Performance of critical care prognostic scoring systems in low and middle-income countries: a systematic review

Rashan Haniffa 1,2,3,4*, Ilhaam Isaam 2,3, A. Pubudu De Silva 2,4, Arjen M. Dondorp 1,5 and Nicolette F. De Keizer 6

Abstract

Background: Prognostic models—used in critical care medicine for mortality predictions, for benchmarking and for illness stratification in clinical trials—have been validated predominantly in high-income countries. These results may not be reproducible in low or middle-income countries (LMICs), not only because of different case-mix characteristics but also because of missing predictor variables. The study objective was to systematically review literature on the use of critical care prognostic models in LMICs and assess their ability to discriminate between survivors and non-survivors at hospital discharge of those admitted to intensive care units (ICUs), their calibration, their accuracy, and the manner in which missing values were handled.

Methods: The PubMed database was searched in March 2017 to identify research articles reporting the use and performance of prognostic models in the evaluation of mortality in ICUs in LMICs. Studies carried out in ICUs in high-income countries or paediatric ICUs and studies that evaluated disease-specific scoring systems, were limited to a specific disease or single prognostic factor, were published only as abstracts, editorials, letters and systematic and narrative reviews or were not in English were excluded.

Results: Of the 2233 studies retrieved, 473 were searched and 50 articles reporting 119 models were included. Five articles described the development and evaluation of new models, whereas 114 articles externally validated Acute Physiology and Chronic Health Evaluation, the Simplified Acute Physiology Score and Mortality Probability Models or versions thereof. Missing values were only described in 34% of studies; exclusion and or imputation by normal values were used. Discrimination, calibration and accuracy were reported in 94.0%, 72.4% and 25% respectively. Good discrimination and calibration were reported in 88.9% and 58.3% respectively. However, only 10 evaluations that reported excellent discrimination also reported good calibration. Generalisability of the findings was limited by variability of inclusion and exclusion criteria, unavailability of post-ICU outcomes and missing value handling.

Conclusions: Robust interpretations regarding the applicability of prognostic models are currently hampered by poor adherence to reporting guidelines, especially when reporting missing value handling. Performance of mortality risk prediction models in LMIC ICUs is at best moderate, especially with limitations in calibration. This necessitates continued efforts to develop and validate LMIC models with readily available prognostic variables, perhaps aided by medical registries.

Keywords: Critical care, Prognostic model, ICU scoring system, Model performance, Low and middle-income countries, Resource-limited settings
Background
Prognostic models used in critical care medicine for mortality predictions, for benchmarking and for illness stratification in clinical trials need to be validated for the relevant setting. An ideal model should have good discrimination (the ability to differentiate between high-risk and low-risk patients) and good calibration (generate risk estimates close to actual mortality) [1]. Acute Physiology and Chronic Health Evaluation (APACHE) or the Simplified Acute Physiology Score (SAPS) and the Mortality Probability Models (MPM) are some common prognostic systems used to predict the outcome of critically ill patients admitted to the intensive care unit (ICU) [2, 3].

The performance of these models has been extensively validated, predominantly in high-income countries (HICs) [4–6]. These results may not be reproducible in low or middle-income countries (LMICs), not only because of different case-mix characteristics but also because of missing predictor variables. Predictor variables that are routinely available in HIC ICUs (e.g. arterial oxygenation) are often not obtainable or reliable where resources are limited [7, 8]. Furthermore, data collection and recording may not be as robust in these settings as in HICs; paper-based recording systems, limited availability of staff and lack of staff training regarding data collection are frequent challenges [9]. The presence of missing values, if imputed as normal as per convention [3, 4, 10–13], will lead to underestimation of the scores and mortality. As part of quality improvement initiatives within ICUs, severity-adjusted mortality rates, which are calculated based on these prognostic systems, are increasingly used as tools for evaluating the impact of new therapies or organisational changes and for benchmarking; therefore, underestimating the risk could result in erroneous admission policies and an underestimation of the quality of care, performance and effectiveness when used for benchmarking [14]. Additionally, the diagnostic categories in these prognostic models may not be suited to capture diagnoses more common in these countries, such as dengue, malaria, snakebite and organophosphate poisoning. Furthermore, hospital discharge outcomes may not be readily accessible [15–17]. These and other factors influence the performance of the models, which may then require adjustment in the form of recalibration (adjustment of the intercept of the model and overall adjustment of the associations (relative weights) of the predictors with the outcome) and/or model revision (adjustment of individual predictor-outcome associations and addition or removal of new predictors) [18–20].

The objective of this article is to systematically review literature on the use of critical care prognostic models in LMICs and assess their ability to discriminate between survivors and non-survivors at hospital discharge of those admitted to ICUs, their calibration and accuracy, and the manner in which missing values are handled.

Methods

Literature search and eligibility criteria
The PubMed database was searched in March 2017, for research articles using the following search strategy: (critical OR intensive) AND (mortality OR survival OR prognostic OR predictive) AND (scoring system OR rating system OR APACHE OR SAPS OR MPM) in the title, abstract and keywords (Additional file 1).

No restrictions were placed on date of publication. Titles and abstracts returned were analysed for eligibility (RH, II). Abstracts reporting the performance of prognostic models were hand searched to identify studies carried out in ICUs in LMICs (as classified by the World Bank [21]) and full-text copies retrieved. Full-text articles were also retrieved when the title or abstract did not provide the country setting. The references of all selected reports were thereafter cross-checked for other potentially relevant articles.

The inclusion criteria for this review were studies carried out in ICUs in LMICs; those evaluating or developing prognostic models in adult ICU patients designed to predict mortality, whether ICU or hospital mortality.

The exclusion criteria for this review were: studies carried out only in ICUs in HICs or in paediatric ICUs; organ failure scoring systems such as SOFA that are not designed for predicting mortality; studies evaluating models in relation to a specific disease (e.g. liver cirrhosis) or limited to trauma patients; those assessing a single prognostic factor (e.g. microalbuminuria); studies published in languages other than English; studies published only as abstracts, editorials, letters and systematic or narrative reviews; and duplicate publications.

Where ICUs in both HICs and LMICs were included in a study, only data from the low/middle-income country were to be extracted. Likewise, where a single-factor or disease-specific scoring system and a non-specialty-specific scoring system were evaluated, only the data pertaining to the latter were extracted. Studies where both adult and paediatric patients were admitted to the same ICU and studies where the age limits of patients were not specified were to be included in this review.

Data extraction and critical appraisal
The full-text articles were reviewed to assess eligibility for inclusion in the report. Disagreements between the two reviewers were resolved by discussion. The list of extracted items was based on the guidance issued by Cochrane for data extraction [22] and critical appraisal for systematic reviews of prediction models (the CHARMS checklist [23]). A second reviewer checked extracted items classed as “not reported” or “unclear.”
or unexpected findings. If an article described multiple models, separate data extraction was carried out for each model.

**Descriptive analyses**

Results were summarised using descriptive statistics. A formal meta-analysis was not planned as it was envisaged that the studies would be too heterogeneous, and a narrative synthesis was undertaken. Discrimination was assessed by the area under the receiver operating characteristic (AUROC) when reported [24]. Discrimination was considered excellent, very good, good, moderate or poor with AUROC values of $0.9–0.99$, $0.8–0.89$, $0.7–0.79$, $0.6–0.69$ and $≤0.6$, respectively [25, 26]. Calibration was assessed by the Hosmer–Lemeshow $C$ statistic (significant departures from perfect calibration were inferred when $p$ values were less than $0.05$ [24, 26]).

**Results**

**Study characteristics**

Of the 2233 studies obtained from PubMed searches, 473 were searched and 43 met the inclusion criteria. Seven further studies were included after cross-checking the reference lists of the selected studies (Fig. 1). Fifty studies met the review criteria and were selected for analysis.

**Quality assessment**

Study quality was assessed in accordance with the CHARMS guidelines [23] and is presented as Additional file 2. Variations existed in the conduct and reporting of the studies, especially with regard to inclusion and exclusion criteria, missing value handling, and performance and outcome measures.

---

![Study selection](image-url)

**Fig. 1 Study selection**
Forty-three of the studies were carried out prospectively. The studies were carried out in 19 different LMICs, with the largest number carried out in India (studies = 11, models evaluated = 22), Thailand (studies = 6, models evaluated = 17) and Brazil (studies = 6, models evaluated = 17) (Table 1). Model adjustment was most frequent in India (n = 4 models). Settings, hospital and ICU characteristics are presented in Additional file 2.

