Original Article
Yonsei Med J 2025 Feb;66(2):121-130
https://doi.org/10.3349/ymj.2024.0126
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Artificial Intelligence-Based Early Prediction of Acute Respiratory Failure in the Emergency Department Using Biosignal and Clinical Data

Changho Han[1*], Yun Jung Jung[2*], Ji Eun Park[2], Wou Young Chung[2], and Dukyong Yoon[1,3,4]

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul;
[2]Department of Pulmonology and Critical Care Medicine, Ajou University School of Medicine, Suwon;
[3]Institute for Innovation in Digital Healthcare (IIDH), Severance Hospital, Seoul;
[4]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Korea.

Purpose: Early identification of patients at risk for acute respiratory failure (ARF) could help clinicians devise preventive strategies. Analyzing biosignals with artificial intelligence (AI) can uncover hidden information and variability within time series. We aimed to develop and validate AI models to predict ARF within 72 h after emergency department admission, primarily using high-resolution biosignals collected within 4 h of arrival.

Materials and Methods: Our AI model, built on convolutional recurrent neural networks, combines biosignal feature extraction and sequence modeling. The model was developed and internally validated with data from 5284 admissions [1085 (20.5%) positive for ARF], and externally validated using data from 144 admissions [7 (4.9%) positive for ARF] from another institution. We defined ARF as the application of advanced respiratory support devices.

Results: Our AI model performed well in predicting ARF, achieving area under the receiver operating characteristic curve (AUROC) of 0.840 and 0.743 in internal and external validations, respectively. It outperformed the Modified Early Warning Score (MEWS) and XGBoost models built only with clinical variables. High predictive ability for mortality was observed, with AUROC up to 0.809. A 10% increase in AI prediction scores was associated with 1.44-fold and 1.42-fold increases in ARF risk and mortality risk, respectively, even after adjusting for MEWS and demographic variables.

Conclusion: Our AI model demonstrates high predictive accuracy and significant associations with clinical outcomes. Our AI model has the potential to promptly aid in triage decisions. Our study shows that using AI to analyze biosignals advances disease detection and prediction.

Key Words: Respiratory failure, artificial intelligence, physiologic monitoring, digital signal processing, vital signs, emergency department

Received: May 21, 2024 **Revised:** July 22, 2024
Accepted: July 25, 2024 **Published online:** November 11, 2024
Co-corresponding authors: Dukyong Yoon, MD, PhD, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 331 Dongbaekjukjeon-daero, Giheung-gu, Yongin 16995, Korea.
E-mail: dukyong.yoon@yonsei.ac.kr and
Wou Young Chung, MD, Department of Pulmonology and Critical Care Medicine, Ajou University School of Medicine, 164 World Cup-ro, Yeongtong-gu, Suwon 16499, Korea.
E-mail: biscut@ajou.ac.kr

*Changho Han and Yun Jung Jung contributed equally to this work.
•The authors have no potential conflicts of interest to disclose.

INTRODUCTION

Acute respiratory failure (ARF) can be caused by various clinical conditions and result in high mortality.[1] The fundamental aspects of its management include early detection and timely treatment.[2] At admission to the emergency department (ED), patients with respiratory risk factors may not develop ARF immediately; however, it can occur later.[3,4] These patients can be undertriaged to the general ward despite eventually requiring intensive care unit (ICU) admission, which may have fatal consequences and lead to increased mortality and morbidity.[5] Therefore, the prediction of ARF occurring days after ED admission and accurate triage of these high-risk patients would be

clinically significant.

Efforts have been made to identify patients who are likely to experience clinical deterioration at an early stage. As a result of these efforts, many early warning scores, disease severity scores, and, more recently, predictive modeling using machine learning or artificial intelligence (AI) technologies have been developed and continue to evolve.[6-8] Many studies have used AI to early predict the ARF risk, showing that mortality and morbidity can be reduced by enabling clinicians to use these predictions for decision-making and close monitoring.[9-11] Since the onset of the COVID-19 pandemic, the high mortality rates associated with COVID-19-related ARF and the effective distribution of medical resources have become important issues. Accordingly, many studies have been conducted on models that predict the risk of ARF or mortality using AI technologies in COVID-19 patients.[12-14]

Analysis of a patient's biosignals, including electrocardiography (ECG), respiratory impedance, and plethysmography, can reveal concealed information about the inherent dynamics and overall variability within the relevant time series.[15] These biosignals hold promise in improving the prediction, detection, and treatment of various diseases, as they are largely non-invasive and widely monitored in various areas of clinical practice, such as the ICU, operating room, and ED.[16,17] There is growing evidence that recent AI technologies provide novel opportunities to reveal hidden information in biosignals that is not apparent in conventional analysis methods.[16] Most previous AI-based ARF prediction studies used electronic medical record (EMR) data, such as demographics, vital signs, laboratory findings, medications, and oxygen therapies, as model inputs. However, no study has predicted the risk of ARF and mortality by analyzing the continuous biosignals monitored in hospitals.

Thus, in this study, we aimed to develop AI models using supervised learning that can predict ARF occurring within the first 72 h after hospital admission using biosignals and clinical data collected within 4 h of admission to the ED. In addition, we evaluated the model's predictive ability for mortality and its association with actual clinical outcomes to assess its potential clinical utility in reducing the adverse consequences of ARF. Moreover, we aimed to assess the applicability of our models in different environments by conducting external validation on a dataset obtained from a different organization.

MATERIALS AND METHODS

Ethics approval

The Institutional Review Boards (IRB) of Ajou University Medical Center (AUMC) and Yongin Severance Hospital (YSH) approved this study [IRB no. AJIRB-MED-MDB-21-387 (AUMC), 9-2021-0177 (YSH)] and waived the requirement for informed consent, as only anonymized data were retrospectively used. The data was accessed during the approved research period [September 27, 2021 to December 31, 2022 for AJIRB-MED-MDB-21-387 (AUMC) and January 6, 2022 to December 31, 2022 for 9-2021-0177 (YSH)].

