JAMIA Open, 2025, **8(2)**, ooaf026
https://doi.org/10.1093/jamiaopen/ooaf026
Research and Applications

Research and Applications

A deep learning model for clinical outcome prediction using longitudinal inpatient electronic health records

Ruichen Rong, PhD[1], Zifan Gu, MS[1], Hongyin Lai, MS[1,2], Tanna L. Nelson, RN, PhD[3], Tony Keller, BA[3], Clark Walker, MPH[3], Kevin W. Jin, BS[1,4,5], Catherine Chen, MD, MSHI[6], Ann Marie Navar, MD, PhD[6], Ferdinand Velasco, MD[3], Eric D. Peterson, MD, MPH[6], Guanghua Xiao, PhD[1,7,8], Donghan M. Yang, PhD[1,*], Yang Xie, PhD[1,7,8,*]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX 75390, United States, [2]Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, United States, [3]Texas Health Resources, Arlington, TX 76011, United States, [4]Program in Computational Biology and Biomedical Informatics, Yale University, New Haven, CT 06511, United States, [5]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT 06511, United States, [6]Department of Internal Medicine, The University of Texas Southwestern Medical Center, Dallas, TX 75390, United States, [7]Department of Bioinformatics, The University of Texas Southwestern Medical Center, Dallas, TX 75390, United States, [8]Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, Dallas, TX 75390, United States

Ruichen Rong and Zifan Gu contributed equally to this work.

*Corresponding authors: Donghan M. Yang, PhD, Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Danciger Research Building, 5323 Harry Hines Blvd. Ste H9.124, Dallas, TX 75390-8821, United States (donghan.yang@utsouthwestern.edu) and Yang Xie, PhD, Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Danciger Research Building, 5323 Harry Hines Blvd. Ste H9.124, Dallas, TX 75390-8821, United States (yang.xie@utsouthwestern.edu)

Abstract

Objectives: Recent advances in deep learning show significant potential in analyzing continuous monitoring electronic health records (EHR) data for clinical outcome prediction. We aim to develop a Transformer-based, Encounter-level Clinical Outcome (TECO) model to predict mortality in the intensive care unit (ICU) using inpatient EHR data.

Materials and Methods: The TECO model was developed using multiple baseline and time-dependent clinical variables from 2579 hospitalized COVID-19 patients to predict ICU mortality and was validated externally in an acute respiratory distress syndrome cohort ($n = 2799$) and a sepsis cohort ($n = 6622$) from the Medical Information Mart for Intensive Care IV (MIMIC-IV). Model performance was evaluated based on the area under the receiver operating characteristic (AUC) and compared with Epic Deterioration Index (EDI), random forest (RF), and extreme gradient boosting (XGBoost).

Results: In the COVID-19 development dataset, TECO achieved higher AUC (0.89-0.97) across various time intervals compared to EDI (0.86-0.95), RF (0.87-0.96), and XGBoost (0.88-0.96). In the 2 MIMIC testing datasets (EDI not available), TECO yielded higher AUC (0.65-0.77) than RF (0.59-0.75) and XGBoost (0.59-0.74). In addition, TECO was able to identify clinically interpretable features that were correlated with the outcome.

Discussion: The TECO model outperformed proprietary metrics and conventional machine learning models in predicting ICU mortality among patients with COVID-19, widespread inflammation, respiratory illness, and other organ failures.

Conclusion: The TECO model demonstrates a strong capability for predicting ICU mortality using continuous monitoring data. While further validation is needed, TECO has the potential to serve as a powerful early warning tool across various diseases in inpatient settings.

Lay Summary

In intensive care units (ICUs), accurately estimating the risk of death is crucial for timely and effective medical intervention. This study developed a new AI algorithm, TECO (Transformer-based, Encounter-level Clinical Outcome model), which uses electronic health records to continuously predict ICU mortality after admission, with the capability to update predictions on an hourly basis. The TECO model was trained on data from over 2500 COVID-19 patients and was designed to analyze multiple types of continuous monitoring data collected during a patient's ICU stay. We tested TECO's performance against a widely used proprietary tool, the Epic Deterioration Index, and other machine learning methods, such as random forest and extreme gradient boosting, across three patient groups: COVID-19, acute respiratory distress syndrome, and sepsis. The TECO model consistently showed better performance and was able to predict death risk earlier than other methods. Additionally, TECO identified key health indicators associated with ICU mortality, making its predictions more interpretable for clinicians. These findings suggest that TECO could become a valuable early warning tool, helping doctors monitor patients' health and take timely action in a range of critical care situations.

Key words: EHR; continuous monitoring data; transformer; ICU; COVID-19.

Received: August 14, 2024; **Revised:** March 7, 2025; **Editorial Decision:** March 17, 2025; **Accepted:** March 20, 2025

Background and significance

Modern medical and information technologies increasingly produce massive amounts of electronic health record (EHR) data. However, there still exists a gap between the rapid digitization of healthcare and the development of analytic tools capable of informing and guiding real-world clinical practices.[1,2] Intensive care unit (ICU) is one area ripe for improved predictive analytics.[3] The COVID-19 pandemic posed unprecedented challenges to the provision of ICU care due to a lack of bed capacity, clinical care staff, and necessary analytics for resource allocation.[4] While the contemporary ICU medical devices and systems collect vast amounts of time-stamped data (eg, vital signs, lab tests, and medication administrations), these rich continuous data streams are often not used to their full extent to develop analytics tools that assist clinicians in predicting clinical outcomes.

Conventional statistical models are limited in processing multivariate, time-dependent datasets and analyzing relations between different variables and different timestamps. Commercial analytics tools often lack technical transparency and interoperability across EHR platforms. For instance, the Epic Deterioration Index (EDI),[5,6] a proprietary metric exclusive to Epic, quantifies patient deterioration in real time. The detailed design and parameters of EDI are not published, and it has not been validated in routine clinical practice.[7] Moreover, the threshold of EDI that calls for intervention has been differentially implemented among healthcare systems, rendering its actual usage dependent on local standards.[7]

Machine learning models have been proposed to predict clinical outcomes or procedures in critically ill patients. For example, researchers employed a random forest (RF) classifier to predict COVID-19 disease severity at hospital admission,[8] applied an extreme gradient boosting (XGBoost)-based algorithm to predict invasive mechanical ventilation,[9] and implemented PICTURE (Predicting Intensive Care Transfers and Other Unforeseen Events) to predict deterioration.[10] However, these algorithms face limitations when dealing with long series of time-dependent data with high dimensionality and irregular time intervals, as commonly encountered in inpatient monitoring data.

