BMJ Open Respiratory Research

Supervised machine learning model to predict mortality in patients undergoing venovenous extracorporeal membrane oxygenation from a nationwide multicentre registry

Haeun Lee,[1] Myung Jin Song,[2] Young-Jae Cho,[2] Dong Jung Kim,[3] Sang-Bum Hong,[4] Se Young Jung,[1,5] Sung Yoon Lim [2]

To cite: Lee H, Song MJ, Cho Y-J, *et al*. Supervised machine learning model to predict mortality in patients undergoing venovenous extracorporeal membrane oxygenation from a nationwide multicentre registry. *BMJ Open Respir Res* 2023;**10**:e002025. doi:10.1136/bmjresp-2023-002025

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjresp-2023-002025).

Received 19 August 2023
Accepted 1 December 2023

For numbered affiliations see end of article.

Correspondence to
Dr Sung Yoon Lim; nucleon727@gmail.com and

Dr Se Young Jung; syjung@snubh.org

ABSTRACT

Background Existing models have performed poorly when predicting mortality for patients undergoing venovenous extracorporeal membrane oxygenation (VV-ECMO). This study aimed to develop and validate a machine learning (ML)-based prediction model to predict 90-day mortality in patients undergoing VV-ECMO.

Methods This study included 368 patients with acute respiratory failure undergoing VV-ECMO from 16 tertiary hospitals across South Korea between 2012 and 2015. The primary outcome was the 90-day mortality after ECMO initiation. The inputs included all available features (n=51) and those from the electronic health record (EHR) systems without preprocessing (n=40). The discriminatory strengths of ML models were evaluated in both internal and external validation sets. The models were compared with conventional models, such as respiratory ECMO survival prediction (RESP) and predicting death for severe acute respiratory distress syndrome on VV-ECMO (PRESERVE).

Results Extreme gradient boosting (XGB) (areas under the receiver operating characteristic curve, AUROC 0.82, 95% CI (0.73 to 0.89)) and light gradient boosting (AUROC 0.81 (95% CI 0.71 to 0.88)) models achieved the highest performance using EHR's and all other available features. The developed models had higher AUROCs (95% CI 0.76 to 0.82) than those of RESP (AUROC 0.66 (95% CI 0.56 to 0.76)) and PRESERVE (AUROC 0.71 (95% CI 0.61 to 0.81)). Additionally, we achieved an AUROC (0.75) for 90-day mortality in external validation in the case of the XGB model, which was higher than that of RESP (0.70) and PRESERVE (0.67) in the same validation dataset.

Conclusions ML prediction models outperformed previous mortality risk models. This model may be used to identify patients who are unlikely to benefit from VV-ECMO therapy during patient selection.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Existing mortality risk models have been developed to estimate the likelihood of in-hospital survival in patients who received extracorporeal membrane oxygenation (ECMO). However, few studies have developed a predictive mortality model combined with machine learning (ML) methods in patients undergoing ECMO therapy. No studies have developed a ML-based model for predicting mortality in patients with venovenous ECMO (VV-ECMO) alone.

WHAT THIS STUDY ADDS

⇒ This is the first study to demonstrate that ML models developed only for patients with VV-ECMO outperform conventional regression-based models such as respiratory ECMO survival prediction (RESP) and predicting death for severe acute respiratory distress syndrome on VV-ECMO (PRESERVE). We developed a more practical model with a readily available electronic health record system without further preprocessing and showed its performance is comparable with those using full features. The ML-based models successfully predicted the risk of 90-day mortality and surpassed the accuracy, precision and sensitivity of the conventional risk-scoring models, RESP and PRESERVE, by 14%, 2.6% and 31%, respectively. External validation with different datasets and decision curve analysis also revealed that our models are transferable to other datasets, and clinicians can achieve positive net benefits across all thresholds for decision.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The ML prediction model for 90-day mortality rate could accurately identify VV-ECMO candidates with a low probability of success, which may facilitate effective utilisation of VV-ECMO by clinicians.

INTRODUCTION

Acute respiratory failure (ARF) is associated with high mortality, exceeding 60% in its most severe forms, despite the various strategies available for reducing ventilator-induced lung injury.[1 2] Extracorporeal membrane oxygenation (ECMO) has emerged as a rescue therapy for managing types of patients.[2] Recent studies with randomised

controlled trials suggest that ECMO should not be delayed and rather should be initiated rapidly in patients with refractory hypoxaemia after optimal conventional management.[3–5] Accordingly, extracorporeal life support organisation (ELSO) guidelines have been developed to help clinicians determine eligibility.[5 6] However, the only absolute contraindication for applying ECMO is for those with anticipated non-recovery without any viable decannulation.

Yet, associated mortality in patients with ECMO therapy still remains very high, and there are several circumstances other than absolute contraindications for ECMO cannulation, a very high-risk group due to numerous clinical conditions.[7 8] Moreover, the demand for ECMO has escalated tremendously among patients with ARF, particularly during the COVID-19 pandemic.[9] Prediction of mortality for ECMO treatment may aid in judicious patient selection for using the finite ECMO resources.[10 11]

To solve this problem, several prognostic scores have been developed to predict survival rates of patients who receive ECMO, such as the PREdiction of Survival on ECMO Therapy Score and predicted death rate for severe ARF on venovenous ECMO (VV-ECMO) (PRESERVE).[12–14] However, these models have relatively poor performance due to the linearity in the studies that were used to develop them.[7] The unique characteristics of VV-ECMO patients, such as the high mortality rates and diverse etiologies of ARF, also were an impediment to accurate mortality prediction. Thus, recent studies of ECMO prediction models discourage using any available scores as a single decision tool.

