OPEN

External Validation of Deep Learning-Based Cardiac Arrest Risk Management System for Predicting In-Hospital Cardiac Arrest in Patients Admitted to General Wards Based on Rapid Response System Operating and Nonoperating Periods: A Single-Center Study

Kyung-Jae Cho, MS[1]
Kwan Hyung Kim, MD[2]
Jaewoo Choi, MS[1]
Dongjoon Yoo, MD[1,3]
Jeongmin Kim, MD, PhD[2,4,5]

OBJECTIVES: The limitations of current early warning scores have prompted the development of deep learning-based systems, such as deep learning-based cardiac arrest risk management systems (DeepCARS). Unfortunately, in South Korea, only two institutions operate 24-hour Rapid Response System (RRS), whereas most hospitals have part-time or no RRS coverage at all. This study validated the predictive performance of DeepCARS during RRS operation and nonoperation periods and explored its potential beyond RRS operating hours.

DESIGN: Retrospective cohort study.

SETTING: In this 1-year retrospective study conducted at Yonsei University Health System Severance Hospital in South Korea, DeepCARS was compared with conventional early warning systems for predicting in-hospital cardiac arrest (IHCA). The study focused on adult patients admitted to the general ward, with the primary outcome being IHCA-prediction performance within 24 hours of the alarm.

PATIENTS: We analyzed the data records of adult patients admitted to a general ward from September 1, 2019, to August 31, 2020.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: Performance evaluation was conducted separately for the operational and nonoperational periods of the RRS, using the area under the receiver operating characteristic curve (AUROC) as the metric. DeepCARS demonstrated a superior AUROC as compared with the Modified Early Warning Score (MEWS) and the National Early Warning Score (NEWS), both during RRS operating and nonoperating hours. Although the MEWS and NEWS exhibited varying performance across the two periods, DeepCARS showed consistent performance.

CONCLUSIONS: The accuracy and efficiency for predicting IHCA of DeepCARS were superior to that of conventional methods, regardless of whether the RRS was in operation. These findings emphasize that DeepCARS is an effective screening tool suitable for hospitals with full-time RRS, part-time RRS, and even those without any RRS.

KEYWORDS: artificial intelligence; clinical deterioration; deep learning; early warning score; heart arrest; hospital rapid response team

DOI: 10.1097/CCM.0000000000006137

KEY POINTS

Question: Can deep learning-based cardiac arrest risk management system (DeepCARS) effectively serve as a screening tool for hospitals during both Rapid Response System (RRS) operating and nonoperating periods?

Findings: This retrospective cohort study demonstrated that DeepCARS exhibited superior performance in predicting in-hospital cardiac arrest (IHCA) as compared with conventional methods, irrespective of whether the RRS was operational.

Meaning: DeepCARS can serve as an effective and efficient IHCA screening tool suitable for hospitals with full-time RRS, part-time RRS, and even those without any RRS.

The Rapid Response System (RRS) is a method for screening patients who have an increased risk for unexpected clinical deterioration. Vital signs frequently become abnormal in the several hours preceding cardiac arrest, which could promote critical adverse events (1, 2). However, the use of an RRS to identify significant abnormal vital sign values from numerous in-hospital patients proves challenging. To decrease RRS burden and detection failure rate for patients with deteriorating status, several early warning scores (EWSs), known as track-and-trigger systems (TTSs), have been devised (3). The Modified Early Warning Score (MEWS) and the National Early Warning Score (NEWS) are the most commonly implemented EWS but have common limitations of low sensitivity and high false alarms, hampering efficient RRS activation (4–6).

Deep learning-based artificial intelligence (AI) could be useful for analyzing sequential data of vital signs and for identifying patients with progressively deteriorating clinical conditions even before the medical staff becomes aware of an emergent situation. The deep learning-based cardiac arrest risk management system (DeepCARS), was first developed in 2018 and was approved for use as a medical device in 2021 by the Ministry of Food and Drug Safety of South Korea. DeepCARS accurately predicts patient deterioration, with a low false-alarm rate, in the general ward and surpasses the conventional EWSs (7–9). Among 47 institutions that enrolled in the Korean government's RRS pilot program, described below, 13 have successfully implemented DeepCARS as a basal and main EWS, replacing conventional EWSs.

Many hospitals have since realized the advantages of using an RRS for improving patient safety. The pilot program for expanding the RRS was initiated by the Korean Health Insurance Review and Assessment Service and the Ministry of Health and Welfare in May 2019. The Ministry of Health and Welfare has started providing daily health insurance fees for inpatients in RRS-operating hospitals with more than 300 beds. However, this support remains insufficient to enable 24-hour RRS operation and maintenance. Approximately 3200 general hospitals, including 45 tertiary hospitals, operate nationwide in South Korea, but only two institutions officially operate a 24-hour RRS.

With a view to facilitating the development of a complementary system for use in healthcare institutions with part-time access to an RRS, we validated the performance of DeepCARS for predicting in-hospital cardiac arrest (IHCA) during RRS operational and nonoperational periods. Furthermore, we explored the potential of DeepCARS for identifying patients who experience health deterioration outside of the RRS operating hours.

MATERIALS AND METHODS

Study Design

We conducted this 1-year, retrospective, single-center cohort study from September 1, 2019, to August 31, 2020, at the Yonsei University Health System Severance Hospital, a tertiary academic hospital with 2454 beds, in the Republic of Korea. All adult patients (≥ 18 yr old) admitted to the general ward during the study period were eligible for study inclusion. We excluded data from patients with admission dates outside the study period, or if admission occurred within 24 hours before study completion only if they did not experience IHCA or an unplanned general ward-to-ICU transfer (UIT), or when no vital signs were recorded 24-hour pre-IHCA or pre-UIT, or no vital signs were recorded throughout the study period (**Additional File Fig. S1**, http://links.lww.com/CCM/H464).

The primary outcome of interest was IHCA (defined as the "cessation of cardiac activity, confirmed by the absence of a detectable pulse, unresponsiveness, and apnea," from the "in-hospital Utstein style" consensus guidelines of the American Heart Association, which was followed by resuscitation attempts) (10). We compared the predictive performance of DeepCARS with that of the conventional EWSs for predicting the primary outcome within 24 hours of the alarm. Assessments were conducted according to whether the RRS was operational or nonoperational.