Sample sizes ranged from 48 to 5780, and participant ages ranged from 1 month to 100 years (Table 1). Of the 33 studies reporting a lower age limit, 17 reported participants under the age of 18 years (Table 1). Model performance is described in the following in terms of discrimination, calibration and accuracy.

Discriminatory ability of models
Discrimination was reported for 104 (91.2%) of the evaluated models (Tables 5, 6 and 7). In three evaluations (two studies [45, 46]) it was reported as sensitivity and specificity only. In 101 model performance evaluations, discrimination was reported as the AUROC; in four of these evaluations AUROC was presented as a figure and a numerical value could not be ascertained [47, 48]. Where the AUROC was reported in numerical form (97 model performance evaluations) a confidence interval was only reported in 63 evaluations.

Where the AUROC was reported as a numerical value, 21 evaluations (21.7%) reported excellent discrimination. For all versions of APACHE II, SAPS II, SAPS 3 and MPM II, excellent discrimination was reported in 16.1%, 11.5%, 47.7% and 36.4% of the model evaluations respectively. Sixty-six (68.0%) model evaluations reported very good or good discrimination; for all versions of APACHE II this was 67.7%, for SAPS II was 80.8%, for SAPS 3 was 58.3% and for MPM II it was 45.5%. Poor discrimination was reported on one occasion only, for an evaluation of SAPS II [49].

Excellent discrimination was reported more frequently when hospital mortality (n = 15, 25%) was the outcome in comparison to when it was ICU mortality (n = 6, 10%). Normal value imputation resulted in better discrimination (n = 4, 25% excellent and n = 9, 56.25% very good) than exclusion (n = 1, 8.33% excellent and n = 3, 25.0% very good) or where missing values were not reported (n = 16, 19.0% excellent and n = 32, 38.1% very good). Discrimination was better for all models with scores calculated further into the ICU stay when compared with those calculated earlier on [32, 48, 50].

Four (n = 2 studies) of the six evaluations with model adjustments compared them to the original model (Table 3). However an independent validation set was employed in only one study (three validations), where the models were recalibrated [51]. For all three modes (APACHE II, SAPS II and SAPS 3), recalibration resulted in the improvement of previously poor
| Study (country) | Scoring systems | Sample size (age in years) | (Outcome) and model performance measures | Study (country) | Scoring system | Sample size (age in years) | (Outcome) and model performance measures |
|----------------|-----------------|---------------------------|----------------------------------------|----------------|----------------|---------------------------|----------------------------------------|
| Abhinandan and Vedavathi, 2013 (India) [60] | APACHE II | 50 (18–90, mean 48.36) | (Unspecified mortality) D: not reported; C: not reported; CM: not reported; comparison of means | Liu et al., 2016 (China) [61] | APACHE II | 137 (mean 69.53 ± 9.28) | (Hospital mortality) D: AUROC; C: not reported; CM: sensitivity, specificity |
| Aggarwal et al., 2006 (India) [7] | APACHE II, SAPS II, MPM II, MPM II 24 | 459 (16–80) | (Hospital mortality) D: AUROC; C: HL; H statistics; CM: sensitivity, specificity, NPV, PPV, accuracy | Mohan et al., 2015 (India) [44] | APACHE II, SAPS II, SAPS 3 | 100 (mean survivors 44.29 ± 15.53, non-survivors 57.37 ± 20.42) | (ICU mortality) D: not reported; C: not reported; CM: sensitivity, specificity; comparison of means |
| Ahluwalia et al., 1999 (India) [29] | APACHE II | 79 (13–80, mean 46) | (Hospital mortality) D: AUROC; C: not reported; CM: sensitivity, specificity | Nair et al., 2016 (India) [62] | SAPS | 48 (mean: survivors 53.56, non-survivors 47.78) | (ICU mortality) D: AUROC; C: not reported |
| Celik et al., 2014 (Turkey) [63] | SAPS II | 545 (>18) | (ICU mortality) D: not reported; C: not reported; CM: sensitivity, specificity | Naqvi et al., 2016 (Pakistan) [33] | APACHE II | 96 (mean 32.93 ± 16.61) | (ICU mortality) D: AUROC; C: HL |
| Chang et al., 2006 (Taiwan) [64] | APACHE II | 1263 (13–92, mean 56) | (Hospital mortality) D: AUROC; C: HL | Nassar et al., 2012 (Brazil) [42] | APACHE IV, SAPS 3, MPM III | 5780 (median 66, IQR 47–79) | (Hospital mortality) D: AUROC; C: HL; C statistic |
| Chiavone and Rasslan, 2005 (Brazil) [65] | APACHE II | 94 (16–97) | (Hospital mortality) D: AUROC; C: calibration curve stratified in 10% risk bands, R² | Naved et al., 2011 (Pakistan) [34] | APACHE II (48 hours) | 253 (15–84, mean 51.26 ± 179) | (ICU mortality) D: not reported; C: chi-square test |
| Eapen et al., 1997 (India) [30] | APACHE II (VA) | 500 (13–84) | (Unspecified mortality) D: AUROC (CI not reported); C: not reported | Nimgaonkar et al., 2004 (India) [35] | APACHE II, ANN22, ANN15 | 2062 (mean 37.5 ± 16.1) | (Hospital mortality) D: AUROC; C: HL |
| Evran et al., 2016 (Turkey) [47] | APACHE II, APACHE IV | 487 (18–96, mean 58.58 ± 18 | (ICU mortality) D: AUROC (as figure); C: not reported; CM: sensitivity, specificity, accuracy, PPV and NPV | Nouira et al., 1998 (Tunisia) [66] | APACHE II, SAPS II, MPM 0, MPM 24 | 1325 (mean 46.6 ± 20.4) | (Hospital mortality) D: AUROC; C: HL; H statistics; CM: NPV, PPV |
| Fadaizadeh et al., 2012 (Iran) [67] | APACHE II, SAPS II | 415 (mean 49.28 ± 0.94) | (ICU mortality) D: AUROC; C: HL; CM: sensitivity, specificity | Ratanarat et al., 2005 (Thailand) [68] | APACHE II | 482 (mean 56 ± 21) | (ICU mortality) D: AUROC; C: not reported; CM: sensitivity, specificity |
| Study (country) | Scoring systems | Sample size (age in years) | (Outcome) and model performance measures | Study (country) | Scoring system | Sample size (age in years) | (Outcome) and model performance measures |
|----------------|----------------|---------------------------|----------------------------------------|----------------|----------------|---------------------------|----------------------------------------|
| Faruq et al., 2013 (Bangladesh) [37] | APACHE II, SAPS II | 194 (mean 61.06 ± 15.42) | ICU mortality: D: AUROC; C: HL; C statistic | Riwello et al., 2016 (Rwanda) [43] | R-MPM (0), Simplified R-MPM | 427 (median 34, IQR 25–47) | Hospital mortality: D: AUROC; C: HL |
| Galal et al., 2013 (Egypt) [49] | APACHE II, SAPS II | 105 (0–88, mean 54.9 ± 15.75) | ICU mortality: D: AUROC (CI not reported); C: HL; CM: sensitivity, specificity, accuracy | Sathe and Bapat, 2014 (India) [69] | APACHE II, SAPS II | 1543 (mean 53.9 ± 16.8) | Hospital mortality: D: AUROC; C: HL |
| Gilani et al., 2014 (Iran) [70] | APACHE II, APACHE III, SAPS II | 202 (14–85, mean 53.1 ± 203) | Hospital mortality: D: AUROC (CI not reported); C: HL; CM: sensitivity, specificity, accuracy | Sekulic et al., 2015 (Serbia) [48] | APACHE II, MPM II (0), MPM II (24), MPM II (48), MPM II (72), MPM II (7 days), SAPS II | 60 (mean 59.0 ± 15.8) | ICU mortality: D: AUROC (as figure for MPM II (24)); CM: sensitivity, specificity |
| Godinjak et al., 2016 (Bosnia and Herzegovina) [31] | APACHE II, SAPS II | 174 (19–87, mean 61.7 ± 163) | Hospital mortality: D: AUROC; C: not reported; CM: sensitivity, specificity | Shoukat et al., 2016 (Pakistan) [71] | APACHE IV | 155 (13–70, mean 38.39 ± 13.61) | (Unspecified mortality): D: not reported; C: not reported; CM: not reported; comparison of means |
| Gupta and Arora, 2004 (India) [72] | APACHE II | 330 (mean 43.32 ± 16.2) | ICU mortality: D: AUROC (CI not reported); C: HL; CM: accuracy | Shrestha et al., 2011 (Nepal) [73] | APACHE III | 117 (16–84, mean 43.18 ± 18.49) | ICU mortality: D: AUROC; C: HL; CM: sensitivity, specificity, NPV, PPV, accuracy |
| Halim et al., 2009 (Indonesia) [75] | APACHE II | 144 (15–85, mean 47.33 ± 17.17) | ICU mortality: D: AUROC (CI not reported); C: HL; CM: sensitivity, specificity, accuracy, NPV, PPV | Soares and Salluh, 2006 (Brazil) [39] | SAPS II, SAPS 3, SAPS 3 (CSA) | 952 (>18, mean 58.3 ± 23.1) | Hospital mortality: D: AUROC; C: HL; C statistic |
| Hamza et al., 2009 (Egypt) [76] | APACHE IV, SAPS II, MPM II | 265 (mean 57.07 ± 16.6) | ICU mortality: D: AUROC; C: HL; H statistic | Soares et al., 2004 (Brazil) [40] | APACHE II, APACHE III, SAPS II, MPM II 0, MPM II 24 | 1257 (18–93, mean 560 ± 16.7) | Hospital mortality: D: AUROC; C: HL; H statistic |
| Hashmi et al., 2016 (Pakistan) [77] | APACHE II, APACHE II (MTA) | 213 (mean 46.31 ± 18.43) | Unspecified mortality: D: AUROC; C: HL; CM: sensitivity, specificity, accuracy | Soares et al., 2010 (Brazil) [41] | SAPS II, SAPS 3, SAPS 3 (CSA), MPM III (24) | 717 (mean 61.2 ± 15.4) | Hospital mortality: D: AUROC; C: HL; C statistic |
| Hernandez et al., 2014 (Philippines) [78] | SAPS 3 | 2426 (mean 62 ± 17) | ICU mortality: D: AUROC (CI not reported); C: HL; C statistic | Surtheechet, 2009 (Thailand) [79] | SAPS II, MPM II (24) | 639 (18–100, mean 569 ± 206) | Hospital mortality: D: AUROC; C: HL; CM: sensitivity, specificity, accuracy, NPV, PPV |

