0.80, 95% CI=0.79-0.81, Brier score=0.15; P<0.001) and APACHE IV model (AUROC=0.84, 95% CI=0.83-0.85, Brier score=0.12; P<0.001) for prediction of hospital mortality. The cross-validation accuracy ranged from 0.98 to 1 (mean=0.99), indicating that our ANN model had good stability. There was no statistically significant difference between our ANN model and an ANN model created by oversampling of minority classes (AUROC=0.89, 95% CI=0.89-0.90; P=0.103).

In the validation set, the ANN model (AUROC=0.88, 95% CI=0.86-0.90, Brier score=0.10, P in Hosmer–Lemeshow test=0.37) was superior to the APACHE II model (AUROC=0.85, 95% CI=0.80-0.85, Brier score=0.14; P<0.001 for both comparisons of AUROCs and Brier scores) but similar to the APACHE IV model (AUROC=0.87, 95% CI=0.85-0.89, Brier score=0.11; P=0.34 for comparison of AUROCs, and P=0.05 for comparison of Brier scores) [Fig 1].

The calibration curve for the validation set showed that the ANN model (Fig 2a) outperformed the APACHE IV model (Fig 2b) and the APACHE II model (Fig 2c).

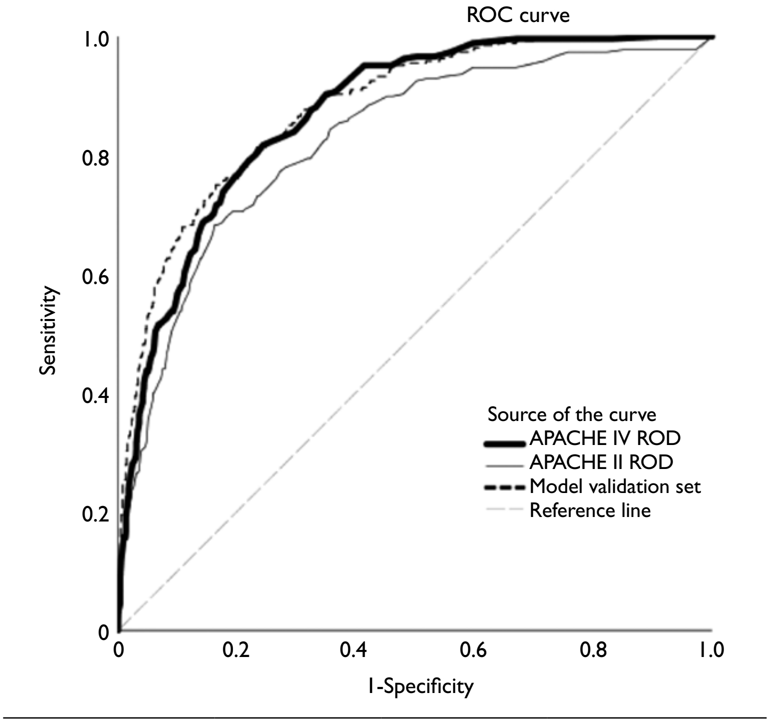
The importances of the predictor variables in predictions of hospital mortality using the ANN model were evaluated. Within 24 hours of ICU admission, the highest sodium level was the most important variable, followed by the highest bilirubin level and the lowest white cell count. Details regarding the normalised importance of each covariate are presented in online supplementary Tables 1 and 2.

Discussion

To our knowledge, this is the first study in Asia to assess the performance of ANN and compare it with the performances of two extensively validated and benchmark scoring systems—APACHE II and APACHE IV—in terms of predicting hospital mortality among critically ill patients. We found that the ANN model provided better discrimination and calibration compared with the APACHE II model. However, the difference between the ANN and APACHE IV models was less prominent. Calibration was slightly better with the ANN model, but discrimination was similar between the ANN and APACHE IV models.

