J Korean Med Sci. 2021 Jul 19;36(28):e187
https://doi.org/10.3346/jkms.2021.36.e187
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article

Emergency & Critical Care Medicine

Prediction of Neurological Outcomes in Out-of-hospital Cardiac Arrest Survivors Immediately after Return of Spontaneous Circulation: Ensemble Technique with Four Machine Learning Models

Received: Mar 5, 2021
Accepted: Jun 14, 2021

Address for Correspondence:
Taegyun Kim, MD, PhD
Department of Emergency Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
E-mail: kimtaegyun@snu.ac.kr

Jonghwan Shin, MD, PhD
Department of Emergency Medicine, Seoul Metropolitan Government Seoul National University Boramae Medical Center, 20 Boramae-ro 5-gil, Dongjak-gu, Seoul 07061, Korea.
E-mail: skycpr@snu.ac.kr

ORCID iDs
Ji Han Heo
https://orcid.org/0000-0001-7859-9058
Taegyun Kim
https://orcid.org/0000-0002-3770-3944
Jonghwan Shin
https://orcid.org/0000-0002-7121-2137
Gil Joon Suh
https://orcid.org/0000-0001-5163-2217
Joonghee Kim
https://orcid.org/0000-0001-5080-7097
Yoon Sun Jung
https://orcid.org/0000-0001-7408-4436

Ji Han Heo,[1,2] **Taegyun Kim**,[1,3] **Jonghwan Shin**,[3,4] **Gil Joon Suh**,[1,3] **Joonghee Kim**,[5] **Yoon Sun Jung**,[6] **Seung Min Park**,[5] **Sungwan Kim**,[7] **and For SNU CARE investigators**

[1]Department of Emergency Medicine, Seoul National University Hospital, Seoul, Korea
[2]Interdisciplinary Program in Bioengineering, Graduate School, Seoul National University, Seoul, Korea
[3]Department of Emergency Medicine, Seoul National University College of Medicine, Seoul, Korea
[4]Department of Emergency Medicine, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul, Korea
[5]Department of Emergency Medicine, Seoul National University Bundang Hospital, Seongnam, Korea
[6]Division of Critical Care Medicine, Seoul National University Hospital, Seoul, Korea
[7]Department of Biomedical Engineering, College of Medicine and Institute of Medical & Biological Engineering, Medical Research Center, Seoul National University, Seoul, Korea

ABSTRACT

Background: We performed this study to establish a prediction model for 1-year neurological outcomes in out-of-hospital cardiac arrest (OHCA) patients who achieved return of spontaneous circulation (ROSC) immediately after ROSC using machine learning methods.
Methods: We performed a retrospective analysis of an OHCA survivor registry. Patients aged ≥ 18 years were included. Study participants who had registered between March 31, 2013 and December 31, 2018 were divided into a develop dataset (80% of total) and an internal validation dataset (20% of total), and those who had registered between January 1, 2019 and December 31, 2019 were assigned to an external validation dataset. Four machine learning methods, including random forest, support vector machine, ElasticNet and extreme gradient boost, were implemented to establish prediction models with the develop dataset, and the ensemble technique was used to build the final prediction model. The prediction performance of the model in the internal validation and the external validation dataset was described with accuracy, area under the receiver-operating characteristic curve, area under the precision-recall curve, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Futhermore, we established multivariable logistic regression models with the develop set and compared prediction performance with the ensemble models. The primary outcome was an unfavorable 1-year neurological outcome.
Results: A total of 1,207 patients were included in the study. Among them, 631, 139, and 153 were assigned to the develop, the internal validation and the external validation datasets, respectively. Prediction performance metrics for the ensemble prediction model in the internal validation dataset were as follows: accuracy, 0.9620 (95% confidence interval [CI],

Seung Min Park
https://orcid.org/0000-0002-3594-8403
Sungwan Kim
https://orcid.org/0000-0002-9318-849X

Disclosure

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Heo JH, Kim T, Shin J, Suh GJ. Formal analysis: Heo JH, Kim T, Kim S. Funding acquisition: not applicable. Investigation: Heo JH, Kim T, Shin J, Suh GJ, Kim J, Jung YS, Park SM, Kim S. Software: Heo JH, Kim T, Kim S. Validation: Heo JH, Kim T, Kim S. Visualization: Heo JH, Kim T, Kim S. Writing - original draft: Heo JH, Kim T, Shin J, Suh GJ, Kim J. Writing - review & editing: Heo JH, Kim T, Shin J, Suh GJ, Kim J, Jung YS, Park SM, Kim S.

0.9352–0.9889); area under receiver-operator characteristics curve, 0.9800 (95% CI, 0.9612–0.9988); area under precision-recall curve, 0.9950 (95% CI, 0.9860–1.0000); sensitivity, 0.9594 (95% CI, 0.9245–0.9943); specificity, 0.9714 (95% CI, 0.9162–1.0000); PPV, 0.9916 (95% CI, 0.9752–1.0000); NPV, 0.8718 (95% CI, 0.7669–0.9767). Prediction performance metrics for the model in the external validation dataset were as follows: accuracy, 0.8509 (95% CI, 0.7825–0.9192); area under receiver-operator characteristics curve, 0.9301 (95% CI, 0.8845–0.9756); area under precision-recall curve, 0.9476 (95% CI, 0.9087–0.9867); sensitivity, 0.9595 (95% CI, 0.9145–1.0000); specificity, 0.6500 (95% CI, 0.5022–0.7978); PPV, 0.8353 (95% CI, 0.7564–0.9142); NPV, 0.8966 (95% CI, 0.7857–1.0000). All the prediction metrics were higher in the ensemble models, except NPVs in both the internal and the external validation datasets.

Conclusion: We established an ensemble prediction model for prediction of unfavorable 1-year neurological outcomes in OHCA survivors using four machine learning methods. The prediction performance of the ensemble model was higher than the multivariable logistic regression model, while its performance was slightly decreased in the external validation dataset.