Table 1 Study description (Continued)
### Table 1: Study description (Continued)

| Study (country) | Scoring systems | Sample size (age in years) | (Outcome) and model performance measures | Study (country) | Scoring system | Sample size (age in years) | (Outcome) and model performance measures |
|-----------------|-----------------|----------------------------|------------------------------------------|-----------------|-----------------|----------------------------|------------------------------------------|
| Hosseini and Ramazani, 2015 (Iran) [80] | APACHE II | 150 (3–97) | ICU mortality | Teoh et al., 1991 (Malaysia) [81] | APACHE II | 100 (1 month–82 years) | Hospital mortality | D: not reported; C: not reported; CM: not reported; correlation |
| Juneja et al., 2012 (India) [1] | APACHE II, APACHE III, APACHE IV, SAPS II, SAPS 3, MPM II (0), MPM III (0) | 653 (mean 58.48 ± 18.6) | ICU mortality | Turner et al., 1989 (South Africa) [36] | APACHE II | 728 (12–88, mean 43) | Unspecified mortality | D: not reported; C: not reported; CM: not reported; correlation |
| Khan et al., 2015 (India) [32] | APACHE II APACHE II (48 hours) SAPS II SAPS II (48 hours) | 85 (mean 39.14 ± 17.3) | Hospital mortality | Wilairatana et al., 1995 (Thailand) [82] | APACHE II APACHE III SAPS II | 209 (16–93, mean 55.36 ± 17.44) | Hospital mortality | D: not reported; C: not reported; CM: sensitivity, specificity, accuracy |
| Khawannimit and Geater, 2007 (Thailand) [38] | APACHE II SAPS II | 1316 (mean 55.6 ± 18.2) | Hospital mortality | Xing et al., 2015 (China) [52] | APACHE II APACHE IV SAPS 3 | 981 (mean 64.8 ± 12.1) | Hospital mortality | D: not reported; C: not reported; CM: sensitivity, specificity, accuracy |
| Khawannimit and Bhurayanontachai, 2011 (Thailand) [51] | APACHE II SAPS 3 SAPS 3 (AUS) SAPS II Customised APACHE II (recalibration) Customised SAPS II (recalibration) Customised SAPS 3 (recalibration) | 2022 (median 62, IQR 49–73) | Hospital mortality | Yamin et al., 2011 (Pakistan) [28] | APACHE IV | 162 (9–90, mean 38.024) | ICU mortality | D: not reported; C: not reported; CM: accuracy; comparison of means |
| Kiatboonsri and Charoenpan, 1995 (Thailand) [45] | APACHE II | 334 (15–98, mean 56.01 ± 18.23) | Hospital mortality | Zhao et al., 2013 (China) [50] | SAPS II (initial) SAPS II (48 hours) SAPS II (24 hours) SAPS II (72 hours) Simplified SAPS II (VA) | 1684 (18–98, mean 58.99 ± 18.30) | ICU mortality | D: not reported; C: not reported; CM: sensitivity, specificity, accuracy; correlation |

---

9 New model development

IA variable adjustment, MTA modelling technique adjustment, D discrimination, AUROC area under the receiver operating characteristic curve, C calibration, HL Hosmer–Lemeshow statistic, CM classification measures (e.g. sensitivity, specificity, PPV positive predictive values, NPV negative predictive values, APACHE Acute Physiology and Chronic Health Evaluation, ANN artificial neural network, SAPS Simplified Acute Physiology Score, SAPS (AUS) Simplified Acute Physiology Score (Australian), SAPS (CSA) Simplified Acute Physiology Score (Central and South American), MPM Mortality Probability Models, R-MPM Rwanda-Mortality Probability Model, ICU intensive care unit, IQR interquartile range, CI confidence interval
calibration; and discrimination which was already excellent remained the same.

**Ability of models to calibrate**

Only 82 (71.9%) evaluations reported calibration (Tables 5, 6 and 7). The Hosmer–Lemeshow test was reported for both $C$ and $H$ statistics 17 (20.7%) times, for $C$ statistic only 21 (25.6%) times, for $H$ statistic only nine (10.9%) times and without further detail 35 (42.7%) times.

A value greater than 0.05 for the Hosmer–Lemeshow statistic was reported by 49 (59.8%) evaluations that reported calibration. For all versions of APACHE II, SAPS II, SAPS 3 and MPM II, $p > 0.05$ was reported in 60.9%, 59%, 66.7% and 50% of model performance evaluations respectively.
| Study | Type of adjustment and changes made | Discrimination (original scoring system) | Discrimination (after adjustment) | Calibration (original scoring system) | Calibration (after adjustment) |
|-------|------------------------------------|------------------------------------------|-----------------------------------|--------------------------------------|-------------------------------|
| APACHE II |                                   |                                          |                                   |                                      |                               |
| Khwanimit and Bhurayanontachai (2011) [51] | Recalibration (first-level customisation): customised APACHE II logit = −7.7206 + (APACHE II score × 0.2013) + new diagnostic category weight (Appendix I [51]) | 0.936 (0.925–0.947) (entire population n = 2022) | 0.936 (0.925–0.947) (validation dataset n = 1011) | C statistic χ² = 1042 (p = 0.001), H statistic χ² = 113.1 (p < 0.001) | C statistic χ² = 16.1 (p = 0.009), H statistic χ² = 14.1 (p = 0.17) |
| Eapen et al. (1997) [30] | Variable adjustment: GCS excluded | Not evaluated | 0.6068 | Not reported | Not reported |
| Hashmi et al. (2016) [77] | Modelling technique adjustments: APACHE II calculated automatically by software which uses manually entered values using the logit equation = −4.063 + (APACHE II × 0.181) | 0.823 (0.76–0.88) (manual calculation) | 0.827 (0.77–0.88) (software calculation) | H statistic χ² = 117.6 (p < 0.001) | H statistic χ² = 5.46 (p = 0.71) |
| Nimgaonkar et al. (2004) [35] | Modelling technique adjustments: Artificial Neural Network (ANN 22) model trained on an Indian patient dataset using all 22 APACHE II variables | 0.77 | 0.87 (p < 0.002) | H statistic χ² = 123.5 (p < 0.05) | H statistic χ² = 22.4 (p < 0.05) |
| Nimgaonkar et al. (2004) [35] | Modelling technique adjustments: Artificial Neural Network (ANN 15) model trained on an Indian patient dataset using 15 APACHE II variables | 0.77 | 0.88 (p < 0.001) | H statistic χ² = 123.5 (p < 0.05) | H statistic χ² = 27.7 (p < 0.05) |
| SAPS II |                                   |                                          |                                   |                                      |                               |
| Khwanimit and Bhurayanontachai (2011) [51] | Recalibration (first-level customisation): customised SAPS II logit = −10.1779 + 0.0719 (SAPS II score) + 1.4891 × ln(SAPS II score + 1) | 0.914 (0.901–0.928) (entire population n = 2022) | 0.919 (0.900–0.938) (validation dataset n = 1011) | C statistic χ² = 1249 (p < 0.001), H statistic χ² = 975 (p < 0.001) | C statistic χ² = 8.6 (p = 0.057), H statistic χ² = 9.6 (p = 0.48) |
| Zhao et al. (2013) [50] | Variable adjustment: 1. Underlying disease variables excluded 2. Admission type variables excluded | 0.776 (95% CI 0.750–0.802) at admission, 0.826 (95% CI 0.803–0.850) at 24 hours | Not reported: correlation was suggested between the simplified SAPS II score at each time point and outcome with OR of 1.109 (p = 0.000), regardless of the diagnosis | Not reported | Not reported |
| SAPS 3 |                                   |                                          |                                   |                                      |                               |
| Khwanimit and Bhurayanontachai (2011) [51] | Recalibration (first-level customisation): customised SAPS 3 logit = −33.4249 + ln(SAPS 3 score + 1) × 7.8699 | 0.913 (0.899–0.924) (entire population n = 2022) | 0.917 (0.897–0.937) (validation dataset n = 1011) | C statistic χ² = 170 (p < 0.001), H statistic χ² = 799 (p < 0.001) | C statistic χ² = 8.2 (p = 0.061), H statistic χ² = 799 (p < 0.001) |
| Riviello et al. (2016) [43] | MPM (0) III | Exclusion of two patients (0.5%) due to lack of discharge vital status Normal values attribution details provided in Supplementary Table 3 of the original paper. Highest proportions of missing values were for GCS (36.30%) followed by chronic renal compromise/insufficiency (7.96%) | Not reported | Not reported | |

