A Multicenter Validation Study of the Deep Learning-based Early Warning Score for Predicting in-hospital Cardiac Arrest in Patients Admitted to General Wards

Yeon Joo Lee
Seoul National University Bundang Hospital
Kyung-Jae Cho
vuno
Oyeon Kwon
VUNO
Hyunho Park
VUNO
Yeha Lee
VUNO
Joon-Myoung Kwon
Sejong Hospital
Jinsik Park
Sejong Hospital
Jung Soo Kim
Inha University College of Medicine
Man-Jong Lee
Inha University Hospital
Ah Jin Kim
Inha University Hospital
Ryoung-Eun Ko
Samsung Medical Center
Kyeongman Jeon
Samsung Medical Center
You Hwan Jo (emdrjyh@gmail.com)
Seoul National University Bundang Hospital https://orcid.org/0000-0002-9507-7603

Research

Keywords: Cardiac arrest, Prediction, Deep learning, Early warning score, Artificial Intelligence, Rapid Response System

Posted Date: August 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-61577/v1

Abstract

Background: The recently developed deep learning (DL)-based early warning score (DEWS) has shown a potential in predicting deteriorating patients. We aimed to validate DEWS in multiple centers and compare the prediction, alarming and timeliness performance with those of the modified early warning score (MEWS) to identify patients at risk for in-hospital cardiac arrest (IHCA).

Methods: This retrospective cohort study included adult patients admitted to the general wards of five hospitals during a 12-month period. We validated DEWS internally at two hospitals and externally at the other three hospitals. The occurrence of IHCA within 24 hours of vital sign observation was the outcome of interest. We used the area under the receiver operating characteristic curve (AUROC) as the main performance metric.

Results: The study population consisted of 173,368 patients (224 IHCAs). The predictive performance of DEWS was superior to that of MEWS in both the internal (AUROC: 0.860 *vs.* 0.754, respectively) and external (AUROC: 0.905 *vs.* 0.785, respectively) validation cohorts. At the same specificity, DEWS had a higher sensitivity than MEWS, and at the same sensitivity, DEWS had a lower mean alarm count than MEWS, with nearly half of the alarm rate in MEWS. Additionally, DEWS was able to predict more IHCA patients in the 24 to 0.5 hours before the outcome.

Conclusion: Our study showed that DEWS was superior to MEWS in the three key aspects (IHCA predictive, alarming, and timeliness performance). This study demonstrates the potential of DEWS as an effective, efficient screening tool in rapid response systems (RRSs) to identify high-risk patients.

Introduction

A rapid response system (RRS) is a strategy for preventing cardiac arrest or deterioration by providing immediate and efficient interventions by tracking and monitoring patients' conditions [1, 2]. The system is composed of four major components: the afferent limb, the efferent limb, the administrative limb and quality monitoring [3]. The afferent limb of an RRS is a sensing structure that identifies "deteriorating patients" in the general ward and properly triggers the efferent limb of the RRS to provide a potential higher level of care [3].

To effectively identify these at-risk patients, several early warning scores (EWSs) have been developed. Because of the limited RRS resources available, an ideal EWS should have high specificity and sensitivity, ensuring the correct identification of at-risk patients while avoiding excessive alarm, which can increase RRS staff desensitization and decrease quality of care [4, 5]. However, a representative EWS, such as the modified EWS (MEWS) and national EWS (NEWS) [6-10], have shown variable accuracy, and the performance of these scores is not satisfactory for the sole use of triggering RRS activation [11-13].

In 2018, a deep learning (DL)-based early warning score, called the DL-based early warning score (DEWS), was consist of only 4 basic vital signs [14]. DEWS measures the real-time risk of cardiac arrest (CA) within

24 hours from vital sign observation. It provides a risk score from zero to 100; the higher the score is, the higher the risk. DEWS showed quite good potential in predicting in-hospital CA (IHCA), higher sensitivity, and a lower false alarm rate than MEWS in the original development study [14]. The original study was performed in 2 hospitals. Each hospital had approximately 300 beds: one was a cardiovascular-specific hospital, and the other was a community general hospital. Therefore, we aimed to validate DEWS in a large multicenter cohort and compare the IHCA predictive performance of DEWS with that of MEWS.

Methods

Study population

A retrospective cohort study was performed in 5 hospitals located in South Korea: Mediplex Sejong Hospital (323 beds), Sejong General Hospital (301 beds), Inha University Hospital (925 beds), Seoul National University Bundang Hospital (1324 beds) and Samsung Medical Center (1989 beds). The two hospitals (hospitals A and B), where the original DEWS was developed [14], were included for internal validation (different periods from the original study), and the other three hospitals (hospitals C, D, and E) were included for external validation. All hospitals have a mature RRS except hospital A. The study population consisted of adult patients (≥18 years old) admitted to the general ward over a 12-month period. The primary outcome of interest was IHCA (defined as lack of a palpable pulse with attempted resuscitation). We collected all the vital signs of the included patients, and DEWS and MEWS were calculated. We compared the performance of DEWS and MEWS to predict IHCA within 24 hours of vital sign measurement. We excluded patients with data recorded less than 30 minutes in the admission no vital sign data measured 24 hours before the CA event, and erroneous patient demographics.