RRS Operation and Characteristics of Yonsei University Hospital

The WhoEver Saves One Life SAVEs World entire is the part-time RRS in our hospital and was first implemented in February 2019. A dedicated team of specialist physicians (four intensivists and five nurses) was exclusively assigned to the RRS team that would operate from 06:00 to 22:00 on nonholiday weekdays and from 07:00 to 12:00 on Saturdays. During the operating period, the RRS included one physician and two nurses. The physician would evaluate patients who were either detected by the RRS screening tools or were identified by ward medical staff. Nurses monitored patients in general wards using screening tools, NEWS, and single-parameter TTS (SPTTS). During the RRS nonoperational period, the centralized screening system was also nonoperational. Even when the centralized screening system was nonoperational, NEWS and SPTTS continued to be generated and recorded. However, due to the limited availability of the clinical team, patient screening was almost entirely the responsibility of the physician and nurse. All patients for whom the RRS activated an alarm, who were transferred from the general ward to the ICU, or who suffered IHCA in the general ward during the preceding 30 days were reviewed during the monthly RRS conference.

Medical personnel composition in the hospital varies temporally. The nursing staff have a rotational schedule of three shifts per day, as follows: day shift (06:00–14:00), evening shift (14:00–22:00), and night shift (22:00–06:00). During night shifts, the number of the nursing staff decreases by 20–25% relative to the number present during day shifts. This nursing staff shortage, combined with the absence of a centralized screening system, complicates patient screening. Most hospital physicians work from 08:00 to 18:00. Therefore, the nighttime responsibility mainly falls on less-experienced doctors, such as resident trainees.

Data Collection and Preprocessing

We collected data on age, sex, IHCA occurrence, and five time-stamped vital signs (systolic blood pressure, diastolic blood pressure, heart rate [HR], respiratory rate [RR], and body temperature [BT]) recorded during hospitalization, the ICU transfer time, surgery time, and the do-not-resuscitate (DNR) status. Erroneous values with extreme deviations from the vital-specific normal ranges and non-numeric values were treated as missing values.

UIT definition varies across reports (11–15). In our study, UIT was defined as unanticipated ICU admission for both medical and surgical patients. For nonsurgical admission, UIT was considered as a transfer that could not be postponed for 24 hours without adverse effects. In contrast, surgical UIT admission was defined as when the patient was transferred to the ICU before surgery or operating room transfer. These criteria align with a recent multinational study on UIT (13).

Ethical Considerations

This study was strictly observational and was conducted based on anonymity. The ethics committee and institutional review board of Yonsei University Hospital approved the study (approval number and date: 2-2021-1353 and November 16, 2021) and waived the need for obtaining informed consent due to the minimal risk using data collected for routine clinical practice. The study was conducted in accordance with the ethical standards of the responsible committee on human experimentation and the Helsinki Declaration of 1975.

Deep Learning-Based Cardiac Arrest Risk Management System

DeepCARS uses only four classic vital signs (HR, BP, BT, and RR) (16), age, and the recorded time of each vital sign, and outputs risk scores on a scale of 0–100. A higher value denotes an augmented propensity for the occurrence of IHCA. The detailed architecture of DeepCARS has been described previously (8, 17).

Performance Evaluation and Statistical Analysis

Question 1: What is the Accuracy of DeepCARS as Compared With Conventional EWSs in Predicting IHCA Both During and Outside of RRS Operating Hours?. We assessed predictive performance using the area under the receiver operating characteristic curve (AUROC). To observe the relationship between sensitivity and specificity at different cutoff values, we plotted the receiver operating characteristic curve. Additionally, we calculated the F1-score (2 × [precision × recall]/[precision + recall]), that is, the harmonic mean of precision and recall, which provides a holistic view of a models' performance, taking into account both its ability to identify positive cases accurately and to avoid making incorrect positive predictions. Additionally, we determined the positive-predictive value (PPV = true positive/[true positive + false positive]), negative-predictive value (NPV = true negative/[true negative + false negative]), sensitivity (true positive/[true positive + false negative]), specificity (true negative/[true negative + false positive]), and daily alarm rate (14, 15). In real clinical practice, an alarm would be considered appropriate if it resulted in active intervention, such as deciding a DNR status, or ICU transfer. We expanded the definition of true positives for PPV (termed PPV+) by including DNR prescriptions and UIT in addition to cardiopulmonary resuscitation. We created a graph illustrating the trends of PPV+, sensitivity, and F-measure. Evaluations were performed separately for different subgroups, stratified first by whether the RRS was operational, to ascertain the consistency of performance, and second by alterations of prediction window.

Question 2: Does DeepCARS Generate Fewer Total Alarms than Conventional Methods During RRS Operating and Nonoperating Periods, Without Compromising Sensitivity?. To assess the alarm performance during both RRS operating and nonoperating periods, we segregated the operating and nonoperating times for each hospitalized patient. We calculated the daily alarm rate by computing the average number of alarms during each time window, dividing this by the total number of beds and the time window, and multiplying the result by 1000. We compared alarm performance by plotting the daily alarm–sensitivity curve, demonstrating the relationship between cutoff values and the daily alarm–sensitivity dyad tendency.

Question 3: What is the Impact of Subsequent Alarms on the Performance of DeepCARS?. During the period when the RRS is nonoperational, availability of clinical resources, such as nursing staff, RRT teams, and dashboard systems, is limited. Therefore, discovering an efficient method to harness DeepCARS to ensure enhanced PPV without requiring additional training is desirable. We hypothesized that maintenance of consistently high triggering scores with DeepCARS would serve as an indicator of a substantial risk of cardiac arrest. Simultaneously, given the difficulty in meeting the conditions for such an event, this approach could potentially lead to an improvement in PPV. Instead of altering the DeepCARS algorithm, we conducted a retrospective data analysis. We found that subsequent alarms could effectively improve the PPV, making them particularly useful during the RRS nonoperating hours. We evaluated the effectiveness by measuring the PPV+, the percentage of IHCA patients who were identified, F-measure, and daily alarms as the number of subsequent alarms increased.