Data sources and labeling

From the AUMC database, we retrospectively extracted clinical and biosignal data that could be collected almost immediately after admission in standard clinical settings to be used as AI model input and output from adult (age ≥18 years) patients who visited the ED acute unit (ED-A) between March 4, 2018 and March 25, 2021 (Fig. 1). Patients underwent continuous bi-

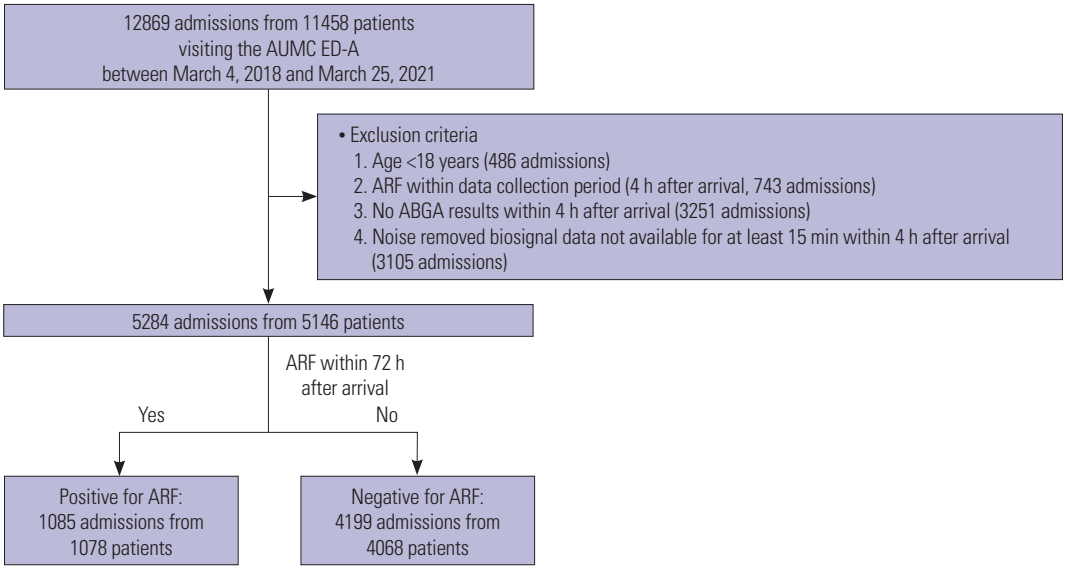

Fig. 1. Patient flow diagram. One thousand seventy-eight patients (1085 admissions) and 4068 patients (4199 admissions) were labeled as positive and negative for ARF, respectively. AUMC, Ajou University Medical Center; ED-A, emergency department acute unit; ABGA, arterial blood gas analysis; ARF, acute respiratory failure.

https://doi.org/10.3349/ymj.2024.0126

osignal monitoring upon admission. Among these biosignal data, we utilized the following: lead II ECG waveform data; and numeric data including heart rate (HR), respiratory rate (RR), SpO_2 (oxygen saturation measured by pulse oximeter), and non-invasive systolic and diastolic blood pressures (NSBP and NDBP). From various tables stored in the EMR database, variables required for model input, labeling, and statistical comparisons were extracted. Details regarding data sources and extraction are described in Supplementary Method 1 (only online).

ARF is a condition in which the lungs cannot adequately exchange oxygen and carbon dioxide, resulting in dangerously low levels of oxygen in the blood (hypoxemia), high levels of carbon dioxide (hypercapnia), or both.[1] This condition can be caused by various respiratory, cardiovascular, or systemic diseases.[1] After an extensive literature review, we found vast heterogeneity in defining ARF, such as defining ARF as an invasive mechanical ventilation (IMV) application, prolonged mechanical ventilation or tracheostomy application, acute respiratory distress syndrome (ARDS), and advanced respiratory support (AdvRS) device application, etc.[8,10,11,13,18,19] Imaging studies are not included in the basic definition of ARF.[1] We determined that setting the task as predicting ARF requiring AdvRS was the most relevant and significant. This is because, among a wide spectrum of respiratory support devices, nasal prongs or oxygen masks are used for low severity in standard clinical practice, and AdvRS devices, including IMV, non-invasive ventilation (NIV), or high flow nasal cannula (HFNC), are used for high severity.[20,21] Patients with higher disease severity require closer monitoring and may be admitted to the ICU, which will benefit more from the deterioration prediction. Therefore, we labeled ARF requiring AdvRS as positive cases in the prediction task.

We aimed to construct AI models to predict ARF using biosignals and clinical data collected early after arrival (within 4 h) in the ED. ARF occurring within the data collection period (within 4 h after arrival) was not predicted since the outcome of interest had already occurred during the data collection period. Thus, patients with ARF occurring within 4 h of arrival at the ED were excluded from the analyses. We also excluded patients whose arterial blood gas analysis (ABGA) results were unavailable within 4 h after arrival and whose continuously monitored biosignals were unavailable for at least 15 min within 4 h after arrival. After applying all the exclusion criteria, cases with ARF occurring up to 72 h after arrival at the ED were defined as positive.

For creating the external validation dataset, we retrospectively extracted biosignal and clinical data from adult (age ≥18 years) patients who visited the YSH ED-A between January 3, 2021 and January 2, 2022 (Supplementary Fig. 1, only online). We were able to extract all relevant data to be used as model input and output and for statistical analyses from the EMR and biosignal databases of the YSH. The same inclusion and exclusion criteria for patient selection and data preprocessing methods applied to the AUMC data were also applied to the YSH data.

Neural network architecture and model training

We constructed our AI models based on convolutional recurrent neural networks (CRNN) (Supplementary Fig. 2, Supplementary Table 1, only online), which integrates feature extraction from the ECG [convolutional neural network (CNN) part] and sequence modeling [recurrent neural network (RNN) part].[22] Details on neural network architecture are described in Supplementary Method 2 (only online). To reflect the most acute condition of the patient as much as possible and to make the prediction time point of the AI model as early as possible, we utilized the initial 15 min of biosignal data upon admission as input. The hidden state (h_t) of the final time step of the CRNN was concatenated with features that could be collected within 4 h of arrival at the ED before being fed into the fully connected layers to produce the final output, and various experiments were conducted with varying numbers of features concatenated to h_t (Supplementary Table 2, only online).

