BMJ Open Respiratory Research

Chest radiograph-based artificial intelligence predictive model for mortality in community-acquired pneumonia

Jessica Quah ,[1] Charlene Jin Yee Liew,[2] Lin Zou,[3] Xuan Han Koh,[4] Rayan Alsuwaigh,[1] Venkataraman Narayan,[5] Tian Yi Lu,[3] Clarence Ngoh,[3] Zhiyu Wang,[3] Juan Zhen Koh,[3] Christine Ang,[3] Zhiyan Fu,[3] Han Leong Goh[3]

To cite: Quah J, Liew CJY, Zou L, *et al*. Chest radiograph-based artificial intelligence predictive model for mortality in community-acquired pneumonia. *BMJ Open Resp Res* 2021;**8**:e001045. doi:10.1136/bmjresp-2021-001045

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjresp-2021-001045).

JQ and CJYL contributed equally.

Received 5 July 2021
Accepted 21 July 2021

For numbered affiliations see end of article.

Correspondence to
Dr Jessica Quah;
jessica.quah.l.s@singhealth.com.sg

ABSTRACT

Background Chest radiograph (CXR) is a basic diagnostic test in community-acquired pneumonia (CAP) with prognostic value. We developed a CXR-based artificial intelligence (AI) model (CAP AI predictive Engine: CAPE) and prospectively evaluated its discrimination for 30-day mortality.

Methods Deep-learning model using convolutional neural network (CNN) was trained with a retrospective cohort of 2235 CXRs from 1966 unique adult patients admitted for CAP from 1 January 2019 to 31 December 2019. A single-centre prospective cohort between 11 May 2020 and 15 June 2020 was analysed for model performance. CAPE mortality risk score based on CNN analysis of the first CXR performed for CAP was used to determine the area under the receiver operating characteristic curve (AUC) for 30-day mortality.

Results 315 inpatient episodes for CAP occurred, with 30-day mortality of 19.4% (n=61/315). Non-survivors were older than survivors (mean (SD)age, 80.4 (10.3) vs 69.2 (18.7)); more likely to have dementia (n=27/61 vs n=58/254) and malignancies (n=16/61 vs n=18/254); demonstrate higher serum C reactive protein (mean (SD), 109 mg/L (98.6) vs 59.3 mg/L (69.7)) and serum procalcitonin (mean (SD), 11.3 (27.8) µg/L vs 1.4 (5.9) µg/L). The AUC for CAPE mortality risk score for 30-day mortality was 0.79 (95% CI 0.73 to 0.85, p<0.001); Pneumonia Severity Index (PSI) 0.80 (95% CI 0.74 to 0.86, p<0.001); Confusion of new onset, blood Urea nitrogen, Respiratory rate, Blood pressure, 65 (CURB-65) score 0.76 (95% CI 0.70 to 0.81, p<0.001), respectively. CAPE combined with CURB-65 model has an AUC of 0.83 (95% CI 0.77 to 0.88, p<0.001). The best performing model was CAPE incorporated with PSI, with an AUC of 0.84 (95% CI 0.79 to 0.89, p<0.001).

Conclusion CXR-based CAPE mortality risk score was comparable to traditional pneumonia severity scores and improved its discrimination when combined.

Key messages

- Can artificial intelligence prognosticate pneumonia mortality based on chest radiographs?
- An artificial intelligence tool was developed based on 2235 chest radiographs from a retrospective cohort of patients with community-acquired pneumonia. Subsequently, its discriminative power for 30-day mortality was determined based on a prospective cohort of 315 inpatient visits and found to be comparable to Pneumonia Severity Index and Confusion of new onset, blood Urea nitrogen, Respiratory rate, Blood pressure-65.
- This is a novel tool using a single data source (chest radiograph) to prognosticate pneumonia mortality. Incorporating this into traditional Pneumonia Severity Scores improved the model discrimination for 30-day mortality.

INTRODUCTION

Community-acquired pneumonia (CAP) is the fourth-leading cause of death globally, with estimates at 2.96 million each year.[1] CAP may result in long-term functional impairment, serious morbidity and mortality, particularly for those who require hospitalisation.[2 3] To aid clinicians, Pneumonia Severity Scores were developed for mortality risk stratification, triaging of appropriate sites-of-care and disease management strategies.[4–6]

Numerous studies have been performed to identify risk factors for adverse outcomes in CAP patients. These serve to build clinical prediction models for stratifying pneumonia severity.[5 6] Two of the most widely used tools are the Pneumonia Severity Index (PSI)[6] and Confusion of new onset, blood Urea nitrogen, Respiratory rate, Blood pressure, 65 (CURB-65) score.[5] PSI consists of 20 variables to derive a weighted score, which is further stratified into five classes of mortality risk.[7] CURB-65 score is calculated using five variables of equal weighting: confusion of new onset; serum urea >7 mmol/L; respiratory rate ≥30 per min; low blood pressure

(systolic blood pressure <90 mm Hg or diastolic blood pressure ≤60 mm Hg); age ≥65 years[5 7 8].

These prognostic scores demonstrate good discrimination for mortality. In a meta-analysis, Chalmers *et al* reported an area under the receiver operating characteristic curve (AUC) of 0.81 and 0.80 for the PSI and CURB-65 scores, respectively.[9] However, the practical use of severity scores has significant challenges. Calculation requires multiple data point acquisition processes (medical history, physical examination, blood sampling, chest imaging), is time-consuming, prone to poor clinician compliance.[10] A recent study in Singapore showed that PSI performed better for mortality prediction than CURB-65, however, its discriminative power decreased with advancing age.[11]

Artificial intelligence (AI) when applied on electronic medical data to support clinical decision-making processes, demonstrate the potential to mitigate some of these challenges.[12 13] AI research in thoracic imaging has focused largely on diagnostic discrimination. Several studies have described convolutional neural network (CNN) models which demonstrate high accuracy rates in predicting chest imaging diagnoses.[14–17] There has also been rapid progress in the use of AI for diagnosis of SARS-CoV-2 pneumonia based on chest computer tomography and radiographs.[18–20]

In addition to discriminating binary diagnoses, AI demonstrates the potential to prognosticate outcomes using chest imaging. Lu *et al* described the use of a single chest radiograph (CXR) in a cancer screening cohort to predict all-cause mortality at 12 years.[21] Similarly, Liu *et al* described the use of AI algorithms to analyse CT changes in SARS-CoV-2 pneumonia to predict disease progression with an AUC of 0.93.[22]

The authors hypothesise that radiological abnormalities present on a single CXR taken at the start of an episode of CAP can aid in prognostication of mortality. Using a retrospective dataset, we developed a CNN named CAPE (CAP AI Predictive Engine). The primary aim of this study is to determine the AUC of the CAPE mortality risk score for 30-day mortality. The secondary aim is to compare the performance of this tool to well-validated pneumonia severity scores—CURB-65 and PSI. The tertiary aim is to investigate the potential additional value of combining CNN with pneumonia severity scores.