In the last decade, advanced modelling and machine learning (ML) techniques have demonstrated promising results in improving the prediction of the prognosis of critically ill patients.[15] Therefore, using nationwide registry data, we aimed to develop ML-based models for 90-day and in-hospital mortality prediction in patients treated with VV-ECMO. The ML prediction model may demonstrate a higher positive gain across different decision threshold probabilities in comparison to traditional scores, such as respiratory ECMO survival prediction (RESP) and PRESERVE scores. The model was further externally validated using an independent dataset to corroborate the classifier's reliability and compare the discrimination performance of the model with conventional prognostic scores. We also developed a derived model with sparse features readily available from the electronic health record (EHR) system.

METHODS

In this retrospective observational cohort study, we used a multicentre registry obtained from 16 tertiary hospitals in South Korea from January 2012 to December 2015. The cohort profile was explained in detail in a previous study.[16 17] The cohort comprised critically ill patients who were at least 16 years old and underwent VV-ECMO for severe ARF. There were no predefined criteria for the indications and contraindications of ECMO use between the participating centres. Decisions were taken at the discretion of the attending physicians at each centre. However, the initiation of ECMO was based on the general recommendations of the ELSO guidelines. Data were collected from each participating hospital using a standardised registry form. Participating hospitals registered a total of 428 patients during the study period. Of them, 60 were excluded for being on the ECMO for less than 48 hours, as patients in severe condition with a mean APACHE score of 30, and septic shock status within the first 48 hours were unlikely to meet the indications for continued ECMO support.[18] We divided the cohort into training (n=257) and test sets (n=111) at a ratio of 7:3. We also obtained another VV-ECMO cohort from the Seoul National University Bundang Hospital (SNUBH) for external validation between January 2016 and December 2021; 78 patients were included into the cohort (online supplemental figure 1). To protect the privacy and confidentiality of research participants' personal information, only anonymised and deidentified data were analysed.

To develop predictive models that can be easily implemented in the EHR systems in clinical practice, we used two sets of input variables: (1) EHR features, which were readily obtainable structured variables from the EHR system without requiring any preprocessing (n=40) and (2) all available manual input features (n=51). The EHR features were reviewed and selected by two attending physicians and an IT technician from the Department of Medical Informatics. All available manual input features included (1) demographic information, (2) anthropometric measurements, (3) laboratory values, (4) vital signs, (5) mechanical ventilator (MV)-related variables, (6) variables on patients' severity of illness before ECMO, (7) hospital-related variables and (8) variables not specified otherwise (online supplemental table 2).

Respiratory diagnoses included viral/bacterial pneumonia, chronic obstructive pulmonary disease/asthma, trauma/burn, asphyxia, acute exacerbation of interstitial lung disease or chronic respiratory failure. Immunocompromised status included solid tumours, haematological malignancies, HIV infection, solid organ transplantation or liver cirrhosis. Central nervous system dysfunction included encephalopathy, neurotrauma, cerebral embolism, stroke, seizures or epileptic syndrome.[13] All the input variables were obtained at the closest value before ECMO insertion. The primary outcome measure was 90-day mortality for a fair comparison with other 90-day mortality models.

To prepare the input features for the model, a comprehensive combination of imputations, outliers and feature scaling methods was implemented to boost the ML models. Extreme outliers with a Z-score greater or less than two were replaced by less extreme values with the 95th percentile, using winsorisation techniques to minimise the influence of outliers. The continuous variables were normalised to transform the varied features for similarity. Random forest-based multivariate imputation by

chained equations for continuous variables and K-nearest neighbours (KNN) for categorical variables were used to substitute missing values based on robust statistics and random forest regression algorithms for reducing bias while increasing precision.[18 19] A bootstrap resampling technique with 1000 replicates was used to compute 95% CIs for areas under the receiver operating characteristic curve (AUROCs). The optimal threshold point of the Youden index was measured to assess the sensitivity and specificity of both mortalities.

Six supervised ML models based on regression, tree ensembles, gradient boosting and neural networks were trained using 10-fold cross-validation to predict 90-day mortality in patients who underwent VV-ECMO. All parameters were tuned with randomised search cross-validation in 30 iterations, and each model's robustness was assessed using AUROC, area under the precision recall curve (AUPRC), sensitivity, specificity, positive predictive values (PPV) and negative predictive values.

Calibration plots were used to assess the reliability of the predictive models, detect biases and ensure that the model's predictions align accurately with the observed outcomes. Additionally, shapley additive explanations (SHAP) analysis was performed to explore the impact of each feature on the response variable and to interpret how a single feature can affect the output of the prediction model.[19] Decision curve analysis (DCA) was performed to evaluate the net benefit of the developed models across different thresholds.

We used two strategies to demonstrate the capabilities of the developed models: (1) comparison of discrimination performance with previously established models such as RESP and PRESERVE and (2) external validation of the developed models using an independent dataset for primary outcome. Therefore, we compared the AUROC of our EHR feature models with those of RESP and PRESERVE. We then validated the models with another dataset from SNUBH. Model development and validation were conducted using Python (Python Software Foundation, Wilmington, Delaware, USA; V.3.8.8) with the Scikit-learn library[20–27]

All statistical analyses were performed by using R studio software (RStudio, Boston, Massachusetts, USA; V.4.1.0). We used a standard two sample t-test for numeric variables and a χ^2 test of independence for categorical variables. Results are present as mean±SD and frequencies and percentages for continuous and categorical variables, respectively. A $p<0.05$ was considered statistically significant.[28]

Patient and public involvement

None.

RESULTS

An overview of the cohorts is summarised in online supplemental figure 1 and the baseline patient characteristics for the training and test (internal and external validation) cohorts are shown in table 1. None of the features differed between the training and test set except for the aetiology of respiratory failure (online supplemental table 3). The 90-day and in-hospital mortality rates were similar between the training and test cohorts (57.2% and 61.9% vs 57.7% and 63.1%, respectively). In the external validation set, the 90-day mortality rate was 48.7%, whereas the in-hospital mortality rate was 51.3%.