Data collection and preprocessing

A series of five basic vital signs, namely, systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate (HR), respiratory rate (RR), and body temperature (BT), were collected during the hospitalization of the patients and were abstracted from the electronic medical records (EMRs). From the initially collected data, erroneous values extremely outside of the normal range of each vital sign or nonnumeric values were excluded and treated as missing values. Missing values were imputed to the most recent previous value. Normalization is important for stabilizing the DL model during training[15]. Therefore, we normalized each vital sign by subtracting the values from the mean and dividing them by the standard deviation.

Deep learning-based early warning system

The DEWS architecture includes three bidirectional long short-term memory (LSTM) layers, three fully connected (FC) layers with the rectified linear unit, and a softmax layer at the end to output a score between 0 and 1. To reflect the vital sign trend for each patient, 20 consecutive series of vital signs are

used as input to the LSTM layer[16]. Dropout and batch normalization are used on each FC layer to stabilize and regularize the DL model[17, 18]. The final IHCA risk score is calculated by multiplying the softmax output by 100. Before passing through the LSTM layer to the FC layer, we use the output of the last time step. The Adam optimizer is used to train the DEWS model with the default parameters and the binary cross-entropy as the loss function[19]. The hyperparameters are tuned with the best performance from 20% of the derivation data. To address the class imbalance problem, we adjusted the ratio of nonevent/event data in the training process by duplicating the data labeled as events.

Performance evaluation and statistical analysis

We compared the performance of DEWS and MEWS in terms of the following three main key questions:

- ***Key question 1: How accurate is DEWS in terms of predicting IHCA compared with MEWS (predictive performance)?***

The predictive performance is measured by comparing the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) [20, 21]. The AUROC is one of the most commonly used metrics and represents the area under the sensitivity versus false positive rate curve. Compared with the AUROC, the AUPRC accounts for the class imbalance data by measuring the area under the plot of the precision versus the sensitivity. Additionally, we compared DEWS with MEWS in terms of the positive predictive value (PPV = true positive / (true positive + false positive)), the negative predictive value (NPV = true negative / (true negative + false negative)), F measure (2 x (precision x recall) / (precision + recall)), the net reclassification index (NRI), the mean alarm count per day per 1000 beds (MACPD), and the number needed to examine (NNE) with the same specificity as MEWS[21, 22]. The study concept is demonstrated graphically in supplemental figure 1.

- ***Key question 2: Does DEWS produce a lower false alarm rate than MEWS with the same sensitivity level (alarming performance)?***

The alarm rate is an important criterion for validating the feasibility of EWS. One of the concerns about implementing automated screening systems such as these EWSs is alarm fatigue [23]. A false alarm that does not need to lead to next level of treatment could exhaust RRS staff, causing alarm fatigue and leading to inappropriate alarm responses. Consequently, excessive false alarms and alarm fatigue can result in staff desensitization and missed responses to alerts of clinical significance, putting patient safety and quality of care at substantial risk [24]. Therefore, an ideal EWS should have high sensitivity and a low false alarm rate, so we compared the alarm rate with the MACPD according to the same sensitivity level.

- ***Key question 3: Does DEWS predict IHCA earlier than MEWS at the same specificity level (timeliness performance)?***

It is already well known that a delayed RRS response is associated with a poor patient outcome[25-27]. Recently, the published “Quality metrics for the evaluation of RRS” study defined predictable IHCA as CAs occurring in hospitalized ward patients who met the hospital’s escalation threshold at least 30 minutes prior to and within 24 hours of the event[28]. In this statement paper, the period between 24 hours and 30 minutes prior to an IHCA was hypothesized to allow the RRS enough time to prevent the event[28]. However, compared to intensive care units (ICUs) where vital signs are measured every hour or are continuously monitored, vital signs are usually measured only 3~4 times a day (every 6 or 8 hours) in the general ward. Therefore, it is important to be aware of the at-risk patients as early as possible so that the RRS staff can prepare in advance and perform suitable action in enough time before the event. Therefore, we compared the cumulative prediction percentage of IHCA at the same time point within 24 hours of the event (supplemental figure 1).

Ethics statement

The Institutional Review Board of each hospital approved the study protocol and waived the requirement of informed consent because of the retrospective study design. The IRB number of each participating hospital is as follows: B-1806-477-002(Seoul National University Bundang Hospital), 2018-054(Mediplex Sejong Hospital), 2018-0689(Sejong General Hospital), 2019-09-001-000(Inha University Hospital), and SMC-2019-09-129(Samsung Medical Center).

Results

Baseline characteristics

During the study period of 12 months, 173,368 patients were included from the five hospitals. An internal validation cohort consisting of 14,365 patients with 23 IHCAs and an external validation cohort consisting of 159,003 patients with 201 IHCAs. The incidence of IHCAs in the overall cohort was 1.29/1000 admissions. We plotted the DEWS and MEWS distributions in the IHCA cases (supplemental Fig. 2) using the average DEWS or MEWS within 24 hours of the event. Among the event cases, only 19 cases had a MEWS greater than five points, which is quite a low number. Many more event cases are distributed in a higher range for DEWS than for MEWS, especially in the external validation cohort. The baseline characteristics of the overall cohort are depicted in Table 1.