Conventional logistic regression–based APACHE systems often lose calibration over time and require regular updates to maintain performance.[6,11,20-22] The original APACHE II model was developed over 30 years ago using data from 13 different hospitals in the US; it was validated in the country before clinical application.[2] Studies in Hong Kong[7] and Singapore[23] have shown that the APACHE II model has good discrimination but poor calibration for ICU patients in Asia. Calibration remained suboptimal regardless of customisation as demonstrated by Lew et al,[23] indicating the need for a new prognostic prediction model. Wong and Young[24] showed that the APACHE II model had equivalent performance status compared with an ANN model that had been trained and validated using the original APACHE II data. In a medical-neurological ICU in India, an ANN model trained on an Indian population (with or without redundant variables) demonstrated better calibration compared with the APACHE II model.[25] The authors speculated that this finding was partly related to differences in standards of care and resources between American and Indian ICUs.[25] Overall, differences in case mix, advances in medical technology, and the use of more recent data may explain the superiority of our ANN model compared with the APACHE II model.

Compared with ICU patients in the US, it is fivefold more common for Hong Kong ICU patients to begin renal replacement therapy.[26] More than 50% of critically ill patients in Hong Kong require mechanical ventilation, compared with 28% in the US.[26,27] A recent population-based study of all patients admitted to adult ICUs in Hong Kong between 2008 and 2018 showed that the APACHE IV standardised mortality ratio decreased from 0.81 to 0.65 during the study period, implying a gradual decline in the performance of the APACHE IV model.[26] This model, which was established using data derived from >100 000 ICU patients in 45 US hospitals between 2002 and 2003,[5] also tends to overestimate hospital mortality among ICU patients in Hong Kong. In contrast to our study population, where Asian ethnicities were



	Area	Standard error	Asymptotic 95% confidence interval
ANN model	0.88	0.01	0.86-0.90
APACHE IV model	0.87	0.01	0.85-0.89
APACHE II model	0.85	0.01	0.80-0.85

FIG 1. Receiver operating characteristic curves for different models (validation set)
Abbreviations: ANN = artificial neural network; APACHE = Acute Physiology and Chronic Health Evaluation; ROC = receiver operating characteristic; ROD = risk of death