Additionally, we plotted the cumulative percentage of patients experiencing IHCA by accumulating the percentage of IHCA patients for whom the DeepCARS presented an alert within 24 hours before IHCA onset. As the MEWS and NEWS exhibited distinct specificities at identical cutoff values, we selected the MEWS to represent the conventional EWSs. Furthermore, we conducted a subgroup analysis by subdividing the period into smaller segments, such as day, evening, and nighttime, with consideration of nurses' rotational work shifts.

RESULTS

Baseline Characteristics

During the 12-month study period from November 1, 2019, to August 31, 2020, we analyzed 95,607 patient admissions, basic demographics, and five routinely collected physiologic variables, generating more than 3,221,077 data records from a single study center. A total of 228 IHCA, 907 DNR orders, and 1994 UIT occurred during the study period.

The data records were divided into two groups based on the RRS operating and nonoperating hours. The difference in baseline characteristics of the data was comparable for the two groups (**Table 1**). However, distinct intergroup differences were observed for the following

TABLE 1.
Baseline Characteristics

Characteristics	Entire Period	Operating Time of RRS	Nonoperating Time of RRS	p
	September 1, 2019, to August 31, 2020			
Number of total admissions, *n*	95,607	95,087	67,614	
Number of data records, *n*	3,221,077	2,191,242	1,040,25	
Age, yr, mean ± SD	57.49 ± 16.16	57.57 ± 16.16	57.96 ± 16.15	< 0.001
Male, sex, % (*n*)	48.87 (46,728)	48.95 (46,552)	48.94 (33,095)	0.586
Length of stay, mean ± SD	6.81 ± 12.30	–	–	–
Variables, mean ± SD				
SBP (mm Hg)	123.86 ± 18.93	124.36 ± 18.89	122.84 ± 18.96	< 0.001
DBP (mm Hg)	75.68 ± 12.17	75.99 ± 12.09	75.05 ± 12.32	< 0.001
HR (/min)	81.48 ± 16.84	80.69 ± 16.40	82.99 ± 17.56	< 0.001
RR (/min)	19.18 ± 3.03	18.99 ± 3.18	19.58 ± 2.61	< 0.001
BT (°C)	37.03 ± 0.50	36.99 ± 0.49	37.08 ± 0.50	< 0.001
Interval of vital sign measurement (hr)	5.28 ± 3.16	4.07 ± 2.27	6.32 ± 4.85	< 0.001
Vital signs within 24 hr before cardiac arrest, mean ± SD				
SBP (mm Hg)	113.69 ± 29.02	115.01 ± 28.64	111.55 ± 29.50	< 0.01
DBP (mm Hg)	68.97 ± 18.53	69.37 ± 18.35	68.32 ± 18.81	0.173
HR (/min)	103.15 ± 29.40	101.91 ± 30.3	105.19 ± 27.73	< 0.01
RR (/min)	22.78 ± 7.30	22.48 ± 7.15	23.27 ± 7.53	< 0.05
BT (°C)	37.09 ± 0.76	37.06 ± 0.77	37.14 ± 0.74x	< 0.05
Number of admissions				
During weekend	64,616	–	64,616	
During day time	92,567	92,567	–	
During evening time	86,050	86,050	–	
During night time	47,089	–	47,089	
Number of admissions with outcomes, *n*				
In-hospital cardiac arrest/1,000 admissions	228 (2.38)	189 (1.98)	184 (2.27)	–
Do-not-resuscitate/1,000 admissions	907 (9.48)	778 (8.18)	856 (1.26)	–
Unplanned ICU transfer/1,000 admissions	1,994 (20.85)	1,892 (19.89)	931 (13.76)	–

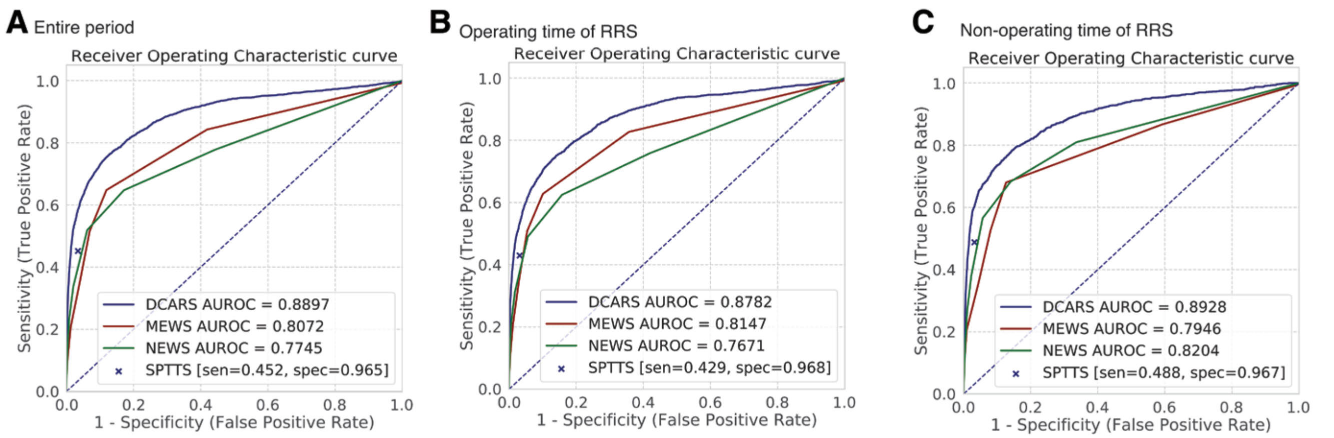
BT = body temperature, DBP = diastolic blood pressure, HR = heart rate, RR = respiratory rate, RRS = Rapid Response System, SBP = systolic blood pressure.

Dashes for number of admissions with outcomes indicate *p* values for are all less than 0.001.

items. The interval of vital sign measurement was longer during RRS nonoperating times than during RRS operating times (6.42 hr vs. 4.07 hr). Furthermore, the number of IHCA patients per 1000 admissions was higher during the RRS nonoperating times (2.27 vs. 1.98). Furthermore, as shown in the severity distribution graph (**Additional File Fig. S2**, http://links.lww.com/CCM/H464), higher NEWS, MEWS, and SPTTS were more prevalent during RRS nonoperating times. We then plotted the cumulative percentage of alarms as the MEWS and NEWS values increased and found that the graph slope became steeper with higher NEWS and MEWS values.

