Available online at www.sciencedirect.com

Resuscitation

journal homepage: www.elsevier.com/locate/resuscitation

Clinical paper

Prediction of good neurological recovery after out-of-hospital cardiac arrest: A machine learning analysis

Jeong Ho Park[a,b], *Sang Do Shin*[b,*], *Kyoung Jun Song*[b], *Ki Jeong Hong*[b], *Young Sun Ro*[c], *Jin-Wook Choi*[a], *Sae Won Choi*[b,d]

[a] *Department of Biomedical Engineering, Seoul National University College of Medicine*
[b] *Department of Emergency Medicine, Seoul National University College of Medicine*
[c] *Laboratory of Emergency Medical Services, Seoul National University Hospital Biomedical Research Institute*
[d] *Office of Hospital Information, Seoul National University Hospital*

Abstract

Background: This study aimed to train, validate and compare predictive models that use machine learning analysis for good neurological recovery in OHCA patients.

Methods: Adult OHCA patients who had a presumed cardiac etiology and a sustained return of spontaneous circulation between 2013 and 2016 were analyzed; 80% of the individuals were analyzed for training and 20% were analyzed for validation. We developed using six machine learning algorithms: logistic regression (LR), extreme gradient boosting (XGB), support vector machine, random forest, elastic net (EN), and neural network. Variables that could be obtained within 24 hours of the emergency department visit were used. The area under the receiver operation curve (AUROC) was calculated to assess the discrimination. Calibration was assessed by the Hosmer–Lemeshow test. Reclassification was assessed by using the continuous net reclassification index (NRI).

Results: A total of 19,860 OHCA patients were included in the analysis. Of the 15,888 patients in the training group, 2228 (14.0%) had a good neurological recovery; of the 3972 patients in the validation group, 577 (14.5%) had a good neurological recovery. The LR, XGB, and EN models showed the highest discrimination powers (AUROC (95% CI)) of 0.949 (0.941–0.957) for all), and all three models were well calibrated (Hosmer–Lemeshow test: $p > 0.05$). The XGB model reclassified patients according to their true risk better than the LR model (NRI: 0.110), but the EN model reclassified patients worse than the LR model (NRI: −1.239).

Conclusion: The best performing machine learning algorithm was the XGB and LR algorithm.

Keywords: Out-of-hospital cardiac arrest, Outcome, Machine learning analysis

Introduction

Out-of-hospital cardiac arrest (OHCA) is a disease with a significant public health burden and has a low survival rate and a high disability rate.[1–3] The outcomes of OHCA depend on multiple variables, which include the patient, community, emergency medical service (EMS), and hospital care. Some of these factors can be changed, while others cannot.[4]

Various prediction models for OHCA outcomes that used traditional biostatistical methods were developed but were either not reliable or were not valid.[5–10] Machine learning analysis is a new

* Corresponding author.
E-mail addresses: timthe@gmail.com (J.H. Park), shinsangdo@gmail.com (S.D. Shin), skciva@gmail.com (K.J. Song), emkjhong@gmail.com (K.J. Hong), Ro.youngsun@gmail.com (Y.S. Ro), jinchoi@snu.ac.kr (J.-W. Choi), saewonchoi@gmail.com (S.W. Choi).

https://doi.org/10.1016/j.resuscitation.2019.07.020
Received 1 March 2019; Received in revised form 28 June 2019; Accepted 16 July 2019

technology for predicting a cardiac event by using a large sample of data with multiple and complex interactions among variables. Previous studies have reported the prediction performance of the incidence of cardiac arrest rather than the outcomes after cardiac arrest events.[11–13] Additionally, various algorithms have been developed to improve the prediction performance. One study compared six machine learning algorithms for the prediction of ROSC and survival. However, a limited number of variables and a small sample of less than 500 patients were used in that study.[14]

The aim of this study was to train, validate and compare predictive models for good neurological recovery by using machine learning algorithms in OHCA patients.

Methods

Study design and setting

This study was a cross-sectional study that used a nationwide, prospective, EMS-based OHCA registry in Korea. The EMS system is exclusively operated by the National Fire Agency. EMS providers can administer CPR with the use of automatic defibrillators at the scene and during transport and can provide limited advanced life support (ALS), including intravenous fluids, endotracheal intubation, or supraglottic airway insertion under direct medical control of a physician.[15] The EMS provider cannot declare a state of death or stop CPR unless the patient regains a pulse. All EMS-assessed patients are transported to the nearest hospital emergency department (ED) by the standard EMS CPR protocol.

Data source

The Korean OHCA registry, which monitors all incident cases of EMS-assessed OHCA in the country, was retrieved from the following four sources: the EMS run sheets for basic ambulance operation information, the EMS cardiac arrest registry, the dispatcher CPR registry for the Utstein factors, and the hospital medical record review registry for hospital care and outcomes. A detailed description of the data acquisition of each registry, as well as the training and quality of the medical record reviewers, are described in previous studies.[16,17]

Study population

Patients with OHCA who had presumed cardiac etiologies, who were aged 18 years or older and who gained sustained ROSC in EDs between 2013 to 2016 were included in the analysis. Patients were excluded if they had missing information regarding their neurological statuses at the time of their hospital discharges or if they had missing information in covariables that included witness status, bystander CPR, initial rhythm, and response time. Patients who had missing information concerning the start time of the post-resuscitation care, including TTM, PCI, and ECMO therapies, were also excluded.