In the first experiment (Experiment 1, E1), no additional features were concatenated with h_t. Thus, this experiment used only biosignals as inputs for the CNN and RNN. In the second experiment (Experiment 2, E2), the initial clinical data [body temperature (BT) upon admission, initial AVPU (an acronym from "alert, verbal responsive, pain responsive, and unresponsive," which is a score measuring a patient's consciousness) score upon admission, age, and sex] were concatenated with h_t. The rationale for using the initial values on admission was the same as described above. In the third experiment (Experiment 3, E3), the initial ABGA results {PaO_2 (partial pressure of oxygen), $PaCO_2$ (partial pressure of carbon dioxide), pH, BE (base excess), bicarbonate [HCO_3^- (bicarbonate), and FiO_2 (fraction of inspired oxygen)]} were concatenated with h_t in addition to the features concatenated in the second experiment. More details on data preprocessing, data augmentation, model training, and hyperparameter tuning are described in Supplementary Methods 1–3 (only online).

A four-fold cross-validation (CV) was performed for training and validation to choose the optimal hyperparameters, and the average performance of each fold was derived. All CV experiments (E1–E3) used the same folds for comparison. Patients were not shared between the training and validation sets. We performed additional internal validation of our model using a temporal validation (TV) approach. In this approach, hospitalizations before August 26, 2020 were assigned to the training set, and hospitalizations after and including that date were assigned to the test set. To ensure that the same patients were not shared between the training and testing sets, we excluded patients included in the training set from the test set. The resulting test set was validated using models trained with all the samples in the training set with the same hyperparameters chosen in

the CV experiment. For external validation using data from the YSH, we used the model developed in the TV, E3 experiment, as it was the best-performing model.

Performance evaluation

For the performance evaluation, the following metrics were used: area under the receiver operating characteristics curve (AUROC), area under the precision recall curve (AUPRC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. For the CV experiments, the average validation performance of the folds was derived. The optimal cut-off value for dichotomizing AI model outputs was defined as the point at which the Youden J statistic was at its highest. This cut-off point was determined in the training set and applied to the validation set.

We then assessed the performance of the AI model in predicting 28-day, in-ICU, and in-hospital mortality, using the same performance metrics aforementioned. For clarification, the AI models were not specifically retrained for this task with the mortality data set as the label. These were evaluated using only the corresponding mortality labels for each experiment. Additionally, we conducted multiple logistic regressions (MLoR) to evaluate the association between AI model predictions and real-world outcomes, with appropriate adjustments. For this task, we used the model developed in the TV, E3 experiment, as it was the best-performing model. Specifically, we utilized the Modified Early Warning Score (MEWS), age, and sex as independent variables, and each outcome label—ARF, 28-day mortality, in-ICU mortality, and in-hospital mortality—as the dependent variable.

Widely used early warning systems, such as the MEWS, use physiological parameters to predict patients at an increased risk of deterioration, resulting in ICU admission or death.[23] We compared the performance of our models with that of MEWS (Supplementary Table 3, only online).[23] Additionally, we compared the performance of our models with XGBoost models built only with initial clinical features that can be collected within 4 h or arrival at the ED, as listed in Supplementary Table 2 (only online). For this task, we used the models developed in the TV, E2 and TV, E3 experiments, and for a fair comparison, the XGBoost models also used the same temporal split, and were built using the corresponding clinical features from each E2 and E3 experiment. The hyperparameters of the XGBoost models were tuned through a grid search with 4-fold CV in the training set of the temporal split, and the XGBoost models trained with the chosen hyperparameters using the entire training set were validated on the test set. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Checklist for Prediction Model Development and Validation was followed (Supplementary Method 4, only online).[24]

RESULTS

Patient characteristics

Fig. 1 shows the flow diagram of data from AUMC. In total, 11458 patients visited the ED-A during the study period. After applying the exclusion criteria, 5146 patients were obtained (5284 admissions). The baseline patient characteristics are shown in Table 1 and Supplementary Table 4 (only online). The study included 1078 patients (1085 admissions) and 4068 patients (4199 admissions) labeled positive and negative for ARF, respectively. In Table 1, the initial numeric biosignal values were calculated as the average of the values measured with the monitoring devices in the first 15 min upon admission. The patients positive for ARF were younger and had higher HR, higher RR, lower SpO_2, lower body temperature, lower level of consciousness as measured by AVPU score, lower oxygenation as identified by the ratio of PaO_2 to FiO_2, higher disease severity (lower Korean Triage and Acuity Scale, KTAS; lower means more severe in KTAS, higher MEWS), and a higher mortality rate compared to those negative for ARF.[25] Details on statistical analyses are described in Supplementary Method 5 (only online).

In total, 995 patients visited YSH ED-A between January 3, 2021 and January 2, 2022 (Supplementary Fig. 1, only online). After applying the exclusion criteria, 144 patients were included. The baseline patient characteristics are shown in Supplementary Table 5 (only online). The sample comprised seven and 137 patients labeled positive and negative for ARF, respectively. Patients positive for ARF had a significantly higher MEWS score compared to those negative for ARF. Supplementary Table 6 (only online) compares the AUMC and YSH cohorts. Patient severity was higher in the AUMC cohort than in the YSH cohort, with a significantly lower KTAS score, higher MEWS, and higher incidence of ARF.

Model performance

Fig. 2 shows the CV and TV experiments' ROC and PR curves. Table 2 shows the performance of the models when the Youden J statistics were at their maximum in the CV and TV experiments. Overall, performance tended to improve as more features were introduced into the models. The AUROCs increased progressively from E1 to E3, although the increase between E2 and E3 was not statistically significant [DeLong test (paired, two-sided): TV, E2 vs. TV, E1, $p<0.001$; TV, E3 vs. TV, E2, $p=0.038$; adjusted significance level (Bonferroni correction) 0.05/2=0.025]. The AUPRCs and Youden J statistics also improved progressively from E1 to E3.