Question 1. Performance in Predicting IHCA

As shown in **Figure 1**, DeepCARS outperformed MEWS, NEWS, and SPTTS in predicting IHCA.



	AUROC (95% CI, p < 0.001)			AUPRC (95% CI, p < 0.001)		
	Entire period	Operating time of RRS	Non-operating time of RRS	Entire period	Operating time of RRS	Non-operating time of RRS
DCARS	0.889 (0.880 – 0.891)	0.878 (0.866 – 0.885)	0.892 (0.881 – 0.903)	0.0213 (0.0201 – 0.0239)	0.0179 (0.0185 – 0.0218)	0.0194 (0.0183 – 0.0206)
MEWS	0.807 (0.797 – 0.812)	0.814 (0.804 – 0.818)	0.794 (0.788 – 0.806)	0.0090 (0.0090 – 0.0100)	0.0095 (0.0086 – 0.0102)	0.0084 (0.0088 – 0.0110)
NEWS	0.774 (0.778 – 0.793)	0.767 (0.765 – 0.779)	0.820 (0.814 – 0.829)	0.0099 (0.0096 – 0.0109)	0.0091 (0.0086 – 0.0105)	0.0108 (0.0098 - 0.0119)

Figure 1. Comparison of predictive performance of deep learning-based cardiac arrest risk management system (DeepCARS) and the conventional Early Warning Score both during Rapid Response System (RRS) operating and nonoperating hours. AUROC = area under the receiver operating characteristic curve, AUPRC = area under the precision–recall curve, MEWS = Modified Early Warning Score, NEWS = National Early Warning Score, Sen = sensitivity, Spec =specificity, SPTTS = Single-Parameter Track-and-Trigger System.

Not only did DeepCARS outperform conventional EWSs in the analysis of the entire dataset (AUROC: 0.889 vs. 0.807 for MEWS vs. 0.774 for NEWS), but it also outperformed them during both RRS operating (AUROC: 0.878 vs. 0.814 for MEWS vs. 0.767 for NEWS) and nonoperating periods (AUROC: 0.892 vs. 0.794 for MEWS vs. 0.820 for NEWS). Although the AUROC of DeepCARS remained consistent across RRS operating and nonoperating periods, the predictive performance of MEWS and NEWS differed between the two periods. Furthermore, the precision–recall graph (**Fig. 2**) showed that DeepCARS demonstrated a higher PPV+ (*solid line*) and F-measure at most corresponding sensitivity points during both RRS operating and nonoperating periods. The results of other metrics are shown in **Additional File Tables S1–S4** (http://links.lww.com/CCM/H464).

Figure 3 illustrates the predictive performance based on the prediction window timeline, ranging from 24 hours to 1 hour before the primary event. It demonstrates that the performance improved for all EWSs as the prediction window narrowed. However, DeepCARS consistently outperformed the other conventional EWSs across all prediction windows, regardless of whether the RRS was operational. Specifically, in the 1-hour prediction window, the AUROC curve for DeepCARS increased to 0.939 and 0.942 during RRS operating and nonoperating periods, respectively.

Question 2. Alarm Performance

In terms of alarm performance, DeepCARS had a lower alarm rate at all corresponding sensitivity values than those of other EWSs (**Fig. 4**). This result was consistent across RRS operating and nonoperating periods. Specifically, assuming a daily alarm rate of conventional EWSs as 100%, the daily alarm rate was reduced by more than half at all main cutoff values of MEWS, NEWS, and SPTTS.

Question 3. Impact of Subsequent Alarm

To increase the PPV, particularly during the RRS nonoperating time, an additional rule—that of a subsequent alarm—was added to the existing alarm