Keywords: Heart Arrest; Cardiopulmonary Resuscitation; Machine Learning

INTRODUCTION

Out-of-hospital cardiac arrest (OHCA) is one of the major health issues worldwide.[1] Less than one-third of OHCA victims achieve return of spontaneous circulation (ROSC), and less than ten percent remain neurologically favorable after OHCA.[2,3]

Current guidelines recommend evaluating neurological outcomes after cardiac arrest at least 72 hours after ROSC to minimize the rate of false-positive results.[4-6] Despite the recommended guidelines, caregivers sometimes request early outcome predictions,[7] which may allow the caregivers and the medical personnel enough time to share information and to discuss the care plan for cardiac arrest survivors.

Machine learning has been widely implemented in recent studies on cardiac arrest. Several studies have shown that prediction models developed with machine learning methods can predict neurological outcomes in cardiac arrest victims.[8-10] These studies mainly used prehospital features for establishing outcome prediction models, except several hospital features such as initial electrocardiography rhythm at emergency department (ED), percutaneous coronary intervention, targeted temperature management and extracorporeal membrane oxygenation.

In recent studies, initial laboratory results at hospital arrival after OHCA, such as arterial pH,[11,12] serum potassium level,[13,14] and serum creatinine level,[15,16] have been reported to be associated with neurological outcomes after cardiac arrest. Machine learning is a crucial component in the establishment of prediction models that include laboratory test results as features since a variety of laboratory tests are performed and conventional statistical techniques have difficulty handling them. Previous machine learning studies did not include laboratory results in their prediction models, which have an important association with neurological outcomes in cardiac arrest survivors.

Few studies have evaluated the prediction of neurological outcomes with machine learning methods in OHCA survivors immediately after ROSC. We performed this study to investigate the long-term neurological outcome prediction performance of several models using machine learning methods in OHCA survivors immediately after ROSC.

METHODS

Study setting and design

We performed a retrospective analysis of prospectively collected data archived in a multicenter registry of OHCA survivors. The registry consists of the data collected from adult OHCA survivors who had visited the EDs of three university hospitals in the Republic of Korea. We analyzed data from patients who had visited the EDs from March 31, 2013, to December 31, 2019. We included all adult (age ≥ 18 years) OHCA patients registered in the registry during the study period. Patients were excluded if their cerebral performance category (CPC) scales before OHCA were between three and five or their 1-year neurological outcomes were missing.

Outcome measures

The primary outcome was neurological status at one year according to the CPC scale. A favorable neurological outcome was defined as a CPC score of one or two, and an unfavorable neurological outcome was defined as a CPC score higher than two (i.e., three to five).

Statistical analysis for demographics

Continuous variables are presented as the mean ± standard deviation and were compared using Student's *t*-test or the Mann-Whitney test as appropriate. Categorical variables are presented as numbers (percentages) and compared using the χ^2 test or Fisher's exact test as appropriate. Two-sided *P* values < 0.05 were statistically significant.

Dataset

After selection of study participants, we first split the whole dataset into two separate datasets: data acquired from March 31, 2013 to December 31, 2018 (dataset 1) and data from January 1, 2019 to December 31, 2019 (dataset 2). Dataset 1 was split again into a develop dataset and an internal validation dataset with an 80:20 ratio, and dataset 2 was reserved as an external validation dataset. Missing values were imputed with means for continuous data and with modes for categorical data. As missing data were not considered missing completely at random, we made new binary variables indicating the missingness of specific variables.

Machine learning models

We implemented four machine learning methods for the prediction of unfavorable neurological outcomes in the develop dataset: random forest (RF), support vector machine, elastic net, and extreme gradient boost. To obtain the best hyperparameters, a grid search was performed for each classifier. After optimization of the hyperparameters, we calculated the following parameters in each model on the develop and the internal validation datasets: accuracy, areas under the receiver operating characteristics curve (AUROCs), areas under the precision-recall curve (AUPRCs), sensitivity, specificity, negative predictive values (NPVs), positive predictive values (PPVs) and F1 scores. We also calculated 95% confidence intervals (CIs) for each value if possible. Five-fold cross validation was implemented to calculate the average prediction performance of each model on the develop dataset. After

model establishment, we implemented the ensemble method with soft voting with the four prediction models and tested the prediction performance of the ensemble prediction model on the internal validation dataset. The cutoff probability score of the ensemble model was selected with which the F1 score was maximized. The F1 score is one of the measures of the overall performance of a prediction model, and it is defined as a harmonic mean of sensitivity and PPV of the prediction model at a certain cutoff probability score. Finally, we tested the prediction performance of the ensemble model in the external validation dataset with the same cutoff probability score and calculated the same prediction performance parameters as in the develop and the internal validation datasets.

Variable selection

Since we aimed to establish prediction models that can be applied immediately after ROSC using machine learning techniques, we selected variables widely available at the time of ROSC. As for the laboratory variables, we used most of initial laboratory test results for model develop. However, we discarded variables 1) that are thought to have strong correlation with other included variables (e.g., total carbon dioxide level, pH, arterial oxygen saturation), 2) that are associated with organ function but represented by other included variables (e.g., aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, activated partial thromboplastin time, creatinine kinase, creatinine kinase MB isoenzyme, pro-B-natriuretic peptide), 3) that are non-classic anion or cation (e.g. ionized calcium, phosphorus), 4) that are not thought to be widely used in general EDs (e.g., red cell distribution width, neuron-specific enolase, S100 protein, central venous oxygen saturation, cortisol, adrenocorticotropic hormone, antidiuretic hormone), and 5) that are not available immediately after ROSC (data from laboratory tests performed at 24 hours and 72 hours after ROSC). We finally used 46 variables for the analysis, including baseline variables, prehospital variables, ED resuscitation variables and laboratory variables. Details of the variables used are described in **Supplementary Table 1**.

Subgroup analysis

We selected patients whose cardiac arrest was presumed to be of cardiac origin as a cardiac subgroup. We performed the same analysis as we performed in the main analysis with the cardiac subgroup dataset, including data splitting, implementation of the four machine learning methods and ensemble technique, selection of cutoff probability scores and calculation of prediction parameters.