METHODS

CNN model development

Model development was based on a single acute tertiary hospital's data. Study consent waiver was obtained. Patients and public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

In the model development of CAPE CNN, a retrospective data set of consisting 2235 CXR from 1966 adult were used. They were identified from electronic medical records by emergency department attendance (with subsequent inpatient admission diagnosis of pneumonia by International Classification of Disease-10 coding) occurring between 1 January 2019 and 30 December 2019. All CXR were deidentified and preprocessed by centre-cropping, resizing to dimensions of 244×244 pixels, followed by histogram equalisation. Inpatient mortality data were used for model development instead of 30-day mortality as this retrospective data were not available from the national death registry.

The retrospective cohort was grouped into three sets: 'training' set for model building; 'validation' set for selection of the optimal model; 'test' set to assess the performance of the selected model. Data from 1 January 2019 to 31 October 2019, were split into 'training' set and 'validation' set with proportion 90% and 10%, respectively (figure 1). Patients admitted from 1 November 2019 to 31 December 2019 were used to create 'test' set which split by calendar month to ensure temporal generalisability of the models. Where there were duplicate inpatient visits by the same patient in the 'training' and 'test' set, the record was excluded to avoid testing based on previously learnt data. These 196 CXRs excluded from the 'test' set was added to the 'training' and 'validation' set for model development.

A deep-learning classifier was developed which combined a pretrained image classification network—Xception. Xception is an extension of the inception architecture, which replaces the standard Inception modules with depth-wise separable convolutions.[23] A transfer learning approach, which uses a predefined model, has the benefit of taking advantage of data from the first setting to extract information that may be useful when learning or even when directly making predictions in the second setting.[24] The models were implemented in Keras, V.1.3.0 29, Scikit-learn, V.0.19.one and Python V.3.7 (Python Software Foundation).

Model training was stopped in April 2020, when AUC for inpatient mortality in the 'test' set was determined to have reached 0.890 and accuracy of 0.899. After accounting for data clustering in the retrospective cohort, CAPE mortality risk score had an AUC of 0.88 (95% CI 0.86 to 0.90, p<0.001). The model showed good internal calibration (calibration intercept=0.00, slope=1.00, Brier score=0.069) (table 1).

Subsequently, CAPE was implemented as a computer application. Independent CXR in Digital Imaging and Communications in Medicine (DICOM) format uploaded into the software may be analysed for determination of an image-based mortality risk score. To aid clinicians in visually interpreting how the predictive score was generated by the deep-learning model, we adopted the use of a gradient-weighted class activation map to generate a heatmap. This demonstrates the neural-network activated in the forward-pass during inference/prediction. Figure 2 demonstrates an AI generated heatmap overlaid on a CXR showing pneumonic consolidation.

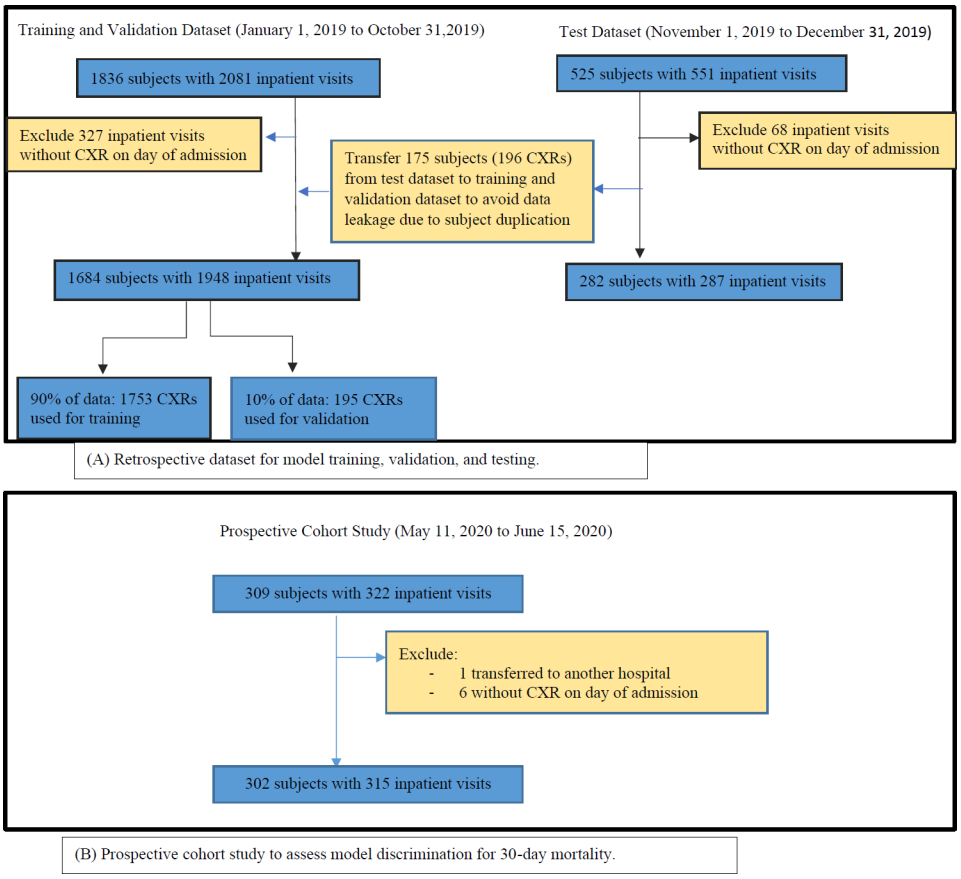


Figure 1 Datasets for CAPE model development and prospective cohort study. (A) Retrospective dataset for model training, validation and testing. (B) Prospective cohort study to assess model discrimination for 30-day mortality. CAPE, CAP AI predictive Engine; CXR, chest radiography.

