Available online at www.sciencedirect.com

Resuscitation

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

Clinical paper

A multicentre validation study of the deep earning-based early warning score for predicting in-hospital cardiac arrest in patients admitted to general wards

Yeon Joo Lee[a,1], Kyung-Jae Cho[b,1], Oyeon Kwon[b], Hyunho Park[b], Yeha Lee[b], Joon-Myoung Kwon[c], Jinsik Park[d], Jung Soo Kim[e], Man-Jong Lee[e], Ah Jin Kim[e], Ryoung-Eun Ko[f], Kyeongman Jeon[f,g], You Hwan Jo[h,i,*]

[a] Division of Pulmonary and Critical Care Medicine, Seoul National University Bundang Hospital, Gyeonggi-do, Republic of Korea
[b] VUNO, Seoul, Republic of Korea
[c] Department of Critical Care and Emergency Medicine, Mediplex Sejong Hospital, Incheon, Republic of Korea
[d] Division of Cardiology, Cardiovascular Center, Mediplex Sejong Hospital, Incheon, Republic of Korea
[e] Division of Critical Care Medicine, Department of Hospital Medicine, Inha College of Medicine, Incheon, Republic of Korea
[f] Department of Critical Care Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[g] Division of Pulmonary and Critical Care Medicine, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[h] Department of Emergency Medicine, Seoul National University Bundang Hospital, Gyeonggi-do, Republic of Korea
[i] Department of Emergency Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

Abstract

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

Method/research design: This retrospective cohort study included adult patients admitted to the general wards of five hospitals during a 12-month period. The occurrence of IHCA within 24 h of vital sign observation was the outcome of interest. We assessed the discrimination using the area under the receiver operating characteristic curve (AUROC).

Results: The study population consists 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 reduced the mean alarm count by nearly half of MEWS. Additionally, DEWS was able to predict more IHCA patients in the 24−0.5 h before the outcome, and DEWS was reasonably calibrated.

Conclusion: Our study showed that DEWS was superior to MEWS in 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.

Keywords: Cardiac arrest, Prediction, Deep learning, Early warning score, Artificial intelligence, Rapid response system

* Corresponding author at: Department of Emergency Medicine, Seoul National University Bundang Hospital, 82, Gumi-ro 173 Beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, Republic of Korea.
E-mail address: emdrjyh@gmail.com (Y.H. Jo).

[1] These two authors contributed equally to this study.

https://doi.org/10.1016/j.resuscitation.2021.04.013
Received 24 August 2020; Received in revised form 28 January 2021; Accepted 1 April 2021
Available online xxx

Introduction

A rapid response system (RRS) is a strategy for preventing cardiac arrest (CA) or deterioration in the general ward by providing immediate and efficient interventions by monitoring patients' conditions.[1,2] To effectively identify these at-risk patients, several early warning scores (EWSs) have been developed. Because of the limited RRS resources, an ideal EWS should have high specificity and sensitivity, ensuring the correct identification of the at-risk patients while avoiding excessive alarm, which can increase RRS staff desensitization and decrease quality of care.[3,4]

However, a representative EWSs, such as the modified EWS (MEWS) and national EWS (NEWS),[5–9] have shown unstable accuracies which is not satisfactory for the sole use of triggering RRS activation.[10–12] In 2018, a DL-based early warning score, called DEWS, was developed which considers only 4 basic vital signs: systolic blood pressure (SBP), heart rate (HR), respiratory rate (RR), and body temperature (BT).[13] We have extended from this version of DEWS by adding diastolic blood pressure (DBP), age, and the recorded time of each vital sign.

DEWS measures the risk of CA within 24 h from vital sign observation. DEWS showed potential in predicting in-hospital CA (IHCA) by showing higher sensitivity, and a lower false alarm rate than MEWS in the original development.[13] The original study was performed in 2 hospitals with approximately 300 beds each; one was a cardiovascular-specific hospital, and the other was a community general hospital. Therefore, we aimed to validate DEWS in a large multicentre cohort and compare the IHCA predictive performance of DEWS with that of MEWS.