Recent advancements in deep learning have demonstrated preliminary success in managing multi-dimensional sequential EHR data. Among these, transformer models have shown promise by leveraging attention-based mechanisms to better capture complex sequential patterns and interrelated features.[11] For example, Wu et al[12] developed a transformer model to forecast influenza prevalence from public health data, while ClinicalBERT predicts 30-day hospital readmissions from clinical notes.[13] Models such as BEHRT[14] and Med-BERT,[15] pre-trained on sequences of diagnosis codes, have been used to predict future diagnoses, with Antikainen et al[16] expanding this approach to incorporate different types of medical events for long-term mortality prediction in cardiovascular patients.

In an ICU setting, MeTra integrates chest radiographs with clinical data for mortality prediction, though it restricts data input to the first 48 hours and may pose a high computational burden due to the Vision Transformer component.[17] Similarly, Cheng et al[18] used transformers for image data to predict COVID-19 mortality, but did not apply transformers to the clinical data component of their model. Song et al[19] proposed the SAnD architecture for ICU tasks, including mortality prediction; however, they restricted data input to the final 24 hours of the ICU stay, requiring prior knowledge of the outcome time, which limits the model's applicability to real-world scenarios. Furthermore, none of these ICU mortality prediction models have been benchmarked against the widely used commercial tool EDI, which limits the assessment of their clinical utility and relevance as a trans-platform tool.

Objective

In this study, we propose the Transformer-based Encounter-level Clinical Outcome (TECO) model, which fully utilizes continuous ICU monitoring data alongside patient-level baseline characteristics for mortality prediction, with the capability to update predictions on an hourly basis. We developed TECO using EHR data from a cohort of COVID-19 patients and validated the model on 2 external non-COVID-19 cohorts. We benchmarked TECO against the EDI, RF, and XGBoost to evaluate its performance in an ICU setting across different disease profiles.

Materials and methods

Study setting and design

In this study, the model development dataset contained EHR data from Texas Health Resources (THR), a large faith-based, nonprofit health system in North Texas, operating 20 acute care hospitals and serving 7 million residents in 16 counties. The THR cohort included 2579 adult patients with laboratory-confirmed COVID-19 (age $\geq$ 18), who were admitted into ICUs during their first COVID-19 hospitalization. Dates of patient hospitalization ranged from March 3, 2020, to August 13, 2021. Patients who had more than one ICU admission during the hospitalization were excluded.

The external validation dataset was extracted from the Medical Information Mart for Intensive Care IV (MIMIC-IV), a large, publicly available, de-identified clinical database for critically ill patients admitted to the emergency department of the Beth Israel Deaconess Medical Center in Boston, MA from 2008 to 2019.[20] We identified 2 patient cohorts admitted to ICU, 1 diagnosed with acute respiratory distress syndrome (ARDS) and the other with sepsis (Supplementary Materials). If a patient had multiple ICU or hospital admissions, only the first ICU visit of the first hospital admission was included. Acute respiratory distress syndrome was defined in accordance with the Berlin definition[21] with the MIMIC-specific positive identification method.[22] Sepsis was defined in accordance with the Third International Consensus Definitions for Sepsis and Septic Shock,[23] quantified by the Sequential Organ Failure Assessment (SOFA) score.[24,25]

In this study, the models were designed to predict a binary outcome: death versus non-death in the ICU. In this context, the non-death outcome refers to conditions not at immediate risk of death and includes a range of states, from requiring continued ICU care to being ready for ICU discharge. We included 2 types of variables in the model: baseline variables and time-dependent variables. Baseline variables include age at hospital admission, sex, ethnicity, and race. Time-dependent variables include ICU monitoring measures, each recorded at a different and irregular pace: body temperature, respiration rate, pulse oximetry (SpO_2), mSOFA (modified Sequential Organ Failure Assessment) overall score,[26] mSOFA respiratory sub-score, and SF ratio (SpO_2/FiO_2, where FiO_2 is the fraction of inspired oxygen). Body mass index (BMI) was included as a time-dependent variable in the COVID-19 dataset. In the MIMIC-IV dataset, where time-dependent BMI was unavailable, the BMI recorded at hospital admission was used as a baseline variable.

The institutional review boards at THR and UT Southwestern Medical Center approved this study (Protocol #STU-2020-0786; activated on August 24, 2020). All patient identifiers were removed before EHR data extraction.

Development of TECO model

The TECO model employs a Transformer-encoder architecture (Figure 1). First, we aligned the time-dependent variables by taking the mean of each variable in every 15-minute interval. If a 15-minute mean value was missing, the value from the previous interval was carried forward. Then, we concatenated the aligned time-dependent variables with baseline variables, creating a feature set for each 15-minute interval, and embedded these features into a 512-dimension vector. The baseline variable input values remained static across the entire time range. The timestamps of these 15-minute intervals were encoded into another representative vector by positional encoding. We applied relative positional encoding to avoid imputing large amounts of default values for time points without measurement. The variable value vector and timestamp vector were fed into a feed-forward network with 6 Transformer-encoder layers. Each layer was equipped with 8 multi-head attention modules. The model outputs the probability of death using a linear classification layer with the Softmax activation function.

We developed the TECO model using the THR COVID-19 dataset. We created 20 different data splits, where in each split the COVID-19 dataset was randomly split into a training set (80%) and a validation set (20%). During the training phase, we included only the data ranges that led to the eventual outcome at the ICU endpoint. This setup ensures a clear representation of both death and survival cases while maintaining a balanced distribution between the two outcomes (36% death, Table 1). In this particular context, the non-death outcome corresponds to ICU discharge. In total, we trained 9 TECO submodels, each designed to predict the outcome at a specific future time point (0, 12, 24, 36, 48, 60, 72, 84, and 96 hours) using data from a specific time interval (Table S1). The algorithm does not require data to be fully available throughout the entire time interval. To select hyperparameters, we performed a grid search on the hyperparameters on one training-validation split. The hyperparameter set yielding the highest area under the receiver operating characteristic (ROC) curve (AUC) from that split was selected, resulting in a model with 6 encoder layers, 8 attention heads, an embedding dimension of 512, and a total of 18 928 130 trainable parameters (Table S2).