Main outcome

The primary outcome of the study was a good neurological recovery at the time of discharge from the hospital. Good neurological recovery was recorded if the patient had a cerebral performance category 1 or 2.

Variables and preprocessing

We collected the patients' demographic, community, EMS and hospital care information. A total of 22 variables, including the primary outcome, were used in the analysis. Detailed descriptions of the variables are presented in Supplementary Table 1. Most of the continuous variables were preprocessed with centering and scaling, and the categorical variables were preprocessed with the one-hot encoding (dummy variable encoding) method. EDs were categorized into either a high-volume ED, which indicated that the annual mean cardiac arrest patient volumes of the EDs were more than or equal to 40, or a low volume ED, which indicated that the annual mean cardiac arrest patient volumes of the EDs were less than 40. The cutoff volume point was derived from previous studies.[18,19] Both no flow time and low flow time were categorized into four quartiles. Patients who had missing information regarding no flow and low flow times were incorporated into the "not calculated" category because we thought that low flow and no flow time data were not randomly missing and provided additional information if they were missing. For example, in patients with prolonged arrests, information might often be missing concerning these variables. Treating missing information as another categorical variable in clinical research that uses machine learning algorithms has been utilized in previous studies.[20,21] For post-resuscitation care, information about care that was provided within 24 h of the ED visit was collected. For each of the post-resuscitation therapies, the variables were categorized according to the quartile of time from ED arrival to the start of each therapy, because the time to post-resuscitation care is known to be important for patient outcomes.[22–24] For patients who did not undergo post-resuscitation care, an additional "not conducted" category was included for each category of post-resuscitation care.

Model development

We developed prediction models for good neurological recovery by using the following six machine learning algorithms: logistic regression (LR), extreme gradient boosting (XGB), support vector machine (SVM), random forest (RF), elastic net (EN), and neural net (NN). The LR algorithm was chosen as the baseline comparison algorithm because it is commonly used in the medical field and has been used for previous prediction model development in the study of OHCA.[5,8,10] The other five algorithms were selected based on their ability to model nonlinear associations, on their relative ease of implementation, and on their general acceptance in the machine learning community.[21,25–29] Details of each Model and hyperparameter tuning processes are described in the Supplementary Methods 1. All algorithms are either a function that map the predictors to a value between 0 and 1 that corresponds to the probability of the outcome occurring or has a method to calculate the probability.

The study population was split into a training cohort from which each of the machine learning prediction models were derived and a validation cohort in which the prediction models were applied and tested. The training cohort was derived from a random sampling of 80% of the entire cohort within each of the outcome categories in order to preserve the overall distribution of the outcomes; the validation cohort comprised the remaining 20%.

Statistical analysis

The demographic findings and survival outcomes of the study population were described in this study. Additionally, the baseline characteristics of the training cohort and the validation cohort were compared. The continuous variables were compared by using Student's T-test or the Wilcoxon rank sum test, and the categorical variables were compared by using the chi-squared test or the Fisher exact test, as appropriate. Unadjusted odds ratios (ORs), with 95% confidence intervals (CIs) for each variable for the study outcome in the training cohort, were calculated by using the LR analysis to assess the unadjusted association between each of the variables and the study outcome.

We assessed the discrimination performance by comparing the area under the receiver operating characteristic curve (AUROC) for each model in the validation cohort. We assessed the calibration power by using the Hosmer–Lemeshow test, the scaled Brier score, and a calibration plot in the validation cohort.[30] The test characteristics of each of the models in the validation cohort, including the sensitivity, specificity, and positive and negative predictive values with 95% CIs, were reported. The cutoff probability used to assess the test characteristics is 0.5 in all models. The added prognostic power of each prediction model compared to the LR model was also evaluated by continuous net reclassification index (NRI). NRI is a statistical method to quantify how well a new model correctly reclassifies study population with the other models. Details of NRI is described in Supplementary method 2. By using a model specific metric, the variable importance of each model was assessed, except for the SVM algorithm. The variable importance was determined by the coefficient effect sizes for the LR model. The XGB and RF models were ranked by variable importance on the selection frequency of the variable as a decision node. The absolute value of the coefficients corresponding to the tuned model were used for the measurement of variable importance in the EN algorithm. The NN algorithm used an overall weighting of the variable within the model.[31]

Two additional analyses were conducted to compare the machine learning algorithms. First, we developed another prediction model by using different variable sets. Two additional variable sets were implemented: 1) a prehospital variable set (set 1), which included all variables except for no flow time, low flow time, and the post-resuscitation care variables (PCI, TTM and ECMO) and represented the variables that could be obtained immediately at the ED visit; 2) an ED variable set (set 2), which included the variables except for the post-resuscitation care variables; and 3) an all variable set (set 3), which was used in our main analyses and represented the variables that could be obtained within 24 hours of ED visit. The discrimination and reclassification improvements were calculated for the validation cohort to assess how each machine learning algorithms adapt additional variables in model to improve performance. Second, by using a cutoff probability of 0.01, we assessed the test characteristics