When the ML models for 90-day mortality were evaluated using AUROC in the internal validation set, the light gradient boosting (LGB) model scored the highest among the ML models using all features in the testing cohort (AUROC of 0.80 (95% CI 0.71 to 0.88); AUPRC of 0.82 (95% CI 0.71 to 0.91)) (online supplemental table 4 and figure 2). The extreme gradient boosting (XGB) model had the second highest scores, with an AUROC of 0.79 (95% CI 0.69 to 0.87) and AUPRC of 0.82 (95% CI 0.72 to 0.91). All the AUROC values in ML models for 90-day mortality were higher than those obtained from PRESERVE and RESP (online supplemental figure 3). When the outcome was defined as in-hospital mortality in the test set, the best model had an AUROC of 0.83 (95% CI 0.74 to 0.91) and AUPRC of 0.88 (95% CI 0.79 to 0.95) (online supplemental table 5). ML models also demonstrated superior performance to conventional models when predicting in-hospital mortality with all available features in the test set (online supplemental figure 4A).

To develop models that use a smaller set of readily available clinical data, we developed ML models comprising only variables readily obtainable from EHR systems without any preprocessing. For the prediction of 90-day mortality, the XGB model had the highest AUROC (0.82; 95% CI 0.73 to 0.89) and AUPRC (0.87; 95% CI 0.79 to 0.93) followed by the LGB model (AUROC, 0.81; 95% CI 0.71 to 0.88) for the test set (table 2, figure 1). The XGB and LGB model achieved a PPV of 0.77 (95% CI 0.65 to 0.87) and 0.74 (95% CI 0.63 to 0.84), respectively. All ML-based models with EHR features achieved a significantly higher AUROC of 0.82 (95% CI 0.73 to 0.89) compared with that of RESP (0.66; 95% CI 0.56 to 0.76) and PRESERVE (0.71; 95% CI 0.61 to 0.81) (table 2, figure 2). Similarly, for the outcome of in-hospital mortality, the predictive effectiveness of XGB models using EHR features was considerably better than the conventional RESP and PRESERVE models (online supplemental figure 4B).

To identify the degree of contribution of each feature in predicting the risk of 90-day mortality, we also described the SHAP summary plot of the top 20 features of the XGB model (all features vs EHR features, online supplemental figure 5 and figure 3, respectively). The parts are sorted in descending order of Shapley values. Consequently, the features that contributed most to the model performance were age, body surface area, blood pressure, blood gas and ventilator parameters. The calibration plots of the XGB model for the 90-day mortality prediction are shown in figure 4 (all features vs EHR features, online supplemental figure 6 and figure 4, respectively).

Table 1 Baseline characteristics of VV-ECMO treated patients

	Model construction data	Internal validation	External validation data
	Training cohort (n=257)	Testing cohort (n=111)	(n=78)
Age (years)	55.9 (15.7)	53.1 (14.7)	58.5 (13.7)
Sex, male (%)	172 (66.9%)	75 (67.6%)	49 (62.8%)
Height (cm)	165 (8.14)	166 (8.5)	164 (8.8)
Weight (kg)	62.1 (11.7)	63.7 (13.3)	66.2 (16.5)
Body mass index (kg/m^2)	22.9 (3.8)	23.1 (4.0)	24.6 (5.20)
Immunocompromised status	58 (22.6%)	30 (27.0%)	20 (25.6%)
CNS dysfunction	10 (3.9%)	6 (5.4%)	10 (12.8%)
Sodium bicarbonate infusion	25 (9.7%)	10 (9.0%)	17 (21.8%)
Cardiac arrest	30 (11.7%)	14 (12.6%)	4 (5.1%)
Pre SOFA score	10.8 (3.9)	11.3 (4.0)	13.9 (2.9)
NMB agent	136 (52.9%)	47 (42.3%)	75 (96.2%)
Aetiology of respiratory failure			
Viral pneumonia	26 (10.1%)	18 (16.2%)	27 (34.6%)
Bacterial pneumonia	73 (28.4%)	31 (27.9%)	4 (5.1%)
COPD/asthma	3 (1.2%)	2 (1.8%)	1 (1.3%)
Trauma/burn	10 (3.9%)	5 (4.5%)	0 (0%)
Asphyxia	0 (0%)	1 (0.9%)	4 (5.1%)
AE-ILD	37 (14.4%)	11 (9.9%)	0 (0%)
Chronic respiratory failure	15 (5.8%)	3 (2.7%)	0 (0%)
Other respiratory failure	93 (36.2%)	40 (36.0%)	42 (53.8%)
Pre-ECMO ventilator settings			
PEEP (cm H2O)	9.52 (4.1)	9.01 (3.6)	7.44 (2.9)
Peak inspiratory pressure (cm H2O)	28.6 (6.4)	29.0 (6.1)	25.9 (9.0)
PF ratio	74.0 (52.9)	78.5 (38.4)	62.9 (31.9)
Minute ventilation (L/min)	10.4 (4.4)	9.96 (3.9)	9.45 (5.0)
Respiratory rate (/min)	24.1 (7.2)	23.2 (6.8)	24.6 (7.3)
MV time before ECMO			
<48 hours	136 (52.9%)	70 (63.1%)	34 (43.6%)
>7 days	74 (28.8%)	22 (19.8%)	23 (29.5%)
48 hours to 7 days	46 (17.9%)	17 (15.3%)	21 (26.9%)
Pre-ECMO blood gases			
PaCO2 (mm Hg)	56.8 (24.9)	57.3 (25.4)	53.1 (26.1)
PaO2 (mm Hg)	65.4 (34.0)	70.0 (22.0)	96.9 (78.5)
SaO2 (%)	83.3 (13.8)	85.4 (12.9)	89.1 (11.6)
Hemoglobin (g/dL)	10.8 (2.4)	10.9 (2.3)	10.9 (2.3)
TBIL (µmol/L)	1.92 (3.7)	1.98 (2.9)	3.90 (13.2)
Creatinine (mg/dL)	1.30 (1.6)	1.36 (1.5)	1.73 (4.9)
Platelet count	166 (113)	155 (114)	174 (105)
Mortality			
90-day mortality	147 (57.2%)	64 (57.7%)	38 (48.7%)
In-hospital mortality	159 (61.9%)	70 (63.1%)	41 (51.3%)