**APACHE** Acute Physiology and Chronic Health Evaluation, **SAPS** Simplified Acute Physiology Score, **MPM** Mortality Probability Models, **GCS** Glasgow Coma Score, **ICU** intensive care unit, **CI** confidence interval, **OR** odds ratio
Ten evaluations that reported excellent discrimination also reported good calibration. Of these, three were for first-level customisations of APACHE II, SAPS II and SAPS 3 (calibration resulted in $p < 0.05$ for the Hosmer–Lemeshow statistic when the non-customised model was used) [51]. The other evaluations that reported excellent discrimination and good calibration were carried out in three studies; Juneja et al. (APACHE III, APACHE IV, MPM II (initial), MPM III (initial) and SAPS 3) [1], Sekulic et al. (MPM II at 7 days) [48] and Xing et al. (SAPS 3) [52].

A $p$ value greater than 0.05 was reported more frequently when ICU mortality was the outcome ($n = 27, 77.1\%$) than when hospital mortality was the outcome ($n = 13, 27.7\%$). A $p$ value greater than 0.05 for the Hosmer–Lemeshow statistic was obtained through exclusion of missing values 100% of the time ($n = 3$), by normal value imputation 40.9% of the time ($n = 9$) or where missing values were not reported 54.7% of the time ($n = 29$).

**Accuracy of models**

Accuracy was reported for 29 evaluations (25.0%) and ranged from 55.20 to 89.7% (Tables 5, 6 and 7).

**New model development**

Three studies reported five new model developments [35, 36, 43]. These are described in Table 8. For all five new models, the AUROC was higher than that obtained with the original prognostic scoring system on which it was based. A good calibration was reported for both R-MPM and Simplified R-MPM; a poor calibration was reported for MPM-III. A poor calibration was reported for both ANN 22 and ANN 15 as well as for the original APACHE II on which they were based.
| Study | Scoring system | Discrimination | Calibration | Sensitivity | Specificity | Accuracy |
|-------|----------------|----------------|-------------|-------------|-------------|----------|
| Khwannimit and Bhurayanontachai (2011) [51] | APACHE II (recalibrated model) | 0.936 (0.925–0.947) | C statistic $\chi^2 = 16.1$ ($p < 0.001$) | NR | NR | NR |
| Khwannimit and Bhurayanontachai (2011) [51] | APACHE II | 0.936 (0.925–0.947) | C statistic $\chi^2 = 104.2$ ($p < 0.001$) | NR | NR | NR |
| Khan et al. (2015) [32] | APACHE II (48 hours) | 0.933 (0.873–0.992) | NR | 94.1% (DC > 9.5) | 86.3% (DC > 9.5) | NR |
| Godinjak et al. (2016) [31] | APACHE II | 0.920 (0.87–0.97) | NR | 74.5% (DC = 27.5) | 93.4% (DC = 27.5) | NR |
| Khwannimit and Geater (2007) [38] | APACHE II | 0.911 (0.891–0.93) | C statistic $\chi^2 = 66.65$ ($p < 0.001$) | 73.87% (95% CI = 65.23–75.66) | 92% (95% CI = 89.66–93.20) | 87% (95% CI = 85.47–89.13) |
| Fadaizadeh et al. (2012) [67] | APACHE II | 0.897 (0.858–0.937) | C statistic $\chi^2 = 3.27$ ($p = 0.916$) | 90% (DC = 13.5) | 75% (DC = 13.5) | NR |
| Juneja et al. (2012) [1] | APACHE II | 0.894 (0.864–0.925) | $\chi^2 = 7.959$ ($p = 0.438$) | 74.8% (DC > 20.5) | 84.9% (DC > 20.5) | NR |
| Soares et al. (2004) [40] | APACHE II | 0.888 (0.868–0.907) | H statistic $\chi^2 = 78.181$ ($p < 0.001$) | NR | NR | NR |
| Xing et al. (2015) [52] | APACHE II | 0.863 (0.804–0.923) | $\chi^2 = 3.486$ ($p = 0.900$) | NR | NR | NR |
| Sathe and Bapat (2014) [69] | APACHE II | 0.86 | $\chi^2 = 12.8$ ($p = 0.03$) | NR | NR | NR |
| Hosseini and Ramaazani (2015) [80] | APACHE II | 0.857 (0.788–0.925) | $\chi^2 = 10.203$ ($p = 0.251$) | 96.6% | 62.80% | 79.70% |
| Naqvi et al. (2016) [33] | APACHE II | 0.835 | $\chi^2 = 3.199$ ($p = 0.866$) | NR | NR | NR |
| Glani et al. (2014) [70] | APACHE II | 0.828 | $\chi^2 = 5.419$ ($p = 0.712$) | 88.2% (DC = 19) | 65.5% (DC = 19) | 27.9% |
| Hashmi et al. (2016) [77] | APACHE II (automatic calculation using software) | 0.827 (0.77–0.88) | $\chi^2 = 5.46$ ($p = 0.71$) | 55.71% | 90.21% | 78.87% |
| Hashmi et al. (2016) [77] | APACHE II | 0.823 (0.76–0.88) | $\chi^2 = 11.76$ ($p = 0.16$) | 51.42% | 90.91% | 77.99% |
| Chang et al. (2006) [64] | APACHE II | 0.82 | $\chi^2 = 9.8$ ($p = 0.28$) | NR | NR | NR |
| Nouroa et al. (1998) [66] | APACHE II | 0.82 | C statistic $\chi^2 = 25.95$ ($p < 0.001$) | NR | NR | NR |
| Liu et al. (2016) [61] | APACHE II | 0.813 ± 0.055 | NR | 89.6% (DC ≥ 15.0) | 74.8% (DC ≥ 15.0) | NR |
| Ratanarat et al. (2005) [68] | APACHE II | 0.788 | NR | 80.9% (DC = 20) | 63.2% (DC = 20) | NR |
| Khan et al. (2015) [32] | APACHE II | 0.785 (0.69–0.88) | NR | 94.4% (DC > 9.5) | 49% (DC > 9.5) | NR |
| Nirmalakondar et al. (2004) [35] | APACHE II | 0.77 | H statistic $\chi^2 = 12.35$ ($p < 0.05$) | NR | NR | NR |
| Faruq et al. (2013) [37] | APACHE II | 0.75 (0.67–0.82) | C statistic $\chi^2 = 8.304$ ($p = 0.40$) | NR | NR | NR |
| Study                                    | Scoring system | Discrimination | Calibration | Sensitivity | Specificity | Accuracy |
|-----------------------------------------|----------------|----------------|-------------|-------------|-------------|----------|
| Ahluwalia et al. (1999) [29]            | APACHE II      | 0.74           | NR          | 93%         | 23.6%       | NR       |
| Chauone and Rasslan (2003) [65]         | APACHE II      | 0.729 (0.63–0.83) | NR          | NR          | NR          | NR       |
| Wilairatana et al. (1995) [82]          | APACHE II      | 0.723           | NR          | 77.4% (DC = 19) | 61.1% (DC = 19) | 70.8%    |
| Aggarwal et al. (2006) [7]              | APACHE II      | 0.713           | C statistic $\chi^2 = 119.3$ (p < 0.001) | 48% (39.9–56.2) (DC = 25%), 84% (79.6–88.2) (DC = 50%), 97.7% (98.1–100) (DC = 75%) | 20.1% (24.1–27.3) (DC = 25%), 96% (92.7–97.7) (DC = 50%), 3.2% (1.1–7.4) (DC = 75%) | 61.1% (DC = 19) |
| Wilairatana et al. (1995) [82]          | APACHE II      | 0.694           | $\chi^2 = 10.627$ (p = 0.014) | 83%         | 55.2%       | 66%      |
| Gupta and Arora (2004) [72]             | APACHE II      | 0.63            | $\chi^2 = 10.34$ (p > 0.05) | NR          | NR          | 89.7% (DC = 70%) |
| Sekulic et al. (2015) [48]              | APACHE II      | 0.623           | 3.05 (p = 0.031) | Presented as a figure | NR          | 81.80% |
| Eapen et al. (1997) [30]                | APACHE II (VA) | 0.6068         | NR          | NR          | NR          | NR       |
| Galal et al. (2013) [49]                | APACHE II      | 0.6             | $\chi^2 = 7.34$, p = 0.39 | 93% (DC = 11) | 24% (DC = 11) | 55.2% |
| Eapen et al. (2016) [47]                | APACHE II      | 0.6             | $\chi^2 = 7.34$, p = 0.39 | 93% (DC = 11) | 24% (DC = 11) | 55.2% |
| Khatouni and Charoenpan (1999) [45]     | APACHE II      | 0.6             | $\chi^2 = 7.34$, p = 0.39 | 93% (DC = 11) | 24% (DC = 11) | 55.2% |
| Juneja et al. (2010) [1]                | APACHE III     | 0.922 (0.894–0.949) | $\chi^2 = 3.674$ (p = 0.088) | 78.6% (DC > 73) | 86% (DC > 73) | NR       |
| Scaes et al. (2004) [40]                | APACHE III     | 0.915 (0.898–0.933) | H statistic $\chi^2 = 117.206$ (p < 0.001) | NR          | NR          | NR       |
| Shekhar et al. (2011) [73]              | APACHE III     | 0.895 (0.839–0.952) | $\chi^2 = 16.904$ (p = 0.031) | 91%         | 73.9%       | 80.34% |
| Gilani et al. (2014) [70]               | APACHE III     | 0.78            | $\chi^2 = 8.442$ (p = 0.392) | 82.3% (DC = 24) | 58% (DC = 24) | NR       |
| Wilairatana et al. (1995) [82]          | APACHE III     | 0.694           | NR          | 79.8% (DC = 60) | 66% (DC = 60) | 72.4% |
| Juneja et al. (2012) [1]                | APACHE IV      | 0.928 (0.903–0.953) | $\chi^2 = 8.790$ (p = 0.360) | 93.2% (cut-off point > 12.5) | 66.9% (cut-off point > 12.5) | NR       |
| Nassar et al. (2012) [42]               | APACHE IV      | 0.883 (0.874–0.891) | $\chi^2 = 5.37$ (p < 0.01) | NR          | NR          | NR       |
| Xing et al. (2015) [52]                 | APACHE IV      | 0.873 (0.813–0.934) | $\chi^2 = 3.756$ (p = 0.087) | NR          | NR          | NR       |
| Harnau et al. (2009) [76]               | APACHE IV      | 0.845 (0.786–0.904) | $\chi^2 = 5.123$ (p = 0.744) | NR          | NR          | NR       |
| Evan et al. (2016) [47]                 | APACHE IV      | Presented as a figure | 86% (DC > 73) | NR          | NR          | 83% (DC = 50%) |