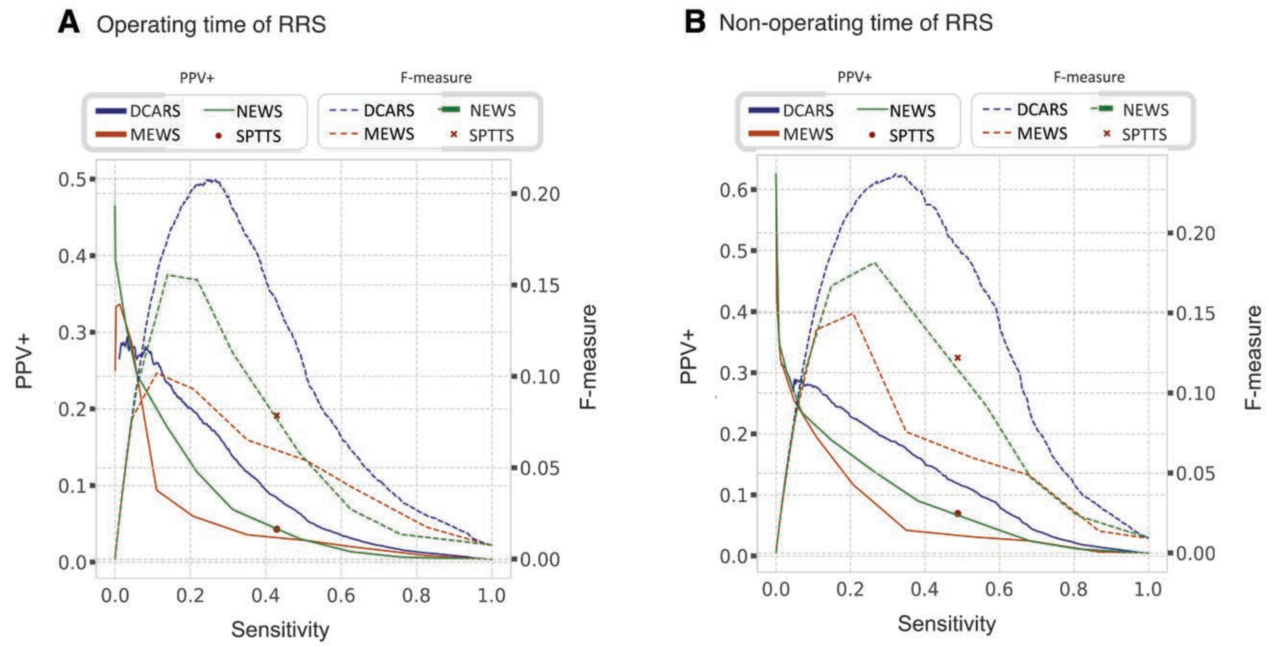
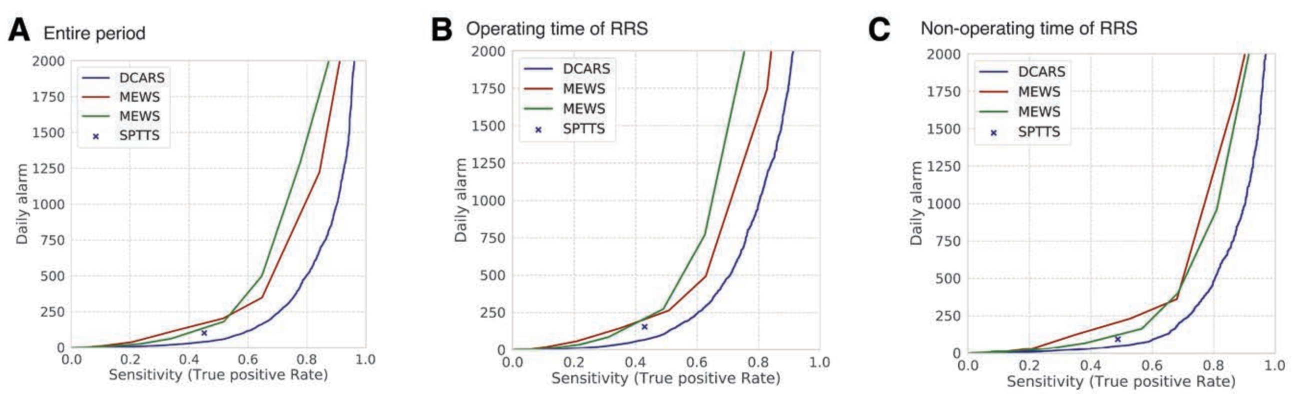


Figure 2. Precision–recall graph of deep learning-based cardiac arrest risk management system (DeepCARS) and Conventional Early Warning Scores. The *solid line* indicates the positive-predictive value. The *dashed line* indicates the F-measure. PPV = Positive-Predictive Value, MEWS = Modified Early Warning Score, NEWS = National Early Warning Score, RRS = Rapid Response System, SPTTS = Single-Parameter Track-and-Trigger System.

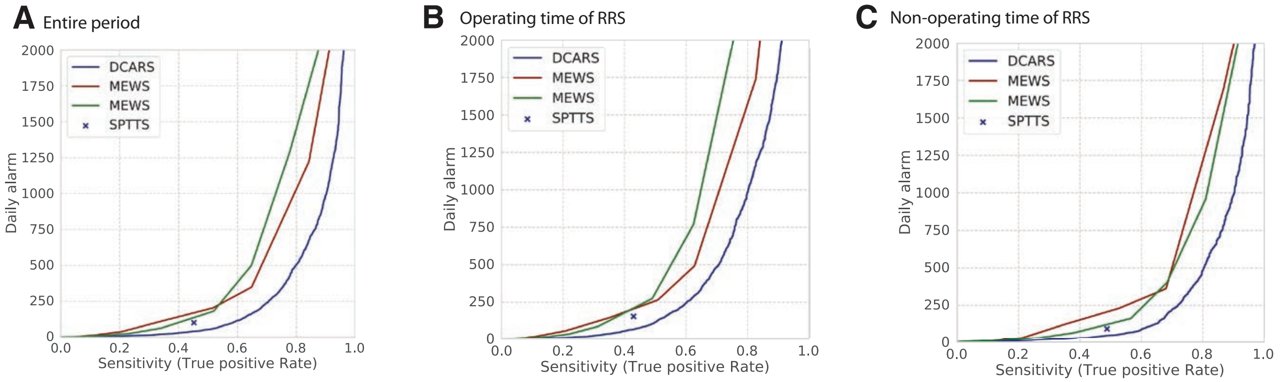


method	cutoff	sensitivity	daily alarm	Alarm Rate
MEWS	3	0.6478	349.3	100%
DCARS	0.8657	0.6478	164.3	47.04
MEWS	4	0.5155	204	100%
DCARS	0.9395	0.5155	59.3	29.07
MEWS	5	0.3509	117.8	100%
DCARS	0.9735	0.3509	21.7	18.42
NEWS	3	0.5191	182.5	100%
DCARS	0.9385	0.5191	60.5	33.15
NEWS	4	0.3388	62.7	100%
DCARS	0.9752	0.3388	20.1	32.06
NEWS	5	0.236	26.9	100%
DCARS	0.9875	0.236	9.7	36.06
SPTTS	1	0.4518	102.5	100%
DCARS	0.9545	0.4518	41.2	40.20

method	cutoff	sensitivity	daily alarm	Alarm Rate
MEWS	3	0.628	492.6	100%
DCARS	0.8806	0.628	274.3	55.68
MEWS	4	0.508	264.9	100%
DCARS	0.9415	0.508	109.9	41.49
MEWS	5	0.352	147.2	100%
DCARS	0.9734	0.352	42.1	28.60
NEWS	3	0.49	274.4	100%
DCARS	0.9458	0.49	99.7	36.33
NEWS	4	0.313	87.1	100%
DCARS	0.9792	0.313	31.8	36.51
NEWS	5	0.218	35.6	100%
DCARS	0.9888	0.218	16.7	46.91
SPTTS	1	0.429	155.4	100%
DCARS	0.9588	0.429	70.4	45.30

method	cutoff	sensitivity	Daily alarm	Alarm Rate
MEWS	3	0.68	362.1	100%
DCARS	0.8842	0.68	180.1	49.74
MEWS	4	0.527	231	100%
DCARS	0.9525	0.527	55.1	23.85
MEWS	5	0.35	123.7	100%
DCARS	0.9779	0.35	21.9	17.70
NEWS	3	0.566	163.4	100%
DCARS	0.9438	0.566	68.6	41.98
NEWS	4	0.381	66.9	100%
DCARS	0.9741	0.381	26.4	39.46
NEWS	5	0.265	30.6	100%
DCARS	0.9857	0.265	13.9	45.42
SPTTS	1	0.488	93.7	100%
DCARS	0.9582	0.488	47.1	50.27