Methods

Study population and design

A retrospective cohort study was performed in 5 hospitals located in South Korea: Mediplex Sejong Hospital (323 beds), Sejong Hospital (301 beds), Inha University Hospital (925 beds), Seoul National University Bundang Hospital (1324 beds) and Samsung Medical Center (1989 beds). The two hospitals (A: Sejong Hospital and B: Mediplex Sejong Hospital), where the original DEWS was developed,[13] were included for internal validation, and the other three hospitals (C: Inha University Hospital, D: Seoul National University Bundang Hospital, and E: Samsung Medical Center) were included for external validation. All hospitals had a mature RRS except hospital A. The structure of the RRS in each hospital are described in supplement Table 1.

The study population included adult patients ($\geq$18 years old) admitted to the general ward over a 12-month period. We excluded patients with data recorded less than 30 min of admission duration, no vital signs measured 24 h before the CA event, and erroneous patient demographics. The specific details on the participant selection process is reported in supplemental Fig. 1. Since there exists no established method in determining sample size for prognostic models using DL methods, we have chosen the sample size appropriate for our experiments.[14]

The primary outcome of interest was IHCA (defined as lack of a palpable pulse with attempted resuscitation). All vital signs used to predict the outcome of interest were collected for every patient. As the vital signs were measured multiple times per patient, the DEWS and MEWS were calculated at each point of measurement. Finally, performances of DEWS and MEWS were compared by the predictive performance of IHCA within 24 h of vital sign measurement.

Data collection and preprocessing

We collected data including age, sex, occurrence of events, time and location of event occurrences, and five time-stamped vital signs (SBP, DBP, HR, RR, and BT) recorded during hospitalization of the patients abstracted from the electronic medical records (EMRs). From the initial data collected, erroneous values with extreme deviations from the vital specific normal ranges and non-numeric values were treated as missing values. The missing values were imputed to the most recent previous value and the missing rates of each variable are presented in supplemental Table 2.

Deep learning-based early warning system

The DEWS architecture includes three long short-term memory (LSTM) layers and three fully connected (FC) layers with the rectified linear unit. To reflect the trend of the vital signs for each patient, 20 consecutive series of vital signs are used as inputs of the LSTM layers.[15] As a regularization technique, dropouts are applied on each FC layer of the model.[16] The DEWS model was trained using 80% of the derivation data and hyperparameter was tuned on the other 20%.[17] To address the class imbalance problem, we adjusted the ratio of nonevent/event data in the training process by duplicating the event data. From the original DEWS model, we have extended the model by adding DBP along with age and the recorded time of each vital sign.

Performance evaluation and statistical analysis

We have 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 was measured by comparing the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC).[18,19] The AUROC is one of the most commonly used metrics and represents the area under the sensitivity-false positive rate curve. Compared with the AUROC, the AUPRC accounts for the class imbalance in data by measuring the area under the plot of the precision-sensitivity curve. 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 × (precision × 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) at the same specificity as MEWS.[19,20] The study concept is demonstrated graphically in supplemental Fig. 2.

- *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 upon implementation of EWS because excessive false alarms can cause alarm fatigue.[21] Excessive false alarms and alarm fatigue can result to staff desensitization and missed responses to alerts of clinical significance, putting patient safety and quality of care at substantial risk.[22] Therefore, an ideal EWS should have high sensitivity and a low false alarm rate, so we compared the alarming rate of DEWS and MEWS using the MACPD at 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.[23–25] Recently, in the study "Quality metrics for the evaluation of RRS" defined predictable IHCA as CAs occurring in hospitalized ward patients who met the hospital's escalation threshold at least 30 min prior to and within 24 h of the event.[26] In this statement paper, it is hypothesized that the period between 24 h and 30 min prior to IHCA is enough time for RRS to prevent the event.[26] However, compared to intensive care units (ICUs) where vital signs are measured continuously, vital signs are usually measured only 3∼4 times a day (every 6 or 8 h) in the general wards. Therefore, it is important that the RRS staff are aware of the at-risk patients as early as possible so that they can prepare in advance and perform suitable action in enough time before the event. In this respect, we compared the cumulative prediction percentage of IHCA at the same time point within 24 h of the event (supplemental Fig. 2).