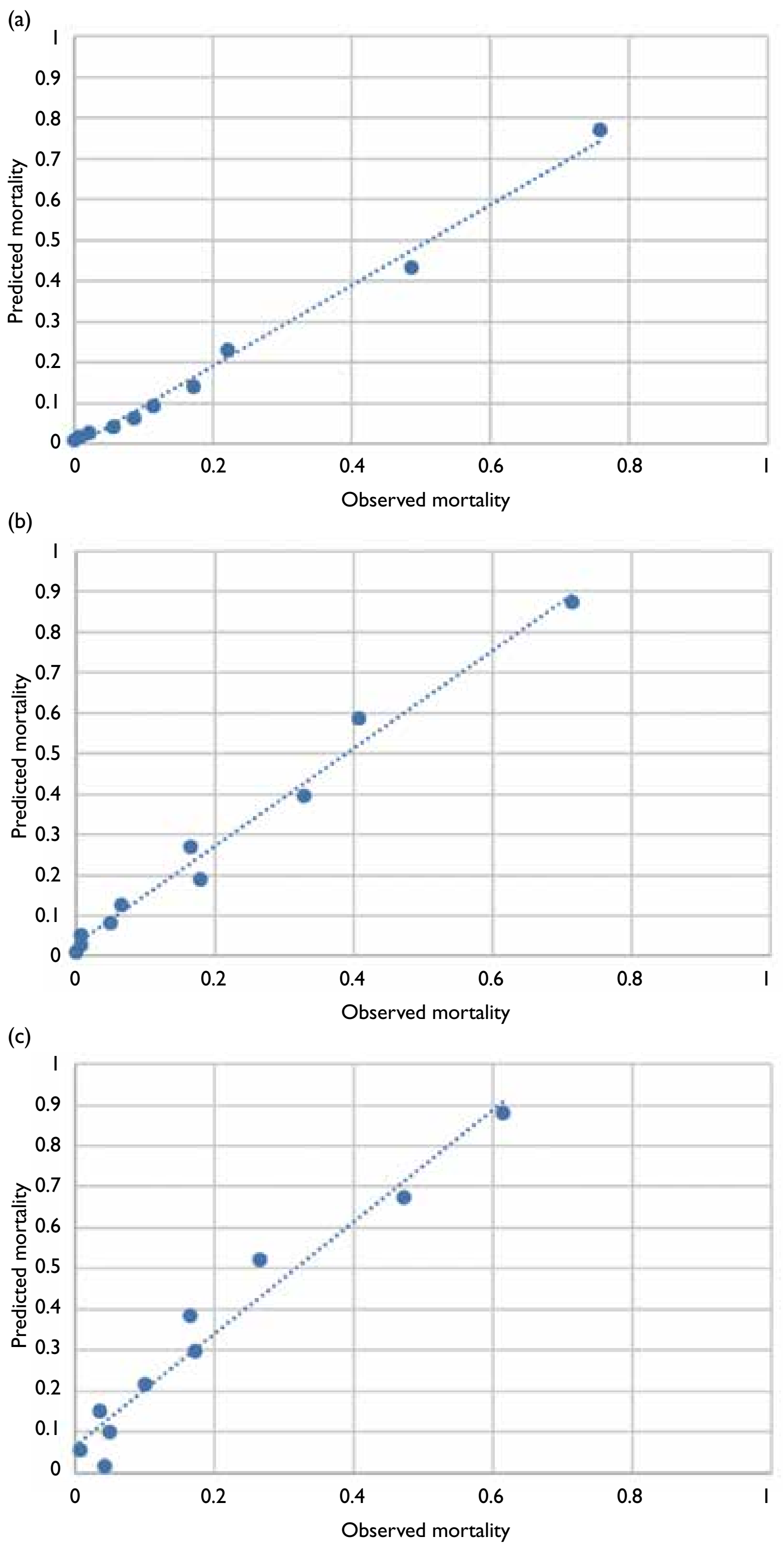


FIG 2. Calibration curves for different models (validation set). (a) Artificial neural network model. (b) Acute Physiology and Chronic Health Evaluation (APACHE) IV model. (c) APACHE II model

most common, 70% of the patients in APACHE IV reference population were Caucasian.[5] The subtle differences in performance between our ANN model and the APACHE IV model could be related to differences in timing during the development of the models. Nevertheless, our ANN model trained on a Hong Kong population was better calibrated for prediction in such a population, compared with the APACHE IV model. This improved calibration could be related to differences in target population (Asian vs Caucasian), epidemiology, and disease profile.

The selection of appropriate variables is a key aspect of model development. The inclusion of additional predictor variables does not necessarily improve a model's overall performance. Redundant variables may result in overfitting and produce a complicated predictive model without additional benefits. A recently published large national cohort study from Sweden showed that a simplified ANN model with eight parameters outperformed the Simplified Acute Physiology Score III model in terms of discrimination and calibration.[28] Among the eight parameters, age and leukocyte count were the most and least important variables, respectively. Notably, leukocyte count was the most important variable in terms of predicting mortality among patients on continuous renal replacement therapy.[29] Similar to the present study, Kang et al[29] found that age was the 12th most important variable. The overall performance of an ANN model trained with APACHE II parameters in an Indian population could be maintained with the 15 highest information gain variables, including serum sodium level and leukocyte count.[25]

Among the 53 parameters in our ANN model, the highest sodium level, highest bilirubin level, and lowest white cell count within 24 hours of ICU admission were the top three most important predictor variables (online supplementary Table 1). The association between acquired hypernatraemia and increased hospital mortality among critically patients has consistently been demonstrated in multiple studies.[30,31] Hyperbilirubinaemia, another complication in patients with sepsis, was associated with the onset of acute respiratory distress syndrome.[32] Sepsis and gastrointestinal/hepatobiliary diseases caused ICU admission in approximately 40% of our patients, possibly explaining the importance of hyperbilirubinaemia in our ANN model. Although the importance of leukocyte count has been demonstrated in other mortality prediction models, the previous models did not specify whether the count was high or low.[25,28,29] In the present study, the lowest white cell count was more important than the highest white cell count. Another intriguing observation was that age constituted the 11th most important predictor in our ANN model (online supplementary Table 1). Age is a predictor of survival in many prognostic models.[3,5,28] Increasing biological age is often associated with multiple co-morbidities and a progressive decline in physiological reserve, leading to increased mortality. However, a recently published systematic review of 129 studies showed large variations in ICU and hospital mortality rates among older ICU patients, ranging from 1% to 51% in single-centre retrospective studies and 6% to 28% in multicentre retrospective studies.[33] These results could be related to differences in admission policies, premorbid functional status, and the intensity of provided to older critically ill patients.