APACHE Acute Physiology and Chronic Health Evaluation, CI confidence interval, NR not reported, DC decision criteria
| Study | Scoring system | Discrimination | Calibration | Sensitivity | Specificity | Accuracy |
|-------|----------------|----------------|-------------|-------------|-------------|----------|
| Haniffa et al. (2018) [60] | SAPS I | 0.746 | NR | 70.2% (DC = 15) | 67.1% (DC = 15) | NR |
| Nair et al. (2016) [62] | SAPS I | 0.742 | NR | 44.4% (DC = 61) | 94.9% (DC = 61) | 85.42% |
| Khwannimit and Buhrayanontachai (2011) [51] | SAPS I (re-calibrated model) | 0.919 (0.899–0.94) | C statistic $\chi^2 = 8.6$ ($p < 0.57$) | NR | NR | NR |
| Khwannimit and Buhrayanontachai (2011) [51] | SAPS II | 0.919 (0.899–0.94) | C statistic $\chi^2 = 12.49$ ($p < 0.001$) | NR | NR | NR |
| Soares et al. (2004) [40] | SAPS II | 0.887 (0.874–0.902) | C statistic $\chi^2 = 70.14$ ($p = 0.522$) | 83% (DC = 86.5) | 7.7% (DC = 86.5) | NR |
| Soares and Salluh (2006) [39] | SAPS II | 0.88 (0.86–0.9) | C statistic $\chi^2 = 26.6$ ($p < 0.001$) | Individual values for each risk level | Individual values for each risk level | Individual values for each risk level |
| Khan et al. (2015) [32] | SAPS II (48 hours) | 0.826 (0.803–0.85) | $\chi^2 = 12.176$ ($p = 0.144$) | 85% | 74.3% | 82.4% |
| Zhao et al. (2013) [50] | SAPS II | 0.821 (0.795–0.848) | $\chi^2 = 11.294$ ($p = 0.186$) | 85% | 74.3% | 83.8% |
| Aggarwal et al. (2006) [7] | SAPS II | 0.781 | C statistic $\chi^2 = 19.56$ ($p < 0.001$) | 46.1% (38.1–54.3) (DC = 25%), 27.35% (20.4–35.0) (DC = 50%), 10.4% (6.1–16.3) (DC = 75%) | 89.3% (85.2–92.5) (DC = 25%), 92.6% (92.7–97.7) (DC = 50%), 98.7% (96.6–99.9) (DC = 75%) | 74.6% (70.3–78.5) (DC = 25%), 72.3% (68.0–76.4) (DC = 50%), 68.6% (64.1–72.8) (DC = 75%) |
| Gani et al. (2014) [70] | SAPS II | 0.78 | $\chi^2 = 8.575$ ($p = 0.379$) | 70.5% (DC = 13) | 63% (DC = 13) | NR |
| Zhao et al. (2013) [50] | SAPS II (initial) | 0.776 (0.75–0.802) | $\chi^2 = 8.332$ ($p = 0.402$) | 85% | 74.3% | 80% |
| Naqvi et al. (2016) [33] | SAPS II | 0.75 | $\chi^2 = 3.724$ ($p = 0.811$) | NR | NR | NR |
| Faruq et al. (2013) [37] | SAPS II | 0.74 (0.66–0.81) | C statistic $\chi^2 = 9.040$ ($p = 0.34$) | NR | NR | NR |
| Khan et al. (2015) [32] | SAPS II | 0.718 (0.696–0.828) | NR | 70.6% (DC > 30) | 60.8% (DC > 30) | NR |
| Wilairatana et al. (1995) [82] | SAPS II | 0.71 | NR | 0.742 (cut-off point = 14) | 0.6 (cut-off point = 14) | 68.40% |
| Sekulic et al. (2013) [48] | SAPS II | 0.69 | $\chi^2 = 4.41$ ($p = 0.732$) | Presented as a figure | Presented as a figure | NR |
| Galal et al. (2013) [49] | SAPS II | 0.59 | $\chi^2 = 7.2$, $p = 0.3$ | 53.4% (DC = 40) | 6.2% (DC = 40) | 57.1% |
| Mohan et al. (2013) [46] | SAPS II | NR | NR | 81.1% (DC > 35) | 59.5% (DC > 35) | NR |
| Xing et al. (2015) [52] | SAPS 3 | 0.948 (0.914–0.982) | $\chi^2 = 4.987$ ($p = 0.759$) | NR | NR | NR |
| Study | Scoring system | Discrimination | Calibration | Sensitivity | Specificity | Accuracy |
|-------|----------------|----------------|-------------|-------------|-------------|----------|
| Khawannimit and Bhurayanontachai (2011) | SAPS 3 (AUS) (recalibrated model) | 0.917 (0.902 – 0.929) | C statistic χ² = 8.2 (p = 0.61) | NR | NR | NR |
| | | | H statistic χ² = 7.9 (p < 0.001) | NR | NR | NR |
| Khawannimit and Bhurayanontachai (2011) | SAPS 3 | 0.917 (0.902 – 0.929) | C statistic χ² = 170 (p < 0.001) | NR | NR | NR |
| | | | H statistic χ² = 79.9 (p < 0.001) | NR | NR | NR |
| Juneja et al. (2012) | SAPS 3 | 0.901 (0.871 – 0.932) | χ² = 13.123 (p = 0.108) | 76.7% (DC > 56.5) | 84.7% (DC > 56.5) | 84.7% (DC > 56.5) |
| Soares and Salluh (2006) | SAPS 3 (CSA) | 0.87 (0.85 – 0.9) | C statistic χ² = 9.132 (p = 0.33) | NR | NR | NR |
| Soares and Salluh (2006) | SAPS 3 | 0.87 (0.85 – 0.9) | C statistic χ² = 13.637 (p = 0.092) | NR | NR | NR |
| Silva Junior et al. (2010) | SAPS 3 | 0.86 (0.83 – 0.88) | C statistic χ² = 10.47 (p = 0.234) | 0.75 (DC = 57) | 0.86 (DC = 57) | 0.86 (DC = 57) |
| Nassar et al. (2012) | SAPS 3 | 0.855 (0.846 – 0.864) | C statistic χ² = 25.66 (p < 0.001) | NR | NR | NR |
| Mohan et al. (2015) | SAPS 3 | 0.84 (0.81 – 0.87) | C statistic χ² = 15.804 (p = 0.045) | NR | NR | NR |
| Hernandez et al. (2014) | SAPS 3 | 0.8 (0.78 – 0.81) | C statistic χ² = 11.5 (p = 0.18) | NR | NR | NR |
| Evran et al. (2016) | SAPS 3 | Presented as a figure | NR | NR | NR | NR |