Data are presented as number (%) or mean (SD), unless otherwise specified.
MV time before ECMO = mechanical Ventilation time before ECMO; PF ratio=PaO2/FiO2 (mm Hg) ratio.
AE-ILD, acute exacerbations of interstitial lung disease; CNS, central nervous system; COPD, chronic obstructive pulmonary disease; MV, mechanical ventilator; NA, not available; NMB, neuromuscular blocking agents; PaCO2, arterial carbon dioxide tension; Pao2, arterial oxygen tension; PEEP, positive end-expiratory pressure; SaO2, arterial oxygen saturation; SOFA, Sequential Organ Failure Assessment; TBIL, total bilirubin; VV-ECMO, venovenous extracorporeal membrane oxygenation.

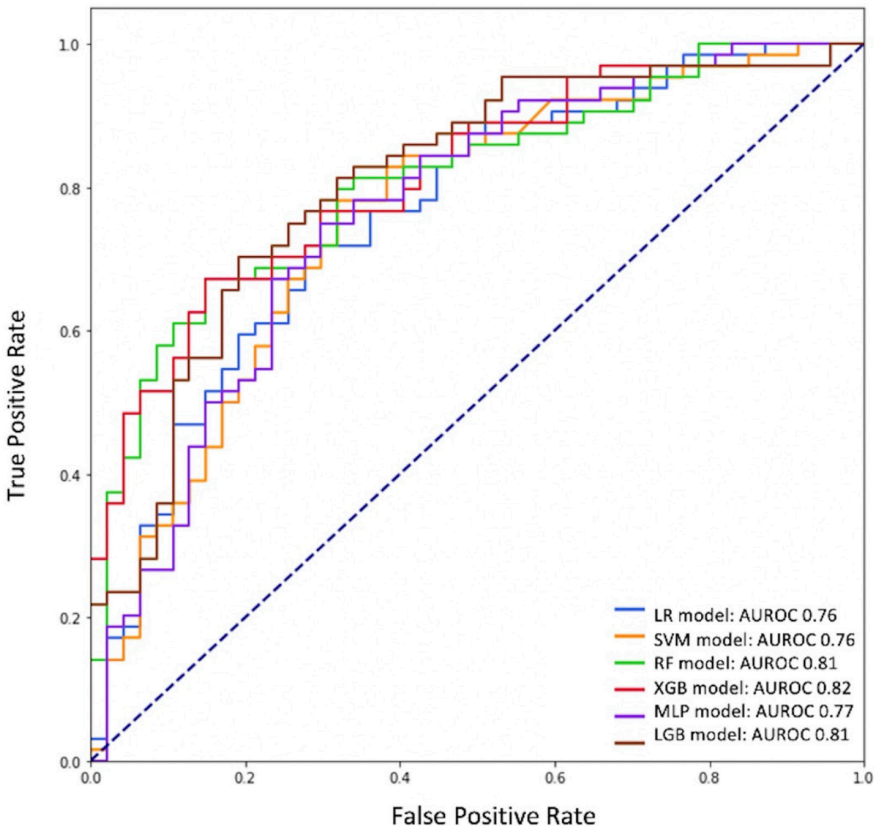

Figure 1 Discrimination performance of prediction models with EHR features for 90-day mortality in the interval validation set. AUROC, area under receiver operating characteristics; EHR, electronic health record; LGB, light gradient boosting; LR, logistic regression; MLP, multilayer perceptron; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting.

Online supplemental figure 7 presents the DCA showing the clinical utility of PRESERVE and RESP, along with ML models using all features and EHR features, to predict 90-day mortality in the test cohort. The results are presented as a plot with the selected risk thresholds (the degree of certainty of mortality per the physicians' decision not to operate) plotted on the x-axis, and the benefits of the prediction model plotted on the

Table 2 Assessment of predictive performance for prediction of 90-day mortality using EHR features in the internal validation set

Models	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F1-score
LR	0.76 (0.67 to 0.85)	0.79 (0.67 to 0.89)	0.72 (0.60 to 0.82)	0.70 (0.57 to 0.83)	0.77 (0.66 to 0.87)	0.65 (0.52 to 0.78)	0.74 (0.64 to 0.82)
SVM	0.76 (0.67 to 0.85)	0.78 (0.65 to 0.89)	0.81 (0.71 to 0.90)	0.62 (0.48 to 0.76)	0.74 (0.63 to 0.84)	0.71 (0.56 to 0.84)	0.78 (0.68 to 0.85)
RF	0.81 (0.72 to 0.88)	0.86 (0.77 to 0.93)	0.83 (0.73 to 0.92)	0.57 (0.44 to 0.72)	0.73 (0.63 to 0.83)	0.71 (0.57 to 0.85)	0.77 (0.69 to 0.85)
XGB	0.82 (0.73 to 0.89)	0.87 (0.79 to 0.93)	0.72 (0.59 to 0.83)	0.70 (0.57 to 0.83)	0.77 (0.65 to 0.87)	0.65 (0.51 to 0.78)	0.74 (0.64 to 0.82)
MLP	0.77 (0.68 to 0.85)	0.77 (0.65 to 0.89)	0.70 (0.58 to 0.81)	0.72 (0.60 to 0.85)	0.78 (0.67 to 0.88)	0.64 (0.52 to 0.76)	0.74 (0.64 to 0.82)
LGB	0.81 (0.71 to 0.88)	0.84 (0.77 to 0.91)	0.86 (0.77 to 0.94)	0.60 (0.46 to 0.74)	0.74 (0.63 to 0.84)	0.76 (0.61 to 0.89)	0.80 (0.71 to 0.86)

All numbers are presented with 95% CI.
AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; LGB, light gradient boosting; LR, logistic regression; MLP, multilayer perceptron; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting.