Prospective cohort study for discrimination of CAPE mortality risk score

The prospective cohort study included adults who required inpatient admission for a physician-determined diagnosis of CAP via the emergency department. This occurred over the period of 11 May 2020 to 15 June 2020. They were identified within 72 hours of admission utilising electronic medical records. Baseline demographic information and health risk factors such as age, smoking status, body mass index and comorbidities were collected by trained research personnel, who were blinded to CAPE. Disease characteristics at initial presentation were recorded, these included vital signs; pneumonia severity scores; self-reported respiratory symptoms; the presence of associated complications such as acute kidney injury, acute myocardial injury and delirium. Laboratory data such as blood indices, biochemistry, infection biomarkers and microbiology tests were collected. Significant treatment data which may affect mortality outcomes, like prior antibiotics usage in the previous 30 days and timing of antibiotic administration, were recorded for analysis.

Patients with pulmonary tuberculosis were not excluded from the cohort as pulmonary tuberculosis is endemic in Singapore. CAP with SARS-CoV-2 as the microbial aetiology is endemic in some countries, hence it was not excluded. Recently, PSI and CURB-65 has been shown to demonstrate good discrimination for SARS-CoV-2 pneumonia.[25–27] The authors discern that a model incorporating all microbial causes of CAP would be practically more useful in healthcare systems where comprehensive testing for microbial aetiologies may be limited by available resources.

The first CXR performed on the day of inpatient admission, was extracted for CAPE analysis. The image is loaded from a standard DICOM file to generate an image-based heatmap and mortality risk score (figure 2). The mortality risk score is expressed in whole numbers from 0 to 100, with higher values indicating greater risk of death. A primary outcome of mortality at 30 days from the time of admission, concluded the data collection for analysis. Sample size calculation was based on a formula reported in Riley *et al*.[28] To estimate the 30-day mortality risk in the prospective dataset with sufficient precision, assuming an anticipated outcome proportion of 0.2 and a margin of error ≤0.05, the required sample size is at least 246.

Statistical analysis

This study was reported in accordance with the Transparent Reporting of a multivariate prediction model for Individual Prognosis or Diagnosis guidelines.[29]

Table 1 Baseline risk factors and pneumonia characteristics in relation to 30-day mortality

Variable	All patients n=315	Survivors n=254	Non-survivors n=61	P value
Age, years	71.4+18.0	69.2+18.7	80.4+10.3	<0.01
Male	179 (56.8)	141 (55.5)	38 (62.3)	0.39
BMI*	22.7+6.1	23.5+6.3	19.9+3.7	0.0003
Smoking†				
Never	177 (56.2)	152 (59.8)	25 (41.0)	<0.01
Former	29 (9.2)	28 (11.0)	1 (1.6)	
Current	24 (7.6)	16 (6.3)	8 (13.1)	
Residing in long-term care facility	63 (20)	45 (17.7)	18 (29.5)	0.05
Hospitalisation in prior 30 days	69 (21.9)	49 (19.3)	20 (32.8)	0.02
Antibiotic therapy in prior 30 days	71 (22.5)	55 (21.7)	16 (26.2)	0.50
Comorbid chronic diseases				
Diabetes mellitus	110 (34.9)	83 (32.7)	27 (39.3)	0.10
Hypertension	190 (60.3)	151 (59.4)	39 (63.9)	0.56
Ischaemic heart disease	86 (27.3)	65 (25.6)	21 (34.4)	0.20
Congestive cardiac failure	17 (5.4)	11 (4.3)	6 (9.8)	0.11
Asthma	22 (7.0)	19 (7.5)	3 (4.9)	0.59
Chronic obstructive lung disease	22 (7.0)	17 (6.7)	5 (8.2)	0.78
Bronchiectasis	22 (7.0)	18 (7.1)	4 (6.6)	1.00
Dementia	85 (27.0)	58 (22.8)	27 (44.3)	0.001
Parkinson's disease	23 (7.3)	18 (7.1)	5 (8.2)	0.79
Chronic kidney disease	48 (15.2)	34 (13.4)	14 (23.0)	0.07
Chronic liver disease	11 (3.5)	8 (3.1)	3 (4.9)	0.45
Active cancers	34 (10.7)	16 (6.3)	18 (29.5)	<0.01
Vital signs on admission				
Fever >38.0°C	191 (60.6)	160 (63.0)	31 (50.8)	0.11
Heart rate/minute	96.9+1.3	95.2+1.3	104.0+3.4	0.006
Respiratory rate/minute	21.3+5.0	20.2+3.7	25.5+7.2	<0.01
Systolic blood pressure (mm Hg)	130.8+26.8	134.5+25.4	115+27.1	<0.01
Diastolic blood pressure (mm Hg)	72.1+14.3	73.1+13.9	67.9+15.3	0.001
Pulse capillary oxygenation (%)	95.0+5.3	95.6+4.7	92.2+6.7	<0.01
Fraction of inspired oxygen required	29.8+20.5	26.2+15.1	44.4+31.1	<0.01
Respiratory symptoms				
Cough	178 (56.6)	150 (59.1)	28 (45.9)	0.08
Dyspnoea	130 (41.3)	88 (34.6)	42 (68.9)	<0.01
Sputum	80 (25.4)	62 (24.4)	18 (29.5)	0.41
Haemoptysis	4 (1.3)	4 (1.6)	0 (0)	1.00
Rhinorrhea	11 (3.5)	10 (3.9)	1 (1.6)	0.70
Throat pain	16 (5.1)	16 (6.3)	0 (0)	0.05
Chest pain	31 (9.8)	27 (10.6)	4 (6.6)	0.47
Wheeze	2 (0.6)	2 (0.8)	0 (0)	1.00
Myalgia	10 (3.2)	9 (3.5)	1 (1.6)	0.69
Lethargy	43 (13.7)	28 (11.0)	15 (24.6)	0.01
Fall	20 (6.3)	17 (6.7)	3 (4.9)	0.78
Nausea or vomiting	178 (56.5)	150 (59.1)	28 (45.9)	0.08
Pneumonia complications				