The Transformer model for developing TECO was implemented in PyTorch (Version 1.8.1)[27] and trained on an NVIDIA Tesla V100 Tensor Core GPU with 32 GB of memory. The TECO model was allowed to train for a maximum of 500 epochs with a batch size of 32. The SGD optimizer with momentum (0.9) was used to update model parameters. The learning rate was set to 0.01 and reduced by a factor of 2 every 50 epochs. We set the dropout rate to 0 and use the Gaussian Error Linear Unit as the activation function in the transformer layers. The training process would stop early if the validation loss did not change by more than 10^{-4} after

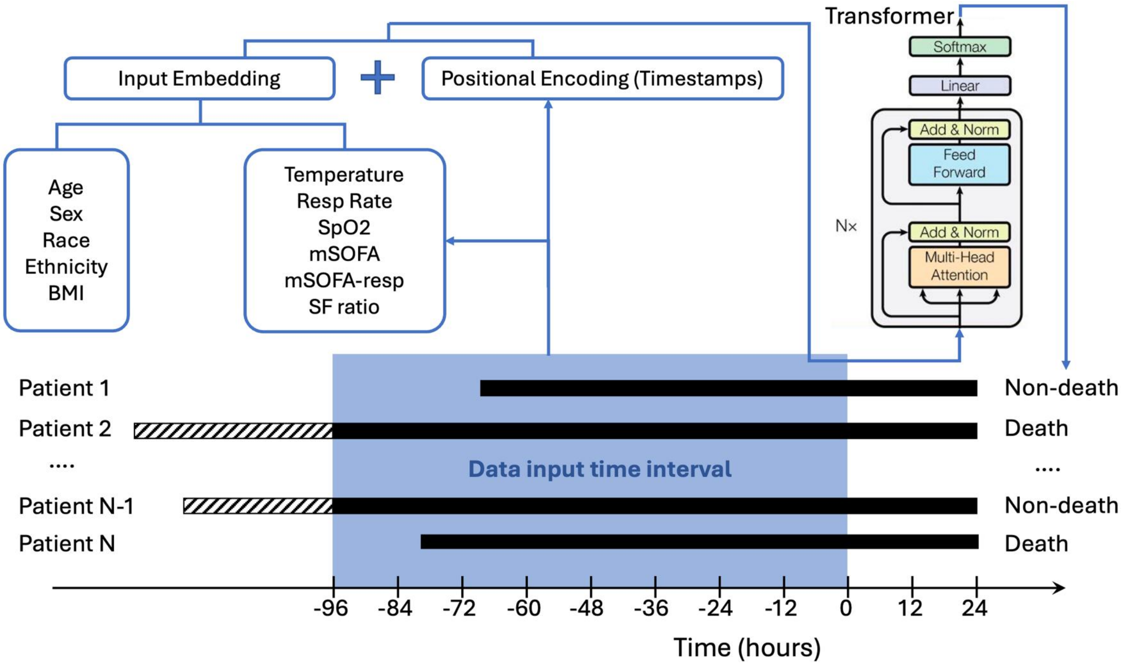

Figure 1. The TECO algorithm design. This figure demonstrates a 24-hour mortality prediction example, utilizing data from the preceding 96 hours to predict the binary outcome (death vs non-death). Time-dependent, ICU monitoring variables were aligned and concatenated with baseline variables, then embedded into 512-dimensional feature vectors. These feature vectors were combined with positional timestamp vectors and fed into a multi-layer Transformer-encoder. Abbreviations: BMI, body mass index; FiO_2, fraction of inspired oxygen; ICU, intensive care unit; mSOFA, modified Sequential Organ Failure Assessment score; mSOFA-resp, mSOFA respiratory sub-score; Resp Rate, respiratory rate; SpO_2, pulse oximetry; SF ratio, SpO_2/FiO_2 ratio; TECO, Transformer-based, Encounter-level Clinical Outcome model. This figure includes a modified version of the Transformer architecture diagram, adapted with the authors' prior permission.[11]

Table 1. Baseline characteristics in the COVID-19, ARDS, and sepsis cohorts.

	COVID-19	MIMIC-ARDS	MIMIC-Sepsis
Number of patients/encounters, n	2579	2799	6622
Age (years)			
Median (Q1, Q3)	63.0 (51.0, 74.0)	66.0 (54.0, 76.0)	66.0 (54.0, 77.0)
Sex, n (%)			
Male	1499 (58.1)	1595 (57.0)	3782 (57.1)
Female	1080 (41.9)	1204 (43.0)	2840 (42.9)
Race, n (%)			
White	1909 (74.0)	1749 (62.5)	4155 (62.7)
Black	401 (15.5)	217 (7.8)	530 (8.0)
Other	125 (4.8)	168 (6.0)	470 (7.1)
Unknown	144 (5.6)	665 (23.8)	1467 (22.2)
Ethnicity, n (%)			
Hispanic	713 (27.6)	–	–
Non-Hispanic	1723 (66.8)	–	–
Unknown	143 (5.5)	–	–
BMI, n (%)			
Underweight	53 (2.1)	53 (1.9)	160 (2.4)
Normal	379 (14.7)	535 (19.1)	1451 (21.9)
Overweight	698 (27.1)	652 (23.3)	1673 (25.3)
Obese	959 (37.2)	792 (28.3)	1540 (23.3)
Unknown	490 (19.0)	767 (27.4)	1798 (27.2)
Outcome, n (%)			
Death	925 (35.9)	471 (16.8)	1031 (15.6)
Discharge	1654 (64.1)	2328 (83.2)	5591 (84.4)

Abbreviations: ARDS, acute respiratory distress syndrome; BMI, body mass index; MIMIC, Medical Information Mart for Intensive Care.

100 epochs. Gradient clipping was set to 1.0 to avoid gradient exploding (Table S2).

Development of other models for comparison

To develop RF and XGBoost models using the same THR COVID-19 dataset, we followed the same data preparation procedure as described above for TECO. The time-dependent variables were aligned and averaged across intervals and concatenated with the baseline variables. The models were trained using the same 20 data splits as for TECO. Hyperparameters were selected through a grid search using the same one training-validation split as for TECO (Table S2). Random forest and XGBoost were implemented in scikit-learn (Version 1.0.2).[28]

To use EDI for outcome prediction, we used the mean EDI value from the preceding 24 hours of each data input time interval. The EDI models predict the binary outcome solely based on a threshold on the continuous EDI values so no hyperparameter tuning was involved.

Internal and external validations

We evaluated the prediction performance of the TECO model based on AUC and compared it with the EDI, RF, and XGBoost models. Performance was evaluated separately for each of the 9 TECO sub-models. For the internal validation based on the THR COVID-19 dataset, we reported the median AUCs on the validation sets across all 20 training-validation splits. To plot ROC for the EDI-based models, we used all possible thresholds within the range of the EDI data.

For external validation based on the MIMIC-IV dataset, we reported the AUCs and real-time probability of death on the ARDS and sepsis cohorts for each involved model, respectively. The EDI was evaluated only in the internal validation due to unavailability of EDI data in the non-Epic-based MIMIC-IV. To validate TECO externally in a manner more akin to a clinical setting, we positioned each model at varying time points after ICU admission and utilized a rolling window of the most recent available data to predict patient outcomes at future time points, as defined by each TECO sub-model's task. For example, the 24-hour sub-model, which utilizes data from the preceding 96 hours (Figure 1), was employed to predict mortality at 120, 132, 144, and up to 240 hours after ICU admission (Figure 2).

Feature importance and ablation study

We analyzed feature importance for TECO and RF in the ARDS cohort. To determine feature importance for TECO, we computed Shapley Additive exPlanations[29] (SHAP) by estimating expected gradients and reformulating them to approximate SHAP values, using 20 pairs of background and batch tensor samples and averaging SHAP values across patient samples and timestamps. For RF, we employed permutation tests, first computing a baseline performance metric with original input features, and then permuting each feature for a recalculated metric. This was repeated 20 times per feature to generate a distribution of raw permutation importance scores.