Our ANN model was trained and internally validated on a large number of representative data samples that included most patients admitted to a tertiary ICU in Hong Kong over the past decade. This approach addressed the small sample size limitation that was common in previous studies.[24,25,34] All data were automatically collected by a computer system, eliminating the risk of human error during data extraction. Healthcare system digitalisation and advances in information technology have enabled effortless generation of abundant clinical data (eg, physiological parameters, laboratory results, and radiological findings), which can facilitate data collection and development of a new risk prediction model via machine learning.[35,36] We hope that generalisability to other ICUs in Asia can be achieved through external validation studies.

Limitations

This study had some limitations. Although the sample size was large, all data were collected from a single centre; in contrast, data for the APACHE scoring system were derived from multiple large centres. Because the primary objective of the present study was comparison of performance between our ANN model and the APACHE II and APACHE IV models using identical parameters, we did not attempt to determine the optimal subset of parameters that would maintain high ANN performance.[25,28] Furthermore, our ANN model may not be applicable to other centres with different case mixes and medical approaches. The lack of external validation may lead to concerns about overfitting, which is a common challenge in ANN model development. Because mortality prediction among ICU patients is a dynamic process, other limitations include the use of static data and the lack of a fixed time point for mortality assessment.

Conclusion

Mortality prediction among critically patients is a challenging endeavour. Our ANN model, which was trained with representative data from a Hong Kong population, outperformed the internationally validated APACHE II model with respect to critically ill patients in Hong Kong. In contrast to the APACHE IV model, our ANN model demonstrated better calibration but similar discrimination performance. External validation studies using data from other hospitals are recommended to confirm our findings. Future studies should explore the feasibility of reducing the number of variables while preserving the discrimination and calibration power of the ANN model. The widespread use of computerised information systems, rather than paper records, in ICU and general ward settings has led to increased data availability. Artificial neural networks, along with other machine learning techniques, are valuable tools that can enhance real-time prognostic prediction.

Author contributions

Concept or design: S Lau, HP Shum, CCY Chan.
Acquisition of data: S Lau, HP Shum, CCY Chan.
Analysis or interpretation of data: S Lau, HP Shum, CCY Chan.
Drafting of the manuscript: S Lau.
Critical revision of the manuscript for important intellectual content: MY Man, KB Tang, KKC Chan, AKH Leung, WW Yan.

All authors had full access to the data, contributed to the research, approved the final version for publication, and take responsibility for its accuracy and integrity.

Conflicts of interest

All authors have disclosed no conflicts of interests.

Declaration

Part of the research was presented at the 34th Annual Congress of the European Society of Intensive Care Medicine (3-6 October 2021, virtual) and the Annual Scientific Meeting 2021 of Hong Kong Society of Critical Care Medicine (12 December 2021, virtual).

Funding/support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics approval

The study protocol complies with the Declaration of Helsinki and was approved by the Hong Kong East Cluster Research Ethics Committee of Hospital Authority, Hong Kong (Ref No.: HKECREC-2021-024). The requirement for patient consent was waived by the Committee due to the retrospective nature of the study.

Supplementary material

The supplementary material was provided by the authors and some information may not have been peer reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by the Hong Kong Academy of Medicine and the Hong Kong Medical Association. The Hong Kong Academy of Medicine and the Hong Kong Medical Association disclaim all liability and responsibility arising from any reliance placed on the content. To view the file, please visit the journal online (https://doi.org/10.12809/hkmj2210235).