The E3 model had a high AUROC of 0.818±0.008 and 0.840 and an AUPRC of 0.560±0.025 and 0.470 in the CV and TV experiments, respectively. The sensitivities were 0.769±0.030 and 0.728, and the PPVs were 0.415±0.021 and 0.315, while maintaining high NPVs of 0.923±0.008 and 0.954 in the CV and TV experiments, respectively. The lower AUPRC and PPV in the TV experiment compared with the CV experiment could be due to

https://doi.org/10.3349/ymj.2024.0126

Table 1. Baseline Patient Characteristics

Characteristics	ARF-positive (n=1085)	ARF-negative (n=4199)	*p* value
Age	67 [55–78]	71 [57–81]	<0.001
Sex, male	644 (59.4)	2456 (58.5)	0.606
Comorbidities			
Diabetes mellitus	211 (19.5)	772 (18.4)	0.430
Chronic lung disease	133 (12.3)	715 (17.0)	<0.001
Cardiovascular disease	347 (32.0)	1318 (31.4)	0.709
Chronic renal disease	192 (17.7)	839 (20.0)	0.094
Solid organ cancer	161 (14.8)	937 (22.3)	<0.001
Initial vital signs			
HR, bpm	99.3 [84.3–117.4]	93.5 [80.0–109.0]	<0.001
RR, bpm	22.0 [18.4–27.1]	21.2 [18.1–25.1]	<0.001
SpO_2, %	96.5 [93.1–98.9]	96.8 [94.7–98.7]	<0.001
NSBP, mm Hg	133.0 [103.8–155.9]	132.0 [110.1–154.0]	0.144
NDBP, mm Hg	80.2 [64.8–96.5]	80.2 [68.0–94.0]	0.289
Body temperature, °C	36.6 [36.0–37.2]	36.8 [36.4–37.3]	<0.001
Level of consciousness			<0.001
Alert	325 (30.0)	1775 (42.3)	
Verbal responsive	187 (17.2)	620 (14.7)	
Pain responsive	394 (36.3)	549 (13.1)	
Unresponsive	159 (14.7)	38 (0.9)	
Initial ABGA results			
PaO_2, mm Hg	96.6 [73.6–162.6]	89.3 [73.8–117.5]	<0.001
FiO_2, %	35 [21–61]	21 [21–37]	<0.001
P/F ratio	337.7 [221.0–435.7]	361.9 [301.4–452.9]	<0.001
$PaCO_2$, mm Hg	33.1 [27.1–43.8]	30.6 [26.3–35.7]	<0.001
pH	7.354 [7.238–7.416]	7.420 [7.376–7.455]	<0.001
BE, mmol/L	-4.8 [-9.6–-1.4]	-2.3 [-5.4–-0.2]	<0.001
HCO_3^-, mmol/L	19.0 [14.7–23.0]	20.4 [17.4–33.0]	<0.001
KTAS			<0.001
1	196 (18.0)	106 (2.5)	
2	655 (60.4)	1470 (35.0)	
3	230 (21.2)	2526 (60.2)	
4	4 (0.4)	90 (2.1)	
5	0 (0.0)	7 (0.2)	
MEWS	4.6±2.2	3.4±1.8	<0.001
ICU referral	557 (51.3)	1441 (34.3)	<0.001
Mortality			
28-day mortality	340 (31.3)	203 (4.8)	<0.001
In-ICU mortality	341 (31.4)	172 (4.1)	<0.001
In-hospital mortality	392 (36.1)	586 (14.0)	<0.001

ARF, acute respiratory failure; Q1, first quantile; Q3, third quantile; HR, heart rate; RR, respiratory rate; SpO_2, oxygen saturation measured by pulse oximeter; NSBP, non-invasive systolic blood pressure; NDBP, non-invasive diastolic blood pressure; ABGA, arterial blood gas analysis; PaO_2, partial pressure or oxygen; FiO_2, fraction of inspired oxygen; P/F ratio, PaO_2/FiO_2 ratio; $PaCO_2$, partial pressure of carbon dioxide; BE, base excess; HCO_3^-, bicarbonate; KTAS, Korean Triage and Acuity Scale; MEWS, Modified Early Warning Score; ICU, intensive care unit.
Data are presented as median [Q1–Q3], n (%), or mean±standard deviation.

the lower percentage of positive samples in the temporally split test set compared with the entire dataset [127 out of 1050 (12.1%) in the temporally split test set and 1085 out of 5284 (20.5%) in the entire dataset]. The biosignal-only model, E1, had a moderate AUROC of 0.681±0.008 and 0.685 in the CV and TV experiments, respectively, and the E2 models typically demonstrated intermediate performance.

Of the patients positive for ARF, 935 developed ARF on the first day, 104 on the second day, and 46 on the third day of admission. When stratified by the time since admission, the models maintained their sensitivity over time after admission (Supplementary Table 7, only online). Therefore, the models retained their ability to predict ARF patients as positive over time after admission. The E3 models generally maintained high sensitivities of over 0.7, regardless of the time after admission, and the E1 models showed similar sensitivities.

The predictive ability of MEWS for ARF was AUROC of 0.657 for the entire cohort. Our AI models had a significantly higher AUROC than MEWS for predicting ARF when compared with the DeLong test (paired, two-sided) in all E2 and E3 experiments (TV experiments and all folds of the CV experiments). In all E1 experiments (TV experiment and all folds of the CV experiment), the AI models had a higher AUROC than MEWS for predicting ARF, although not statistically significant when compared with the DeLong test (paired, two-sided). The predictive ability of the XGBoost model, built only with clinical features, was AUROC of 0.682 and 0.746 for the TV, E2 and TV, E3 experiments, respectively, and was outperformed by the AI models, as compared using the DeLong test (paired, two-sided). High predictive ability for mortality was observed, with AUROC of up to 0.809 for predicting in-ICU mortality with the E3 model and AUROC of up to 0.667±0.012 for predicting in-ICU mortality even with the biosignal-only E1 model (Supplementary Tables 8 and 9, only online).