Continued

Table 1 Continued

Variable	All patients n=315	Survivors n=254	Non-survivors n=61	P value
Acute myocardial injury	22 (7.0)	12 (4.7)	10 (6.4)	0.004
Acute kidney injury	73 (23.2)	43 (16.9)	30 (9.1)	<0.001
Delirium	26 (8.3)	11 (4.3)	15 (24.6)	<0.001
Critical care admission	12 (3.8)	11 (4.3)	1 (1.6)	0.34
Laboratory indices				
Serum C reactive protein (mg/L)	68.9+78.5	59.3+69.7	109.0+98.6	<0.001
Serum procalcitonin (µg/L)	3.2+13.5	1.4+5.9	11.3+27.8	<0.001
White cell count (x10∧9 cells/L)	12.8+14.5	11.3+4.9	18.9+30.7	0.0002
Microbial aetiology				
SARS-CoV-2	18 (5.7)	18 (7.1)	0 (0)	1.00
Mycobacterium tuberculosis	11 (3.5)	11 (4.3)	0 (0)	1.00

Variable	All patients n=315	Survivors n=254	Non-survivors n=61
Pneumonia Severity Scores			
CURB-65 score			
0	62 (19.7)	61 (24.0)	1 (1.6)
1	99 (31.4)	90 (35.4)	9 (14.8)
2	103 (32.7)	73 (28.7)	30 (49.2)
3	40 (12.7)	27 (10.6)	13 (21.3)
4	9 (2.9)	3 (1.2)	6 (9.8)
5	2 (0.6)	0 (0)	2 (3.3)
Pneumonia Severity Index	101.4+40.2	92.7+36.1	137.5+36.3
Pneumonia Severity Index Class			
I	34 (10.8)	34 (13.4)	0 (0)
II	36 (11.4)	36 (14.2)	0 (0)
III	54 (17.1)	48 (18.9)	6 (9.8)
IV	115 (36.5)	96 (37.8)	19 (31.1)
V	76 (24.1)	40 (15.7)	36 (59.0)
CAPE Mortality Risk Score	49.6+29.1	44.0+27.8	72.8+22.4
Outcomes			
Received antibiotics within 24 hours of presentation	288 (91.4)	227 (89.4)	61 (100)
Corticosteroids (oral or intravenous)	34 (10.8%)	30 (11.8)	4 (6.6)
Hospitalisation days‡	8.9+9.1 5.5(3,12)	9.5+0.6 6(3,13)	6.8+0.8 4(2,10)

Data are presented as number (%), mean±SD, median (IQR).
*Data missing for 102 subjects.
†Data missing for 85 subjects.
‡Data missing for 1one subject.
BMI, body mass index; CAPE, community-acquired pneumonia artificial intelligence predictive engine.

Logistic regression was used to model 30-day mortality, with the CAPE mortality risk score as a predictor in the retrospective dataset, and PSI or CURB-65 as a predictor in the prospective dataset. We assessed if the log odds of 30-day mortality was linearly associated with CAPE mortality risk score and PSI using Box-Tidwell tests. We then modelled CAPE mortality risk scores and PSI values using a restricted cubic spline function with five knots at the 5th, 27.5th, 50th, 72.5th and 95th ercentiles before testing if the coefficient of the non-linear spline terms jointly equal 0.[30] Cluster-robust SEs were used to account for clustering by subjects. Model discrimination was assessed using the AUC, with 95% CI calculated by clustered bootstrap resampling (1000 replications).

Model calibration was assessed graphically using calibration plots with locally weighted scatterplot smoothing to examine the agreement between predicted and observed mortality risk across deciles.[31] The calibration-in-the-large

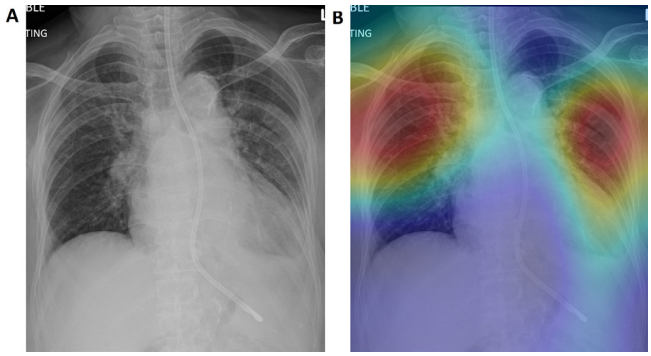


Figure 2 AI generated Grad-CAM heatmap of a CXR with community-acquired pneumonia. Frontal chest radiograph (A) of a patient presenting with acute respiratory failure secondary to pneumonia, performed in the emergency department. Grad-CAM heatmap (B) highlights areas of greatest class activation by the AI model, which corresponds to areas of pulmonary consolidation, with the extent and intensity of activation mirroring the severity of pneumonia. AI, artificial intelligence; CXR, chest radiography; Grad-CAM, gradient-weighted class activation map.

(CITL) and the calibration slope were assessed. An ideal CITL and calibration slope would have values of 0 and 1 respectively.[32] A CITL of <0 (or >0) indicates that the model overestimates (or underestimates) risk on average, while a calibration slope of <1 (or >1) indicates that the predicted risks are too extreme (or too moderate).[32] Where there was miscalibration, a more parsimonious method of model recalibration was adopted as the prospective dataset was small relative to the retrospective dataset. The intercept and slope were updated using the prospective dataset by using the linear predictor in the original model as the only covariable (logistic calibration).[33 34]

Overall goodness-of-fit was assessed using the Brier score,[31] which measures the accuracy of predictions. The score ranges from 0 to 1, with a lower score indicating better model performance.

The high mortality risk cut-off was selected at ≥20%, similar with commonly used cut-offs in the literature.[5] Model performance in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were assessed across potential mortality risk cut-offs.

To quantify the incremental value of CAPE, the net reclassification improvement (NRI) for the addition of CAPE to PSI and CURB-65, respectively, were calculated.[35]

Spearman's r was used to measure the strength of correlation between CAPE mortality risk score and PSI; CAPE mortality risk score and CURB-65. Weighted Cohen's kappa was used to estimate the degree of agreement between CAPE mortality risk score and PSI class[36]; CAPE mortality risk score and CURB-65. For this analysis, CAPE mortality risk score (range 0–100) was divided into five categories (0–20, 21-40, 41–60, 61–80, 81–100); while CURB-65 (range 0–5), had scores 4 and 5 collapsed into one severity band, to make five categories.