Logistic regression analysis

To explain the variable importance indirectly and to compare performance metrics of the ensemble models with that of classic prediction models, we established multivariable logistic regression models for unfavorable neurological outcomes with the same variables used in the machine learning analysis. We set the cutoff probability score of 0.5 for logistic regression analyses. Same performance metrics used in the machine learning analysis, such as accuracy, AUROC, AUPRC, sensitivity, specificity, PPV, NPV, and F1 scores were calculated.

Tools for analysis

All statistical analyses for demographics, data splitting and logistic regression analysis were performed with R version 4.0.2 (R Foundation, Vienna, Austria). All codes for machine learning analyses and calculation of performance metrics were written in Python 3.7 (Python Software Foundation, Wilmington, DE, USA).

https://doi.org/10.3346/jkms.2021.36.e187

Ethics statement

Study protocols for collecting data for the registry and for the main analyses were approved by the Institutional Review Boards (IRBs) of participating hospitals (Seoul National University Hospital, IRB No. 1408-012-599; Seoul Metropolitan Government-Seoul National University Boramae Hospital, IRB No. 16-2013-157; Seoul National University Bundang Hospital, IRB No. B-1401/234-402) and the IRB of Seoul National University Hospital (IRB No. 2012-016-117), respectively. Informed consent was waivered by the IRB of Seoul National University Hospital, according to the retrospective nature of the study.

RESULTS

Patient selection and baseline demographics

During the study period, 1,214 patients were registered in the registry of which 1,061 comprised dataset 1 and 153 comprised dataset 2 (**Fig. 1**). A total of 1,054 of the 1,061 patients in dataset 1 met the inclusion criteria. After excluding patients meeting the prespecified exclusion criteria, 789 patients were included in the final analysis. Six hundred thirty-one patients were assigned to the develop dataset, and the rest were assigned to the internal validation dataset. Four hundred ninety-two (78.0%) patients in the develop dataset remained with unfavorable neurological outcomes at the 1-year follow-up. Thirty-nine patients from dataset 2 were excluded, and the remaining 114 patients were included in the external validation dataset. The baseline characteristics of the develop dataset are described in **Table 1**.

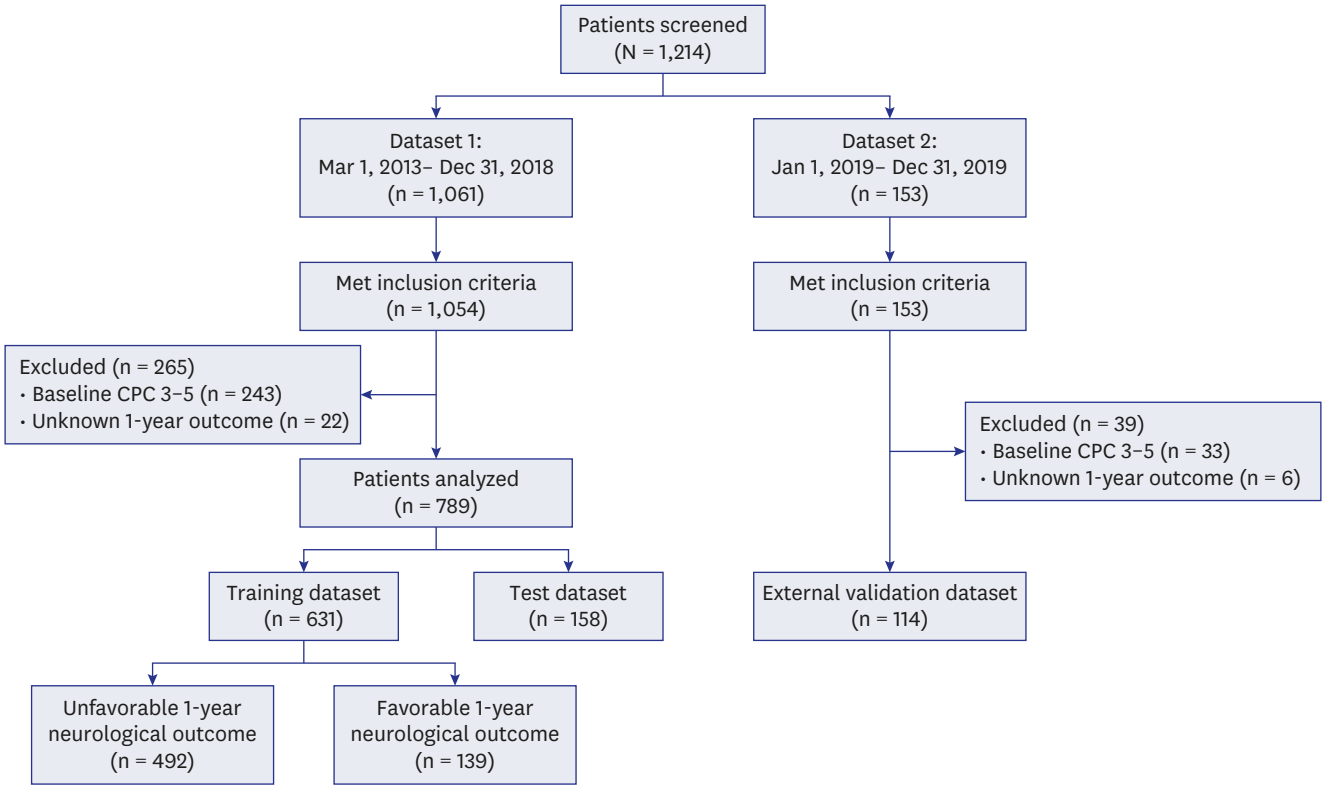

Fig. 1. Study flow chart.
CPC = cerebral performance category.