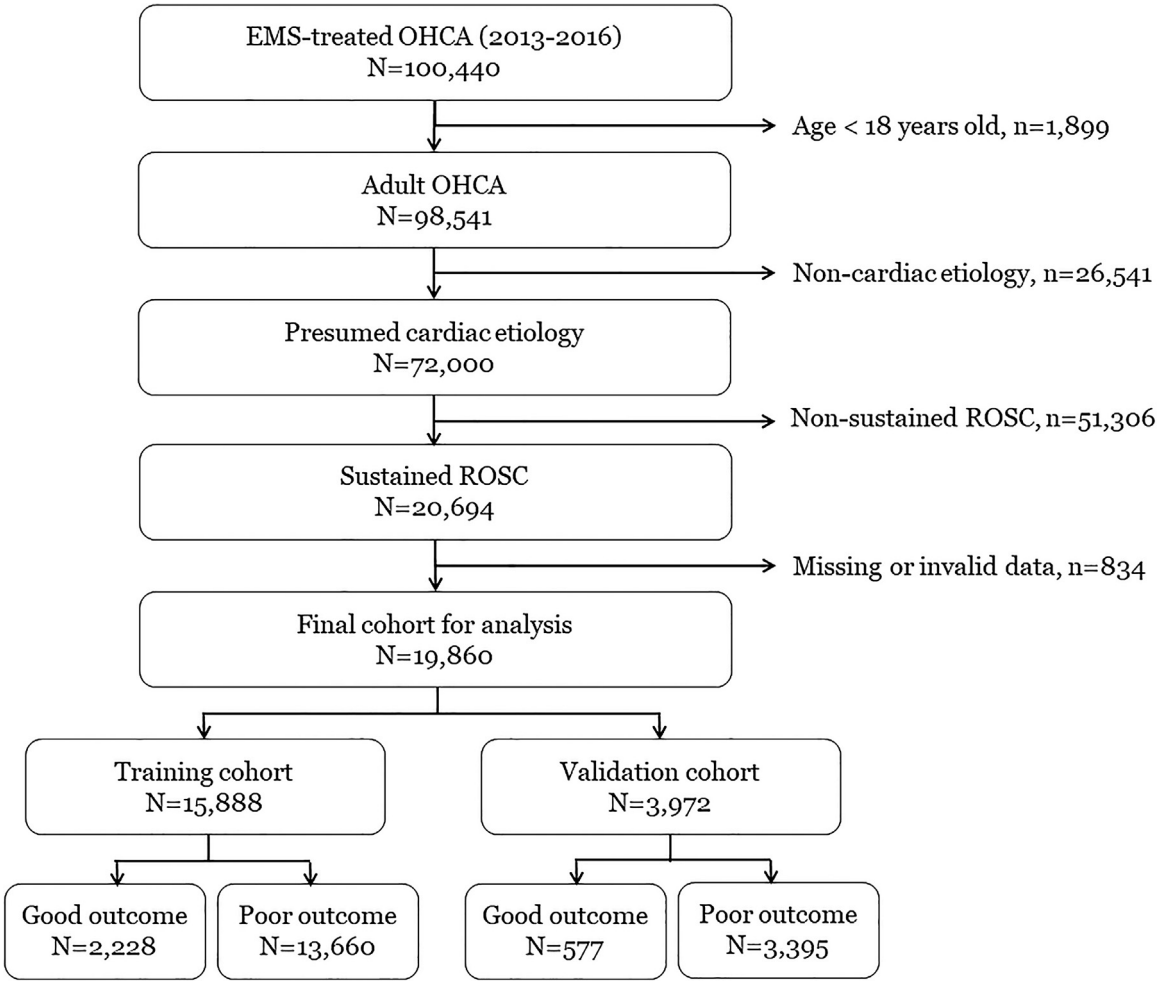

Fig. 1 – Flow diagram. OHCA: out of hospital cardiac arrest; ROSC: return of spontaneous circulation.

Table 1 – Distribution and univariable logistic regression analyses for good neurological recovery in the training cohorts.