To assess the impact of baseline variables on model performance, we conducted an ablation study by comparing the AUCs of the full models to those of models using only time-dependent variables.

Results

A total of 2579 patients were included in the THR model development cohort. All enrolled patients had COVID-19, and of these, 925 (35.9%) expired in the ICU. The characteristics of the baseline variables are presented in Table 1. Among these patients, the median age was 63.0 years, and a majority were male (1499, 58.1%), white (1909, 74.0%),

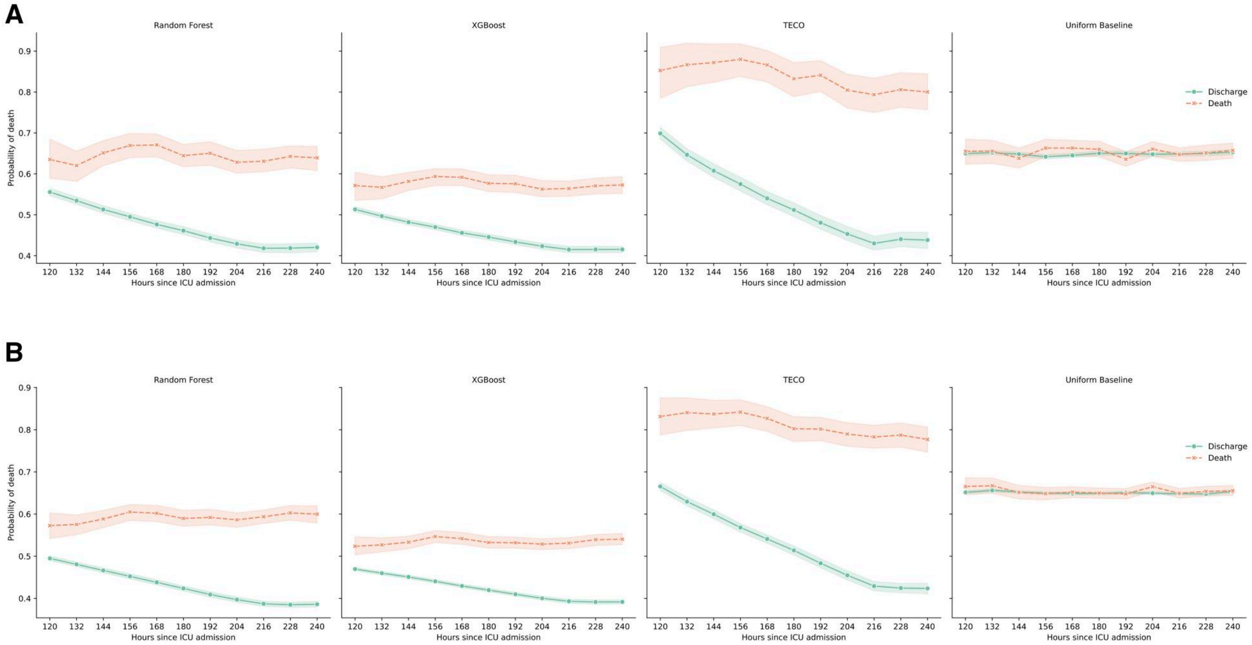

Figure 2. Model-derived mortality probability in (A) ARDS and (B) sepsis cohorts. The green line represents patients who were eventually discharged alive from ICU, while the orange line represents patients who died in ICU. Probability of mortalities is aggregated over repeated hours since admission to show the mean and 95% confidence interval. Models presented from left to right: random forest, XGBoost, TECO, and Uniform Baseline. The uniform baseline model assigns probabilities to patients by sampling from a uniform distribution between the minimum and maximum Softmax probabilities of the 3 models. Abbreviations: ARDS, acute respiratory distress syndrome; ICU, intensive care unit; TECO, Transformer-based, Encounter-level Clinical Outcome; XGBoost, extreme gradient boosting.

non-Hispanic (1723, 66.8%), and overweight to obese (BMI ≥ 25) (1657, 64.3%). The MIMIC ARDS validation cohort included 2799 patients, of whom 471 (16.8%) expired in the ICU. The MIMIC sepsis validation cohort included 6622 patients, of whom 1031 (15.6%) expired in the ICU. These 2 external validation cohorts presented similar trends in the distribution of baseline variables, with the majority being elderly, male, and white (Table 1).

Internal validation

In the COVID-19 model development cohort, AUCs on the validation sets across the 20 data splits are summarized in Figure S1 and Table S1. In general, all models' performance improved as the targeted prediction time window was shortened, with the median AUCs ranging from 0.86 to 0.97. The median AUC of the TECO model, ranging from 0.89 to 0.97, was higher than that of EDI (0.86-0.95), RF (0.87-0.96), and XGBoost (0.88-0.96) at most prediction time windows, demonstrating its overall advantages. On the other hand, the median AUC of EDI-based prediction was consistently lower than that of the other models at every prediction time window. It is noteworthy that the median AUC achieved by TECO when predicting 60-hour mortality (0.93) matched with that based on EDI when predicting 12-hour mortality (Table S1), showcasing the advantage in early warning capability for TECO.

External validation

In the 2 external validation cohorts, similar trends of AUC were observed across different time intervals for each model (Table 2; Figure S2). Using the sub-model for prediction of 24-hour mortality as a demonstration, for the ARDS cohort, AUCs for all 3 models improved from the earliest time point of prediction (120 hours since ICU admission), where the AUCs were 0.66 for TECO, 0.61 for RF, and 0.61 for XGBoost, to the latest time point (240 hours), where the AUCs were 0.76 (TECO), 0.74 (RF), and 0.74 (XGBoost), respectively. Similarly, for the sepsis cohort, AUCs of the same 24-hour sub-model improved from 0.65 (TECO), 0.59 (RF), and 0.59 (XGBoost) at the earliest prediction time point to 0.75 (TECO), 0.75 (RF), and 0.74 (XGBoost) at the latest time point (Table 2). Based on AUC, the TECO model consistently outperformed RF and XGBoost throughout most of the 5-day monitoring period in both cohorts. This advantage of TECO was particularly apparent at earlier lookout time points (eg, 120 through 216 hours since ICU admission).

The performance of all models in the external validation was generally lower than that in the internal validation. The TECO's performance in the ARDS cohort was slightly better than that in the sepsis cohort, especially at earlier lookout time points (Figure S2).