https://doi.org/10.3346/jkms.2021.36.e187

Table 1. Baseline characteristics of the original develop dataset

Variables	One-year neurological outcome		*P* value
	Unfavorable (n = 492)	Favorable (n = 139)	
Baseline variables			
Age	63.4 ± 16.5	53.6 ± 13.2	< 0.001
Male, sex	317 (64.4)	111 (79.9)	0.001
Antiplatelet	90 (19.7)	27 (20.6)	0.914
Anticoagulant	28 (6.1)	10 (7.6)	0.681
Antihypertensive agent	155 (34.1)	57 (42.9)	0.079
Diabetes mellitus	148 (30.5)	21 (15.2)	0.001
Hypertension	210 (43.3)	55 (39.9)	0.532
Dyslipidemia	29 (6.0)	14 (10.1)	0.130
Prehospital variables			
Witnessed	328 (66.7)	116 (83.5)	< 0.001
Bystander CPR	197 (40.2)	89 (65.0)	< 0.001
Bystander AED use	6 (1.2)	6 (4.3)	0.043
Prehospital initial rhythm			< 0.001
Shockable	87 (20.7)	91 (82.0)	
Asystole	206 (48.9)	7 (6.3)	
PEA	122 (29.0)	13 (11.7)	
Unknown	6 (1.4)	0 (0.0)	
Prehospital defibrillation by EMS	91 (18.8)	99 (72.3)	< 0.001
EMS CPR	440 (89.6)	119 (86.9)	0.450
EMS airway			0.020
Endotracheal tube	31 (7.2)	4 (3.6)	
Supraglottic airway	155 (36.0)	29 (26.4)	
Oral airway	102 (23.7)	24 (21.8)	
None	142 (33.0)	53 (48.2)	
ROSC before EMS arrival	13 (2.6)	6 (4.3)	0.451
ROSC by EMS	64 (13.0)	94 (68.6)	< 0.001
Hospital variables			
ED airway	398 (99.5)	71 (97.3)	0.220
Initial GCS, eye	1.2 ± 0.7	2.3 ± 1.4	< 0.001
Initial GCS, verbal	0.2 ± 0.6	1.5 ± 2.1	< 0.001
Initial GCS, motor	1.5 ± 1.2	3.6 ± 2.1	< 0.001
Initial light reflex, right			< 0.001
Prompt	107 (24.9)	105 (82.7)	< 0.001
Sluggish	28 (6.5)	14 (11.0)	
Fixed	295 (68.6)	8 (6.3)	
Initial light reflex, Left			< 0.001
Prompt	105 (24.4)	104 (81.9)	
Sluggish	33 (7.7)	13 (10.2)	
Fixed	292 (67.9)	10 (7.9)	
Initial pupil size, right, mm	4.5 ± 1.9	3.6 ± 1.3	< 0.001
Initial pupil size, left, mm	4.5 ± 1.9	3.6 ± 1.3	< 0.001
Initial corneal reflex, left			< 0.001
Yes	3 (0.6)	11 (7.9)	
No	20 (4.1)	3 (2.2)	
Not checked	469 (95.3)	125 (89.9)	
Etiology			< 0.001
Medical (cardiac)	137 (27.8)	113 (81.3)	
Medical (noncardiac)	180 (36.6)	15 (10.8)	
Medical (unknown)	33 (6.7)	5 (3.6)	
Nonmedical	142 (28.9)	6 (4.3)	
Laboratory variables			
White blood cells, 1,000/μL	13.7 ± 6.7	14.4 ± 8.0	0.378
Hemoglobin, g/dL	11.9 ± 3.7	14.1 ± 2.7	< 0.001
Platelets, 1,000/μL	185.0 ± 90.9	227.8 ± 74.9	< 0.001
Na, mmol/L	139.2 ± 7.0	138.7 ± 4.2	0.322
K, mmol/L	5.1 ± 2.0	3.9 ± 1.0	< 0.001

(continued to the next page)

Table 1. (Continued) Baseline characteristics of the original develop dataset

Variables	One-year neurological outcome		*P* value
	Unfavorable (n = 492)	Favorable (n = 139)	
Cl, mmol/L	102.6 ± 11.2	103.8 ± 4.9	0.076
Blood urea nitrogen, mg/dL	28.1 ± 21.6	20.6 ± 13.9	< 0.001
Creatinine, mg/dL	2.0 ± 1.8	1.6 ± 1.9	0.015
Total bilirubin, mg/dL	1.1 ± 2.3	0.7 ± 0.4	0.001
Glucose, mg/dL	267.2 ± 165.2	219.2 ± 95.3	< 0.001
Albumin, g/dL	3.1 ± 0.7	3.8 ± 0.5	< 0.001
PT INR	2.1 ± 7.3	1.2 ± 0.4	0.015
Troponin I, ng/mL	1.7 ± 5.9	8.2 ± 50.3	0.139
D-dimer, μg/dL	18.6 ± 22.7	8.9 ± 15.7	< 0.001
pH	6.9 ± 0.2	7.2 ± 0.2	< 0.001
PCO_2, mmHg	75.7 ± 28.0	49.4 ± 23.6	< 0.001
PO_2, mmHg	73.9 ± 79.0	108.8 ± 110.4	0.057
HCO_3^-, mmol/L	23.4 ± 30.0	25.0 ± 26.6	0.745
Lactate, mmol/L	13.3 ± 6.3	10.5 ± 5.2	0.005

CPR = cardiopulmonary resuscitation, AED = automated external defibrillator, PEA = pulseless electrical activity, EMS = emergency medical service, ROSC = return of spontaneous circulation, ED = emergency department, GCS = glasgow coma scale, PT INR = prothrombin time international normalized ratio, PCO_2 = partial pressure of carbon dioxide, PO_2 = partial pressure of oxygen, HCO_3^- = bicarbonate ion.