Figure 3. Predictive performance based on varying prediction time. AUROC = area under the receiver operating characteristic curve, DeepCARS = deep learning-based cardiac arrest risk management system, MEWS = Modified Early Warning Score, NEWS = National Early Warning Score, RRS = Rapid Response System.



method	cutoff	sensitivity	daily alarm	alarm rate
MEWS	3	0.6478	349.3	100%
DCARS	0.8657	0.6478	164.3	47.04
MEWS	4	0.5155	204	100%
DCARS	0.9395	0.5155	59.3	29.07
MEWS	5	0.3509	117.8	100%
DCARS	0.9735	0.3509	21.7	18.42
NEWS	3	0.5191	182.5	100%
DCARS	0.9385	0.5191	60.5	33.15
NEWS	4	0.3388	62.7	100%
DCARS	0.9752	0.3388	20.1	32.06
NEWS	5	0.236	26.9	100%
DCARS	0.9875	0.236	9.7	36.06
SPTTS	1	0.4518	102.5	100%
DCARS	0.9545	0.4518	41.2	40.20

method	cutoff	sensitivity	daily alarm	alarm rate
MEWS	3	0.628	492.6	100%
DCARS	0.8806	0.628	274.3	55.68
MEWS	4	0.508	264.9	100%
DCARS	0.9415	0.508	109.9	41.49
MEWS	5	0.352	147.2	100%
DCARS	0.9734	0.352	42.1	28.60
NEWS	3	0.49	274.4	100%
DCARS	0.9458	0.49	99.7	36.33
NEWS	4	0.313	87.1	100%
DCARS	0.9792	0.313	31.8	36.51
NEWS	5	0.218	35.6	100%
DCARS	0.9888	0.218	16.7	46.91
SPTTS	1	0.429	155.4	100%
DCARS	0.9588	0.429	70.4	45.30

method	cutoff	sensitivity	Daily alarm	alarm rate
MEWS	3	0.68	362.1	100%
DCARS	0.8842	0.68	180.1	49.74
MEWS	4	0.527	231	100%
DCARS	0.9525	0.527	55.1	23.85
MEWS	5	0.35	123.7	100%
DCARS	0.9779	0.35	21.9	17.70
NEWS	3	0.566	163.4	100%
DCARS	0.9438	0.566	68.6	41.98
NEWS	4	0.381	66.9	100%
DCARS	0.9741	0.381	26.4	39.46
NEWS	5	0.265	30.6	100%
DCARS	0.9857	0.265	13.9	45.42
SPTTS	1	0.488	93.7	100%
DCARS	0.9582	0.488	47.1	50.27

Figure 4. Comparison of the alerting performance (alarms) of deep learning-based cardiac arrest risk management system (DeepCARS) and conventional early warning system both during Rapid Response System (RRS) operating and nonoperating periods. MEWS = Modified Early Warning Score, NEWS = National Early Warning Score, SPTTS = Single-Parameter Track-and-Trigger System.

standard. When the number of subsequent alarms increased, the PPV (*solid line*) increased to a certain level and then decreased (**Fig. 5*A***). Similarly, the F-measure (*dashed line*) increased as the number of subsequent alarms increased and started to decrease faster than the PPV. For the F-measure, a cutoff value of 99 with three consequent alarms had the highest value (23.35%). The results obtained during the RRS operating period were similar to those obtained during the RRS nonoperating period (**Fig. 5*B***), except that the optimal number (6–8) of subsequent alarms was higher.

Early Prediction Performance

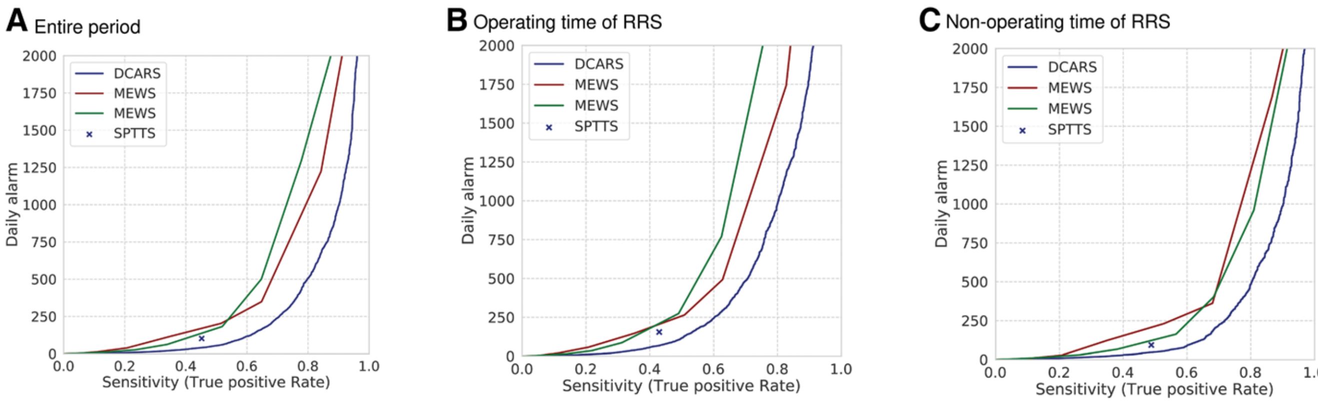
We constructed a plot of the cumulative percentage of IHCA patients versus the prediction time. **Additional File Figure S4** (http://links.lww.com/CCM/H464) shows that DeepCARS consistently detected more IHCA patients at all time points than MEWS. The mean prediction time for MEWS was 10.71, whereas that of DeepCARS was 13.34, indicating DeepCARS predicted IHCA onset by 3 hours earlier than MEWS, on average.