SAPS: Simplified Acute Physiology Score, SAPS (AUS): Simplified Acute Physiology Score (Australian), SAPS (CSA): Simplified Acute Physiology Score (Central and Southern American), CI: confidence interval, DC: decision criterion, NR: not reported.
| Study                  | Scoring system | Discrimination | Calibration         | Sensitivity       | Specificity       | Accuracy          |
|------------------------|----------------|----------------|---------------------|-------------------|-------------------|-------------------|
| Nouira et al. (1998)   | MPM (24 hours) | 0.88           | C statistic $\chi^2 = 29.59$ ($p < 0.001$) | NR                | NR                | NR                |
|                        |                |                | H statistic $\chi^2 = 19.9$ ($p < 0.05$)          |                   |                   |                   |
|                        | MPM (initial)  | 0.85           | C statistic $\chi^2 = 36.66$ ($p < 0.001$) | NR                | NR                | NR                |
|                        |                |                | H statistic $\chi^2 = 38$ ($p < 0.05$)           |                   |                   |                   |
| Sekulic et al. (2015)  | MPM II (7 days)| 1.00           | $\chi^2 = 0.00$ ($p = 1.000$)                    | 100%              | Presented as figure | NR                |
|                        |                |                | $\chi^2 = 86.27$ ($p = 0.375$)                   |                   |                   |                   |
| Juneja et al. (2012)   | MPM II (initial) | 0.928 (0.904–0.952) | $\chi^2 = 8.627$ ($p = 0.375$)                   | 95.1% (DC > 27)   | 68.2% (DC > 27)   | NR                |
|                        |                | 0.836          | $\chi^2 = 11.37$ ($p = 0.181$)                   | Presented as figure | Presented as figure | NR                |
|                        |                | 0.817          | $\chi^2 = 6.04$ ($p = 0.534$)                    | Presented as figure | Presented as figure | NR                |
| Hamza et al. (2009)    | MPM II (initial) | 0.81 (0.738–0.882) | $\chi^2 = 8.625$ ($p = 0.357$)                   | NR                | NR                | NR                |
| Aggarwal et al. (2006) | MPM II (24 hours) | 0.733         | C statistic $\chi^2 = 73.1$ ($p < 0.001$)    | 48.1% (59.9–56.2) | 24.0% (17.5–31.6) | 14.3% (9.2–20.8) |
|                        |                |                | H statistic $\chi^2 = 69.7$ ($p = 0.001$)       | (DC = 25%)        | (DC = 50%)        | (DC = 75%)        |
|                        | MPM II (initial) | 0.665         | C statistic $\chi^2 = 59.92$ ($p < 0.001$)  | 27.9% (0.210–0.355) | 114% (0.069–0.174) | 25% (0.007–0.064) |
|                        |                |                | H statistic $\chi^2 = 45.64$ ($p < 0.001$)     | (DC = 25%)        | (DC = 50%)        | (DC = 75%)        |
|                        | MPM II (initial) | 0.654         | $\chi^2 = 6.90$ ($p = 0.548$)                   | 85.2%             | Presented as figure | NR                |
|                        | MPM II (24 hours) | 0.836         | $\chi^2 = 16.23$ ($p = 0.039$)                   | Presented as figure | Presented as figure | NR                |
| Juneja et al. (2012)   | MPM II (initial) | 0.947 (0.927–0.967) | $\chi^2 = 13.835$ ($p = 0.086$)               | 94.2% (DC > 19)   | 80.2% (DC > 19)   | NR                |
|                        | MPM II (initial) | 0.84 (0.838–0.849) | C statistic $\chi^2 = 134.2$ ($p < 0.01$) | NR                | NR                | NR                |
| Nassar et al. (2012)   | MPM II (initial) | 0.72           | $\chi^2 = 17.66$ (0.024)                        | NR                | NR                | NR                |
| Riviello et al. (2016) | MPM II (initial) | 0.71 (0.67–0.75) | C statistic $\chi^2 = 14.242$ ($p = 0.076$)    | NR                | NR                | NR                |