Prediction performance

The average prediction performance of each model in the original develop dataset calculated by five-fold cross validation for each model with cutoff probability scores of 0.5 is described in **Supplementary Table 2**. The cutoff probability score of the ensemble model was set as 0.605, and the prediction performance of the ensemble model in the internal validation dataset is described in **Table 2**, **Fig. 2A and B**. When the ensemble model was implemented in the external validation dataset, overall prediction performance metrics such as accuracy, AUROC, AUPRC, and F1 score were all decreased by certain degrees compared with those in the internal validation dataset (**Table 2**, **Fig. 2C and D**). The average prediction performance of each model in the cardiac subgroup develop dataset calculated by five-fold cross validation for each model with cutoff probability scores of 0.5 is described in **Supplementary Table 3**. In the cardiac subgroup analysis, the cutoff probability score of the ensemble model was set as 0.525. Prediction performance was decreased in the cardiac subgroup internal validation dataset compared with that in the original internal validation dataset (**Table 2**, **Fig. 2E and F**), and prediction performance in the cardiac subgroup external validation dataset was also decreased (**Table 2**, **Fig. 2G and H**).

Table 2. Accuracy, AUROC, AUPRC, sensitivity, specificity, PPVs, NPVs, and F1 scores for the ensemble model in the internal validation dataset, the external validation dataset and the cardiac subgroups of the internal validation and the external validation datasets

Population	Dataset	Accuracy (95% CI)	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score
Original analysis	Internal validation	0.9620 (0.9352–0.9889)	0.9800 (0.9612–0.9988)	0.9950 (0.9860–1.0000)	0.9594 (0.9245–0.9943)	0.9714 (0.9162–1.0000)	0.9916 (0.9752–1.0000)	0.8718 (0.7669–0.9767)	0.9752
	External validation	0.8509 (0.7825–0.9192)	0.9301 (0.8845–0.9756)	0.9476 (0.9087–0.9867)	0.9595 (0.9145–1.0000)	0.6500 (0.5022–0.7978)	0.8353 (0.7564–0.9142)	0.8966 (0.7857–1.0000)	0.8931
Cardiac subgroup analysis	Internal validation	0.9661 (0.9191–1.0000)	0.9954 (0.9781–1.0000)	0.9959 (0.9797–1.0000)	1.0000 (1.0000–1.0000)	0.9286 (0.8332–1.0000)	0.9394 (0.8580–1.0000)	1.0000 (1.0000–1.0000)	0.9688
	External validation	0.6600 (0.5021–0.8179)	0.8917 (0.7906–0.9928)	0.8968 (0.7980–0.9956)	0.9000 (0.7685–1.0000)	0.5000 (0.3211–0.6789)	0.5455 (0.3756–0.7153)	0.8824 (0.7292–1.0000)	0.6792

The cutoff probability scores for unfavorable neurological outcomes were set at 0.605 and 0.525 for the ensemble models in the original and the cardiac subgroup analyses, respectively.
AUROC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve, PPV = positive predictive value, NPV = negative predictive value, CI = confidence interval.

https://doi.org/10.3346/jkms.2021.36.e187

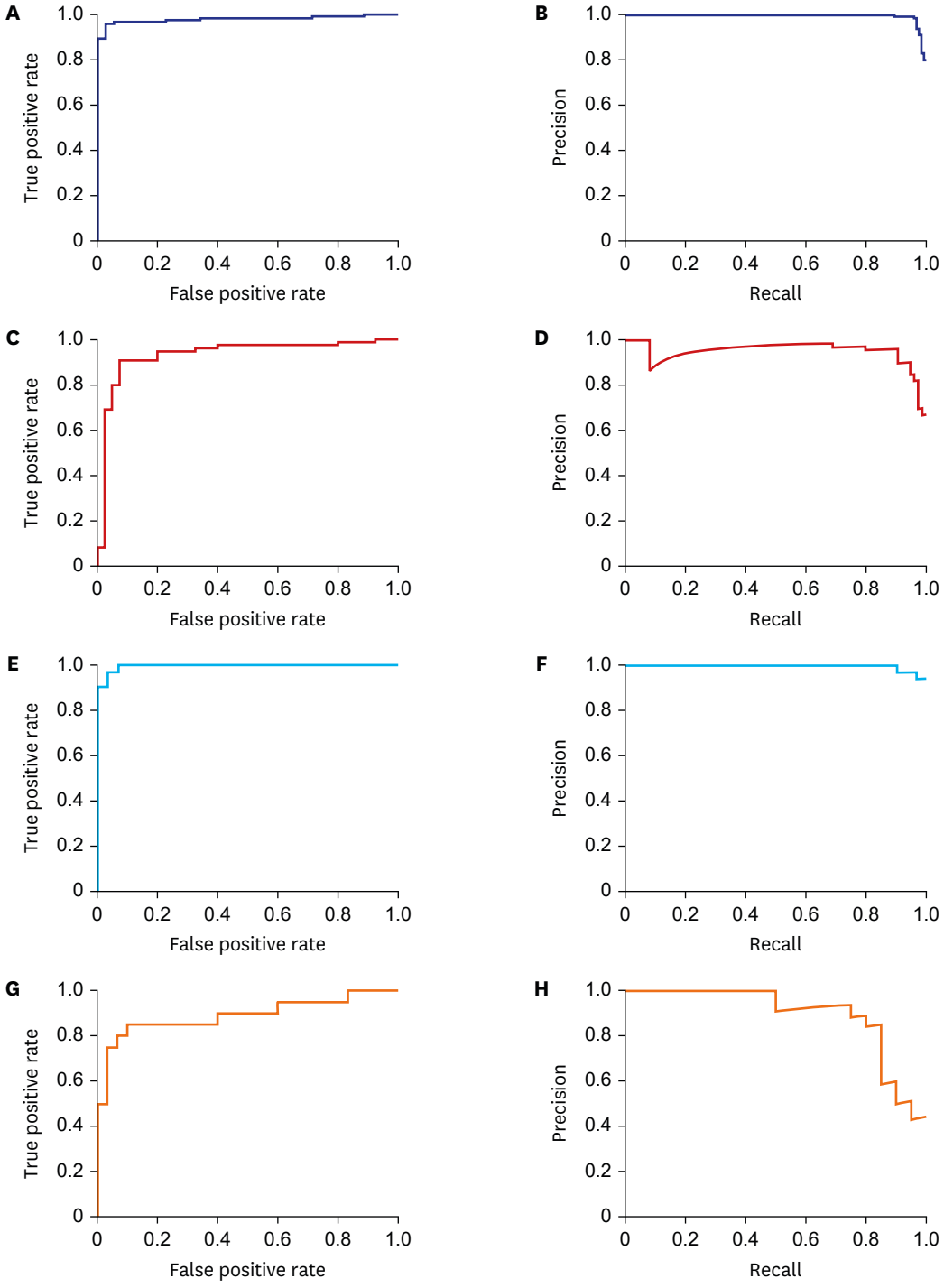

Fig. 2. Receiver operating characteristic curves and precision-recall curves for the ensemble prediction model in various datasets. (**A**) and (**B**) in the original internal validation dataset, (**C**) and (**D**) in the external validation dataset, (**E**) and (**F**) in the cardiac subgroup internal validation dataset, (**G**) and (**H**) in the cardiac subgroup external validation dataset.