Table 1
Baseline Characteristics

Characteristics	Overall cohort (hospital A,B,C,D,E)	Internal validation (hospital A,B)	External validation (hospital C,D,E)	P-value
Total admitted patients, n	173,368	14,365	159,003	
Age, y, mean ± SD	57.50 ± 15.82	59.93 ± 16.43	57.30 ± 15.76	< 0.001
Length of stay, median (IQR)	3.01 (1.61–6.74)	3.08 (1.54–7.60)	3.01 (1.63–6.72)	< 0.001
Male, sex, n (%)	86,198 (49.7%)	7,260 (50.5%)	78,938 (49.6%)	0.040
Initial vital signs, mean ± SD				
SBP (mmHg)	126.60 ± 19.92	126.71 ± 18.94	126.60 ± 20.00	0.521
DBP (mmHg)	74.50 ± 12.39	76.15 ± 12.59	74.36 ± 12.36	< 0.001
HR (/min)	77.94 ± 14.50	76.21 ± 15.01	78.22 ± 14.39	< 0.001
RR (/min)	18.11 ± 2.07	17.93 ± 2.01	18.13 ± 2.07	< 0.001
BT (℃)	36.64 ± 0.57	36.72 ± 0.46	36.64 ± 0.87	< 0.001
Vital signs within 24hours before cardiac arrest in cardiac arrest patients, mean ± SD				
SBP (mmHg)	113.82 ± 26.02	111.03 ± 24.55	114.39 ± 26.28	0.180
DBP (mmHg)	66.27 ± 17.24	72.51 ± 17.27	65.05 ± 16.98	< 0.001
HR (/min)	101.24 ± 22.94	100.15 ± 23.36	101.40 ± 22.87	0.569
RR (/min)	21.53 ± 5.44	21.15 ± 5.99	21.63 ± 5.29	0.424
BT (℃)	36.76 ± 0.85	37.03 ± 0.55	36.72 ± 0.87	< 0.001
Initial mental status, No. (%)				
Alert	36,294 (96.4%)	463 (82.5%)	35,831 (96.6%)	< 0.001
Reacting to Voice	796 (2.1%)	22 (3.9%)	774 (2.0%)	

SD standard deviation, IQR interquartile range, SBP systolic blood pressure, DBP diastolic blood pressure, HR heart rate, RR respiratory rate, BT body temperature, IHCA in-hospital cardiac arrest, ICU intensive care unit.

Characteristics	Overall cohort (hospital A,B,C,D,E)	Internal validation (hospital A,B)	External validation (hospital C,D,E)	P-value
Reacting to Pain	189 (0.5%)	14 (2.4%)	175 (0.4%)	
Unresponsive	159 (0.4%)	62 (11.0%)	97 (0.2%)	
Not alert	1,341 (3.5%)	98 (17.4%)	1,243 (3.3%)	
Mental status within 24hours before cardiac arrest, No. (%)				
Alert	129 (71.2%)	7 (100.0%)	122 (70.1%)	
Reacting to Voice	8 (4.4%)	0 (0.0%)	8 (4.5%)	
Reacting to Pain	1 (0.5%)	0 (0.0%)	1 (0.5%)	
Unresponsive	5 (2.7%)	0 (0.0%)	5 (2.8%)	
Not alert	52 (28.7%)	0 (0.0%)	52 (29.8%)	
Number of admissions with outcomes, n				
IHCA	224	23	201	0.329
IHCA/1000 admission	1.29	1.60	1.26	

SD standard deviation, IQR interquartile range, SBP systolic blood pressure, DBP diastolic blood pressure, HR heart rate, RR respiratory rate, BT body temperature, IHCA in-hospital cardiac arrest, ICU intensive care unit.

Key Question 1. Predictive performance of IHCA

As shown in Fig. 1, the performance of DEWS for predicting IHCA was superior to that of MEWS in both the internal (AUROC: 0.860 *vs.* 0.754, respectively) and external (AUROC: 0.905 *vs.* 0.785, respectively) validation cohorts. Additionally, the AUPRC for DEWS was higher than that of MEWS in both the internal (0.012 *vs.* 0.003, respectively) and external (0.017 *vs.* 0.005 respectively) validation cohorts. We validated MEWS at the most commonly used cut-off scores of 3, 4, and 5 in terms of the sensitivity, specificity, PPV, F measure, NPV, NNE, NRI and compared these values to those of DEWS at the same specificity[28, 29]. As shown in Table 2, DEWS achieved higher sensitivity for all the cut-off scores and achieved at most a 230% and 68.7% higher sensitivity than MEWS in the internal validation and external validation cohorts, respectively. The predictive performance of each hospital is shown in supplemental Fig. 3, and DEWS outperformed MEWS in each of five hospitals.