MPM Mortality Probability Models, CI confidence interval, NR not reported, DC decision criteria.
| Study | Model | Source | Participants | Outcomes | Predictors | Sample size | Missing data | Model development | Model performance |
|-------|-------|--------|--------------|----------|------------|-------------|--------------|-------------------|------------------|
| Ahluwalia et al. (1999) [29] | New score | Prospective cohort | Consecutive admissions (>13 years) to eight-bed medical ICU, India; inclusion period NR; participant age range 13-80, mean = 46 | Hospital mortality | 1. pH (at admission); 2. serum albumin (at admission); 3. heart rate (at 48 hours); 4. GCS (at 48 hours); 5. bilirubin (at 48 hours) | 79 | Not reported | Based on APACHE II (Knaus et al. 1985 [10]) and 11 other clinical and laboratory parameters. Backward step method used to remove non-significant (p > 0.05) variables (of univariate analysis) | Discrimination measured in terms of AUROC, sensitivity and specificity. Multivariate and univariate regression |
| Rivello et al. (2016) [43] | Rwanda MPM (R-MPM) | Prospective cohort | Consecutive patients (>15 years) admitted to two ICUs in different hospitals; exclusion criteria: not specified; August 2013–October 2014; participant age range 34 years (IQR 25–47) (median) | Hospital mortality | Only the following five variables were included: 1. age; 2.confirmed or suspected infection within 24 hours of ICU admission; 3. hypotension or shock as a reason for ICU admission; 4. heart rate at ICU admission; 5. GSC at time of admission | 427 | | Based on the 16 MPM III (initia) and additional variables. Variables for inclusion in model selected from the univariate analyses, based on their predictive power (as determined by p < 0.05) as well as their ease of capture based on experience, the proportion of missing values in the dataset, and their clinical significance. | Discrimination measured in terms of AUROC. Calibration measured as Hosmer–Lemeshow. Multivariate and univariate regression |
| Rivello et al. (2016) [43] | Simplified R-MPM | Prospective cohort | All consecutive patients (>12 years) admitted to 17-bed medical–neurological ICU, tertiary referral hospital, India; January 1996–May 1998 | Hospital mortality | Altered mental status on ICU admission (present vs not present) used in place of the GCS score in the R-MPM (see previous model) | 2962 | Normal values attributed as in original study; two patients excluded due to lack of discharge status | Based on the 16 MPM III (initial) and additional variables. Variables for inclusion in model selected from the univariate analyses, based on their predictive power (as determined by p < 0.05) as well as their ease of capture based on experience, the proportion of missing values in the dataset, and their clinical significance. | Discrimination measured as AUROC. Calibration measured as Hosmer–Lemeshow. Multivariate and univariate regression |
| Nimgaonkar et al. (2004) [35] | Artificial Neural Network (ANN 22) | Prospective cohort | All consecutive patients (>12 years) admitted to 17-bed medical–neurological ICU, tertiary referral hospital, India; August 2000–June 2000 | Hospital mortality | 22 APACHE II variables | 2962 | Not reported | Artificial Neural Network trained on an Indian patient dataset using the 15 APACHE II variables with the highest information gain (measured by calculation of entropy) | Discrimination measured as AUROC. Calibration measured as Hosmer–Lemeshow. Multivariate and univariate regression |
| Nimgaonkar et al. (2004) [35] | Artificial Neural Network (ANN 15) | Prospective cohort | All consecutive patients (>12 years) admitted to 17-bed medical–neurological ICU, tertiary referral hospital, India; August 2000–June 2000 | Hospital mortality | 15 APACHE II variables with the highest information gain (measured by calculation of entropy) | 2962 | Not reported | Artificial Neural Network trained on an Indian patient dataset using the 15 APACHE II variables with the highest information gain (measured by calculation of entropy) | Discrimination measured as AUROC. Calibration measured as Hosmer–Lemeshow. Multivariate and univariate regression |
| Study                      | Ahluwalia et al. (1999) [29] | Rivello et al. (2016) [43] | Rivello et al. (2016) [43] | Nimgaonkar et al. (2004) [35] | Nimgaonkar et al. (2004) [35] |
|---------------------------|-----------------------------|---------------------------|---------------------------|-----------------------------|-----------------------------|
| Model evaluation          | Developmental dataset only, no further evaluation (compared with APACHE II at 48 hours) | Internal validation with bootstrapping (compared with MPM III (initial)) | Internal validation with bootstrapping (compared with MPM III (initial)) | Data from 1962 patients were used to train the neural network using a back-propagation algorithm. Data from the remaining 1000 patients were used for testing this model and comparing it with APACHE II | Data from 1962 patients were used to train the neural network using a back-propagation algorithm. Data from the remaining 1000 patients were used for testing this model and comparing it with APACHE II |
| Results                   | New score ROC: 0.90, sensitivity: 98.2%, specificity: 66.6%, APACHE II (after 48 hours) ROC: 0.74, sensitivity: 92.8%, specificity: 23.6% | Rwanda MPM (R-MPM) AUROC: 0.81 (0.77–0.86), MPM III (initial) AUROC: 0.72, HL: $\chi^2 = 17.66$ (p = 0.024) | Simplified R-MPM AUROC: 0.76, MPM III (initial) AUROC: 0.72, HL: $\chi^2 = 17.66$ (p = 0.024) | ANN. 22 AUROC: 0.87, HL $\chi^2 = 22.4$ (p < 0.05), MPM III (initial) AUROC: 0.77, HL $\chi^2 = 123.5$ (p < 0.05) | ANN. 15 AUROC: 0.88, HL $\chi^2 = 27.7$ (p < 0.05), APACHE II AUROC: 0.77, HL $\chi^2 = 123.5$ (p < 0.05) |

APACHE Acute Physiology and Chronic Health Evaluation, MPM Mortality Probability Models, ICU intensive care unit, GCS Glasgow Coma Score, IQR interquartile range, HL Hosmer–Lemeshow statistic, AUROC area under the receiver operating characteristic.
Discussion

This systematic review of critical care prognostic models in LMICs reports good to excellent discrimination in 88.9% of evaluations between survivors and non-survivors of ICU admission and good calibration in 58.3% of those reporting calibration. In keeping with findings in HICs [3, 53], this review found good discrimination to be more frequently reported than good calibration; although good discrimination and good calibration were rarely (11.9%) reported together in the same evaluation [1, 48, 51, 52]. Three of the 10 evaluations reporting both excellent discrimination and good calibration were from recalibrated models [51], and in two [48] the sample size was small (n = 60). It is known that a calibration measure such as the Hosmer–Lemeshow goodness-of-fit test might demonstrate high p values in these circumstances, simply as a consequence of the test having lower power and not necessarily as an indication of a good fit [53].

Differences in predictors in the different models (e.g. acute diagnosis is a variable in APACHE II but not SAPS II) and the differences in the datasets used in the various studies may have contributed to the discrepancies seen in performances of the models. Three major findings, with special relevance to the LMIC settings, limit generalisability and can affect performance: post-ICU outcomes were not available for 40.5% where ICU mortality was the outcome; only 44.8% reported a lower age limit, with 55.8% of these including patients who were aged younger than 18 years; and missing values and their handling. The original models being evaluated were developed to assess hospital mortality. Therefore, the lack of post-ICU outcome may impact on their performance, particularly as discharge from the ICU (especially in these settings) may be influenced by non-clinical discharge decisions such as shortage of ICU beds. However, post-ICU follow-up may not always be feasible in these settings due to the lack of established follow-up systems (e.g. medical registries, electronic records). Patient age may affect model performance and could be another cause for the heterogeneity seen between studies. The lower age limit for admission to adult ICUs varies between settings, perhaps resulting in the admission of paediatric patients into adult ICUs (and their subsequent use in the datasets for the validation of adult prognostic models). Twenty-three studies did not report a lower age limit for patient admission and 17 studies included patients younger than the age of 18 years; the variation in both age criteria for inclusion and for reporting make unfeasible a complete exclusion of paediatric patients from this review of adult prognostic models. Missing value handling, which can lead to bias and thus influence model performance especially in LMIC settings [53], was only reported infrequently. Where reported, imputation by normal values (which is less justifiable in LMIC settings [9]) and exclusion of incomplete records (leading to inefficient use of the dataset) were the methods frequently utilised. Research into the utility of other techniques of imputation (e.g. multiple imputation) for missing values may reduce bias and increase the interpretability of model performance. However, missing values in prognostic models in LMIC settings are likely to be a persistent problem. Some of these difficulties may be alleviated by increasing efforts to improve the availability and recording of measures such as GCS and saturations or by effecting substitutions for the measurements that are more inaccessible in LMIC settings (e.g. urea for creatinine and saturations for PaO₂). Although two studies in this review reported the exclusion of variables [30, 50], the effect of the modifications could not be ascertained: in one case, no comparison was made with the original APACHE II model [30]; and in the second, discrimination was not reported for the simplified version of SAPS II [50]; calibration was not reported for either of these models.

Validation studies of prognostic models in LMIC settings are becoming more common; 16 of the 50 studies included were published in 2015, 2016 or 2017 and additional studies, for example Moralez et al. in Brazil [54] and Haniffa et al. [9] in Sri Lanka, have been published/awaiting publication subsequent to the literature search for this review. Consequently it is important for investigators to adhere to reporting standards, such as CHARMS—especially with regard to performance measures, outcomes and missing values—to enable better interpretation.

For a critical care prognostic model to be effective it needs to be calibrated to the target setting and have an acceptable data collection burden. However, in this review, first-level customisation was carried out in only one study [51]; the calibration of APACHE II, SAPS II and SAPS 3 models improved from poor to good and the discrimination remained excellent before and after recalibration. In HIC, medical registries enable standardised, centralised, often automated, electronic data gathering, which can then be validated; thus reducing the burden of data collection. These registries include mechanisms for providing feedback on critical care unit performance and also enable regular recalibration of prognostic models, thus minimising the incorrect estimation of predicted mortalities due to changes in case mix and treatment. The absence of such registries in LMIC settings, with important exceptions (e.g. in Brazil, Malaysia and Sri Lanka), is a significant barrier for the validation and recalibration of existing models, and the development of models tailored to these settings. Accordingly, none of the validation studies included in this review is an output from a medical registry, no studies reported on model performance from different time
points in the same setting and only three studies were conducted in two or more hospitals [41, 43, 55].

The use of prognostic models in practice is thought to be influenced by the complexity of the model, the format of the model, the ease of use and the perceived relevance of the model to the user [56]. The development of models with fewer and more commonly available measures perhaps in conjunction with medical registries promoting research may also be effective in improving mortality prediction in these settings; for example, the simplified Rwanda MPM [43] and TropICS [57]. Introducing simple prognostic models like those already mentioned and emphasising their usefulness by providing output that is relevant to clinicians, administrators and patients is therefore more likely to result in the collection of required data and their application in a clinical context.