Multivariable logistic regression models derived from the original develop dataset and the cardiac subgroup develop dataset are described in the **Supplementary Tables 4** and **5**, respectively. Most of the performance metrics were decreased in the logistic regression models compared with that in the ensemble models (**Table 3**). Only following metrics were

https://doi.org/10.3346/jkms.2021.36.e187

Table 3. Accuracy, AUROC, AUPRC, sensitivity, specificity, PPVs, NPVs, and F1 scores for the multivariable logistic regression model in the internal validation dataset, the external validation dataset and the cardiac subgroups of the internal validation and the external validation datasets

Population	Dataset	Accuracy (95% CI)	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score
Original analysis	Internal validation	0.9051 (0.8590–0.9512)	0.8777 (0.8239–0.9315)	0.9374 (0.9015–0.9734)	0.7632 (0.6280–0.8983)	0.9500 (0.9110–0.9890)	0.8286 (0.7037–0.9534)	0.9268 (0.8808–0.9729)	0.9383
	External validation	0.8421 (0.7716–0.9126)	0.8095 (0.7318–0.8871)	0.8337 (0.7613–0.9061)	0.8235 (0.6954–0.9517)	0.8500 (0.7718–0.9283)	0.7000 (0.5580–0.8420)	0.9189 (0.8567–0.9811)	0.8831
Cardiac subgroup analysis	Internal validation	0.8305 (0.7258–0.9352)	0.8318 (0.7275–0.9361)	0.7969 (0.6832–0.9106)	0.8000 (0.6569–0.9431)	0.8621 (0.7366–0.9876)	0.8571 (0.7275–0.9868)	0.8065 (0.6674–0.9455)	0.8333
	External validation	0.8000 (0.6678–0.9322)	0.8083 (0.6784–0.9383)	0.6621 (0.5044–0.8197)	0.8846 (0.7618–1.0000)	0.7083 (0.5265–0.8902)	0.7667 (0.6153–0.9180)	0.8500 (0.6935–1.0000)	0.7727

The cutoff probability scores for unfavorable neurological outcomes were set at 0.5 for the logistic regression model.
AUROC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve, PPV = positive predictive value, NPV = negative predictive value, CI = confidence interval.

better in the logistic regression models: NPV in the original internal validation dataset, NPV in the original external validation dataset, accuracy, specificity, PPV and F1 score in the cardiac subgroup external validation dataset. Receiver operating characteristics curves and precision-recall curves for the multivariable logistic regression models in each dataset are presented in the **Supplementary Fig. 1**.

DISCUSSION

In the present study, we established and validated a prediction model using an ensemble technique with four machine learning methods for the prediction of unfavorable 1-year neurological outcomes in OHCA survivors. The overall prediction performance of the ensemble model in the external validation set was favorable, with an AUROC of 0.9301 (95% CI, 0.8845–0.9756) and an AUPRC of 0.9476 (95% CI, 0.9087–0.9867). The prediction performance of the ensemble model in the cardiac subgroup external validation set was also good but not comparable with that in the original external validation set, with an AUROC of 0.8917 (95% CI, 0.7906–0.9928) and an AUPRC of 0.8968 (0.7980–0.9956). Performance metrics of the ensemble models were higher than that of the multivariable logistic regression models in general.

The prediction performance of certain machine learning methods is decreased when class imbalance is present in the develop dataset.[17] In the develop dataset of our cardiac subgroup, the number of patients with favorable 1-year neurological outcomes was 114 (48.3%) among 236, which means that the classes in the cardiac subgroup develop dataset were more balanced than those in the original develop dataset. In the present study, however, prediction performance in terms of the AUPRC was decreased in general in the cardiac subgroup compared with the original group. Despite relatively balanced classes in the cardiac subgroup, a smaller sample size might have carried a higher risk of model overfitting, which might have resulted in slightly decreased prediction performance.

Early neurologic prognostication after cardiac arrest is important to avoid obvious futile treatment or inappropriate withdrawal of postcardiac arrest care. Current international guidelines recommend that neurologic prognostication be performed using multiple modalities, including clinical examination findings, serum biomarkers and electrophysiological tests.[4-6] It is also recommended that the timing of prognostication be delayed for at least 72 hours after ROSC.[4-6]

https://doi.org/10.3346/jkms.2021.36.e187

Several studies have evaluated the prediction performance of machine learning-based models for neurological outcomes after OHCA. Kwon et al.[9] used national OHCA registry data to develop a deep learning-based prediction model, the prediction performance of which was better than conventional machine learning-based models. These authors' model did not include hospital variables except the ED visit to ROSC time, and the study endpoints were short-term neurological outcome and survival discharge. Seki et al.[8] used the RF model to predict 1-year survival in OHCA patients with presumed cardiac etiology without predicting long-term functional outcomes. Park et al.[10] also developed machine learning-based prediction models for neurological outcomes at discharge in OHCA patients; however, long-term neurological outcome was not the scope of the study.