Table 2
Comparison of Accuracy of In-hospital Cardiac Arrest Prediction Model with Same Specificity Point

Characteristics	Sensitivity	Specificity	PPV	NPV	F-measure	NRI	MACPD	NNE
Internal validation cohort								
MEWS ≥ 3	0.484	0.932	0.0011	1	0.002		104	391
DEWS ≥ 53.1	0.548	0.932	0.0029	1	0.006	0.0011	103	342
MEWS ≥ 4	0.419	0.953	0.0032	1	0.006		71	308
DEWS ≥ 60.5	0.484	0.953	0.0037	1	0.007	0.0015	71	269
MEWS ≥ 5	0.234	0.992	0.0106	1	0.007		12	94
DEWS ≥ 87.5	0.306	0.992	0.0136	1	0.026	0.0032	12	73
External validation cohort								
MEWS ≥ 3	0.551	0.908	0.0033	1	0.007		335	302
DEWS ≥ 69.9	0.700	0.908	0.0042	1	0.008	0.0014	334	236
MEWS ≥ 4	0.386	0.958	0.0050	1	0.010		154	191
DEWS ≥ 83.2	0.560	0.958	0.0073	1	0.032	0.0024	154	137
MEWS ≥ 5	0.230	0.989	0.0117	1	0.022		39	85
DEWS ≥ 94.1	0.338	0.989	0.0166	1	0.032	0.0052	41	60

PPV positive predictive value, NPV negative predictive value, NRI net reclassification improvement, MACPD mean alarm count per day per 1000 beds, NNE number needed to examine, MEWS modified early warning score, DEWS deep learning-based early warning score

Key Question 2. Alarming performance

We compared DEWS and MEWS by the MACPD at the same sensitivity level. As shown in Fig. 2, at the same sensitivity level, DEWS achieved a lower alarm rate than MEWS. This result indicates that DEWS can detect the same number of deteriorating patients with a much lower false alarm rate than MEWS. Specifically, at MEWS cut-offs of 3, 4, and 5, DEWS produced at most 66.7% and 63% fewer alarms than MEWS in the internal and external validation cohorts, respectively.

Key Question 3: Timeliness performance

We validated DEWS and MEWS by enrolling IHCA patients at the time point where the early warning score first triggered the alarm from 24 to 0.5 hours before the CA occurred. As shown in Fig. 3, DEWS detected more patients with CA in this period than MEWS. Especially in the external validation cohort, DEWS

detected 10 and 20 more IHCA patients 20 and 15 hours before the event, respectively. This finding indicates that DEWS can not only predict more IHCA patients within 24 hours but can also detect more patients in advance and thus save time for the medical team to effectively manage patients at risk.

Discussion

We evaluated the ability of DEWS to predict IHCA in general ward-admitted patients in a large multicenter cohort. The results of all three key questions (predictive performance of IHCA, alarming performance, timeliness performance) for DEWS were superior to those of MEWS. In both cohorts, DEWS achieved better performance in predicting IHCA within 24 hours of vital sign observation than MEWS: DEWS achieved 14.0% (300%) and 15.2% (240%) higher AUROCs (AUPRCs) than MEWS, respectively. Alarms are a very sensitive issue for RRS teams because they are eventually associated with the team's workload. In this study, the alarm rate of DEWS was 44.2% that of MEWS for a cut-off score of 3, 37.0% that of MEWS for a cut-off score of 4, and 48.7% that of MEWS for a cut-off score of 5 in the external validation cohort. DEWS has nearly half of the alarm rate of MEWS. The third key question was the timeliness of the prediction. At every time point from 24 hours to 30 minutes before the event, DEWS detected more IHCA cases at the same time point than MEWS. It enables RRSs to evaluate and assess deteriorating patients with more time to respond. Therefore, better prediction with fewer alarms and earlier prediction indicate that DEWS has the potential to be an effective alternative screening tool for conventional early warning systems.

Various studies have attempted to predict mortality in critically ill patients (i.e., those in ICUs) using machine learning (ML) or DL [30–34]. ICUs, in particular, have many databases for continuous vital sign monitoring and large numbers of diagnostic tests, including laboratory tests, imaging tests, microbiologic reports, medical history panels, patient demographics, ordered fluids, drugs, transfusions, etc. This large database enables ICUs to be a setting for which to conduct artificial intelligence (AI)-based studies. Most AI-based ICU studies have studied mortality or major event prediction (such as hypotension, sepsis, readmission), and generally, algorithm-based prediction achieved better performance compared to conventional prognostic systems [35, 36]. Furthermore, a study using reinforcement learning in sepsis patients showed the potential to solve a complex medical problem and suggest individualized and clinically interpretable treatment strategies for sepsis [37].

However, few studies have focused on deteriorating patients admitted to general wards. In 2016, Churpek et al's study [38] showed that an ML (i.e., random forest) algorithm (AUROC 0.80, 95% confidence interval (CI) [0.80–0.80]) predicted clinical deterioration more accurately than MEWS (AUROC 0.70, 95% CI [0.70–0.70]) in general ward patients. In Churpek et al's study, an ML method was used to develop the prediction algorithm. Both ML and DL methods analyze data through self-learning to solve the task or problem. ML requires feature engineering, whereas DL does not; rather, it tries to learn the representation of the raw data in multiple levels of abstractions by itself, which is the essence of why DL methods achieve higher

accuracy than ML methods [15]. Alvin Rajkomar et al. demonstrated the effectiveness of DL models in a wide variety of predictive problems and settings [39]. However, the study did not focus on general ward patients and sudden CA but rather on the entire length of stay, including the general ward and the ICU. The outcomes of interest were inpatient mortality, readmission, length of stay and discharge diagnoses. Thus, to the best of our knowledge, our study is the first to apply DL to detect deteriorating patients in general wards in a large multicenter cohort.