Table 3 presents the results of the MLoR analysis. To facilitate a more intuitive interpretation of the data, we adjusted the scale of the AI prediction scores and the age variable used in the analysis. The AI prediction scores, originally presented on a scale from 0 to 1, were rescaled to a new range of 0 to 10 by multiplying them by a factor of 10. Consequently, the odds ratios for the AI prediction scores, as presented in Table 3, now reflect the change in odds associated with a 10% absolute increase in the AI prediction score. For age, we set the unit to 10 years. Therefore, the odds ratios for age, as presented in Table 3, reflect the change in odds associated with an increase of 10 years in age. MLoR revealed significant positive associations between the AI prediction score and the likelihood of all clinical outcomes. For example, for every 10% absolute increase in the AI prediction score, the odds of developing ARF and of in-ICU mortality were 1.44 [95% confidence interval (CI); 1.27–1.63] and 1.42 (95% CI; 1.24–1.63) higher, respectively, after adjusting for the MEWS and demographic variables. Supplementary Fig. 3 (only online) shows the calibration plot of the best-per-

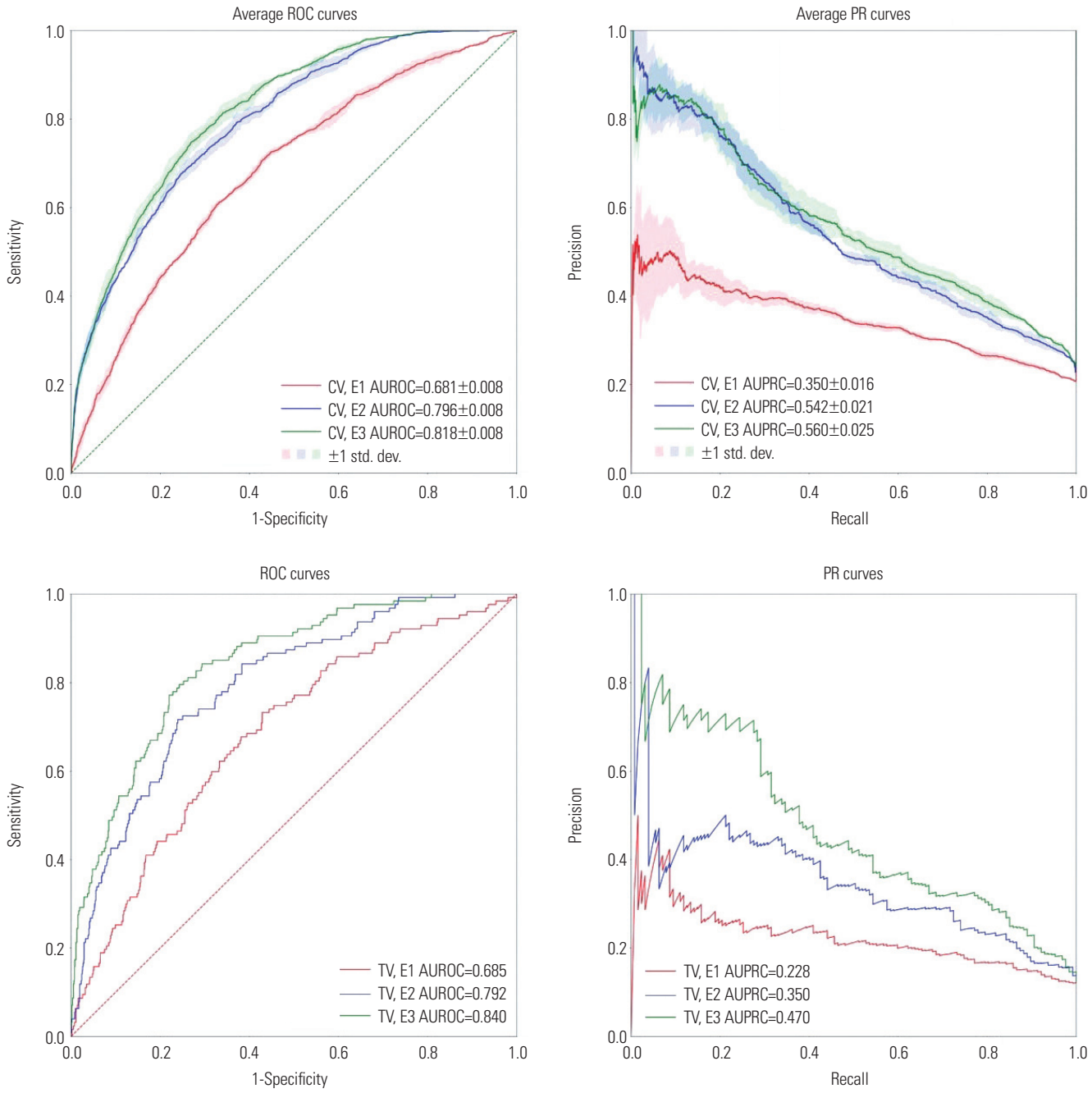

Fig. 2. The ROC and PR curves of CV and TV experiments. The AUROCs and AUPRCs increased from E1 to E3. Solid lines depict the average curves of the CV, and shaded areas depict ±1 standard deviation of the curves. ROC, receiver operating characteristic; PR, precision-recall; CV, cross-validation; std. dev., standard deviation; TV, temporal validation; AUROC, area under the ROC curve; AUPRC, area under the PR curve.

forming model (TV, E3), demonstrating relatively good agreement between the AI prediction score and the fraction of ARF positives.

Overall, a robust performance was observed in the external validation compared to the internal TV: AUROC 0.743 vs. 0.840 [DeLong test (unpaired, two-sided), p=0.257], sensitivity 0.571 vs. 0.728, specificity 0.759 vs. 0.782 (Supplementary Fig. 4, Supplementary Table 10, only online).