Aside from the overall performance of the prediction models, one of the most important issues is minimizing false positive prediction for unfavorable neurological outcomes when predicting neurological outcomes of cardiac arrest survivors. False positive prediction can lead to withdrawal of intensive postcardiac arrest care from patients who otherwise may fully or nearly fully recover and return to daily life. To exclude the possibility of false positive prediction, recent guidelines recommended the use of prognostic measures with false positive rates lower than or equal to 1%, i.e., with specificity higher than 99%.[4-6,18] We set cutoff probability scores in each prediction model with which the F1 score is maximized. Although the specificity of the ensemble model in the original internal validation dataset scored 0.9714 (0.9162–1.0000) with a cutoff probability score of 0.605, which is not over 99% but is acceptable, the specificity was significantly reduced (0.6500 [95% CI, 0.5022–0.7978]) when it was implemented in the external validation dataset. Although we trained the prediction models comprising the ensemble model in a separate develop dataset, the prediction performance was different in the internal validation and external validation datasets. Both datasets were hold-out datasets, which had never been involved in model training. However, the internal validation dataset was collected in the same period in which the develop dataset was acquired, and the external validation dataset was collected thereafter. The internal validation dataset was more likely to be similar to the develop dataset than the external validation dataset, and the difference in similarity between the two datasets might have resulted in different prediction performances.

The previously reported specificity of machine learning methods for the prediction of unfavorable outcomes in cardiac arrest victims ranged from 66.7% to 95.3%.[9,10] and the ensemble prediction models in the present study outperformed the previous models in terms of specificity in the internal validation dataset. The major difference between our study and previous studies is that we included laboratory variables to train and to establish prediction models. Initial laboratory data immediately after ROSC have a significant association with neurological outcomes in cardiac arrest survivors.[11-16] Establishing prediction models by adding widely available laboratory data might have contributed to the improvement of model performance, despite a smaller sample size than those of previous studies.

One of the strengths of our study is that we developed a neurological outcome prediction model that can be implemented immediately after ROSC in OHCA survivors. Earlier timing of prognostication than currently recommended by guidelines[4-6] may aid medical personnel and the guardians of the OHCA victims in shared decision on implementation of intensive care or withdrawal of life-sustaining treatment. We used laboratory variables that had been initially obtained at the timing of ED arrival. Previous studies using machine learning models for neurological outcomes in OHCA patients did not include laboratory values in the

prediction models,[8-10] which might have improved prediction performance if they had been included. We defined 1-year neurological outcomes as the primary outcome, which was not focused on previous studies.[8-10] As the hospital cost of caring for cardiac arrest survivors is considerable.[19-21] our study may have a role in reducing the socioeconomic burden associated with potentially futile treatment.

There are a couple of factors to consider before implementation of our prediction model in clinical field to aid clinical decisions. First, prognostic measures that are considered reasonable for neuroprognostication in the guidelines showed specificity higher than or equal to 99%.[4-6,18] As our prognostic model could not reach such high specificity for unfavorable neurological outcomes, performance improvement is essential before clinical implementation, especially in terms of specificity. We hope organizing dataset with a large number of medical centers may improve specificity of the prediction model, without compromising sensitivity. Second, prognostic measures that are available immediately after ROSC, such as gray-white matter ratio,[22] may improve prognostic performance of the model when added. Furthermore, as guidelines recommend multimodal approach for neuroprognostication, our prediction model may help clinical decision by providing outcome probability as one of the prognostic measures, not by simply discriminating the prognosis into favorable outcomes or unfavorable outcomes.

Our study has several limitations. First, the small sample size compared with previous studies reduced the statistical power of the results.[8-10] Considering that the rate of survival to ED arrival in OHCA patients is approximately one-fourth,[2,3] the number of participants in our study may be larger than it was thought needed to be. Second, although we performed an external validation with the ensemble prediction model, the external validation dataset was too small. Moreover, the prediction performance of the model in the external validation set showed a potential risk of overfitting, which may impede the generalizability of the study results. However, we performed the analyses with a multicenter registry, and the multicenter nature of the study may attenuate this weakness. Finally, we did not include several prognostic tools that are suggested in the current guidelines, such as neuron-specific enolase or quantitative pupillometry. These tests are not always routinely performed in small centers; therefore, the exclusion of those variables from the models is reasonable in view of practical use.

In conclusion, we established an ensemble prediction model for prediction of unfavorable 1-year neurological outcomes in OHCA survivors using four machine learning methods. The prediction performance of the ensemble model was higher than the multivariable logistic regression model, while its performance was slightly decreased in the external validation dataset.

ACKNOWLEDGMENTS

This study was conducted based on the OHCA registry constructed by SNU CARE investigators.

SUPPLEMENTARY MATERIALS

Supplementary Table 1

Variables included in the machine learning analysis

Click here to view

Supplementary Table 2

Accuracy, AUROC, AUPRC, sensitivity, specificity, PPV, and NPV for each classifier in the original develop dataset

Click here to view

Supplementary Table 3

Accuracy, AUROC, AUPRC, sensitivity, specificity, PPV, and NPV for each classifier in the cardiac subgroup develop dataset

Click here to view

Supplementary Table 4

Multivariable logistic regression analysis for unfavorable neurological outcomes in the original develop dataset

Click here to view

Supplementary Table 5

Multivariable logistic regression analysis for unfavorable neurological outcomes in the cardiac subgroup develop dataset

Click here to view

Supplementary Fig. 1

(A) Receiver operating characteristic curve and **(B)** precision-recall curve for the multivariable logistic regression model in the original internal validation dataset, **(C)** receiver operating characteristic curve and **(D)** precision-recall curve for the multivariable logistic regression model in the external validation dataset, **(E)** receiver operating characteristic curve and **(F)** precision-recall curve for the multivariable logistic regression model in the cardiac subgroup internal validation dataset and **(G)** receiver operating characteristic curve and **(H)** precision-recall curve for the multivariable logistic regression model in the cardiac subgroup external validation dataset.