The strength of DEWS is that it consists of a limited number of basic vital signs as predictor variables. In this validation study, DEWS used only five basic vital signs: SBP, DBP, HR, RR and BT. The two previous AI-based studies [38, 39] in general ward patients used a variety of predictor variables, including demographics, vital signs, laboratory values, etc. Prediction models with more variables would have better predictability, but there are significant limitations to the scalability and applicability of models with many variables. The predictor variables used in DEWS are basic essential vital signs that are almost always checked in admitted patients and lack missing data. Therefore, DEWS can be applied worldwide without any difficulties in technical implementation. Additionally, a DL-based algorithm enables each institution to have tailored approach by adding one or two main variables depending on the specific features of the hospital. [40].

Five hospitals in South Korea participated in this validation study. The characteristics of each hospital are quite different in terms of the locations, hospital sizes, admitted patients and operating policies. The two hospitals involved in the internal validation have approximately 300 beds; one is a cardiovascular-specific hospital, and the other is community general hospital. The hospitals in the external validation have more than 900 ~ 1000 beds, and all three hospitals are tertiary teaching hospitals, which are affiliated with each of the three different medical universities. Since the original DL model was developed and trained from the two hospitals with 300 beds, the results of the external validation cohort are important in terms of generalization. As a result, DEWS achieved superior performance in the external validation cohort (AUROC 0.905, 95% CI [0.901–0.910]) compared to the internal validation cohort (AUROC 0.860, 95% CI [0.832–0.888]), which suggests that DEWS is robust across multiple hospitals.

Our study has several limitations. First, DL is known as a “black-box” method, as it tries to find the relationship between the training data and the labels rather than creating rules using domain knowledge. Although, most of DEWS alarm can be interpretable in clinical practice through patient review, there would be some cases that RRS staff does not know the exact reason for the alarm. Therefore, RRS staff would need a certain amount of time to react. In this study, DEWS reduces the number of false alarms and increases the sensitivity at the same time. Thus, the rapid response team can spare enough time to verify the alarms, and the staff can intuitively speculate the reason. Second, we consider only the first CA for each patient admission, although second and third CAs are also important. Nonetheless, the first CA is the highest priority because the rapid response team focuses on patients after CA. Last, this study was performed in a retrospective manner. To apply DEWS in clinical practice as an alternative to other triggering score systems in RRS, a well-designed prospective clinical trial is necessary.

Conclusion

We compared DEWS and MEWS in multiple centers via extensive experiments. The results showed that DEWS not only predicts deteriorating patients more accurately than MEWS but also reduces the false alarm rate. Additionally, DEWS was able to predict more CA patients in the period from 24 hours to 0.5 hours before the outcome than MEWS. This finding demonstrates the potential of DEWS as an effective screening tool in RRSs that can be efficiently applied to identify high-risk patients.

Declarations

Ethical Approval and Consent to participate:

The study was conducted in accordance with the guiding principles of the Declaration of Helsinki and was approved by the local institutional review boards of each participating hospital. Written informed consent was waived because of the retrospective manner of the study.

Consent for publication:

Not applicable

Availability of data and materials:

The datasets used for the analysis in the current study are available from the corresponding author on reasonable request.

Competing interests:

None

Funding:

None

Author Contributions:

Dr YHJ and YJL had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: YJL, KJC, OK, HP, YL, JMK, JP, JSK, MJL, AJK, REK, KJ, YHC

Acquisition, analysis, or interpretation of data: KJC, OK, YJL, YHJ

Drafting of the manuscript: YJL, KJC

Critical revision of the manuscript for important intellectual content: YJL, KJC, OK, HP, YL, JMK, JP, JSK, MJL, AJK, REK, KJ, YHC

Statistical analysis: KJC, OK, HP, YL

Administrative, technical, or material support: YJL, JMK, JP, JSK, MJL, AJK, REK, KJ

Supervision: OK, HP, YL, JMK, JP, JSK, MJL, AJK, REK, KJ, YHC

Image analysis: YJL, KJC, OK, HP

Acknowledgements:

Not applicable

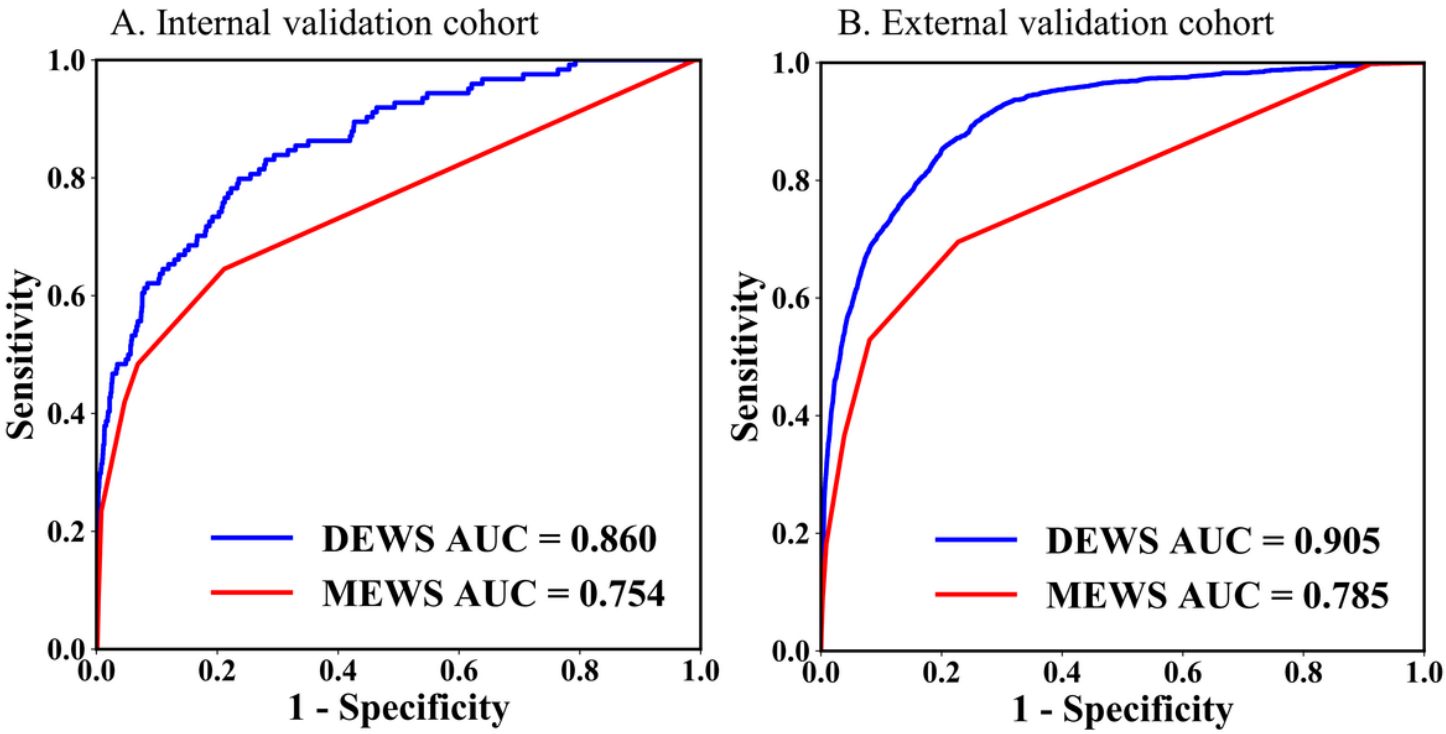
References

1. Jones DA, DeVita MA, Bellomo R. Rapid-response teams. N Engl J Med. 2011;365:139–46.
2. Kim Y, Lee DS, Min H, et al. Effectiveness Analysis of a Part-Time Rapid Response System During Operation Versus Nonoperation. Crit Care Med. 2017;45:e592-e9.
3. Winters BD, Pham JC, Hunt EA, Guallar E, Berenholtz S, Pronovost PJ. Rapid response systems: a systematic review. Crit Care Med. 2007;35:1238–43.
4. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs. Crit Care Med. 2012;40:2102.
5. Lee BY, Hong S-B. Rapid response systems in Korea. Acute Critical Care. 2019;34:108.
6. Duckitt R, Buxton-Thomas R, Walker J, et al. Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. An observational, population-based single-centre study. Br J Anaesth. 2007;98:769–74.
7. Paterson R, MacLeod D, Thetford D, et al. Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. Clin Med (Lond). 2006;6:281.
8. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS—towards a national early warning score for detecting adult inpatient deterioration. Resuscitation. 2010;81:932–7.
9. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation. 2013;84:465–70.

10. Subbe C, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM. 2001;94:521–6.
11. Romero-Brufau S, Huddleston JM, Naessens JM, et al. Widely used track and trigger scores: are they ready for automation in practice? Resuscitation. 2014;85:549–52.
12. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI. Review and performance evaluation of aggregate weighted 'track and trigger'systems. Resuscitation. 2008;77:170–9.
13. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI, Higgins B. A review, and performance evaluation, of single-parameter "track and trigger" systems. Resuscitation. 2008;7:11–21.
14. Kwon Jm, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. Journal of the American Heart Association. 2018;7:e008678.
15. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521:436–44.
16. Hochreiter S, Schmidhuber J. Long short-term memory. **Neural Comput**. 1997;9:1735–80.
17. Santurkar S, Tsipras D, Ilyas A, Madry A. How does batch normalization help optimization? 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).
18. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. JMLR. 2014;15:1929–58.
19. Kingma DP, Ba J. Adam. A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
20. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015;68:855–9.
21. Weng CG, Poon J, editors. A new evaluation measure for imbalanced datasets. Proceedings of the 7th Australasian Data Mining Conference-Volume 87; 2008.
22. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med. 2014;160:122–31.
23. Welch J, Kanter B, Skora B, et al. Multi-parameter vital sign database to assist in alarm optimization for general care units. J Clin Monit Comput. 2016;30:895–900.
24. Nguyen J, Davis K, Guglielmello G, Stawicki SP. Combating Alarm Fatigue: The Quest for More Accurate and Safer Clinical Monitoring Equipment. Vignettes in Patient Safety - Volume 4, 2019 DOI:10.5772/intechopen.84783.
25. Barwise A, Thongprayoon C, Gajic O, Jensen J, Herasevich V, Pickering BW. Delayed rapid response team activation is associated with increased hospital mortality, morbidity, and length of stay in a tertiary care institution. Crit Care Med. 2016;44:54–63.
26. Boniatti MM, Azzolini N, Viana MV, et al. Delayed medical emergency team calls and associated outcomes. Crit Care Med. 2014;42:26–30.
27. Chen J, Bellomo R, Flabouris A, et al. The relationship between early emergency team calls and serious adverse events. Crit Care Med. 2009;37:148–53.