Monitoring of patient deterioration

The TECO model can be used to provide real-time estimation of mortality probability during the ICU stay. An illustration of this feature is shown in Figure 2, where results from external validations of the 24-hour sub-model are presented. The TECO model demonstrated that, at the cohort level, patients who ultimately survived their ICU stay exhibited a consistently lower probability of mortality throughout the 5-day monitoring period compared to those who eventually died in the ICU. In contrast, RF and XGBoost showed weaker separation, while uniform baseline model showed no separation (Figure 2). Moreover, TECO displayed a decreasing trajectory in mortality probability for the surviving patients, a trend consistently observed across 2 external validation cohorts. Similar trajectories were observed with all 9 TECO sub-models (Figure S3).

Table 2. Model performance on the external validation cohorts to predict 24-hour mortality.

Cohort	Hours since ICU admission	TECO	RF	XGBoost
ARDS	120	0.66 [0.60-0.71]	0.61 [0.54-0.67]	0.61 [0.54-0.68]
	132	**0.70 [0.65-0.74]**	0.61 [0.56-0.66]***	0.62 [0.57-0.67]***
	144	**0.73 [0.69-0.76]**	0.67 [0.63-0.71]*	0.67 [0.63-0.72]*
	156	**0.75 [0.72-0.78]**	0.71 [0.67-0.75]*	0.71 [0.67-0.75]*
	168	**0.76 [0.72-0.79]**	0.73 [0.69-0.77]	0.73 [0.69-0.76]*
	180	**0.76 [0.72-0.78]**	0.72 [0.68-0.75]*	0.71 [0.68-0.75]*
	192	**0.77 [0.74-0.80]**	0.73 [0.70-0.77]*	0.73 [0.69-0.76]***
	204	**0.76 [0.73-0.79]**	0.72 [0.69-0.76]*	0.72 [0.68-0.75]***
	216	**0.76 [0.74-0.79]**	0.73 [0.70-0.76]*	0.73 [0.69-0.76]*
	228	**0.76 [0.73-0.79]**	0.75 [0.71-0.77]	0.74 [0.71-0.77]*
	240	0.76 [0.72-0.79]	0.74 [0.71-0.78]	0.74 [0.71-0.78]
Sepsis	120	**0.65 [0.62-0.70]**	0.59 [0.55-0.63]***	0.59 [0.55-0.63]***
	132	**0.68 [0.65-0.71]**	0.62 [0.58-0.65]***	0.61 [0.58-0.65]***
	144	**0.71 [0.68-0.73]**	0.65 [0.62-0.68]***	0.64 [0.61-0.67]***
	156	**0.72 [0.70-0.75]**	0.69 [0.66-0.71]***	0.68 [0.66-0.70]***
	168	**0.73 [0.71-0.75]**	0.70 [0.67-0.72]*	0.69 [0.66-0.72]***
	180	**0.73 [0.70-0.75]**	0.70 [0.67-0.72]*	0.69 [0.66-0.71]***
	192	**0.74 [0.72-0.76]**	0.71 [0.69-0.74]*	0.70 [0.68-0.72]***
	204	**0.75 [0.73-0.77]**	0.72 [0.70-0.74]*	0.71 [0.68-0.73]***
	216	**0.76 [0.74-0.78]**	0.74 [0.72-0.76]*	0.72 [0.70-0.74]***
	228	**0.76 [0.74-0.78]**	0.75 [0.73-0.77]	0.73 [0.71-0.75]*
	240	0.75 [0.73-0.78]	0.75 [0.72-0.77]	0.74 [0.71-0.76]

At each time point after ICU admission, the models use the most recent 96 hours of data to predict mortality in the next 24 hours. For example, at 156 hours after ICU admission, the models use data from 60 to 156 hours to predict outcomes at 180 hours after admission. Model performances are presented as AUC [95% CI]. Confidence intervals (CIs) are estimated by bootstrapping with 500 iterations, sampling the whole dataset with replacement. Statistical significance was assessed using DeLong's test to compare RF or XGBoost with TECO. The top-performing results are shown in bold.

* $P < .05$,

*** $P < .001$.

Abbreviations: ARDS, acute respiratory distress syndrome; AUC, area under the receiver operating characteristic curve; RF, random forest; TECO, Transformer-based Encounter-level Clinical Outcome; XGBoost, extreme gradient boosting.

For an illustration of TECO-generated deterioration monitoring at the individual patient level, the mortality probability projection of 8 representative patients from the ARDS and sepsis cohorts is shown in Figure S4.

Feature importance and ablation study

Among all features, mSOFA appeared to be most important for TECO, especially at later time points (Figure 3). Body mass index, SF ratio, and temperature also showed considerable importance. Similarly, for RF, mSOFA, and its respiratory sub-score were among the most important features. Overall, RF's feature importance appeared more evenly distributed, whereas TECO SHAP values exhibited more distinct importance patterns.

Removing the contribution of baseline variables led to a performance decrease across all models (TECO, RF, and XGBoost) in the 2 external validation cohorts (Table S3). This performance drop was consistent across models and at different time points when the predictions were made. Notably, the contribution of these baseline variables appears to be independent of the timing of prediction. Importantly, even without baseline variables, TECO remained the top-performing model, particularly at earlier prediction time points, underscoring its intrinsic capability to effectively handle dynamic, time-dependent data.

Discussion

In this study, we developed and validated a novel deep learning algorithm, TECO, for mortality prediction in the ICU. Some existing methods for ICU mortality prediction also utilize transformer architecture and continuous monitoring data, particularly those available from the MIMIC-III and -IV databases.[17,19] In contrast to these approaches, TECO is a lightweight transformer model specifically tailored to handle time-dependent, irregularly recorded features and time-independent baseline features in a joint manner. Unlike MeTra or Song et al. models, TECO does not presume a fixed time range of either the input data or the outcome.[17,19] Instead, it can leverage the most recent ICU data to make predictions at future time points.

Our work benchmarks TECO against EDI, a proprietary metric exclusive to Epic platforms. In contrast, TECO is not restricted to a single EHR system provider. The EDI was developed prior to the COVID-19 pandemic but was widely used for clinical decision support and ICU triage during the pandemic.[6,7] We demonstrate that the EDI had relatively lower predictive performance beyond the 24-hour window prior to the outcome. In contrast, all 3 non-proprietary models—RF, XGBoost, and TECO—showed advantages especially at time points further from the ICU outcome. According to limited public information, the EDI model does not appear to use SF ratio in its development.[5,6] Our feature importance analysis demonstrates that SF ratio could be of high importance, which may explain the limited performance of EDI. Besides SF ratio, mSOFA also had a high impact on the RF and TECO. This is consistent with the findings that SOFA is a reliable indicator for mortality among COVID-19 patients,[30–32] and that mSOFA has an equivalent performance in mortality prediction.[26]

We selected ARDS and sepsis for external validation for several reasons. First, both conditions are leading causes of ICU mortality, with severe sepsis and septic shock having a conservative estimated mortality rate of 30%-50%.[33–37]