Subgroup Analysis

We assessed the performance of DeepCARS, MEWS, and NEWS across different periods (day, evening, and night on weekdays and weekends) by analyzing the cohort (**Additional File Fig. S5**, http://links.lww.com/CCM/H464). DeepCARS demonstrated consistent and superior performance compared with conventional EWSs across all periods. In contrast, MEWS and NEWS performance was unstable across different periods, with the AUROC of MEWS fluctuating from 0.806 to 0.759. A similar tendency was observed for NEWS.

DISCUSSION

To our knowledge, no previous study has undertaken performance validation of an AI-based cardiac arrest prediction system in a center with a part-time RRS, by comparing performance during the RRS operating and nonoperating periods. Vital signs are crucial predictors of IHCA. Yet, vital sign data collected during RRS operating and nonoperating periods varied in terms of measurement frequency and interval, for various reasons, including the intention to avoid disturbing



method	cutoff	sensitivity	daily alarm	alarm rate
MEWS	3	0.6478	349.3	100%
DCARS	0.8657	0.6478	164.3	47.04
MEWS	4	0.5155	204	100%
DCARS	0.9395	0.5155	59.3	29.07
MEWS	5	0.3509	117.8	100%
DCARS	0.9735	0.3509	21.7	18.42
NEWS	3	0.5191	182.5	100%
DCARS	0.9385	0.5191	60.5	33.15
NEWS	4	0.3388	62.7	100%
DCARS	0.9752	0.3388	20.1	32.06
NEWS	5	0.236	26.9	100%
DCARS	0.9875	0.236	9.7	36.06
SPTTS	1	0.4518	102.5	100%
DCARS	0.9545	0.4518	41.2	40.20

method	cutoff	sensitivity	daily alarm	alarm rate
MEWS	3	0.628	492.6	100%
DCARS	0.8806	0.628	274.3	55.68
MEWS	4	0.508	264.9	100%
DCARS	0.9415	0.508	109.9	41.49
MEWS	5	0.352	147.2	100%
DCARS	0.9734	0.352	42.1	28.60
NEWS	3	0.49	274.4	100%
DCARS	0.9458	0.49	99.7	36.33
NEWS	4	0.313	87.1	100%
DCARS	0.9792	0.313	31.8	36.51
NEWS	5	0.218	35.6	100%
DCARS	0.9888	0.218	16.7	46.91
SPTTS	1	0.429	155.4	100%
DCARS	0.9588	0.429	70.4	45.30

method	cutoff	sensitivity	Daily alarm	alarm rate
MEWS	3	0.68	362.1	100%
DCARS	0.8842	0.68	180.1	49.74
MEWS	4	0.527	231	100%
DCARS	0.9525	0.527	55.1	23.85
MEWS	5	0.35	123.7	100%
DCARS	0.9779	0.35	21.9	17.70
NEWS	3	0.566	163.4	100%
DCARS	0.9438	0.566	68.6	41.98
NEWS	4	0.381	66.9	100%
DCARS	0.9741	0.381	26.4	39.46
NEWS	5	0.265	30.6	100%
DCARS	0.9857	0.265	13.9	45.42
SPTTS	1	0.488	93.7	100%
DCARS	0.9582	0.488	47.1	50.27

Figure 5. F-measure and positive-predictive value graph based on varying numbers of subsequent alarms. PPV = positive-predictive value, RRS = Rapid Response System.

the patient's sleep or burdening the medical staff with unnecessary alarms when the RRS is nonoperational. Therefore, curiosity about the predictive power of the DeepCARS during these periods prompted the current study. Consequently, we assessed the ability of DeepCARS to predict patient deterioration and IHCA within 24 hours during both RRS operating and non-operating periods and compared this ability with that of conventional EWSs. We found that DeepCARS outperformed conventional EWSs in predicting IHCA in both periods, with consistent AUROCs across RRS operating and nonoperating periods.

AI techniques have been used in several studies to predict adverse events in patients, mainly in the ICU setting, by using continuous vital signs and numerous diagnostic tests (18–22). However, only a few studies have specifically focused on patients with deteriorating status in general wards, and have mostly relied on machine learning algorithms using a large number of variables, including demographics, vital signs, and laboratory test results (23, 24). In contrast, we previously demonstrated that deep-learning techniques are more effective than conventional EWSs (7, 8). Furthermore, previous investigations (7, 8, 17, 25, 26) did not examine the ability of AI to predict IHCA by distinguishing between the RRS operating and non-operating periods. In this study, we explored the use of deep learning-based AI techniques for predicting IHCA by comparing the occurrences during periods with and without an operating RRS.

We identified an additional rule for DeepCARS that can enhance its F-measure, particularly when clinical resources are scarce. This rule requires DeepCARS to yield a score above a certain threshold, which reduces the number of alarms presented, while maintaining sensitivity, and results in an improved F-measure. However, there is a tradeoff between early detection and performance, and thus it would be better applied only during the RRS nonoperating period. Even when DeepCARS encountered a high score, the AI system had to wait for the rule to be fulfilled before generating an alarm. However, instances of incorrect alarms and the resulting desensitization to alarms pose significant risks to patient safety, potentially leading to delayed responses in critical situations. Thus, adoption of an improved AI system with an additional rule for reducing false alarms will help to decrease the RRS workload and facilitate proper decision-making. Further studies are required to confirm the effectiveness of this method.

Our study had some limitations that should be considered. First, this study was conducted retrospectively, and a well-designed prospective clinical trial is required to demonstrate the effectiveness of DeepCARS further as a screening tool in clinical practice. Second, our findings were obtained from a single tertiary care hospital affiliated with a university; thus, it may not be reasonable to expect similar benefits from implementing DeepCARS in all hospitals. Consequently, the generalizability of our results is limited.

CONCLUSIONS

The predictive performance of DeepCARS was consistently superior across both RRS operating and nonoperating periods as compared with MEWS, NEWS, and SPTTS, and thereby demonstrated potential effective use in hospitals with various types of RRS. Furthermore, DeepCARS demonstrated superior performance when the RRS was nonoperational, indicating its potential for use in hospitals without an RRS.