Characteristics	*N* (%)	Good neurological recovery	
		(%)	OR (95% CI)
All	15,888	14.0	
Age			
mean (SD) [years]	66.3 (15.3)	N/A	0.95 (0.95–0.95)*
Gender			
Female	5363 (33.8%)	8.3	Reference
Male	10,525 (66.2%)	17.0	2.26 (2.03–2.52)
Diabetes			
No	11,847 (74.6%)	15.6	Reference
Yes	4041 (25.4%)	9.3	0.55 (0.49–0.62)
Hypertension			
No	9678 (60.9%)	14.4	Reference
Yes	6210 (39.1%)	13.4	0.92 (0.83–1.00)
Heart disease			
No	12,886 (81.1%)	12.9	Reference
Yes	3002 (18.9%)	19.1	1.6 (1.44–1.77)
Stroke			
No	14,481 (91.1%)	14.6	Reference
Yes	1407 (8.9%)	7.9	0.5 (0.41–0.61)
Cancer			
No	14,209 (89.4%)	15.0	Reference
Yes	1679 (10.6%)	6.1	0.37 (0.30–0.46)
Place of arrest			
Public	3515 (22.1%)	25.8	Reference
Private	10,278 (64.7%)	9.6	0.31 (0.28–0.34)
Ambulance	1875 (11.8%)	14.5	0.49 (0.42–0.57)
Others	220 (1.4%)	28.2	1.13 (0.83–1.53)
Witness status			
Not witnessed	5490 (34.6%)	7.4	Reference
Layperson, or first responders	8070 (50.8%)	18.4	2.83 (2.52–3.17)
EMS-provider	2328 (14.7%)	14.6	2.15 (1.84–2.50)
Bystander CPR			
No	8199 (51.6%)	10.6	Reference
Without dispatcher assistance	1810 (11.4%)	18.2	1.88 (1.64–2.16)
With dispatcher assistance	5879 (37.0%)	17.6	1.80 (1.63–1.98)
Bystander AED use			
No	15,386 (96.8%)	14.0	Reference
Yes	502 (3.2%)	14.3	1.03 (0.80–1.32)
Bystander defibrillation			
No	15,747 (99.1%)	13.9	Reference
Yes	141 (0.9%)	30.5	2.72 (1.9–3.91)
Response time, median (IQR), min	6.0 (5.0;9.0)	N/A	0.94 (0.93–0.95)†
Initial rhythm			
Shockable	3355 (21.1%)	51.9	Reference
PEA	2616 (16.5%)	8.2	0.08 (0.07–0.10)
Asystole	9917 (62.4%)	2.8	0.03 (0.02–0.03)
Prehospital IV line placement			
No	12,896 (81.2%)	12.8	Reference
Yes	2992 (18.8%)	19.3	1.62 (1.46–1.80)
Prehospital advanced airway management			
No	963 (6.1%)	13.5	Reference
Endotracheal intubation	3808 (24.0%)	14.4	1.08 (0.88–1.32)
Supraglottic airway	11,117 (70.0%)	13.9	1.04 (0.86–1.26)
Prehospital mechanical CPR device use			
No	15,688 (98.7%)	14.1	Reference
Yes	200 (1.3%)	11.5	0.79 (0.51–1.23)
No flow time			
0 min	4150 (26.1%)	17.4	
1–4 min	2915 (18.3%)	22.4	1.37 (1.22–1.55)
4–9 min	3168 (19.9%)	14.4	0.27 (0.22–0.32)
11– min	3257 (20.5%)	5.3	0.80 (0.70–0.91)
Not calculated	2398 (15.1%)	9.3	0.49 (0.42–0.57)

Table 1 *(continued)*

Characteristics	*N* (%)	Good neurological recovery	
		(%)	OR (95% CI)
Low flow time			
0–20 min	3746 (23.6%)	42.2	Reference
21–34 min	3600 (22.7%)	9.4	0.14 (0.12–0.16)
35–54 min	3564 (22.4%)	2.9	0.04 (0.03–0.05)
55- min	3630 (22.8%)	1.7	0.02 (0.02–0.03)
Not calculated	1348 (8.5%)	11.1	0.17 (0.14–0.21)
Annual volume of ED			
<40	741 (4.7%)	4.2	Reference
≥40	15,147 (95.3%)	14.5	3.89 (2.70–5.58)
PCI			
Not done	14,673 (92.4%)	11.1	Reference
0–70 min	305 (1.9%)	60.7	12.34 (9.75–15.61)
71–92 min	303 (1.9%)	51.5	8.49 (6.74–10.70)
93–137 min	298 (1.9%)	38.5	5.01 (3.96–6.34)
138- min	309 (1.9%)	46.3	6.90 (5.46–8.72)
TTM			
Not done	14,411 (90.7%)	12.2	Reference
0–90 min	378 (2.4%)	33.6	3.65 (2.93–4.54)
91–153 min	362 (2.3%)	34.3	3.75 (3.01–4.69)
154–240 min	369 (2.3%)	32.2	3.43 (2.74–4.29)
241- min	368 (2.3%)	27.7	2.76 (2.19–3.49)
ECMO			
Not done	15,495 (97.5%)	14.1	Reference
0–50 min	101 (0.6%)	14.9	1.06 (0.61–1.84)
51–82 min	101 (0.6%)	3.0	0.19 (0.06–0.59)
83–153 min	93 (0.6%)	9.7	0.65 (0.33–1.3)
154- min	98 (0.6%)	15.3	1.10 (0.63–1.91)

OR, odds ratio; 95% CI, 95% confidence interval; CPR, cardiopulmonary resuscitation; AED, automated electrical defibrillation; IV, intravenous; ED, emergency department; PCI, percutaneous coronary intervention; TTM, targeted temperature management; ECMO, extracorporeal membrane oxygenation.
†Response time: odd ratios were calculated per 1-min increase.
* Age: odds ratios were calculated per 1-year increase.

of each prediction model in the validation cohort. The cutoff probability of 0.01 was chosen because 99% sensitivity for good neurological recovery is a commonly used threshold for the termination of resuscitation rules (TOR). Therefore, this analysis was used to compare the performance of each of the models as a TOR tool.[32]

All statistical analyses were performed by using R version 3.5.1. with packages included caret, e1071, xgboost, randomForest, glmnet, and nnet for the analysis of the machine learning algorithms.

Ethical statements

This study complied with the Declaration of Helsinki, and its protocol was approved by the Institutional Review Board on the study site with a waiver of informed consent.