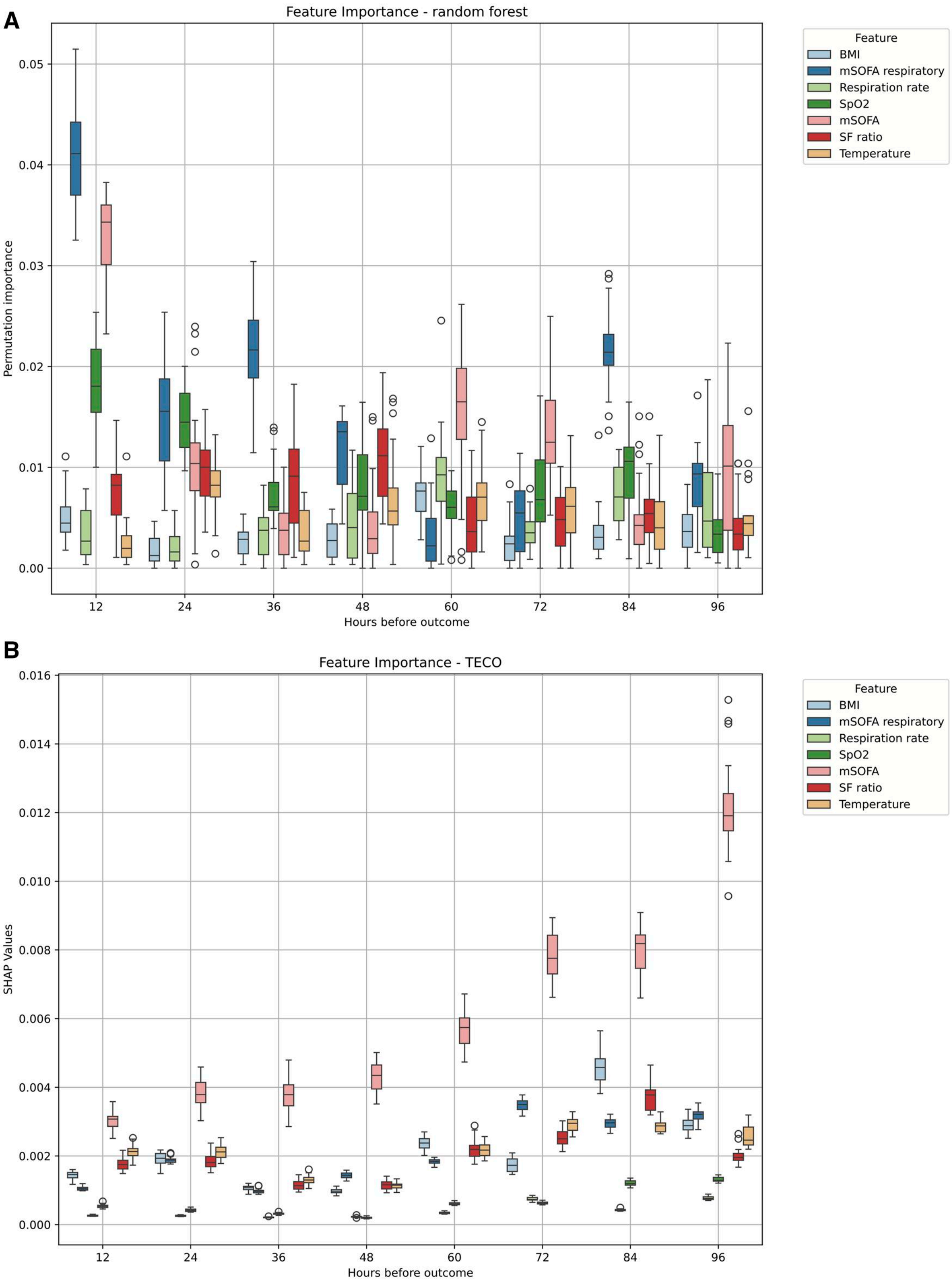

Figure 3. Feature importance analysis in the ARDS cohort using (A) random forest and (B) TECO. Random forest feature importance was assessed using permutation importance, with each feature permuted 20 times. Transformer-based, Encounter-level Clinical Outcome feature importance was estimated via expected gradients and reformulated to approximate SHAP values. Shapley Additive exPlanations values were computed using 20 different background—batch tensor pairs and averaged across patient samples and timestamps. A zoomed-in version of TECO importance is available in Figure S6. Abbreviation: ARDS, acute respiratory distress syndrome.

Additionally, multi-organ dysfunction syndrome, including respiratory failure and ARDS, accounts for a significant proportion of ICU deaths.[38] Second, while ARDS primarily involves acute lung injury and respiratory failure,[39,40] sepsis triggers a systemic inflammatory response leading to multi-organ dysfunctions.[41–43] By including both conditions, we aim to validate TECO across a broader yet connected spectrum of critical illness. Third, both conditions are well-defined using clinical and laboratory criteria (eg, Berlin criteria for ARDS,[21] Sepsis-3 definition for sepsis[23]) and can be reliably retrieved from MIMIC-IV.[20,22] This standardized cohort formation supports robust model evaluation. Finally, because COVID-19 frequently leads to ARDS[44–47] and sepsis,[48–50] these 2 conditions are particularly relevant in external validations.

The TECO model outperformed non-transformer models across all external cohorts over the 5-day monitoring period, particularly in the early ICU days (Table 2). It also exhibited superior discrimination between ICU mortality and survival compared to the other models (Figure 2; Figure S5). The TECO model's multi-head attention modules overcome a bottleneck in traditional recurrent neural networks to learn long-range dependencies in sequences by linearly projecting the dimensions and queries of the input embedding.[11] It is also worth noting that TECO's performance is comparable to several well-established ICU and in-hospital mortality prediction algorithms.[51–57] Some of these algorithms, while showing better predictive capabilities,[53,54,56] are limited by some essential factors such as a lack of external validation, a significantly smaller development sample size, challenges in real-time monitoring implementation, or a combination of these. We used a Softmax activation function, which generates a score output that does not represent a true posterior probability in the Bayesian sense. Therefore, model calibration may become unreliable in external validation. Future studies should investigate methods to optimize model calibration across diverse cohorts.

We demonstrate TECO's practical utility in monitoring patient deterioration in the ICU, as the TECO-estimated risk score was consistently elevated among patients who eventually expired (Figure 2). To exclude the possibility that such trends were systemically introduced due to model artifacts, we examined and compared the mortality probability of individual patients (Figure S4). We observed different patterns among these patients, especially in the earlier days in ICU. Some of these patterns (eg, Patients D, F, I) differ significantly from the aggregate trends observed at the group level (Figure 2), highlighting TECO's ability to capture real-time, patient-specific details.