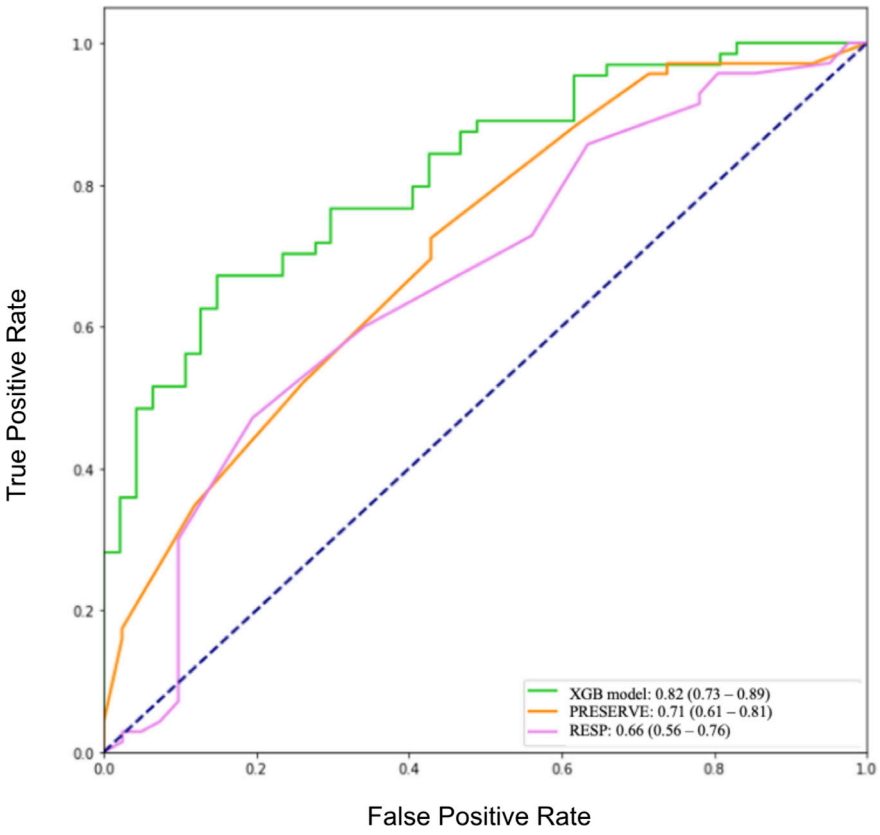

Figure 2 ROC comparing 90-day mortality prediction models using EHR features with the RESP and PRESERVE scores in the internal validation set. ECMO, extracorporeal membrane oxygenation; PRESERVE, predicting death for severe acute respiratory distress syndrome on VV-ECMO; RESP, respiratory ECMO survival prediction; ROC, receiver operating characteristics; XGB, extreme gradient boosting.

y-axis.[15] The benefit of the ML model is greater than that of PRESERVE and RESP, particularly above 50% of the probability threshold.

In the external validation cohort, the predictive ability of the XGB model with EHR features to predict 90-day mortality showed the highest performance with an AUROC of 0.75 (95% CI 0.64 to 0.85) and AUPRC of 0.74 (95% CI 0.58 to 0.86) (table 3). Models based on ML with EHR features also achieved a significantly higher AUROC than those of RESP (0.70; 95% CI 0.58 to 0.82) and PRESERVE (0.67; 95% CI 0.56 to 0.78) (online supplemental figure 8). The XGB model showed an overall good calibration and clinical utility on the external validation dataset, as illustrated in online supplemental figures 9 and 10, respectively.

DISCUSSION

In this multicentre registry study, we developed ML algorithms to predict 90-day mortality in patients undergoing VV-ECMO. The ML-based models, such as XGB and LGB, successfully predicted the risk of 90-day mortality and in-hospital mortality and outperformed conventional risk-scoring models, such as RESP and PRESERVE. The XGB model had the best performance among all models and a higher PPV and AUPRC than conventional scoring methods. This indicated that ML algorithms could accurately identify VV-ECMO candidates with a higher likelihood of death. Moreover, the developed models were validated using an external validation cohort and were further developed using readily available EHR data to implement the models in clinical practice quickly.

Critically ill patients with ARF come in various complex clinical situations, frequently impeding clinical outcome predictions. ML may overcome the difficulty in decision-making during these difficult situations.[29] Kang *et al* proved that ML algorithms increase the accuracy of mortality prediction for patients undergoing continuous renal replacement therapy when compared with those of conventional models such as Acute Physiology and Chronic Health Evaluation or Sequential Organ Failure Assessment.[30] Regarding mortality prediction for patients undergoing ECMO, Ayers *et al* reported the potential for ML models to augment clinical decision-making for patients undergoing venoarterial-ECMO.[18] However, there are no ML-based mortality prediction models for patients undergoing VV-ECMO (online supplemental table 6). To the best of our knowledge, this is the first study to use ML for mortality prediction in patients undergoing VV-ECMO.