Results

Demographic findings

Among the 100,440 EMS-treated OHCA patients, 19,860 patients were included in the final analysis. This cohort was split into 2 samples: an 80% sample, consisting of 15,888 patients in the training cohort, and a remaining sample of 3972 patients who was used for validation (Fig. 1).

From the total training cohort of 15,888 patients, the survival-to-discharge rate was 23.4%, and the good neurological recovery rate was 14.0%. There was no significant difference in the baseline characteristics between the training cohort and the validation cohort (Supplementary Table 2). Table 1 shows the association between the predictor variables and good neurological recovery (Table 1).

Main analyses

The classification results of the machine learning models on the validation cohorts are presented in Table 2 and Supplementary Fig. 1. The LR, XGB, and EN models had the highest AUROC (AUROC, 95% CI) at 0.949 (0.941–0.957), 0.949 (0.941–0.957), and 0.949 (0.941–0.957), respectively. There were no significant differences in the AUROC among the LR, XGB, and EN models. The XGB and RF models resulted in continuous reclassification improvement compared to the LR model (NRI, 95% CI): 0.110 (0.022–0.198) and 0.305 (0.218–0.393), respectively. All other algorithms showed worse reclassifications than the LR model especially in the EN model (NRI, 95% CI): −1.239 (−1.309 to −1.169) (Table 2). Supplementary Fig. 2 shows the calibration plot for each of the machine learning models. Calibration was poor for the SVM and RF models (Hosmer–Lemeshow test: all p values <0.001). Supplementary Table 3 shows the top 10 most important variables for each of the prediction models.

Table 2 – Discrimination, reclassification and test characteristics of good neurological recovery prediction models on validation cohorts.

Model	Discrimination		Reclassification		Test characteristics			
	AUROC (95% CI)	*p*-value*	NRI (95% CI)	*p*-value*	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
LR	0.949 (0.941, 0.957)	N/A	N/A	N/A	66.7 (62.7, 70.6)	96.6 (95.9, 97.2)	77.0 (73.1, 80.6)	94.5 (93.7, 95.2)
XGB	0.949 (0.941, 0.957)	0.975	0.110 (0.022, 0.198)	0.014	66.0 (62.0, 69.9)	96.7 (96.0, 97.3)	77.3 (73.3, 80.9)	94.4 (93.5, 95.1)
SVM	0.943 (0.934, 0.952)	0.009	−0.585 (−0.671, −0.499)	$p < 0.001$	62.9 (58.8, 66.9)	96.8 (96.2, 97.4)	77.2 (73.2, 80.9)	93.9 (93.0, 94.7)
RF	0.937 (0.926, 0.948)	<0.001	0.305 (0.218, 0.393)	$p < 0.001$	66.7 (62.7, 70.6)	96.6 (95.9, 97.1)	76.7 (72.7, 80.3)	94.5 (93.7, 95.2)
EN	0.949 (0.941, 0.957)	0.639	−1.239 (−1.309, −1.169)	$p < 0.001$	66.6 (62.5, 70.4)	96.6 (96.0, 97.2)	77.1 (73.2, 80.7)	94.4 (93.6, 95.2)
NN	0.942 (0.931, 0.952)	0.025	−0.382 (−0.468, −0.295)	$p < 0.001$	70.4 (66.5, 74.1)	95.8 (95.1, 96.5)	74.2 (70.3, 77.8)	95.0 (94.2, 95.7)

LR, logistic regression; XGB, extreme gradient boosting; SVM, support vector machine; RN, random forest; EN, elastic net; NN, neural network; AUROC, area under the receiver operating characteristic curve; NRI, net reclassification index; 95% CI, 95% confidence interval; PPV, positive predictive value; NPV, negative predictive value

* *p*-value for comparison with LR algorithms.

Table 3 shows the performance change of good neurological recovery prediction among the models that used different variable sets. The AUROC was increased by more than 3.0% when no flow time and low flow time were added to the variables obtained immediately after ROSC. Although the magnitude was less than 1%, the AUROC was also significantly increased when the post-resuscitation care variables were added to the variable list, except for the NN model. The NRI significantly increased in most of the cases when variables were added. However, the NN model with the post-resuscitation care variables showed a worse reclassification than the model without the hospital variables (Table 4).

Table 4 shows the test characteristics of each prediction model when using the cutoff probability of 0.01 for good neurological recovery in the validation cohort. The positive predictive value was the highest in the RF model and the smallest in the SVM model (positive predictive value [%] (95% CI): 25.0 (23.2–26.8) vs 16.8 (15.6–18.1), respectively). The negative predictive value was the highest in the SVM model and the lowest in the RF model (negative predictive value [%] (95% CI): 100.0 (99.3–100.0) vs 99.3 (98.8–99.6), respectively).