Click here to view

REFERENCES

1. Holmberg MJ, Ross CE, Fitzmaurice GM, Chan PS, Duval-Arnould J, Grossestreuer AV, et al. Annual incidence of adult and pediatric in-hospital cardiac arrest in the United States. *Circ Cardiovasc Qual Outcomes* 2019;12(7):e005580.
PUBMED | CROSSREF
2. Hawkes C, Booth S, Ji C, Brace-McDonnell SJ, Whittington A, Mapstone J, et al. Epidemiology and outcomes from out-of-hospital cardiac arrests in England. *Resuscitation* 2017;110:133-40.
PUBMED | CROSSREF
3. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. *Circulation* 2020;141(9):e139-596.
PUBMED | CROSSREF

4. Nolan JP, Soar J, Cariou A, Cronberg T, Moulaert VR, Deakin CD, et al. European resuscitation council and European society of intensive care medicine guidelines for post-resuscitation care 2015: section 5 of the European resuscitation council guidelines for resuscitation 2015. *Resuscitation* 2015;95:202-22.
PUBMED | CROSSREF
5. Panchal AR, Bartos JA, Cabanas JG, Donnino MW, Drennan IR, Hirsch KG, et al. Part 3: adult basic and advanced life support: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* 2020;142(16_suppl_2):S366-468.
PUBMED | CROSSREF
6. Kim YM, Park KN, Choi SP, Lee BK, Park K, Kim J, et al. Part 4. post-cardiac arrest care: 2015 Korean guidelines for cardiopulmonary resuscitation. *Clin Exp Emerg Med* 2016;3(Suppl):S27-38.
PUBMED | CROSSREF
7. Dale CM, Sinuff T, Morrison LJ, Golan E, Scales DC. Understanding early decisions to withdraw life-sustaining therapy in cardiac arrest survivors. a qualitative investigation. *Ann Am Thorac Soc* 2016;13(7):1115-22.
PUBMED | CROSSREF
8. Seki T, Tamura T, Suzuki MSOS-KANTO 2012 Study Group. Outcome prediction of out-of-hospital cardiac arrest with presumed cardiac aetiology using an advanced machine learning technique. *Resuscitation* 2019;141:128-35.
PUBMED | CROSSREF
9. Kwon JM, Jeon KH, Kim HM, Kim MJ, Lim S, Kim KH, et al. Deep-learning-based out-of-hospital cardiac arrest prognostic system to predict clinical outcomes. *Resuscitation* 2019;139:84-91.
PUBMED | CROSSREF
10. Park JH, Shin SD, Song KJ, Hong KJ, Ro YS, Choi JW, et al. Prediction of good neurological recovery after out-of-hospital cardiac arrest: a machine learning analysis. *Resuscitation* 2019;142:127-35.
PUBMED | CROSSREF
11. Daou O, Winiszewski H, Besch G, Pili-Floury S, Belon F, Guillon B, et al. Initial pH and shockable rhythm are associated with favorable neurological outcome in cardiac arrest patients resuscitated with extracorporeal cardiopulmonary resuscitation. *J Thorac Dis* 2020;12(3):849-57.
PUBMED | CROSSREF
12. Kiehl EL, Amuthan R, Adams MP, Love TE, Enfield KB, Gimple LW, et al. Initial arterial pH as a predictor of neurologic outcome after out-of-hospital cardiac arrest: a propensity-adjusted analysis. *Resuscitation* 2019;139:76-83.
PUBMED | CROSSREF
13. Bender PR, Debehnke DJ, Swart GL, Hall KN. Serum potassium concentration as a predictor of resuscitation outcome in hypothermic cardiac arrest. *Wilderness Environ Med* 1995;6(3):273-82.
PUBMED | CROSSREF
14. Choi DS, Shin SD, Ro YS, Lee KW. Relationship between serum potassium level and survival outcome in out-of-hospital cardiac arrest using CAPTURES database of Korea: Does hypokalemia have good neurological outcomes in out-of-hospital cardiac arrest? *Adv Clin Exp Med* 2020;29(6):727-34.
PUBMED | CROSSREF
15. Tamura T, Suzuki M, Hayashida K, Sasaki J, Yonemoto N, Sakurai A, et al. Renal function and outcome of out-of-hospital cardiac arrest - multicenter prospective study (SOS-KANTO 2012 Study). *Circ J* 2018;83(1):139-46.
PUBMED | CROSSREF
16. D'Arrigo S, Cacciola S, Dennis M, Jung C, Kagawa E, Antonelli M, et al. Predictors of favourable outcome after in-hospital cardiac arrest treated with extracorporeal cardiopulmonary resuscitation: a systematic review and meta-analysis. *Resuscitation* 2017;121:62-70.
PUBMED | CROSSREF
17. Li DC, Hu SC, Lin LS, Yeh CW. Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. *PLoS One* 2017;12(8):e0181853.
PUBMED | CROSSREF
18. Callaway CW, Donnino MW, Fink EL, Geocadin RG, Golan E, Kern KB, et al. Part 8: post-cardiac arrest care: 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* 2015;132(18 Suppl 2):S465-82.
PUBMED | CROSSREF
19. Graf J, Mühlhoff C, Doig GS, Reinartz S, Bode K, Dujardin R, et al. Health care costs, long-term survival, and quality of life following intensive care unit admission after cardiac arrest. *Crit Care* 2008;12(4):R92.
PUBMED | CROSSREF
20. Efendijev I, Folger D, Raj R, Reinikainen M, Pekkarinen PT, Litonius E, et al. Outcomes and healthcare-associated costs one year after intensive care-treated cardiac arrest. *Resuscitation* 2018;131:128-34.
PUBMED | CROSSREF

21. Damluji AA, Al-Damluji MS, Pomenti S, Zhang TJ, Cohen MG, Mitrani RD, et al. Health care costs after cardiac arrest in the United States. *Circ Arrhythm Electrophysiol* 2018;11(4):e005689.
PUBMED | CROSSREF

22. Hong JY, Lee DH, Oh JH, Lee SH, Choi YH, Kim SH, et al. Grey-white matter ratio measured using early unenhanced brain computed tomography shows no correlation with neurological outcomes in patients undergoing targeted temperature management after cardiac arrest. *Resuscitation* 2019;140:161-9.
PUBMED | CROSSREF