DISCUSSION

In this study, we developed and validated predictive AI models using supervised learning that incorporate biosignal and clinical data collected within 4 h of admission to the ED to identify patients at high risk of developing ARF within the first 72 h after admission. Our AI model, constructed based on a CRNN, integrates feature extraction from biosignals and sequence modeling. Our AI models performed well in predicting ARF, achiev-

https://doi.org/10.3349/ymj.2024.0126

Table 2. Performances of Models When the Youden J Statistics Were Highest in CV and TV Experiments

Performances	CV			TV		
	E1	E2	E3	E1	E2	E3
Accuracy	0.637±0.032	0.711±0.030	0.726±0.020	0.540	0.580	0.776
Sensitivity	0.656±0.045	0.734±0.039	0.769±0.030	0.748	0.847	0.728
Specificity	0.633±0.052	0.705±0.047	0.715±0.031	0.511	0.543	0.782
PPV	0.320±0.019	0.396±0.027	0.415±0.021	0.174	0.206	0.315
NPV	0.876±0.007	0.911±0.007	0.923±0.008	0.937	0.963	0.954
F1 score	0.428±0.009	0.513±0.016	0.538±0.016	0.282	0.330	0.440
Youden J statistics	0.289±0.012	0.439±0.015	0.484±0.020	0.259	0.389	0.511

CV, cross-validation; TV, temporal validation; PPV, positive predictive value; NPV, negative predictive value.

Table 3. Results of Multiple Logistic Regression Analyses

Variables	Odds ratio			
	ARF (95% CI)	28-day mortality (95% CI)	In-ICU mortality (95% CI)	In-hospital mortality (95% CI)
Age, 10 years	0.80 (0.71–0.92)	0.89 (0.77–1.04)	0.97 (0.83–1.14)	1.15 (1.02–1.30)
Sex (male)	1.02 (0.65–1.61)	1.31 (0.71–2.05)	1.17 (0.68–2.04)	1.07 (0.73–1.56)
MEWS	1.27 (1.11–1.44)	1.10 (0.95–1.27)	1.12 (0.96–1.30)	1.08 (0.97–1.21)
AI prediction score*10	1.44 (1.27–1.63)	1.39 (1.21–1.59)	1.42 (1.24–1.63)	1.28 (1.15–1.43)

ARF, acute respiratory failure; CI, confidence interval; ICU, intensive care unit; MEWS, Modified Early Warning Score; AI, artificial intelligence.

ing higher AUROCs than the MEWS and XGBoost models built only with clinical features. Overall, performance increased as more features were introduced into the models. When stratified by the time since admission, our model maintained its sensitivity over time. Moderate external validation performance was achieved, and we also verified the capability of our AI model to effectively predict mortality. MLoR revealed significant positive associations between the AI prediction score and the likelihood of ARF and mortality.

Numerous studies have compared the ARF prediction performance of AI models with that of MEWS, a commonly used early warning scoring system.[8,12,18] We also compared the performance of our models with that of MEWS, and our best-performing model significantly outperformed MEWS in ARF prediction (AUROC 0.840 vs. 0.657), confirming the proficiency of our AI model for ARF prediction. Our AI model showed a robust performance in external validation using data from the YSH. Patient severity was higher in the AUMC cohort than in the YSH cohort, with a significantly lower KTAS score, a higher MEWS, and especially a higher incidence of ARF (20.5% at AUMC compared to 4.9% at YSH). Our AI model demonstrated meaningful performance in ARF prediction, even in a relatively low-severity population group.

Our study had numerous implications. We demonstrated the value of continuously monitored biosignals in disease prediction. To the best of our knowledge, this is the first study to use continuously monitored biosignals for ARF prediction. Patient vital sign surveillance, performed in various hospital areas, is integral to promptly detecting adverse events.[26] Numerous systems have been developed to collect and store biosignals in databases, and an increasing number of hospitals are adopting these systems for disease prediction.[27,28] Researchers at the Massachusetts Institute of Technology collected biosignals from ICU patients and integrated them with clinical data to create a publicly available database known as the Medical Information Mart for Intensive Care.[29] Our study highlights the value of incorporating AI techniques with novel high-resolution data sources, such as continuously monitored biosignals, aligning with growing evidence from numerous studies suggesting that AI technologies have the potential to identify hidden information in biosignals that is not apparent with conventional methods.[16,30-32] For example, Hatib, et al.[31] showed that an AI-based analysis of continuously monitored intraoperative arterial pressure waveforms could predict intraoperative hypotension, and a randomized controlled trial conducted by Wijnberge, et al.[30] found that it reduced intraoperative hypotension compared to standard care. In our study, besides the E3 model demonstrating high performance, the biosignal-only E1 model also showed moderate performance, with an AUROC of 0.681–0.685 and a sensitivity of 0.656–0.748, while maintaining a high NPV of 0.876–0.937. When stratified by admission time, the E1 model showed similar sensitivities to the best-performing model, the E3 model. Our AI models achieved higher AUROCs compared to the XGBoost models built only with clinical features. Our results also showed that using AI to analyze biosignals, either independently or in conjunction with other available and relevant clinical data, yields new opportunities for disease detection, prediction, and management.

Our AI model can operate promptly after admission to the ED as it is designed to use data acquired almost immediately after admission. All the clinical data we used as model input were obtained immediately after arrival at the ED in standard

clinical practice: age, sex, level of consciousness (using the AVPU score), and BT were identified and recorded immediately upon arrival. ABGA results can also be acquired directly via point-of-care testing. Patients were also attached to a biosignal monitoring sensor as soon as they were admitted, and the biosignal data collection system stored the data in real time.[27] Most other studies predicting ARF with AI used laboratory findings that required at least a few hours to analyze and retrieve the test results.[8-11] On the other hand, our model can operate almost immediately after admission, having the potential to be incorporated into the ED triage system.

Our target was ED patients, unlike many previous studies that targeted ICU patients or those in the general ward, introducing an additional rationale for its potential inclusion in the ED triage system. To the best of our knowledge, this is the first study to predict ARF in patients in the ED using AI. ED triage is a complex decision-making process that requires balancing benefits and risks, not only for a specific patient but also for all patients admitted, as resources are limited.[33] Patients initially admitted to the general ward from the ED who subsequently underwent an unplanned transfer to the ICU showed higher morbidity and mortality rates compared to those admitted directly from the ED to the ICU.[5] Therefore, it is important to accurately identify patients who need or will need ICU treatment so they are not initially admitted to the general ward. Our AI system can assist with triage decisions in the ED, making it clinically significant.