As for discriminatory performance, the AUROC of the XGB (0.82) model for the prediction of 90-day mortality was 15.5% and 24.2% higher than that of

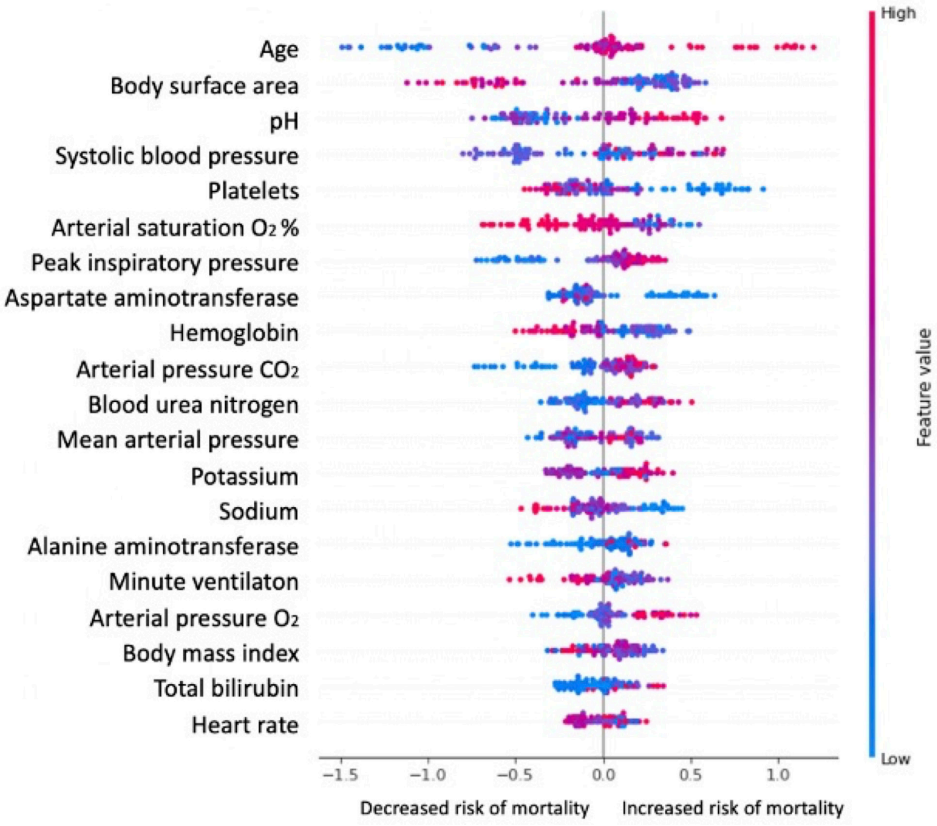

Figure 3 SHAP analysis of 90-day mortality prediction with EHR features in the internal validation set. The colour scheme in the plot uses red to represent higher features values and blue to represent lower feature values. On the x-axis, positive values indicate an increased risk of mortality, while negative values represent a decreased risk of mortality. EHR, electronic health record; SHAP, shapley additive explanations.

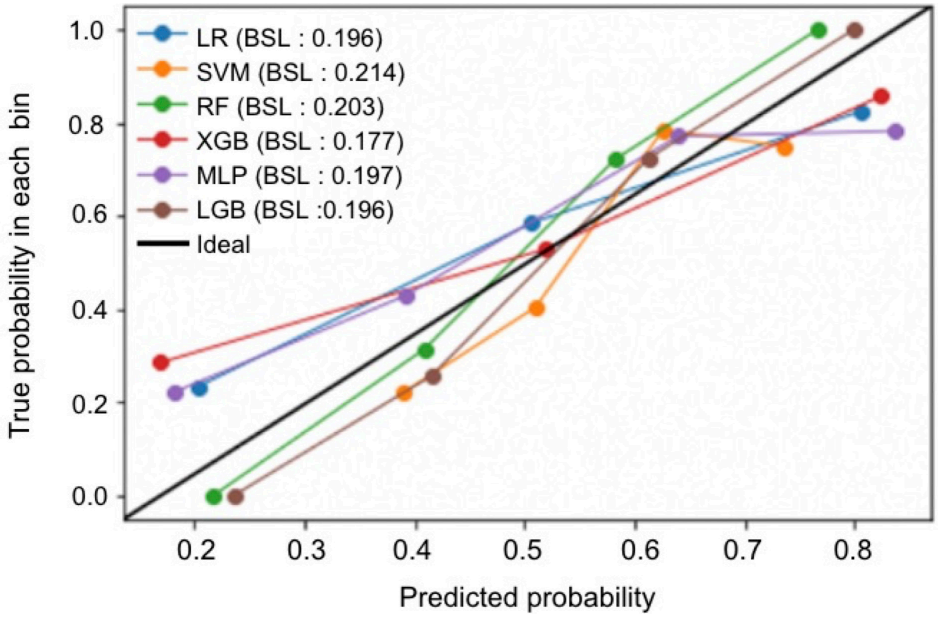

Figure 4 Calibration performance of 90-day mortality prediction models with EHR features in the internal validation set. BSL, Brier Score Loss; EHR, electronic health record; LGB, light gradient boosting; LR, logistic regression; MLP, multilayer perceptron; RF, random forest; SVM, support vector machine.