Missing data were present in some variables such as body mass index and laboratory data. No missing data were present in the calculation of Pneumonia Severity Scores of CURB-65, PSI and CAPE mortality risk score. As the primary and secondary outcome analysis did not require use of variables with missing data, no treatment of missing data using imputation methods was necessary.

All statistical analyses were conducted using Stata V.15.0 (StataCorp).

RESULTS

A total of 315 inpatient visits for CAP were included for analysis over the prospective cohort period between 11 May 2020 and 15 June 2020 (figure 1). This comprised 302 subjects, of whom two had three inpatient visits; nine had two inpatient visits. Statistical analysis was performed for all 315 inpatients visits for the following four reasons: subjects who had more than one visit had returned to community prior to a second or third inpatient episode of pneumonia; the CAPE mortality score generated based on the CXR at the start of each visit, were varied due to different severities of CAP at presentation; the 61 subjects who did not survive to 30 days were unique with no duplication primary outcome in the analysis; the authors sought to validate the model in a real-world situation, where some patients experience early hospital readmission after discharge.

Baseline demographics and comorbidities

A total of 315 inpatient visits for CAP over the prospective cohort study period had 30-day mortality of 19.4% (n=61/315). 56.8% (n=179/315) were male. Baseline demographics, health risk factors and comorbidities are presented in table 1.

Non-survivors were older than survivors (mean (SD) age, 80.4 (10.3) vs 69.2 (18.7)). They had lower body mass index (mean (SD), 19.9 (3.7) vs 23.5 (6.3)); more residing in long-term care facilities (n=18/61 vs n=45/254); had prior hospitalisations in previous 30 days (n=20/61 vs n=49/254). Comorbidities found more commonly in non-survivors were dementia (n=27/61 vs n=58/254); active malignancies (n=16/61 vs n=18/254) and chronic kidney disease (n=14/61 vs n=34/254).

CAP data

At initial disease presentation, non-survivors were more likely to have higher heart rates (mean (SD), higher respiratory rates, lower blood pressures, lower pulse oximetry readings, require high oxygen supplementation. Symptoms of dyspnoea (n=42/61 vs n=88/254) and delirium (n=15/61 vs n=11/254) were more common in the non-survivors. As were CAP complications of delirium (n=15/61 vs n=11/254); acute kidney injury (n=30/61

vs n=43/254); acute myocardial injury (n=10/61 vs n=43/254).

Significantly, non-survivors demonstrated higher inflammatory biomarkers such as serum C-reactive protein (mean (SD), 109 mg/L (98.6) vs 59.3 mg/L (69.7)); serum procalcitonin (mean (SD), 11.3 (27.8) µg/L vs 1.4 (5.9) µg/L); white cell count (mean (SD), 18.9x10∧9 cells/L(30.7) vs 11.3x10∧9 cells/L(4.9)).

The spectrum of microbiological data captured in the prospective study cohort was influenced by limited microbiological testing for non-severe CAP, and microbial aetiologies apart from SAR-CoV-2, over the period of study recruitment. This was to preserve local laboratory capacity for population SAR-CoV-2 screening. All patients received SARS-CoV-2 PCR testing, of which 5.7% (n=18/315) were positive. Twenty-nine mycobacterium tests were performed at physician discretion, yielded 3.5% (n=11/315) of pulmonary tuberculosis as the underlying microbial cause of CAP. No patients with COVID-19 or pulmonary tuberculosis demised.

All non-survivors received appropriate antibiotics in the first 24 hours of presentation, while 91.4% (n=288/315) of the total study cohort received the same.

CAPE mortality risk score and pneumonia severity scores: CURB-65, PSI

The AUC of CAPE mortality risk score for 30-day mortality was determined to be 0.79 (95% CI 0.73 to 0.85, p<0.001). The AUC of CURB-65 for 30-day mortality was 0.76 (95% CI 0.70 to 0.81, p<0.001); while that of PSI was 0.80 (95% CI 0.74 to 0.86, p<0.001) (table 2, figure 3A).

There was evidence of miscalibration of CAPE mortality risk score with a CITL of 0.84, calibration slope of 0.58 and Brier score 0.14 (online supplemental figure 1). We recalibrated the model by updating both the intercept and slope using the prospective dataset. The recalibrated model had a CITL of 0.00, calibration slope of 1.00 and Brier score 0.13 (table 2).

We incorporated CAPE with PSI, and CAPE with CURB-65, respectively. We then assessed if there were differences in the AUCs between (1) CURB-65 +CAPE, and (2) PSI+CAPE models, based on a method described by DeLong *et al*[37] CURB-65 +CAPE (AUC 0.83, 0.77 to 0.88, p<0.001) had a larger AUC than CURB-65 (χ^2=8.66, p=0.003), while PSI+CAPE (AUC 0.84, 95% CI 0.79 to 0.89, p<0.001) had a larger AUC than PSI (χ^2=3.79, p=0.052) (figure 3B,C). Calibration performed in the prospective dataset is presented in online supplemental figure 2.

The NRI for the addition of CAPE to PSI at a 30-day mortality risk threshold of 0.20 was 4.6% (95% CI 3.9% to 5.3%); while the NRI for the addition of CAPE to CURB-65 was 4.5% (95% CI 3.8% to 5.2%), presented in online supplemental table 1.

The performance of the CAPE mortality risk score in clinically relevant metrics across different risk cut-offs are described in table 3. At the 30-day mortality risk cut-off of 0.20, sensitivity was 0.77 (n=47/61), specificity 0.67 (n=169/254), PPV 0.36 (n=47/132), NPV of 0.92 (n=169/183).