Discussion

This study developed, validated, and compared several prediction models by using machine learning algorithms and data from the national OHCA registry. The best performance machine learning algorithms for predicting good neurological recovery in OHCA patients were the LR and XGB model. Comparing to the LR model, discrimination was significantly lower in the SVM, RF and NN model. Although NRI was the highest in the RF model, the RF model was poorly calibrated (Hosmer–Lemeshow test: $p < 0.001$). Discrimination was the highest with the same value in the LR, XGB and EN models (AUCs (95% CI): 0.949 (0.941–0.957)) and all three models were well calibrated (Hosmer–Lemeshow test: $p > 0.05$). The XGB model reclassified the patients according to their true risk better than the LR model (NRI (95% CI): 0.110 (0.022, 0.198)), but the EN model reclassified the patients worse than the LR model did (NRI (95% CI): −1.239 (−1.309 to −1.169)) (Table 2 and Supplementary figure 2). Although there was a significant difference of reclassification between the LR and XGB model, the LR model showed almost similar test characteristics to the XGB model. In addition, when prediction models using only prehospital variable sets, the AUC of the LR model was higher than that of the XGB model (0.915 vs 0.913, respectively) (Table 3).

We found that the discrimination power of the developed models was higher than that observed in previous studies.[5–10] We investigated novel machine learning algorithms, such as the XGB model, which have not been evaluated in previous OHCA studies. However, our findings showed that the LR model, which has been the predominant algorithm for developing a prediction model for OHCA patients, had a similar discrimination compared to the XGB model. Therefore, the discrimination improvement in our models was mainly caused by the variables that we used rather than the models that we applied. We added no flow time, low flow time and post-resuscitation care variables into our analyses. No flow time and low flow time are known to be strongly associated with survival outcomes in OHCA.[33] However, these factors were not fully utilized in prediction model development in previous studies. The post-resuscitation care variables also contributed to additional performance improvement in our study (Table 3). Our data preprocessing methods and the

Table 3 – Performance change of the good neurological recovery prediction models in the validation cohort among models using different sets of variables.

Models	AUROC (95% CI) for less variables	AUROC (95% CI) for more variables	Absolute change of AUROC; *p*-value	NRI (95% CI); *p*-value
Comparison between variable: set 1 and set 2				
LR	0.915 (0.903, 0.926)	0.945 (0.936, 0.954)	+3.0%; $p<0.001$	0.900 (0.824, 0.975); $p<0.001$
XGB	0.913 (0.900, 0.925)	0.945 (0.937, 0.954)	+3.2%; $p<0.001$	1.006 (0.934, 1.078); $p<0.001$
SVM	0.903 (0.889, 0.917)	0.934 (0.923, 0.945)	+3.1%; $p<0.001$	0.778 (0.700, 0.856); $p<0.001$
RF	0.898 (0.884, 0.913)	0.930 (0.918, 0.942)	+3.2%; $p<0.001$	0.491 (0.414, 0.569); $p<0.001$
EN	0.914 (0.902, 0.926)	0.945 (0.937, 0.954)	+3.1%; $p<0.001$	1.049 (0.975, 1.122); $p<0.001$
NN	0.907 (0.894, 0.920)	0.942 (0.933, 0.951)	+3.5%; $p<0.001$	0.870 (0.797, 0.943); $p<0.001$
Comparison between variable: set 2 and set 3				
LR	0.945 (0.936, 0.954)	0.949 (0.941, 0.957)	+0.4%; $p<0.001$	0.251 (0.165, 0.338); $p<0.001$
XGB	0.945 (0.937, 0.954)	0.949 (0.941, 0.957)	+0.4%; $p=0.011$	0.360 (0.274, 0.447); $p<0.001$
SVM	0.934 (0.923, 0.945)	0.943 (0.934, 0.952)	+0.9%; $p=0.002$	0.829 (0.745, 0.913); $p<0.001$
RF	0.930 (0.918, 0.942)	0.937 (0.926, 0.948)	+0.7%; $p=0.012$	0.071 (-0.012, 0.155); $p=0.095$
EN	0.945 (0.937, 0.954)	0.949 (0.941, 0.957)	+0.4%; $p<0.001$	0.276 (0.189, 0.362); $p<0.001$
NN	0.942 (0.933, 0.951)	0.942 (0.931, 0.952)	+0.0%; $p=0.927$	−0.438 (-0.5228, -0.353); $p<0.001$

LR, logistic regression; XGB, extreme gradient boosting; SVM, support vector machine; RN, random forest; EN, elastic net; NN, neural network; AUROC, area under the receiver operating curve; NRI, net reclassification index.
Variable set 1: All variables except no flow time, low flow time and hospital care variables (targeted temperature management, percutaneous coronary intervention, and extracorporeal membrane oxygenation)
Variable set 2: All variables except hospital care variables (targeted temperature management, percutaneous coronary intervention, and extracorporeal membrane oxygenation)
Variable set 3: All variables

Table 4 – Validation cohort test characteristics of prediction models using cutoffs probability of 0.01 for good neurological recovery.