Our AI model consistently maintained high sensitivity, accurately predicting patients with ARF as positive for ARF for up to 72 h after admission. Many prospective multicenter studies have shown that among patients presenting with respiratory risk factors at the time of admission to the ED, ICU, or general ward, those who developed acute lung injury experienced its onset within an interquartile range of 0–3 days after admission.[3,4] Therefore, the selected prediction window of 72 h was clinically relevant and significant. Moreover, patients already showing signs of ARF at the time of admission (within 4 h) were not subjects for prediction, and it would be clinically more significant to predict more unapparent cases where ARF occurs a few days after admission. Thus, we excluded cases of ARF that had already occurred within 4 h of arrival at the ED.

We designated the task as predicting ARF requiring AdvRS, including IMV, NIV, and HFNC, judging this definition to be the most clinically relevant and significant. In standard clinical practice, among a wide spectrum of respiratory support devices, healthcare professionals utilize nasal prongs or oxygen masks for patients with relatively low-severity respiratory issues, whereas patients with high severity are typically provided with AdvRS devices.[20,21] High-severity conditions are deemed more clinically significant and necessitate prioritized prediction. Furthermore, although IMV may be applied directly, it is often used in cases of NIV or HFNC failure.[34,35] Therefore, including all three devices in the prediction task provides more preemptive therapeutic opportunities than simply predicting the application of IMV. Specifically, predicting patients with ARF needing AdvRS devices may be particularly useful during the COVID-19 pandemic. Numerous AI models have been developed to predict the risk of ARF or mortality in COVID-19 patients.[12-14] The increase in COVID-19 patients has also led to a rise in ARF patients, making it critical to predict which of them will need AdvRS devices for allocating medical resources, including medical staff, ICU beds, and medical equipment, as well as for transfers to superior hospitals, especially in lower-resource hospitals or countries.[36,37] Our predictive model is clinically significant not only for COVID-19 but also for future endemic or epidemic situations.

We demonstrated that the predictive ability of our AI model for mortality was high, with an AUROC of 0.809 for in-ICU mortality in the best-performing model. In our study, the mortality rate of patients with ARF was considerable (28-day, in-ICU, and in-hospital mortality rates were 31.3%, 31.4%, and 36.1%, respectively), indicating that ARF is highly lethal. This aligns with previous studies reporting an ARF hospital mortality rate of approximately 30%.[38] Thus, identifying patients at high risk of ARF can lower mortality by activating intensive monitoring and prompt therapeutic management.[39] The high predictive ability of our AI model for mortality, even though it was not specifically trained to predict mortality, demonstrates the potential clinical utility of this model in predicting and reducing adverse consequences. We found significant positive associations between AI prediction scores and the occurrence of actual clinical outcomes. An absolute increase of 10% in the AI prediction scores was linked to a 1.44-fold and 1.42-fold elevation in the risk of developing ARF and experiencing in-ICU mortality, respectively, even after adjusting for the MEWS and demographic variables. Consequently, this indicates a critical need for closer surveillance of patients with higher AI prediction scores to prevent adverse events.

Our study had some limitations. It is difficult to interpret the predictions of our model. Continuous research is underway to develop approaches that enhance the interpretability of end-to-end deep learning models. For example, for CNN models, methods such as gradient-weighted class activation mapping have been proposed.[40] However, we could not find a suitable explanatory algorithm for CRNN models with multimodal inputs. The external validation cohort may have been too small to adequately validate the robustness of our AI model in new environments. Only 144 patients were included in the external validation cohort, with only seven patients positive for ARF. Of the 1020 admissions at YSH ED-A shown in Supplementary Fig. 1 (only online), a significant number were excluded as they lacked ABGA results or had missing noise-removed biosignal data. Future studies should extend this external validation period or use multicenter prospective designs to confirm the reliability of our AI model. An additional platform that can automatically extract the collected biosignal and clinical data, preprocess the data as needed, and apply our AI model would be necessary for real-time applications. Our task was limited to predicting ARF requiring

https://doi.org/10.3349/ymj.2024.0126

AdvRS, as we considered this definition to be the most clinically relevant and significant. However, using a different definition of ARF could yield different results and implications, and future studies should explore model development and validation with various definitions of ARF.

In conclusion, we developed an AI model that can effectively identify patients at risk of developing ARF within 72 h after ED arrival using biosignal and clinical data. We have demonstrated the clinical relevance and utility of our model by confirming its high predictive capability for mortality and by showing significant positive associations between AI prediction scores and the occurrence of actual clinical outcomes. Our AI model is of high clinical significance as it has the potential to promptly aid triage decisions and reduce adverse consequences. Our study indicates that using AI to analyze biosignals, alone or in combination with other pertinent medical data, opens up new possibilities for the detection, prediction, and treatment of diseases.

ACKNOWLEDGEMENTS

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711196067, RS-2020-KD000095). This research was supported by a grant of the MD-Phd/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea.

AUTHOR CONTRIBUTIONS

Conceptualization: Wou Young Chung and Dukyong Yoon. Data curation: Yun Jung Jung, Ji Eun Park, and Changho Han. Formal analysis: Changho Han and Yun Jung Jung. Funding acquisition: Dukyong Yoon. Investigation: Changho Han and Yun Jung Jung. Methodology: Changho Han, Yun Jung Jung, Wou Young Chung, and Dukyong Yoon. Project administration: Wou Young Chung and Dukyong Yoon. Resources: Wou Young Chung and Dukyong Yoon. Software: Changho Han. Supervision: Dukyong Yoon and Wou Young Chung. Validation: Wou Young Chung, Dukyong Yoon, and Ji Eun Park. Visualization: Changho Han. Writing—original draft: Changho Han and Yun Jung Jung. Writing—review & editing: Wou Young Chung, Dukyong Yoon, and Ji Eun Park. Approval of final manuscript: all authors.