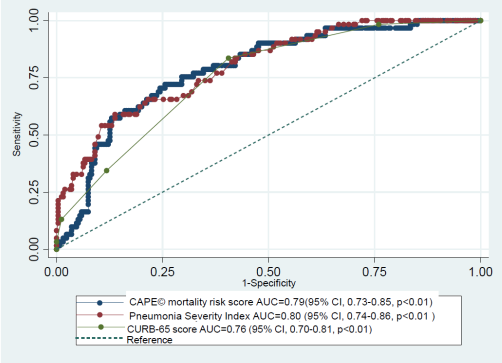
Logistic regression of CAPE mortality risk score for the binary outcome of 30-day mortality yielded an unadjusted OR of 1.04 (95% CI 1.03 to 1.05, p<0.01), indicating a 4% increase in the odds of death for every 1-point increase. The unadjusted OR of CURB-65 (0–5 scale) for 30-day mortality was 2.66 (95% CI 1.94 to 3.65, p<0.01); while that of PSI (ranged 16 to 210 in cohort) was 1.03 (95% CI 1.02 to 1.04, p<0.01), respectively. The

Table 2 Model discrimination and calibration in retrospective and prospective datasets

		Retrospective dataset (n=1948)		Prospective cohort (n=315)			
Model	Outcome	AUROC (95% CI)	Brier score	AUROC (95% CI)	CITL (95% CI)	Calibration slope (95% CI)	Brier score
CAPE	Inpatient mortality	0.88 (0.86 to 0.90)	0.069	0.81 (0.73 to 0.88)	0.14 (–0.26 to 0.53)	0.67 (0.39 to 0.95)	0.092
	30-day mortality	–	–	0.79 (0.73 to 0.85)	0.84 (0.46 to 1.22)	0.58 (0.39 to 0.76)	0.137
Recalibrated CAPE	30-day mortality	–	–	0.79 (0.73 to 0.85)	0.00 (–0.31 to 0.31)	1.00 (0.67 to 1.33)	0.130
PSI	30-day mortality	–	–	0.80 (0.74 to 0.86)	0.00 (–0.32 to 0.32)	1.00 (0.72 to 1.28)	0.121
CURB-65	30-day mortality	–	–	0.76 (0.70 to 0.81)	0.00 (–0.30 to 0.30)	1.00 (0.66 to 1.34)	0.131
CAPE+PSI	30-day mortality	–	–	0.84 (0.79 to 0.89)	0.00 (–0.33 to 0.33)	1.00 (0.73 to 1.27)	0.116
CAPE+CURB-65	30-day mortality	–	–	0.83 (0.77 to 0.88)	0.00 (–0.32 to 032)	1.00 (0.71 to 1.29)	0.121

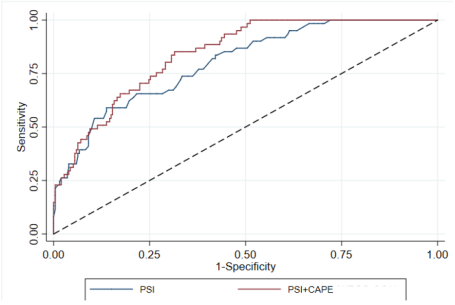
AUROC, area under the receiver operating characteristic curve; CAPE, community-acquired pneumonia artificial intelligence predictive engine; CITL, calibration-in-the-large; CURB-65, Confusion of new onset, blood Urea nitrogen, Respiratory rate, Blood pressure, 65 years old; PSI, Pneumonia Severity Index.

Figure 3a. CAPE© mortality risk score and pneumonia severity scores receiver operator characteristic curves for 30-day mortality.



Comparing receiver operator characteristic curves for CAPE© mortality risk score, Pneumonia Severity Index, CURB-65, for 30-day mortality, using DeLong's method, p=0.32

Figure 3b. PSI+CAPE© mortality risk score and PSI receiver operator characteristic curves for 30-day mortality.

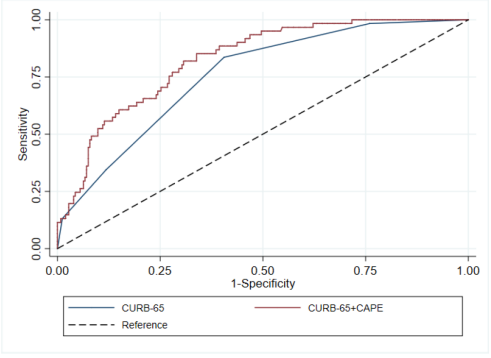


AUROC of PSI: 0.80 (95% CI 0.74 to 0.86, P<0.01)

AUROC of PSI+CAPE©: 0.84 (95% CI 0.79 to 0.89, P<0.01)

Wald's test using the DeLong, DeLong, and Clarke-Pearson[1] covariance estimates for the AUROC estimate: χ^2=3.79, P=0.052

Figure 3c. CURB-65+CAPE© mortality risk score and CURB-65 receiver operator characteristic curves for 30-day mortality.



AUROC of CURB-65: 0.76 (95% CI 0.70 to 0.81, P<0.01)

AUROC of CURB-65+CAPE: 0.83 (95% CI 0.77 to 0.88, P<0.01)

Wald's test using the DeLong, DeLong, and ClarkePearson[1] covariance estimates for the AUROC estimate: χ^2=8.66, P=0.003

Figure 3 (A) CAPE mortality risk score and pneumonia severity scores receiver operator characteristic curves for 30-day mortality. (B) PSI+CAPE mortality risk score and PSI receiver operator characteristic curves for 30-day mortality. (C) CURB-65 + CAPE mortality risk score and CURB-65 receiver operator characteristic curves for 30-day mortality. AUROC, area under the receiver operating characteristic curve; CAPE, CAP AI predictive Engine; CURB-65, Confusion of new onset, blood Urea nitrogen, Respiratory rate, Blood pressure, 65 years old; PSI, Pneumonia Severity Index.

Box-Tidwell test did not show evidence of departure from linearity for CAPE mortality risk score (p=0.63) or PSI values (p=0.80). This was consistent with tests of the coefficients of the non-linear spine terms for CAPE mortality risk scores (χ2=4.31, p=0.230) and PSI values (χ2=6.18, p=0.103).

To assess for correlation between CAPE mortality risk score and PSI, the Spearman's r was 0.50 (p<0.01), while that of CAPE mortality risk score and CURB-65 was 0.44 (p<0.01), indicating moderate and low positive correlation, respectively.

To assess for agreement, weighted Cohen's kappa was used for analysis. There was moderate agreement between CAPE mortality risk score and PSI, with kappa determined to be 0.46; while that of CAPE mortality risk score and CURB-65 had a kappa of 0.38, showing fair agreement.

The characteristics of the patients with discordant and concordant 30-day mortality risk categories based on CAPE and PSI are presented in online supplemental table 2.

DISCUSSION

In this study, we demonstrated that an AI model based on first CXR image performed during the assessment for CAP can prognosticate 30-day mortality with an AUC of 0.79. This is comparable to that of currently used, well-validated pneumonia severity risk scores, with an AUC of 0.80 for PSI and 0.77 for CURB-65 demonstrated in the same study cohort. The AUCs for PSI and CURB-65 in this study are similar with that of a prior meta-analysis.[9]

We showed that CAPE mortality risk score had moderate positive correlation and agreement with PSI; low positive correlation and fair agreement with CURB-65. This suggests that while all three prognostic tools displayed similar AUCs, the CAPE mortality risk score can do so by using imaging parameters captured by CNN, independent of the need for descriptive medical data.