28. Subbe CP, Bannard-Smith J, Bunch J, et al. Quality metrics for the evaluation of Rapid Response Systems: Proceedings from the third international consensus conference on Rapid Response Systems. Resuscitation. 2019;141:1–12.
29. Subbe C, Davies R, Williams E, Rutherford P, Gemmell L. Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions. Anaesthesia. 2003;58:797–802.
30. Harutyunyan H, Khachatrian H, Kale DC, Steeg GV, Galstyan A. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:170307771. 2017.
31. Johnson AE, Pollard TJ, Mark RG, editors. Reproducibility in critical care: a mortality prediction case study. Machine Learning for Healthcare Conference; 2017.
32. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: customized mortality prediction. Journal of personalized medicine. 2012;2(4):138–48.
33. Gupta P, Malhotra P, Vig L, Shroff G, editors. Using Features From Pre-trained TimeNET For Clinical Predictions. KHD@ IJCAI; 2018.
34. Kaji DA, Zech JR, Kim JS, et al. An attention based deep learning model of clinical events in the intensive care unit. PloS One. 2019;14:e0211057.
35. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3:42–52.
36. Kamio T, Van T, Masamune K. Use of machine-learning approaches to predict clinical deterioration in critically ill patients: a systematic review. Int J Med Res Health Sci. 2017;6:1–7.
37. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24:1716–20.
38. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med. 2016;44:368.
39. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18.
40. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018;6:905–14.

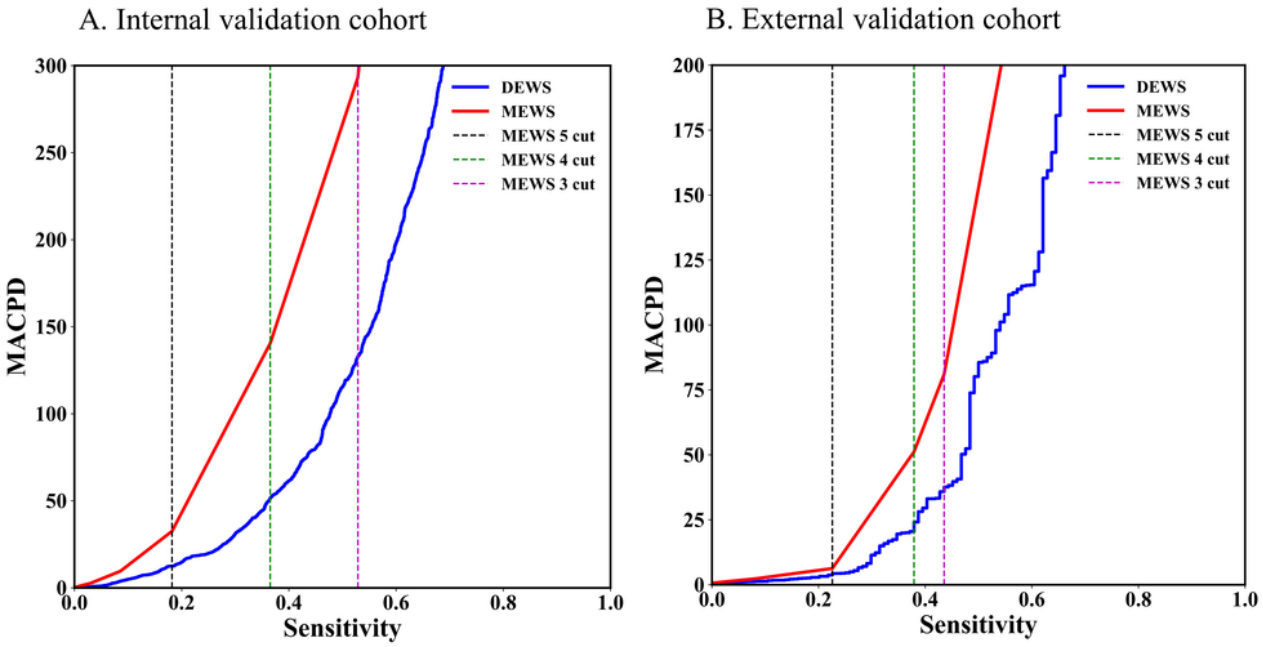
Figures



	AUC	(95% CI)	AUPRC	(95% CI)
Internal Validation				
DEWS	0.860	(0.832 - 0.888)	0.012	(0.007 - 0.019)
MEWS	0.754	(0.716 - 0.789)	0.003	(0.002 - 0.005)
External Validation				
DEWS	0.905	(0.901 - 0.910)	0.017	(0.016 - 0.020)
MEWS	0.785	(0.784 - 0.799)	0.005	(0.005 - 0.007)

Figure 1

Performance of the early warning scores for predicting in-hospital cardiac arrest. DEWS indicates the deep learning-based early warning score, MEWS indicates the modified early warning score, AUROC indicates the area under the receiver operating characteristic curve, AUPRC indicates the area under precision-recall curve, and CI indicates the confidence interval.