Table 3 Assessment of predictive performance for prediction of 90-day mortality using EHR features in the external validation set

Models	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F1-score
LR	0.62 (0.49 to 0.74)	0.6 (0.43 to 0.75)	0.71 (0.56 to 0.85)	0.48 (0.33 to 0.63)	0.56 (0.42 to 0.71)	0.63 (0.47 to 0.80)	0.63 (0.50 to 0.74)
SVM	0.64 (0.52 to 0.76)	0.63 (0.46 to 0.78)	0.71 (0.56 to 0.85)	0.43 (0.28 to 0.57)	0.54 (0.39 to 0.67)	0.61 (0.43 to 0.78)	0.61 (0.48 to 0.72)
RF	0.69 (0.58 to 0.80)	0.66 (0.50 to 0.81)	0.84 (0.72 to 0.94)	0.43 (0.28 to 0.56)	0.58 (0.45 to 0.70)	0.74 (0.56 to 0.90)	0.69 (0.57 to 0.78)
XGB	0.75 (0.64 to 0.85)	0.74 (0.58 to 0.86)	0.82 (0.69 to 0.92)	0.6 (0.45 to 0.75)	0.66 (0.52 to 0.79)	0.77 (0.62 to 0.91)	0.73 (0.61 to 0.83)
MLP	0.71 (0.59 to 0.82)	0.70 (0.53 to 0.84)	0.84 (0.72 to 0.95)	0.33 (0.19 to 0.47)	0.54 (0.41 to 0.67)	0.68 (0.56 to 0.75)	0.66 (0.54 to 0.76)
LGB	0.66 (0.54 to 0.78)	0.66 (0.49 to 0.80)	0.89 (0.79 to 0.97)	0.48 (0.33 to 0.63)	0.62 (0.49 to 0.75)	0.83 (0.67 to 0.96)	0.73 (0.62 to 0.82)

All numbers are presented with 95% CI.
AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; LGB, light gradient boosting; LR, logistic regression; MLP, multilayer perceptron; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting.

PRESERVE (0.71) and RESP (0.66), respectively. Similar to our results, discrimination between survivors and non-survivors with PRESERVE scores was only moderate (AUROC of approximately 0.6) for most trials.[7] The RESP score also has moderate discrimination between survivors and non-survivors, although slightly better than the PRESERVE score (AUROC of approximately 0.7–0.75) in other studies.[7] Enger *et al* developed a mortality prediction model for VV-ECMO based on a hospital study of 304 patients with an AUROC of 0.75–0.79, but no external validation has been reported.[11] On the contrary, the AUROC of our XGB model achieved only a 7% decrease in performance when validated in the external validation cohort.

To develop models with readily available clinical data, a more practical model was devised comprising only available features from EHR systems without any preprocessing. ML classifiers with sparse features achieved better performance than the conventional RESP or PRESERVE scoring models and even better performance than models with all features. The findings of our study show the potential of the model to be incorporated into existing EHR systems to serve as a prognostic tool and aid in the decision-making for ECMO initiation in patients with severe respiratory failure.

In response to the most recent data and ECMO trials, indications for the initiation of VV-ECMO are straightforward, and the list of contraindications has decreased considerably.[8] However, several conditions outside the list for contraindications constitute very high-risk patients with a low likelihood of success with ECMO therapy.[8] Thus, each centre and provider involved in identifying the contraindication for ECMO initiation should take them into account in a separate manner. Our ML models could identify high-risk groups unlikely to survive even with ECMO therapy. The PPV of the XGB model was 0.77 for 90-day mortality, where 77% of predicted mortality cases were confirmed at a 0.61 threshold (online supplemental table 7). The high precision of the developed model might help improve clinical judgement for rejecting high-risk ECMO candidates.[31]

Furthermore, the DCA helped clinicians to assess the potential clinical benefits of ECMO therapy and rule out patients with a low likelihood of success in the range of clinical threshold probabilities.[32] If physicians want to sacrifice sensitivity and increase specificity to gain a maximum PPV, they could change the probability threshold from 40% to 70%. The developed model showed better effectiveness than PRESERVE and RESP while maintaining a positive net gain, particularly above 50% of the probability threshold. The ML-based approach could be advantageous in identifying which patients would benefit from ECMO cannulation, particularly during a pandemic when resources become more constrained, calling for more stringent contraindications.

Despite these advantages, our ML-based model has several limitations. First, the prediction model was not trained on different ethnic groups. The study has only been validated with a predominantly Northeast Asian population, which may depreciate the model performance when applied to another ethnicity. Future research should involve different populations to improve and validate the model performance. Second, although our sample size was relatively large compared with previous studies, our cohort size was still insufficient to extrapolate the results to a certain extent. However, the model was developed using multi-institutional registry data from 16 tertiary hospitals, in which patients with different characteristics were included. Additionally, we demonstrated the validity and reliability of predictive mortality algorithms in an external validation cohort to avoid inflated results due to overfitting. Finally, this study did not measure the

developed model's impact on clinical practice enhancement. Further research is needed to evaluate the model's usefulness in clinical environments.

Conclusions

The ML prediction model for 90-day mortality could accurately identify VV-ECMO candidates with a low probability of success. This model could provide valuable prognostic information and help decision-making, particularly with efficiently allocating the very limited number of ECMO machines. A larger dataset would improve the performance and validation of our current models in future studies.

Author affiliations
[1]Department of Digital Healthcare, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[2]Devision of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University College of Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[3]Department of Cardiovascular and Thoracic Surgery, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[4]Department of Pulmonary and Critical Care Medicine, Asan Medical Center, Seoul, Republic of Korea
[5]Department of Family Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

Acknowledgements We thank all the medical staff and ECMO centres participating in the ECMO registry for their contribution: Chi Ryang Chung, Jae-Seung Jung, Jin Young Oh, Jung-Hyun Kim, Jung-Wan Yoo, Sang-Min Lee, Seung Yong Park, So Hee Park, So-My Koo, Sunghoon Park, Woo Hyun Cho, Youjin Chang and Yun Su Sim.