Developing an algorithm that can leverage the longitudinal time course of inpatient EHR data may improve prediction accuracy and enable earlier detection. In this study, the median AUC achieved by TECO using data up to 60 hours before the outcome (0.93) matched with that based on EDI at 12 hours before the outcome. With the successful validation of the external cohorts, this may suggest that TECO could signal a deterioration alert a full 48 hours before EDI. For ICUs with heavy workloads such as those observed during the COVID-19 pandemic, this improvement could substantially facilitate hospital resource planning, clinician communication with patient families, and play a vital role in future public health emergencies.

Limitations and future directions

The usage of COVID-19 data to develop the TECO model was primarily motivated by the large sample size accumulated over the pandemic in our EHR systems. In addition, this epic-based dataset included the proprietary EDI as a benchmark to evaluate our model—an option not available in other public ICU data sources. Ideally, external validation could have been conducted on a COVID-19 cohort from a different health system. Unfortunately, such data were not available at the time of this study. When validating TECO in the 2 external, non-COVID-19 cohorts, we found the model's performance decreased compared with that in the COVID-19 cohort. While a performance drop from the training setting onto the independent testing setting is frequently observed,[58,59] it is important to note that the 2 external validation cohorts in this study represent 2 distinct conditions that differ from COVID-19 and the models were not trained on these diseases. Another potential contributor to the observed performance drop may stem from the fact that, when training TECO using the COVID-19 cohort, we only utilized patient outcomes at the ICU endpoint (ie, death or discharge). This approach was intended to ensure the representation of the 2 extreme scenarios across a broad spectrum of patients' physiological conditions in the ICU. However, when evaluating TECO in the external cohorts over a moving time scale (120-240 hours since ICU admission), patients who are not at immediate risk of death may not necessarily present a health status that is ready for discharge. While TECO demonstrates proof of concept as a potential ICU monitoring tool, further validation across a broader patient population, including those meeting critical care criteria beyond COVID-19, widespread inflammation, respiratory failure, and other organ dysfunctions, could significantly improve its performance and generalizability.

Secondly, the current version of TECO incorporates only 11 variables. Future work could investigate whether more complex models using additional variables could influence the models' performance and generalizability. However, operational costs associated with more complex and computationally intensive models must be thoroughly evaluated when considering these potential improvements. The TECO model in this study remains a lightweight transformer, which should not present significantly greater implementation challenges than the more conventional RF and XGBoost models.

Third, not all patients in the external validation cohorts developed sepsis or ARDS prior to ICU admission. Considering documentation and lab delays, we found that 73% of patients likely acquired sepsis before ICU admission (Supplementary Info 1). While we tested TECO in these cohorts from as early as 120 hours post-ICU admission, its utility may be limited to only after a patient developed the conditions. Future work could explore potential performance differences between patients with pre-ICU versus ICU-acquired conditions.

Conclusion

We developed TECO, a transformer-based model, to analyze multi-dimensional, continuous monitoring data for ICU mortality prediction. In internal validation, TECO outperformed EDI-based prediction and other conventional machine learning methods. In 2 external validation cohorts (where EDI was not available), TECO outperformed other conventional machine learning methods. The TECO model could be further developed into an early warning tool for ICU or inpatient settings across various conditions.

Author contributions

Ruichen Rong (Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Zifan Gu (Conceptualization, Data

curation, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Hongyin Lai (Data curation), Tanna Nelson (Data curation), Tony Keller (Data curation), Clark Walker (Data curation), Kevin W. Jin (Writing—original draft), Catherine Chen (Formal analysis, Methodology, Writing—original draft), Ann Marie Navar (Methodology, Writing—original draft), Ferdinand Velasco (Data curation, Writing—original draft), Eric D. Peterson (Methodology, Writing—original draft), Guanghua Xiao (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Donghan M. Yang (Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), and Yang Xie (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This study was supported in part by the National Institutes of Health under award number R01GM140012 (G.X.), R01DE030656 (G.X.), R01GM115473 (G.X.), U01CA249245 (G.X.), U01AI169298 (Y.X.), R35GM136375 (Y.X.), the Cancer Prevention and Research Institute of Texas (CPRIT RP180805: Y.X.; CPRIT RP230330: G.X.), and the Texas Health Resources Clinical Scholars Program (D.M.Y.).

Conflict of interests

The authors have no potential conflict of interest to disclose.

Data availability

The COVID-19 dataset could not be shared publicly due to data and privacy protection policies at Texas Health Resources and UT Southwestern Medical Center. The MIMIC dataset is publicly available at https://physionet.org/content/mimiciv/2.2/.

References

1. Awrahman BJ, Aziz Fatah C, Hamaamin MY. A review of the role and challenges of big data in healthcare informatics and analytics. *Comput Intell Neurosci*. 2022;2022:5317760.
2. Peterson ED. Machine learning, predictive analytics, and clinical practice: can the past inform the present? *JAMA*. 2019;322:2283-2284.
3. Dieteren CM, van Hulsen MAJ, Rohde KIM, et al. How should ICU beds be allocated during a crisis? Evidence from the COVID-19 pandemic. *PLoS One*. 2022;17:e0270996.
4. Craxi L, Vergano M, Savulescu J, et al. Rationing in a pandemic: lessons from Italy. *Asian Bioeth Rev*. 2020;12:325-330.
5. Systems E. *Saving Lives with AI: Using the Deterioration Index Predictive Model to Help Patients Sooner*; 2022. Accessed March 1, 2025. https://www.epicshare.org/share-and-learn/saving-lives-with-ai
6. Systems E. *Artificial Intelligence Triggers Fast, Lifesaving Care for COVID-19 Patients*; 2020. https://www.epic.com/epic/post/artificial-intelligence-epic-triggers-fast-lifesaving-care-covid-19-patients/
7. Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Ann Am Thorac Soc*. 2021;18:1129-1137.
8. Raman G, Ashraf B, Demir YK, et al. Machine learning prediction for COVID-19 disease severity at hospital admission. *BMC Med Inform Decis Mak*. 2023;23:46.
9. Bendavid I, Statlender L, Shvartser L, et al. A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19. *Sci Rep*. 2022;12:10573.
10. Cummings BC, Ansari S, Motyka JR, et al. Predicting intensive care transfers and other unforeseen events: analytic model validation study and comparison to existing methods. *JMIR Med Inform*. 2021;9:e25066.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv, arXiv:1706.03762, preprint: not peer reviewed.
12. Wu N, Green, B Ben, X, et al. 2020. Deep transformer models for time series forecasting: the influenza prevalence case, arXiv, arXiv:2001.08317, preprint: not peer reviewed.
13. Huang K Altosaar J, Ranganath R, 2019. Clinicalbert: modeling clinical notes and predicting hospital readmission, arXiv, arXiv:1904.05342, preprint: not peer reviewed.
14. Li Y, Rao S, Solares JRA, et al. BEHRT: transformer for electronic health records. *Sci Rep*. 2020;10:7155.
15. Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*. 2021;4:86.
16. Antikainen E, Linnosmaa J, Umer A, et al. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. *Sci Rep*. 2023;13:3517.
17. Khader F, Kather JN, Muller-Franzes G, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Sci Rep*. 2023;13:10666.
18. Cheng J, Sollee J, Hsieh C, et al. COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data. *Eur Radiol*. 2022;32:4446-4456.
19. Song H, Rajan, D, Thiagarajan, J, et al. Attend and diagnose: clinical time series analysis using attention models. *Proceedings of the AAAI Conference on Artificial Intelligence*; 2018.
20. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:219.
21. Ranieri VM, Rubenfeld GD, Thompson BT, et al.; ARDS Definition Task Force. Acute respiratory distress syndrome: the Berlin definition. *JAMA*. 2012;307:2526-2533.
22. Yang P, Wu T, Yu M, et al. A new method for identifying the acute respiratory distress syndrome disease based on noninvasive physiological parameters. *PLoS One*. 2020;15:e0226962.
23. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*. 2016;315:801-810.
24. Vincent JL, Moreno R, Takala J, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med*. 1996;22:707-710.
25. Jones AE, Trzeciak S, Kline JA. The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Crit Care Med*. 2009;37:1649-1654.
26. 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;4:277-284.
27. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inform Process Syst*. 2019;32.
28. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Machine Learn Res*. 2011;12:2825-2830.

29. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inform Process Syst*. 2017;30.
30. Liu S, Yao N, Qiu Y, et al. Predictive performance of SOFA and qSOFA for in-hospital mortality in severe novel coronavirus disease. *Am J Emerg Med*. 2020;38:2074-2080.
31. Yang Z, Hu Q, Huang F, et al. The prognostic value of the SOFA score in patients with COVID-19: a retrospective, observational study. *Medicine (Baltimore)*. 2021;100:e26900.
32. Esmaeili Tarki F, Afaghi S, Rahimi FS, et al. Serial SOFA-score trends in ICU-admitted COVID-19 patients as predictor of 28-day mortality: a prospective cohort study. *Health Sci Rep*. 2023;6:e1116.
33. Bauer M, Gerlach H, Vogelmann T, et al. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019—results from a systematic review and meta-analysis. *Crit Care*. 2020;24:239.
34. Tzimas KN, Papadakos PJ. An updated review of sepsis for the anesthesiologist. *Semin Cardiothorac Vasc Anesth*. 2013;17:262-268.
35. Blanco J, Muriel-Bombin A, Sagredo V, Grupo de Estudios y Análisis en Cuidados Intensivos; et al. Incidence, organ dysfunction and mortality in severe sepsis: a Spanish multicentre study. *Crit Care*. 2008;12:R158.
36. Martin GS, Mannino DM, Eaton S, et al. The epidemiology of sepsis in the United States from 1979 through 2000. *N Engl J Med*. 2003;348:1546-1554.
37. Cecconi M, Evans L, Levy M, et al. Sepsis and septic shock. *Lancet*. 2018;392:75-87.
38. Mayr VD, Dunser MW, Greil V, et al. Causes of death and determinants of outcome in critically ill patients. *Crit Care*. 2006;10:R154.
39. Xu H, Sheng S, Luo W, et al. Acute respiratory distress syndrome heterogeneity and the septic ARDS subgroup. *Front Immunol*. 2023;14:1277161.
40. Yildirim F, Karaman I, Kaya A. Current situation in ARDS in the light of recent studies: classification, epidemiology and pharmacotherapeutics. *Tuberk Toraks*. 2021;69:535-546.
41. Srzic I, Nesek Adam V, Pejak DT. Sepsis definition: what's new in the treatment guidelines. *Acta Clin Croat*. 2022;61:67-72.
42. Rello J, Valenzuela-Sanchez F, Ruiz-Rodriguez M, et al. Sepsis: a review of advances in management. *Adv Ther*. 2017;34:2393-2411.
43. Oczkowski S, Alshamsi F, Belley-Cote E, et al. Surviving sepsis campaign guidelines 2021: highlights for the practicing clinician. *Pol Arch Intern Med*. 2022;132:7-8.
44. Makkar P, Pastores SM. Respiratory management of adult patients with acute respiratory distress syndrome due to COVID-19. *Respirology*. 2020;25:1133-1135.
45. Gosangi B, Rubinowitz AN, Irugu D, et al. Correction to: COVID-19 ARDS: a review of imaging features and overview of mechanical ventilation and its complications. *Emerg Radiol*. 2022;29:225.
46. Narota A, Puri G, Singh VP, et al. COVID-19 and ARDS: update on preventive and therapeutic venues. *Curr Mol Med*. 2022;22:312-324.
47. Lin SH, Zhao YS, Zhou DX, et al. Coronavirus disease 2019 (COVID-19): cytokine storms, hyper-inflammatory phenotypes, and acute respiratory distress syndrome. *Genes Dis*. 2020;7:520-527.
48. Widjaja G, Turki Jalil A, Sulaiman Rahman H, et al. Humoral immune mechanisms involved in protective and pathological immunity during COVID-19. *Hum Immunol*. 2021;82:733-745.
49. Piccioni A, Franza L, Rosa F, et al. The role of SARS-COV-2 infection in promoting abnormal immune response and sepsis: a comparison between SARS-COV-2-related sepsis and sepsis from other causes. *Infect Med (Beijing)*. 2023;2:202-211.
50. Fang C, Ma Y. Peripheral blood genes crosstalk between COVID-19 and sepsis. *Int J Mol Sci*. 2023;24:2591.
51. Ye Z, An S, Gao Y, et al. The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models. *Eur J Med Res*. 2023;28:33.
52. Li F, Xin H, Zhang J, et al. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ Open*. 2021;11:e044779.
53. Pang K, Li L, Ouyang W, et al. Establishment of ICU mortality risk prediction models with machine learning algorithm using MIMIC-IV database. *Diagnostics (Basel)*. 2022;12:1068.
54. Iwase S, Nakada TA, Shimada T, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. *Sci Rep*. 2022;12:12912.
55. Jamshidi E, Asgary A, Tavakoli N, et al. Using machine learning to predict mortality for COVID-19 patients on day 0 in the ICU. *Front Digit Health*. 2021;3:681608.
56. Villar J, Gonzalez-Martin JM, Hernandez-Gonzalez J, et al.; Predicting Outcome and STratifiCation of severity in ARDS (POSTCARDS) Network. Predicting ICU mortality in acute respiratory distress syndrome patients using machine learning: the predicting outcome and STratifiCation of severity in ARDS (POSTCARDS) study. *Crit Care Med*. 2023;51:1638-1649.
57. Jeon ET, Lee HJ, Park TY, et al. Machine learning-based prediction of in-ICU mortality in pneumonia patients. *Sci Rep*. 2023;13:11527.
58. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14:49-58.
59. Siontis GC, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68:25-34.

JAMIA Open, 2025, 8, 1–10
https://doi.org/10.1093/jamiaopen/ooaf026
Research and Applications