	Sensitivity	MACPD	Alarm rate
Internal validation			
MEWS >=3	0.484	104	100%
DEWS >= 62.5		65	62.55
MEWS >=4	0.419	71	100%
DEWS >= 74.8		33	46.50%
MEWS >= 5	0.234	12	100%
DEWS >= 94.7		4	33.30%
External validation			
MEWS >=3	0.551	335	100%
DEWS >= 83.7		148	44.20%
MEWS >=4	0.386	154	100%
DEWS >= 92.3		57	37.00%
MEWS >=5	0.23	39	100%
DEWS >=96.7		19	48.70%

Figure 2

Comparison of the mean alarm count per day per 1000 beds at the same sensitivity point for predicting in-hospital cardiac arrest. MACPD indicates the mean alarm count per day per 1000 beds, DEWS indicates the deep learning-based early warning score, and MEWS indicates the modified early warning score.

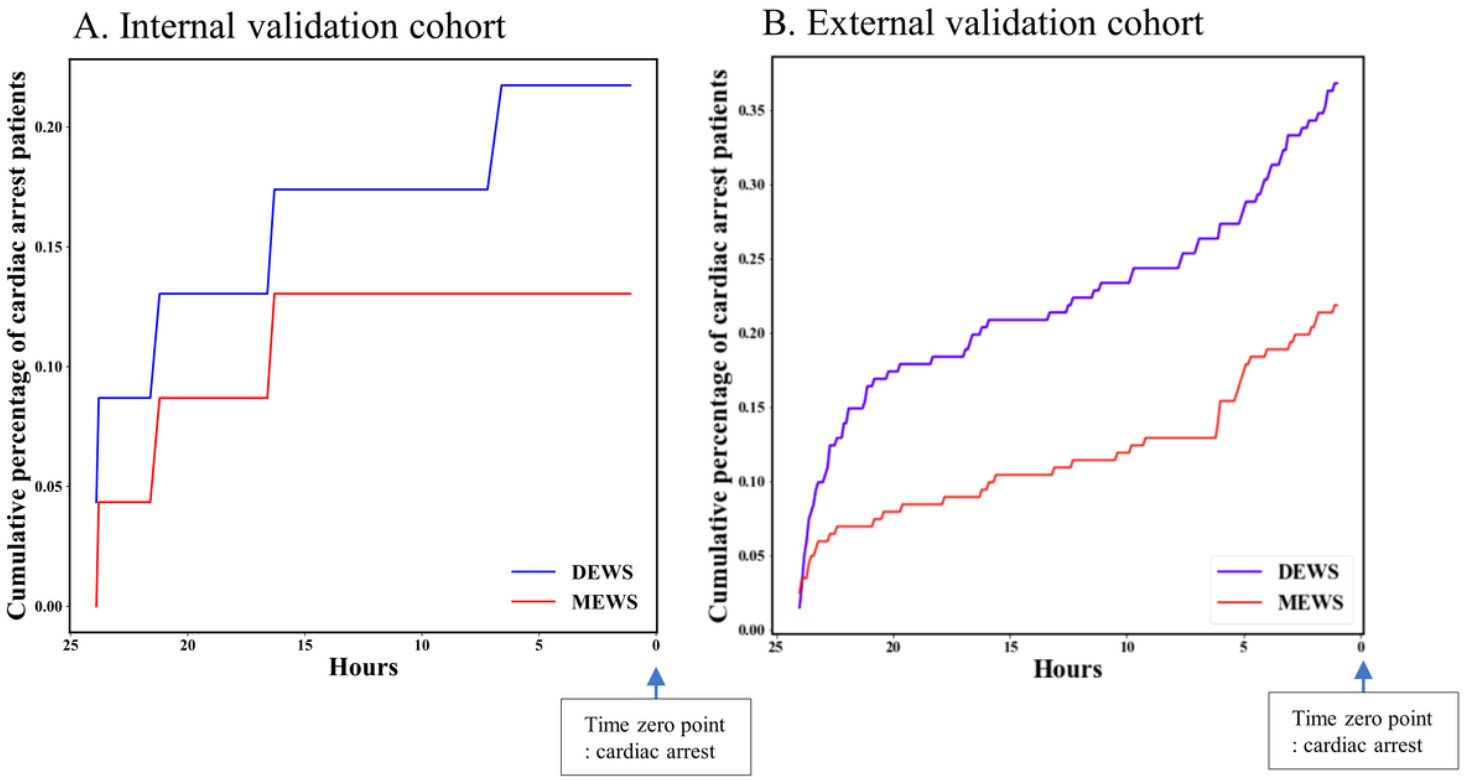


Figure 3

Comparison of the cumulative percentages of cardiac arrest patients. DEWS indicates the deep learning-based early warning score, and MEWS indicates the modified early warning score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supfig3.tif
- supfig2.tif
- supfig1.tif