Contributors HEL contributed to data cleansing, data analysis, statistical analysis, and machine learning, and drafting of the manuscript. DJK, YJC and SBH contributed to data collection, data curation, and data interpretation, and verified the integrity of data. All authors had access to the data and reviewed the results of the study. SYJ and SYL conceptualised and oversaw the research and drafting of the manuscript. MJS contributed to the critical review of the final version of the manuscript. All authors read and approved the final manuscript. SYL is responsible for the overall content.

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.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the institutional review board of each participating hospital (online supplemental table 1), including the Seoul National University Bundang Hospital (B-1704-391-109), and was in accordance with the Declaration of Helsinki of 1975. The requirement for informed consent was waived owing to the retrospective nature of the study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Aggregated data available by request.

ORCID iD
Sung Yoon Lim http://orcid.org/0000-0003-3161-8711

REFERENCES

1 Bellani G, Laffey JG, Pham T, *et al*. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA* 2016;315:788–800.
2 Thompson BT, Chambers RC, Liu KD. Acute respiratory distress syndrome. *N Engl J Med* 2017;377:1904–5.
3 Combes A, Hajage D, Capellier G, *et al*. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *N Engl J Med* 2018;378:1965–75.
4 Li X, Hu M, Zheng R, *et al*. Delayed initiation of ECMO is associated with poor outcomes in patients with severe COVID-19: A multicenter retrospective cohort study. *Front Med (Lausanne)* 2021;8:716086.
5 Tonna JE, Abrams D, Brodie D, *et al*. Management of adult patients supported with Venovenous Extracorporeal membrane oxygenation (VV ECMO): guideline from the Extracorporeal life support Organization (ELSO). *ASAIO J* 2021;67:601–10.
6 Badulak J, Antonini MV, Stead CM, *et al*. Extracorporeal membrane oxygenation for COVID-19: updated 2021 guidelines from the Extracorporeal life support organization. *ASAIO J* 2021;67:485–95.
7 Harnisch LO, Moerer O. n.d. Contraindications to the initiation of veno-venous ECMO for severe acute respiratory failure in adults: A systematic review and practical approach based on the current literature. *Membranes*;11:584.
8 MacLaren G. When to initiate ECMO with low likelihood of success. *Crit Care* 2018;22:217.
9 Shaefi S, Brenner SK, Gupta S, *et al*. Extracorporeal membrane oxygenation in patients with severe respiratory failure from COVID-19. *Intensive Care Med* 2021;47:208–21.
10 Tabatabai A, Ghneim MH, Kaczorowski DJ, *et al*. Mortality risk assessment in COVID-19 Venovenous Extracorporeal membrane oxygenation. *Ann Thorac Surg* 2021;112:1983–9.
11 Enger T, Philipp A, Videm V, *et al*. Prediction of mortality in adult patients with severe acute lung failure receiving veno-venous Extracorporeal membrane oxygenation: a prospective observational study. *Crit Care* 2014;18:R67.
12 Schmidt M, Zogheib E, Rozé H, *et al*. The PRESERVE mortality risk score and analysis of long-term outcomes after Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *Intensive Care Med* 2013;39:1704–13.
13 Schmidt M, Bailey M, Sheldrake J, *et al*. Predicting survival after Extracorporeal membrane oxygenation for severe acute respiratory failure. The respiratory Extracorporeal membrane oxygenation survival prediction (RESP) score. *Am J Respir Crit Care Med* 2014;189:1374–82.
14 Hilder M, Herbstreit F, Adamzik M, *et al*. Comparison of mortality prediction models in acute respiratory distress syndrome undergoing Extracorporeal membrane oxygenation and development of a novel prediction score. *Crit Care* 2017;21:301.
15 Nielsen AB, Thorsen-Meyer H-C, Belling K, *et al*. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish national patient Registry and electronic patient records. *Lancet Digit Health* 2019;1:e78–89.
16 Baek MS, Lee S-M, Chung CR, *et al*. Improvement in the survival rates of Extracorporeal membrane oxygenation-supported respiratory failure patients: a multicenter retrospective study in Korean patients. *Crit Care* 2019;23:1.
17 Baek MS, Chung CR, Kim HJ, *et al*. Age is major factor for predicting survival in patients with acute respiratory failure on Extracorporeal membrane oxygenation: a Korean multicenter study. *J Thorac Dis* 2018;10:1406–17.
18 Ayers B, Wood K, Gosev I, *et al*. Predicting survival after Extracorporeal membrane oxygenation by using machine learning. *Ann Thorac Surg* 2020;110:1193–200.
19 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;30.
20 Chen T, Guestrin C. Xgboost: A Scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016
21 Ke G, Meng Q, Finley T, *et al*. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;30.

22 Waskom M, Botvinnik O, O'Kane D, *et al*. Mwaskom/Seaborn: V0. 8.1 (September 2017). Zenodo. 2017.
23 Virtanen P, Gommers R, Oliphant TE, *et al*. Scipy 1.0: fundamental Algorithms for scientific computing in python. *Nat Methods* 2020;17:352:261–72.:.
24 VanG. The python library reference, release 3.8. 2. *Python Software Foundation* 2020;16.
25 Hunter JD. Matplotlib: A 2d Graphics environment. *Comput Sci Eng* 2007;9:90–5.
26 Reback J, McKinney W, Van Den Bossche J, *et al*. Pandas-Dev/Pandas: Pandas 1.0. 5. Zenodo. 2020.
27 Harris CR, Millman KJ, van der Walt SJ, *et al*. Array programming with Numpy. *Nature* 2020;585:357–62.
28 Team RC. R: A language and environment for statistical computing. 2013.
29 Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. *Crit Care* 2022;26:75.
30 Kang MW, Kim J, Kim DK, *et al*. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. *Crit Care* 2020;24:42.
31 Ozenne B, Subtil F, Maucort-Boulch D. The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol* 2015;68:855–9.
32 Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA* 2015;313:409–10.