1 Department of Research and Development, VUNO, Seoul, Republic of Korea.

2 Department of Anesthesiology and Pain Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea.

3 Department of Critical Care Medicine and Emergency Medicine, Inha University Hospital, Incheon, Republic of Korea.

4 Anesthesia and Pain Research Institute, Yonsei University College of Medicine, Seoul, Republic of Korea.

5 Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Drs. Cho and Kim contributed equally.

Drs. Yoo and Kim supervised this work.

Drs. Cho, K.H. Kim, Yoo, and J. Kim participated in the study design and data acquisition, analysis, or interpretation of data. Drs. Cho, K.H. Kim, Yoo, and J. Kim participated in the drafting of the article. Drs. Cho, Choi, and Yoo participated in statistical analysis. Drs. Yoo and J. Kim participated in supervision. All authors read and approved the final article.

Drs. Yoo and Kim's institutions received funding from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HI20C2125). Dr. Choi disclosed work for hire. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: Anesjeongmin@yuhs.ac

This study was strictly observational and conducted based on anonymity. The ethics committee and institutional review board of Yonsei University Hospital approved the study (approval number and date: 2-2021-1353 and November 16, 2021) and waived the need to obtain informed consent, due to the minimal risk when using data collected for routine clinical practice. The study was conducted in accordance with the ethical standards of the responsible committee on human experimentation and the Helsinki Declaration of 1975.

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

REFERENCES

1. Andersen LW, Kim WY, Chase M, et al; American Heart Association's Get With the Guidelines(®)–Resuscitation Investigators: The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest. *Resuscitation* 2016; 98:112–117
2. Benjamin EJ, Blaha MJ, Chiuve SE, et al; American Heart Association Statistics Committee and Stroke Statistics Subcommittee: Heart disease and stroke statistics-2017 update: A report from the American Heart Association. *Circulation* 2017; 135:e146–e603
3. Jones DA, DeVita MA, Bellomo R: Rapid-response teams. *N Engl J Med* 2011; 365:139–146
4. Smith GB, Prytherch DR, Schmidt PE, et al: Review and performance evaluation of aggregate weighted "track and trigger" systems. *Resuscitation* 2008; 77:170–179
5. Kramer AA, Sebat F, Lissauer M: A review of early warning systems for prompt detection of patients at risk for clinical decline. *J Trauma Acute Care Surg* 2019; 87(1S Suppl 1):S67–S73
6. Romero-Brufau S, Huddleston JM, Naessens JM, et al: Widely used track and trigger scores: Are they ready for automation in practice? *Resuscitation* 2014; 85:549–552
7. Cho KJ, Kwon O, Kwon JM, et al: Detecting patient deterioration using artificial intelligence in a Rapid Response System. *Crit Care Med* 2020; 48:e285–e289
8. Kwon JM, Lee Y, Lee Y, et al: An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc* 2018; 7:e008678
9. Churpek MM, Yuen TC, Park SY, et al: Derivation of a cardiac arrest prediction model using ward vital signs*. *Crit Care Med* 2012; 40:2102–2108
10. Jacobs I, Nadkarni V, Bahr J, et al; International Liaison Committee on Resuscitation: Cardiac arrest and cardiopulmonary resuscitation outcome reports: Update and simplification of the Utstein templates for resuscitation registries: A statement for healthcare professionals from a task force of the International Liaison Committee on Resuscitation (American Heart Association, European Resuscitation Council, Australian Resuscitation Council, New Zealand Resuscitation Council, Heart and Stroke Foundation of Canada, InterAmerican Heart Foundation, Resuscitation Councils of Southern Africa). *Circulation* 2004; 110:3385–3397

11. van Galen LS, Struik PW, Driesen BE, et al: Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: A root cause analysis of unplanned ICU admissions. *PLoS One* 2016; 11:e0161393
12. van der Zee EN, Termorshuizen F, Benoit DD, et al: One-year mortality of cancer patients with an unplanned ICU admission: A cohort analysis between 2008 and 2017 in the Netherlands. *J Intensive Care Med* 2022; 37:1165–1173
13. Orosz J, Bailey M, Udy A, et al: Unplanned ICU admission from hospital wards after rapid response team review in Australia and New Zealand. *Crit Care Med* 2020; 48:e550–e556
14. Miles AH, Spaeder MC, Stockwell DC: Unplanned ICU transfers from inpatient units: Examining the prevalence and preventability of adverse events associated with ICU transfer in pediatrics. *J Pediatr Intensive Care* 2016; 5:21–27
15. Kurita T, Nakada TA, Kawaguchi R, et al; IHER-J collaborators: Impact of increased calls to Rapid Response Systems on unplanned ICU admission. *Am J Emerg Med* 2020; 38:1327–1331
16. Kellett J, Sebat F: Make vital signs great again—a call for action. *Eur J Intern Med* 2017; 45:13–19
17. Lee YJ, Cho KJ, Kwon O, et al: A multicentre validation study of the deep learning-based early warning score for predicting in-hospital cardiac arrest in patients admitted to general wards. *Resuscitation* 2021; 163:78–85
18. Hyland SL, Faltys M, Huser M, et al: Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* 2020; 26:364–373
19. Nemati S, Holder A, Razmi F, et al: An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 2018; 46:547–553
20. Gutierrez G: Artificial intelligence in the intensive care unit. *Crit Care* 2020; 24:101
21. Gong KD, Lu R, Bergamaschi TS, et al: Predicting intensive care delirium with machine learning: Model development and external validation. *Anesthesiology* 2023; 138:299–311
22. Pirracchio R, Petersen ML, Carone M, et al: Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir Med* 2015; 3:42–52
23. Churpek MM, Yuen TC, Winslow C, et al: Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014; 190:649–655
24. Churpek MM, Yuen TC, Winslow C, et al: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44:368–374
25. Lee Y, Kwon JM, Lee Y, et al: Deep learning in the medical domain: Predicting cardiac arrest using deep learning. *Acute Crit Care* 2018; 33:117–120
26. Park SJ, Cho KJ, Kwon O, et al: Development and validation of a deep-learning-based pediatric early warning system: A single-center study. *Biomed J* 2022; 45:155–168