Model	TP	FN	TN	FP	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
LR	575	2	1508	1887	99.7 (98.8–100.0)	44.4 (42.7–46.1)	23.4 (21.7–25.1)	99.9 (99.5–100.0)
XGB	572	5	1555	1840	99.1 (98.0–99.7)	45.8 (44.1–47.5)	23.7 (22.0–25.5)	99.7 (99.3–99.9)
SVM	577	0	535	2860	100.0 (99.4–100.0)	15.8 (14.5–17.0)	16.8 (15.6–18.1)	100.0 (99.3–100.0)
RF	565	12	1698	1697	97.9 (96.4–98.9)	50.0 (48.3–51.7)	25.0 (23.2–26.8)	99.3 (98.8–99.6)
EN	576	1	1487	1908	99.8 (99.0–100.0)	43.8 (42.1–45.5)	23.2 (21.5–24.9)	99.9 (99.6–100.0)
NN	571	6	1295	2100	99.0 (97.8–99.6)	38.1 (36.5–39.8)	21.4 (19.8–23.0)	99.5 (99.0–99.8)

LR, logistic regression; XGB, extreme gradient boosting; SVM, support vector machine; RN, random forest; EN, elastic net; NN, neural network; 95% CI, 95% confidence interval; TP, true positive; FN, false negative; TN, true negative; FP, false positive; PPV, positive predictive value; NPV, negative predictive value.

relatively large amount of data considering the number of variables can be also another reason for the high and similar performance of the LR model with other machine learning models.

Initial rhythm and low flow time were highly weighted in all machine learning models; however, there was a notable variation in the importance of the variables depending on the machine learning algorithms. Although the LR, XGB, and EN models showed similar discriminations, the EN placed a different pattern of variable importance than the LR and XGB models did. The EN model placed more weight on post-resuscitation cares, such as ECMO and PCI, than community factors, such as witness status, the location of arrest, and bystander defibrillation. The variable type, the prevalence of each category, the interaction with the other variables, and the method used to calculate variable importance could all affect the variable importance differently in each of the machine learning algorithms. A further exploration to find robust predictors of good outcomes and to investigate factors that were not recognized as important for a good outcome would be helpful to develop more accurate prediction models.

Each of the machine learning algorithms showed similarities and differences according to the specific task. When using prediction models as the TOR tool, the SVM model relatively overestimates the probability of good neurological recovery, and the RF model relatively underestimates the probability of good neurological recovery within the prediction models. When changing the available variables in the developing models, the discrimination and reclassification improvements were similar among the LR, XGB, and EN models. However, when comparing the models with and without the post-resuscitation care variables, the NN model showed an insignificant discrimination improvement and a worse reclassification when the hospital care variables were added. The LR, XGB, and EN models showed the highest discrimination regardless of the variables used.

Reliable and accurate prediction models can help identify patients who need specific care for good recovery. We found that different

machine learning algorithms result in different performance, and that the LR model showed similar performance to the XGB model, which has been recently developed and has shown satisfactory results in machine learning competitions.[34] We also found that more relevant variables significantly affect the performance of the prediction models.

Limitation

This study had several limitations. First, we could not adapt commonly used biomarker and vital sign variables in our model.[5,9] Second, we could not fully interpret each of the machine learning methods because of the inherent complexity of the algorithms. However, we compared a variety of aspects of each model, and the findings of the different characteristics of each model helped us to gain more insight into each model. Third, there is the possibility of underfitting or overfitting effects of each of the models. We addressed this possibility by examining the appropriate choice of pretraining methods and by conducting a rigorous search and a careful selection of the hyper-parameters. Lastly, this study was performed in an EMS system with an intermediate service level. The generalization of these study findings should be made with caution.

Conclusion

In this study, which involved the development, validation, and comparison of models for the prediction of good neurological recovery after OHCA, the best performance machine learning algorithms were XGB and LR.

Author Contributions:

Drs. Shin SD and Park JH had full access to all of the data in the study and take responsibility for the integrity of the data, as well as for the accuracy of the data analysis.

Study concept and design: Dr. Shin SD

Acquisition, analysis, and interpretation of the data: Drs. Park JH and Shin SD

Drafting of the manuscript: Dr. Park JH

Critical revision of the manuscript for important intellectual content: Drs. Song KJ, Hong KJ, Ro YS, Choi JW, and Choi SW

Statistical analysis: Dr. Park JH

Obtained funding: Dr. Shin SD

Administrative, technical, or material support: Drs. Song KJ, Hong KJ, and Ro YS

Study supervision: Drs. Shin SD and Choi JW

Manuscript approval: all authors.

Conflict of interest statement

There are no potential conflicts of interest for any of the authors in this study.

Acknowledgments

The study was funded for the Korean OHCA Registry by the Korea Centers for Disease Control and Prevention (2013–2017).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.resuscitation.2019.07.020.