References

1. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. Crit Care Med 2011;39:163-9.
2. Breslow MJ, Badawi O. Severity scoring in the critically ill: part 1—interpretation and accuracy of outcome prediction scoring systems. Chest 2012;141:245-52.
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818-29.
4. APACHE III study design: analytic plan for evaluation of severity and outcome [editorial]. Crit Care Med 1989;17:S169-221.
5. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med 2006;34:2517-29.
6. Zimmerman JE, Kramer AA. Outcome prediction in critical care: the Acute Physiology and Chronic Health Evaluation models. Curr Opin Crit Care 2008;14:491-7.
7. Tan IK. APACHE II and SAPS II are poorly calibrated in a Hong Kong intensive care unit. Ann Acad Med Singap 1998;27:318-22.
8. Gupta R, Arora VK. Performance evaluation of APACHE II score for an Indian patient with respiratory problems. Indian J Med Res 2004;119:273-82.
9. Choi JW, Park YS, Lee YS, et al. The ability of the Acute Physiology and Chronic Health Evaluation (APACHE) IV score to predict mortality in a single tertiary hospital. Korean J Crit Care Med 2017;32:275-83.
10. Ghorbani M, Ghaem H, Rezaianzadeh A, Shayan Z, Zand F, Nikandish R. A study on the efficacy of APACHE-IV for predicting mortality and length of stay in an intensive care unit in Iran. F1000Res 2017;6:2032.
11. Ko M, Shim M, Lee SM, Kim Y, Yoon S. Performance of APACHE IV in medical intensive care unit patients: comparisons with APACHE II, SAPS 3, and MPM0 III. Acute Crit Care 2018;33:216-21.
12. Xie J, Su B, Li C, et al. A review of modeling methods for predicting in-hospital mortality of patients in intensive care unit. J Emerg Crit Care Med 2017;1:18.
13. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49:1225-31.
14. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001;91(8 Suppl):1636-42.
15. Meiring C, Dixit A, Harris S, et al. Optimal intensive care outcome prediction over time using machine learning. PLoS One 2018;13:e0206862.
16. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18.
17. Norgaard M, Ravn O, Poulsen NK, Hansen LK. Neural Networks for Modelling and Control of Dynamic Systems: A Practitioner's Handbook. London: Springer; 2000.
18. Ludermir TB, Yamazaki A, Zanchettin C. An optimization methodology for neural network weights and architectures. IEEE Trans Neural Netw 2006;17:1452-9.
19. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.
20. Lam KW, Lai KY. Evaluation of outcome and performance of an intensive care unit in Hong Kong by APACHE IV model: 2007-2014. J Emerg Crit Care Med 2017;1:16.
21. Paul E, Bailey M, Van Lint A, Pilcher V. Performance of APACHE III over time in Australia and New Zealand: a retrospective cohort study. Anaesth Intensive Care 2012;40:980-94.
22. Mann SL, Marshall MR, Holt A, Woodford B, Williams AB. Illness severity scoring for intensive care at Middlemore Hospital, New Zealand: past and future. N Z Med J 2010;123:47-65.
23. Lew CC, Wong GJ, Tan CK, Miller M. Performance of the Acute Physiology and Chronic Health Evaluation II (APACHE II) in the prediction of hospital mortality in a mixed ICU in Singapore. Proc Singapore Healthc 2019;28:147-52.
24. Wong LS, Young JD. A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. Anaesthesia 1999;54:1048-54.
25. Nimgaonkar A, Karnad DR, Sudarshan S, Ohno-Machado L, Kohane I. Prediction of mortality in an Indian intensive care unit. Comparison between APACHE II and artificial neural networks. Intensive Care Med 2004;30:248-53.
26. Ling L, Ho CM, Ng PY, et al. Characteristics and outcomes of patients admitted to adult intensive care units in Hong Kong: a population retrospective cohort study from 2008 to 2018. J Intensive Care 2021;9:2.
27. Wunsch H, Angus DC, Harrison DA, Linde-Zwirble WT, Rowan KM. Comparison of medical admissions to intensive care units in the United States and United Kingdom. Am J Respir Crit Care Med 2011;183:1666-73.
28. Holmgren G, Andersson P, Jakobsson A, Frigyesi A. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. J Intensive Care 2019;7:44.
29. Kang MW, Kim J, Kim DK, et al. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. Crit Care 2020;24:42.
30. O'Donoghue SD, Dulhunty JM, Bandeshe HK, Senthuran S, Gowardman JR. Acquired hypernatraemia is an independent predictor of mortality in critically ill patients. Anaesthesia 2009;64:514-20.
31. Olsen MH, Møller M, Romano S, et al. Association between ICU-acquired hypernatremia and in-hospital mortality: data from the medical information mart for intensive care III and the electronic ICU collaborative research database. Crit Care Explor 2020;2:e0304.
32. Zhai R, Sheu CC, Su L, et al. Serum bilirubin levels on ICU admission are associated with ARDS development and mortality in sepsis. Thorax 2009;64:784-90.
33. Vallet H, Schwarz GL, Flaatten H, de Lange DW, Guidet B, Dechartres A. Mortality of older patients admitted to an ICU: a systematic review. Crit Care Med 2021;49:324-34.
34. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 2001;29:291-6.
35. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17:232-43.
36. Bulgarelli L, Deliberato RO, Johnson AE. Prediction on critically ill patients: the role of "big data". J Crit Care 2020;60:64-8.