We further combined CAPE mortality risk score with well-validated pneumonia severity risk scores. The AUC of CURB-65 improved from 0.76 (95% CI 0.70 to 0.81) to 0.83 (95% CI 0.77 to 0.88), while PSI improved from 0.80 (95% CI 0.74 to 0.86) to 0.84 (95% CI 0.79 to 0.89). This indicates that the additional of an imaging CNN model to traditional pneumonia severity scoring has value in improving the discrimination of the model for mortality.

To our knowledge, this is the first report describing deep learning of CXRs to predict 30-day mortality in CAP. Future research may be conducted to accurately quantify the degree to which CNN analysis of CXRs correlate with commonly used pneumonia severity markers, such as oxygenation and sepsis indices, to understand the additional value that CNN brings to pneumonia prognostication.

The authors suggest that CAPE has the potential be a clinical decision support tool incorporated into emergency department or inpatient clinical workflows, for the purposes of triaging of CAP. One such example could be a

Table 3 Performance of CAPE mortality risk score in predicting 30-day mortality at different risk cut-offs in the prospective dataset

30-day mortality risk cut-off	CAPE mortality risk score	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
≥0.05	≥22.2	57	97	157	4	93.4	38.2	26.6	96.0
≥0.10	≥40.6	53	135	119	8	86.9	53.2	30.8	94.4
≥0.15	≥52.0	49	155	99	12	80.3	61.0	33.1	92.8
≥0.20	≥60.6	47	169	85	14	77.1	66.5	35.6	92.4
≥0.25	≥67.7	44	183	71	17	72.1	72.1	38.3	91.5
≥0.30	≥73.9	43	191	63	18	70.5	75.2	40.6	91.4
≥0.35	≥79.5	37	211	43	24	60.7	83.1	46.3	89.8
≥0.40	≥84.8	28	225	29	33	45.9	88.6	49.1	87.2

CAPE, community-acquired pneumonia artificial intelligence predictive engine; FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.

triaging clinician indicating the diagnosis of CAP on a CXR request. The DICOM image can be processed through the CAPE software by trained personnel to generate the CAPE mortality risk score. If the CAPE mortality risk score threshold is below 5%, this would correspond to an NPV of 0.96. This information may be communicated on a radiology report to the clinician to encourage outpatient care for low-risk CAP or early discharge strategies. Conversely, if the CAPE mortality risk score threshold is above 20%, this would correspond to a PPV of 0.36 for mortality. This information may be communicated to the clinician to strongly consider the need for critical care monitoring, with advanced care plans in place. In centres where PSI and CURB-65 are scored routinely and available electronically, the combined PSI- CAPE or CURB-65- CAPE model can provide greater discrimination. This may subsequently be applied in pre-existing institution-specific workflows for CAP.

A potential advantage of an AI prognostic model is the flexibility of rapidly customising cut-off points or risk thresholds to CAP epidemiology and healthcare resource availability over time and space. This would maximise efficiency of healthcare resources. For example, during pandemic hospital bed shortages, (in conjunction with real-time vital signs) the model risk threshold can be increased to identify patients at higher risk of deterioration to prioritise care. (Data scientists) can remotely calibrate the model by CXR extraction, based on recent mortality data, and adjust the risk threshold of the model accordingly.

The authors recognise that while robust prognostic models may exist, further studies are needed to assess the effectiveness of real-world implementation. Currently, there is a paucity of data on quality improvement outcomes using AI as clinician decision support tools.[38 39] In addition, the real-world implementation of existing disease prognostication tools may or may not contribute to improving clinical outcomes ultimately. Factors such as clinician acceptance and the availability of effective clinical support systems to incorporate these tools are likely to play a greater role in improving care outcomes.[40–42]

Limitations

The model was developed using radiological and health data from a single institution with prospective validation performed at the same place. Hence, we have not demonstrated generalisability. The study authors are currently in the process of performing a multicentre study for this purpose and welcome any collaborators who may be interested in developing, validating, and using this tool. While CAPE is proprietary (intellectual property belonging to Singapore Health Services and Integrated Health Information Systems), the authors have collectively agreed for free use of this software with acknowledgements, for research purposes, over the duration of COVID-19 pandemic.

A second limitation would be the lack of comprehensive or standardised microbiological data collection during the study period. This was due to manpower and laboratory resources being diverted to prioritise pandemic planning and SARS-CoV-2 testing. Despite this, the study authors suggest that radiological severity can be more predictive of mortality than microbiological aetiology. While impact of microbiological data on the AUC of CAPE mortality risk score is yet uncertain, the authors suggest that there may be minimal effect on the discrimination of the model.

A third limitation is that the authors have yet to ascertain if CAPE mortality risk score would have higher discriminative power if combined with non-imaging medical data, apart from PSI and CURB-65. Further model development in combination with known CAP mortality predictors is in progress.

Lastly, the authors acknowledge that the outcome of mortality in CAP, while important, may be less clinically useful than other outcome indicators such as risk of critical care admissions and estimated length of inpatient stay. The authors are currently working on AI models to address these clinical questions.

CONCLUSION

We have shown that AI can be used to build a mortality prognostic model for CAP based on CXR. The AUC for

30-day mortality is comparable to conventional pneumonia severity scores such as PSI and CURB-65, with further potential to improve its discrimination for mortality.

Author affiliations
[1]Department of Respiratory and Critical Care Medicine, Changi General Hospital, Singapore
[2]Department of Radiology, Changi General Hospital, Singapore
[3]Integrated Health Information Systems Pte Ltd, Singapore
[4]Health Services Research, Changi General Hospital, Singapore
[5]Data Management and Informatics, Changi General Hospital, Singapore

Acknowledgements The authors thank Mr. Koh Tzan Tsai, Ms. Sandhiya Ramanathan, Dr. Angeline Poh, Dr. Jansen Koh, Dr. Perry Liew, Dr. Oh Hong Choon, Dr. Srinath Sridharan and research coordinators of Changi General Hospital, Clinical Trials & Research Unit, for their dedicated help in data preparation and analysis; We also thank the physicians of General Medicine and Respiratory departments, radiologists of Diagnostic Radiology department, for contributing data through their daily clinical work.