REFERENCES

1. Grasner JT, Lefering R, Koster RW, et al. EuReCa ONE-27 Nations, ONE Europe ONE Registry: A prospective one month analysis of out-of-hospital cardiac arrest outcomes in 27 countries in Europe. Resuscitation 2016;105:188–95.
2. Berdowski J, Berg RA, Tijssen JG, Koster RW. Global incidences of out-of-hospital cardiac arrest and survival rates: Systematic review of 67 prospective studies. Resuscitation 2010;81:1479–87.
3. Myat A, Song KJ, Rea T. Out-of-hospital cardiac arrest: current concepts. Lancet 2018;391:970–9.
4. Tanaka H, Ong MEH, Siddiqui FJ, et al. Modifiable factors associated with survival after out-of-hospital cardiac arrest in the Pan-Asian resuscitation outcomes study. Ann Emerg Med 2018;71:608–17 e15.
5. Maupain C, Bougouin W, Lamhaut L, et al. The CAHP (Cardiac Arrest Hospital Prognosis) score: a tool for risk stratification after out-of-hospital cardiac arrest. Eur Heart J 2016;37:3222–8.
6. Kiehl EL, Parker AM, Matar RM, et al. C-GRApH: A validated scoring system for early stratification of neurologic outcome after out-of-hospital cardiac arrest treated with targeted temperature management. J Am Heart Assoc 2017;6.
7. Goto Y, Maeda T, Goto Y. Decision-tree model for predicting outcomes after out-of-hospital cardiac arrest in the emergency department. Crit Care 2013;17:R133.
8. Rea TD, Cook AJ, Stiell IG, Powell J, Bigham B, Callaway CW, et al. Predicting survival after out-of-hospital cardiac arrest: role of the Utstein data elements. Ann Emerg Med 2010;55:249–57.
9. Adrie C, Cariou A, Mourvillier B, et al. Predicting survival with good neurological recovery at hospital admission after successful resuscitation of out-of-hospital cardiac arrest: the OHCA score. Eur Heart J 2006;27:2840–5.
10. Aschauer S, Dorffner G, Sterz F, Erdogmus A, Laggner A. A prediction tool for initial out-of-hospital cardiac arrest survivors. Resuscitation 2014;85:1225–31.
11. Liu N, Koh ZX, Goh J, et al. Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. BMC Med Inform Decis Mak 2014;14:75.
12. Layeghian Javan S, Mehdi Sepehri M, Aghajani H. Toward analyzing and synthesizing previous research in early prediction of cardiac arrest using machine learning based on a multi-layered integrative framework. J Biomed Inform 2018.
13. Kwon JM, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. J Am Heart Assoc 2018;7.
14. Krizmaric M, Verlic M, Stiglic G, Grmec S, Kokol P. Intelligent analysis in predicting outcome of out-of-hospital cardiac arrest. Comput Methods Programs Biomed 2009;95:S22–32.
15. Kim C, Choi HJ, Moon H, et al. Prehospital advanced cardiac life support by EMT with a smartphone-based direct medical control for nursing home cardiac arrest. Am J Emerg Med 2018.
16. Kim YT, Shin SD, Hong SO, et al. Effect of national implementation of utstein recommendation from the global resuscitation alliance on ten steps to improve outcomes from Out-of-Hospital cardiac arrest: a ten-year observational study in Korea. BMJ Open 2017;7:e0169.
17. Ro YS, Shin SD, Song KJ, et al. A trend in epidemiology and outcomes of out-of-hospital cardiac arrest by urbanization level: a nationwide observational study from 2006 to 2010 in South Korea. Resuscitation 2013;84:547–57.
18. Ro YS, Shin SD, Song KJ, et al. A comparison of outcomes of out-of-hospital cardiac arrest with non-cardiac etiology between emergency

departments with low- and high-resuscitation case volume. Resuscitation 2012;83:855–61.
19. Cudnik MT, Sasson C, Rea TD, et al. Increasing hospital volume is not associated with improved survival in out of hospital cardiac arrest of cardiac etiology. Resuscitation 2012;83:862–8.
20. Maslove DM, Podchiyska T, Lowe HJ. Discretization of continuous features in clinical datasets. J Am Med Inform Assoc. 2013;20:544–53.
21. Taylor RA, Moore CL, Cheung KH, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS One 2018;13:e0194.
22. Jentzer JC, Scutella M, Pike F, et al. Early coronary angiography and percutaneous coronary intervention are associated with improved outcomes after out of hospital cardiac arrest. Resuscitation 2018;123:15–21.
23. Yukawa T, Kashiura M, Sugiyama K, Tanabe T, Hamabe Y. Neurological outcomes and duration from cardiac arrest to the initiation of extracorporeal membrane oxygenation in patients with out-of-hospital cardiac arrest: a retrospective study. Scand J Trauma Resusc Emerg Med 2017;25:95.
24. Kim KH, Shin SD, Song KJ, et al. Scene time interval and good neurological recovery in out-of-hospital cardiac arrest. Am J Emerg Med 2017;1682–90.
25. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989;2:359–66.
26. Breiman L. Random Forests. Machine Learning 2001;45:5–32.
27. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc: Series B (Statistical Methodology) 2005;67:301–20.
28. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intel Syst Appl 1998;13:18–28.
29. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM;. p. 785–94.
30. Fenlon C, O'Grady L, Doherty ML, Dunnion J. A discussion of calibration techniques for evaluating binary and categorical predictive models. Prev Vet Med 2018;149:107–14.
31. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12:e0174944.
32. Drennan IR, Case E, Verbeek PR, Reynolds JC, Goldberger ZD, Jasti J, et al. A comparison of the universal TOR Guideline to the absence of prehospital ROSC and duration of resuscitation in predicting futility from out-of-hospital cardiac arrest. Resuscitation 2017;111:96–102.
33. Adnet F, Triba MN, Borron SW, Lapostolle F, Hubert H, Gueugniaud PY, et al. Cardiopulmonary resuscitation duration and survival in out-of-hospital cardiac arrest patients. Resuscitation 2017;111:74–81.
34. Chen T, He T. Higgs boson discovery with boosted trees. In: Glen C, Cécile G, Isabelle G, Balázs K, David R, editors. Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning. Proceedings of Machine Learning Research: PMLR;. . p. 69–80.