ORCID iDs

Changho Han https://orcid.org/0000-0003-4121-5465
Yun Jung Jung https://orcid.org/0000-0002-8887-0881
Ji Eun Park https://orcid.org/0000-0002-3035-6353
Wou Young Chung https://orcid.org/0000-0002-5435-2787
Dukyong Yoon https://orcid.org/0000-0003-1635-8376

REFERENCES

1. Roussos C, Koutsoukou A. Respiratory failure. Eur Respir J Suppl 2003;47:3s-14.
2. Serin SO, Karaoren G, Esquinas AM. Delayed admission to ICU in acute respiratory failure: critical time for critical conditions. Am J Emerg Med 2017;35:1571-2.
3. Gajic O, Dabbagh O, Park PK, Adesanya A, Chang SY, Hou P, et al. Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. Am J Respir Crit Care Med 2011;183:462-70.
4. Ferguson ND, Frutos-Vivar F, Esteban A, Gordo F, Honrubia T, Peñuelas O, et al. Clinical risk conditions for acute lung injury in the intensive care unit and hospital ward: a prospective observational study. Crit Care 2007;11:R96.
5. Liu V, Kipnis P, Rizk NW, Escobar GJ. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. J Hosp Med 2012;7:224-30.
6. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. Emerg Med J 2006;23:841-5.
7. Yu S, Leung S, Heo M, Soto GJ, Shah RT, Gunda S, et al. Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. Crit Care 2014;18:R132.
8. Dziadzko MA, Novotny PJ, Sloan J, Gajic O, Herasevich V, Mirhaji P, et al. Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. Crit Care 2018;22:286.
9. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018;320:2199-200.
10. Martín-González F, González-Robledo J, Sánchez-Hernández F, Moreno-García MN. Success/failure prediction of noninvasive mechanical ventilation in intensive care units. Using multiclassifiers and feature selection methods. Methods Inf Med 2016;55:234-41.
11. Zeiberg D, Prahlad T, Nallamothu BK, Iwashyna TJ, Wiens J, Sjoding MW. Machine learning for patient risk stratification for acute respiratory distress syndrome. PLoS One 2019;14:e0214465.
12. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. J Med Internet Res 2021;23:e24246.
13. Ferrari D, Milic J, Tonelli R, Ghinelli F, Meschiari M, Volpi S, et al. Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia-challenges, strengths, and opportunities in a global health emergency. PLoS One 2020;15:e0239172.
14. Bendavid I, Statlender L, Shvartser L, Teppler S, Azullay R, Sapir R, 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.
15. Kennedy HL. Heart rate variability--a potential, noninvasive prognostic index in the critically ill patient. Crit Care Med 1998;26:213-4.
16. Yoon D, Jang JH, Choi BJ, Kim TY, Han CH. Discovering hidden information in biosignals from patients using artificial intelligence. Korean J Anesthesiol 2020;73:275-84.
17. Park JE, Kim TY, Jung YJ, Han C, Park CM, Park JH, et al. Biosignal-based digital biomarkers for prediction of ventilator weaning success. Int J Environ Res Public Health 2021;18:9229.
18. Wong AI, Kamaleswaran R, Tabaie A, Reyna MA, Josef C, Robichaux C, et al. Prediction of acute respiratory failure requiring advanced respiratory support in advance of interventions and treatment: a multivariable prediction model from electronic medical record data. Crit Care Explor 2021;3:e0402.

19. Parreco J, Hidalgo A, Parks JJ, Kozol R, Rattan R. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. J Surg Res 2018;228:179-87.
20. Slattery M, Vasques F, Srivastava S, Camporota L. Management of acute respiratory failure. Medicine 2020;48:397-403.
21. Yuste ME, Moreno O, Narbona S, Acosta F, Peñas L, Colmenero M. Efficacy and safety of high-flow nasal cannula oxygen therapy in moderate acute hypercapnic respiratory failure. Rev Bras Ter Intensiva 2019;31:156-63.
22. Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell 2017;39:2298-304.
23. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. QJM 2001;94:521-6.
24. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015;13:1.
25. Park JB, Lee J, Kim YJ, Lee JH, Lim TH. Reliability of Korean triage and acuity scale: interrater agreement between two experienced nurses by real-time triage and analysis of influencing factors to disagreement of triage levels. J Korean Med Sci 2019;34:e189.
26. Storm-Versloot MN, Verweij L, Lucas C, Ludikhuize J, Goslings JC, Legemate DA, et al. Clinical relevance of routinely measured vital signs in hospitalized patients: a systematic review. J Nurs Scholarsh 2014;46:39-49.
27. Yoon D, Lee S, Kim TY, Ko J, Chung WY, Park RW. System for collecting biosignal data from multiple patient monitoring systems. Healthc Inform Res 2017;23:333-7.
28. Lee HC, Park Y, Yoon SB, Yang SM, Park D, Jung CW. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Sci Data 2022;9:279.
29. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
30. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA 2020;323:1052-60.
31. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology 2018;129:663-74.
32. Han C, Kang KW, Kim TY, Uhm JS, Park JW, Jung IH, et al. Artificial intelligence-enabled ECG algorithm for the prediction of coronary artery calcification. Front Cardiovasc Med 2022;9:849223.
33. Blanch L, Abillama FF, Amin P, Christian M, Joynt GM, Myburgh J, et al. Triage decisions for ICU admission: report from the task force of the World Federation of Societies of Intensive and Critical Care Medicine. J Crit Care 2016;36:301-5.
34. Oczkowski S, Ergan B, Bos L, Chatwin M, Ferrer M, Gregoretti C, et al. ERS clinical practice guidelines: high-flow nasal cannula in acute respiratory failure. Eur Respir J 2022;59:2101574.
35. Rochwerg B, Brochard L, Elliott MW, Hess D, Hill NS, Nava S, et al. Official ERS/ATS clinical practice guidelines: noninvasive ventilation for acute respiratory failure. Eur Respir J 2017;50:1602426.
36. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. JAMA 2020;323:2052-9.
37. Goyal P, Choi JJ, Pinheiro LC, Schenck EJ, Chen R, Jabri A, et al. Clinical characteristics of COVID-19 in New York city. N Engl J Med 2020;382:2372-4.
38. Vincent JL, Akça S, De Mendonça A, Haji-Michael P, Sprung C, Moreno R, et al. The epidemiology of acute respiratory failure in critically ill patients. Chest 2002;121:1602-9.
39. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, 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.
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336-59.

https://doi.org/10.3349/ymj.2024.0126