ICU risk prediction models need to exhibit good calibration before they can be used for quality improvement initiatives [58, 59]. Setting-relevant models such as TropICS [57], which are well calibrated, can be used for stratification of critically ill patients according to severity, which is a pre-requisite for impact assessment of training and other quality improvement initiatives. However, models that show poor calibration but have a good discriminatory ability may still be of benefit if their intended use is for identifying high-risk patients for diagnostic testing or therapy and/or for inclusion criteria or covariate adjustment in a randomised controlled trial [58, 59].

Limitations
This review was limited to a single database (PubMed). There is no MeSH for LMIC (non-HIC) and hence a hand search strategy was deployed. No attempt was made to distinguish between upper and lower middle-income countries which are very heterogeneous in terms of provision, resources and access to healthcare. The review was intended to be for adult prognostic models used only in adult patients; however, due to the manner in which the studies were reported it was not possible to exclude paediatric patients.

Conclusion
Performance of mortality risk prediction models for ICU patients in LMICs is at most moderate, especially with limitations in calibration. This necessitates continued efforts to develop and validate LMIC models with readily available prognostic variables, perhaps aided by medical registries. Robust interpretations of their applicability are currently hampered by poor adherence to reporting guidelines, especially when reporting missing value handling.

Additional files

Additional file 1: A table presenting the search terms used. (XLSX 27 kb)
Additional file 2: A table presenting the checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies. (XLSX 41 kb)

Acknowledgements
Not applicable.

Funding
None.

Availability of data and materials
All data generated or analysed during this study are included in this published article (and its Additional files 1 and 2).

Authors’ contributions
RH and AMD conceived the study and developed the study design. RH and II performed the primary study search, extracted data, carried out statistical analysis, drafted the manuscript, and revised the manuscript. NFDK, AMD and APDS improved the idea and design, and revised the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that there are no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. 2Network for Improving Critical Care Systems and Training, Colombo, Sri Lanka. 3AA (Ltd), London, UK. 4National Intensive Care Surveillance, Ministry of Health, Amsterdam, Netherlands. 5Mahidol–Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand. 6Department of Medical Informatics, Academic Medical Center, Amsterdam, Netherlands.

Received: 4 August 2017 Accepted: 21 December 2017
Published online: 26 January 2018

References
1. Juneja D, Singh O, Nasa P, et al. Comparison of newer scoring systems with the conventional scoring systems in general intensive care population. Minerva Anestesiologica. 2012;78(2):194–200. https://www.minervamedica.it/en/journals/minerva-aneStesiologica/article.php?cod=R02Y2012N02A0194. Accessed 4 Oct 2016.
2. Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: a compendium. Indian J Crit Care Med. 2014;18(4):220–8. https://doi.org/10.4103/0972-5229.130573.
3. Vincent JJ-L, Moreno RR, Moreno RR, et al. Clinical review: scoring systems in the critically ill. Crit Care. 2010;14(2):207. https://doi.org/10.1186/cc8204.
4. Knaus WA, Zimmerman JE, Wagner DP, et al. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med. 1981;9(8):591–7. http://journals.lww.com/ccmjournal/Abstract/1981/08000/APACHE_acute_physiology_and_chronic_health_6.aspx. Accessed 4 Oct 2016.
5. Grissom CK, Brown SM, Kuttler KG, et al. A modified sequential organ failure assessment score for critical care triage. Disaster Med Public Health Prep. 2010(4)(4):277–84. https://doi.org/10.1017/dmp.2010.40.
46. Mohan A, Shrestha P, Guleria R, et al. Development of a mortality prediction formula due to sepsis/severe sepsis in a medical intensive care unit. Lung India. 1974;3(4):313–19. https://doi.org/10.4103/0970-2113.19553.

47. Evin T, Serin S, Gürses E, et al. Various scoring systems for predicting mortality in Intensive Care Unit. Niger J Clin Pract. 2016;19(4):530–4. https://doi.org/10.1016/j.njcp.2015.11.007.183307.

48. Sekulic AD, Trpkovic SV, Pavlovic AP, et al. Scoring systems in assessing survival of critically ill ICU patients. Med Sci Monit. 2015;21:2621–9. https://doi.org/10.21659/MSM.S94153.

49. Galal I, Kassem E, Mansour M. Study of the role of different severity scores in respiratory ICU. Egypt J Bronchol. 2013;7(2):53. https://doi.org/10.1016/j.ejbr.2013.03.005.

50. Zhao X-X, Su Y-Y, Wang M, et al. Evaluation of neuro-intensive care unit performance in China: predicting outcomes of Simplified Acute Physiology Score II or Glasgow Coma Scale. Chin Med J (Engl). 2013;126(6):1132–7. https://doi.org/10.4269/j.issn.1671-0296.2012.12.010.

51. Khwanimit B, Bhurayanontachai R. A comparison of the performance of Simplified Acute Physiology Score 3 with old standard severity scores and customized scores in a mixed medical-coronary care unit. Minerva Anestesiol. 2011;77(3):305–12. http://www.minervamedica.it/en/journals/minerva-ane-stesiologica/article.php?article=R0212011N3A0305. Accessed 7 Nov 2015.

52. Xing X, Gao Y, Wang H, et al. Performance of three prognostic models in patients with cancer in need of intensive care in a medical center in China. PLoS One. 2015;10(6):e0131329. https://doi.org/10.1371/journal.pone.0131329. Accessed 7 Nov 2015.

53. Silva Junior JM, Malbouisson LMS, Nuevo HL, et al. Aplicabilidade do escore fisiológico agudo simplesificado (SAPS 3) em hospitais brasileiros. Rev Bras Estud Saúde. 2010;60(1):20–31. https://doi.org/10.1590/S0034-70942010000100003.

54. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS): 1. a framework for researching clinical outcomes. BMJ. 2013;346:e5595. https://doi.org/10.1136/bmj.e5595.

55. Hanifah R, Mukaka M, Munasinghe SB, et al. Simplified prognostic model for critically ill patients in resource limited settings in South Asia. Crit Care. 2017;21:250. doi.org/10.1186/s13054-017-1943-6.

56. Nassar AP, Malbouisson LMS, Moreno R. Evaluation of Simplified Acute Physiology Score 3 performance: a systematic review of external validation studies. Crit Care. 2014;18(3):R1.17. https://doi.org/10.1186/cc13911.

57. Steyerberg EW, Moons KGM, Vickers AJ, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381.

58. Abhinandan KS, Vedavathi R. Usefulness of Sequential Organ Failure Assessment (SOFA) and Acute Physiological and Chronic Health Evaluation II (APACHE II) score in predicting mortality in surgical patients with multiple organ dysfunction syndrome in sepsis. J Evol Med Dent Sci. 2013;2(49):5931-605. https://jems.org.com/data/pdf/180abhinandan620.pdf. Accessed 1 Apr 2017.

59. Liu X, Shen Y, Li Z, et al. Prognostic significance of APACHE II score and plasma sTnInR in Chinese patients with sepsis: a prospective observational study. BMC Anesthesiol. 2016;16:46. https://doi.org/10.1186/s12871-016-0212-3.

60. Nair R, Bhandary NM, D’Souza AD. Initial Sequential Organ Failure Assessment score versus Simplified Acute Physiology score to analyze multiple organ dysfunction in infectious diseases in intensive care unit. Indian J Crit Care Med. 2016;20(4):210–5. https://doi.org/10.4103/0972-2229.180041.

61. Celik S, Sahin D, Korkmaz C, et al. Potential risk factors for patient mortality during admission to the intensive care units. Saudi Med J. 2014;35(2):159–64. https://www.smj.org.sa/index.php/smj/article/view/15599.85f5b8a5dc6b291dbae9660086b6473a.pdf. Accessed 6 Mar 2016.

62. Chang L, Hong C-F, Huang Y-CT, et al. Prognostic accuracy of Acute Physiology and Chronic Health Evaluation II scores in critically ill cancer patients. Am J Crit Care. 2006;15(1):57–63. https://ajcc.aacnjournals.org/article_text.php?article_id=2828. Accessed 7 Nov 2015.

63. Haniffa et al. Critical Care (2018) 22:18 Page 21 of 22
Malaysia. 1991;46(1):72–81. http://www.e-njm.org/1991/V46n1/APACHE_II.pdf. Accessed 6 Mar 2016.

82. Wilairatana P, Noan NS, Chinprasatsak S, et al. Scoring systems for predicting outcomes of critically ill patients in northeastern Thailand. Southeast Asian J Trop Med Public Health. 1995;26(1):66–72. http://www.tm.mahidol.ac.th/seameo/1995-26-1/1995-26-1-66.pdf. Accessed 7 Nov 2015.