Contributors All listed authors have substantial contributions to the conception or design of the work; or the acquisition, analysis or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content; and dinal approval of the version to be published; and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Disclaimer We grant BMJ Open Respiratory Research exclusive license for publication of this work.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethics approval was given by SingHealth Centralised Institutional Review Board (CIRB 2020/2100).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Deidentified data are available from the corresponding author on reasonable request subjected to institutional approval.

ORCID iD
Jessica Quah http://orcid.org/0000-0002-2227-9810

REFERENCES

1 World Health Organisation. Global health Observatory data. Available: http://www.who. int/gho/mortality_burden_disease/causes_death/top_10/en [Accessed 14 Aug 2020].
2 Fine MJ, Smith MA, Carson CA, *et al*. Prognosis and outcomes of patients with community-acquired pneumonia. A meta-analysis. *JAMA* 1996;275:134–41.
3 Pieralli F, Vannucchi V, De Marzi G, *et al*. Performance status and in-hospital mortality of elderly patients with community acquired pneumonia. *Intern Emerg Med* 2018;13:501–7.
4 Bont J, Hak E, Hoes AW, *et al*. Predicting death in elderly patients with community-acquired pneumonia: a prospective validation study Reevaluating the CRB-65 severity assessment tool. *Arch Intern Med* 2008;168:1465–8.
5 Lim WS, van der Eerden MM, Laing R, *et al*. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* 2003;58:377–82.
6 España PP, Capelastegui A, Gorordo I, *et al*. Development and validation of a clinical prediction rule for severe community-acquired pneumonia. *Am J Respir Crit Care Med* 2006;174:1249–56.
7 Niederman MS. Making sense of scoring systems in community acquired pneumonia. *Respirology* 2009;14:327–35.
8 Singanayagam A, Chalmers JD, Hill AT. Severity assessment in community-acquired pneumonia: a review. *QJM* 2009;102:379–88.
9 Chalmers JD, Singanayagam A, Akram AR, *et al*. Severity assessment tools for predicting mortality in hospitalised patients with community-acquired pneumonia. systematic review and meta-analysis. *Thorax* 2010;65:878–83.
10 Lee RWW, Lindstrom ST. A teaching hospital's experience applying the Pneumonia Severity Index and antibiotic guidelines in the management of community-acquired pneumonia. *Respirology* 2007;12:754–8.
11 Zhang ZX, Yong Y, Tan WC, *et al*. Prognostic factors for mortality due to pneumonia among adults from different age groups in Singapore and mortality predictions based on psi and CURB-65. *Singapore Med J* 2018;59:190–8.
12 Lynn LA. Artificial intelligence systems for complex decision-making in acute care medicine: a review. *Patient Saf Surg* 2019;13:6.
13 Coorevits P, Sundgren M, Klein GO, *et al*. Electronic health records: new opportunities for clinical research. *J Intern Med* 2013;274:547–60.
14 Kermany DS, Goldbaum M, Cai W, *et al*. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122–31.
15 Stephen O, Sain M, Maduh UJ, *et al*. An efficient deep learning approach to pneumonia classification in healthcare. *J Healthc Eng* 2019;2019:1–7.
16 Heckerling PS, Gerber BS, Tape TG, *et al*. Prediction of community-acquired pneumonia using artificial neural networks. *Med Decis Making* 2003;23:112–21.
17 Hwang EJ, Park S, Jin K-N, *et al*. Development and validation of a deep Learning-Based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019;2:e191095.
18 Ito R, Iwano S, Naganawa S. A review on the use of artificial intelligence for medical imaging of the lungs of patients with coronavirus disease 2019. *Diagn Interv Radiol* 2020;26:443–8.
19 Li L, Qin L, Xu Z, *et al*. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology* 2020;296:E65–71.
20 Ozturk T, Talo M, Yildirim EA, *et al*. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med* 2020;121:103792.
21 Lu MT, Ivanov A, Mayrhofer T, *et al*. Deep learning to assess long-term mortality from chest radiographs. *JAMA Netw Open* 2019;2:e197416.
22 Liu F, Zhang Q, Huang C, *et al*. Ct quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients. *Theranostics* 2020;10:5613–22.
23 Chollet F. Xception: deep learning with Depthwise separable Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition* 2017:1251–8.
24 Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge, MA: The MIT Press, 2016.
25 Satici C, Demirkol MA, Sargin Altunok E, *et al*. Performance of pneumonia severity index and CURB-65 in predicting 30-day mortality in patients with COVID-19. *Int J Infect Dis* 2020;98:84–9.
26 Peng J, Qi D, Yuan G, *et al*. Diagnostic value of peripheral hematologic markers for coronavirus disease 2019 (COVID-19): a multicenter, cross-sectional study. *J Clin Lab Anal* 2020;34:e23475.
27 Fan G, Tu C, Zhou F, *et al*. Comparison of severity scores for COVID-19 patients with pneumonia: a retrospective study. *Eur Respir J* 2020;56:2002113.
28 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
29 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.

30 Harrell Jr FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. New York: Springer, 2001.
31 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.
32 Van Calster B, McLernon DJ, van Smeden M. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.
33 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, *et al*. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567–86.
34 Su T-L, Jaki T, Hickey GL, *et al*. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res* 2018;27:185–97.
35 Pencina MJ, D'Agostino RB, D'Agostino RB, *et al*. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
36 Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. *Perspect Clin Res* 2017;8:187–91.
37 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
38 Cresswell K, Callaghan M, Khan S, *et al*. Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: a systematic review. *Health Informatics J* 2020;26:2138–47.
39 Ramkumar PN, Haeberle HS, Bloomfield MR, *et al*. Artificial intelligence and arthroplasty at a single institution: real-world applications of machine learning to big data, value-based care, mobile health, and remote patient monitoring. *J Arthroplasty* 2019;34:2204–9.
40 Bedoya AD, Clement ME, Phelan M, *et al*. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Crit Care Med* 2019;47:49–55.
41 Escobar GJ, Liu VX, Schuler A, *et al*. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020;383:1951–60.
42 Kollef MH, Chen Y, Heard K, *et al*. A randomized trial of real-time automated clinical deterioration alerts sent to a rapid response team. *J Hosp Med* 2014;9:424–9.