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
Number of total admissions, *n*	173,368	14,365	159,003	
Number of observation sets, *n*	5,875,253	342,854	5,532,399	
Number of admissions on telemetry[a], No. (%)	59,567 (34.3%)	694 (4.8%)	58,873 (37.0%)	
Number of observation sets on telemetry, *n*	1,178,270	5310	1,172,960	
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%)	7260 (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 (°C)	36.64 ± 0.57	36.72 ± 0.46	36.64 ± 0.87	<0.001
Vital signs within 24 h 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 (°C)	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%)	
Reacting to Pain	189 (0.5%)	14 (2.4%)	175 (0.4%)	
Unresponsive	159 (0.4%)	62 (11.0%)	97 (0.2%)	
Not alert	1341 (3.5%)	98 (17.4%)	1243 (3.3%)	
Mental status within 24 h 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	
Number of observation sets with outcomes, *n*	3190	124	3066	
Number of admissions with outcome on telemetry, No. (%)	25 (11.1%)	1 (0.8%)	24 (11.9%)	
Number of observation sets with outcomes with telemetry, *n*	186	4	182	

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.

[a] We assumed admissions on telemetry with less than 5 min of vital sign measurement interval.

We assessed the calibration of DEWS using a calibration plot and the average absolute error between the actual outcome and the estimated probabilities.[27,28] The x-axis of the calibration plot is the means of decile-binned predictions, and the y-axis is the means of the observed outcomes in each bin so that well calibrated model will fall close to the diagonal. Additionally, to mitigate the black-box prediction problem, we applied a Shapley additive explanations (SHAP) algorithm to our prediction model to obtain interpretability of the features that drive predictions.[29] SHAP is a game theoretic approach designed to explain the output of a machine learning model where the influence of each feature on a prediction is described using Shapley values.

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).

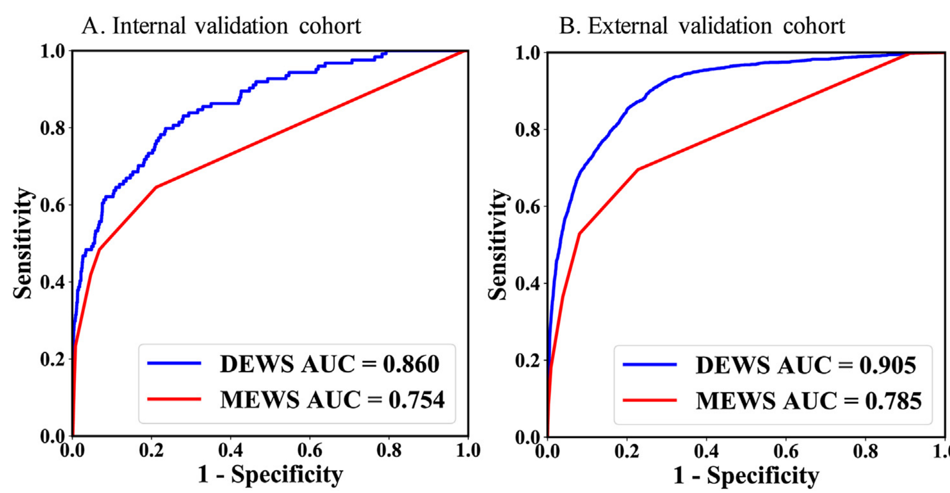
Results

Baseline characteristics

During the study period of 12 months, 173,368 patients from the five hospitals was examined. The internal validation cohort contained 14,365 patients with 23 IHCAs, and the external validation cohort contained 159,003 patients with 201 IHCAs. The incidence rate of IHCAs in the overall cohort was 1.29 per 1000 admissions. We plotted the DEWS and MEWS distributions in the IHCA cases (supplemental Fig. 3) using the average DEWS or MEWS within 24 h of the event. Among the event cases, only 19 cases had a MEWS greater than five points, which was quite a low number. When the number of event cases is compared, more cases are distributed at higher score ranges for DEWS than for MEWS, especially in the external validation cohort. The baseline characteristics of the overall cohort are depicted in Table 1.

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,



	AUC	(95% CI)	AUPRC	(95% CI)	P-value	power
Internal Validation						
DEWS	0.860	(0.832 - 0.888)	0.012	(0.007 - 0.019)	< 0.001	0.999
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)	< 0.001	1.000
MEWS	0.785	(0.784 - 0.799)	0.005	(0.005 - 0.007)		

Fig. 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. P-value was calculated using Delong test. Power was calculated according to the formula by Obuchowski and McClish, 1997.

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, 5, and 6 in terms of the sensitivity, specificity, PPV, F measure, NPV, NNE, NRI and compared these values to those of DEWS at the same specificity.[26,30] As shown in Table 2, DEWS achieved higher sensitivity for all the cut-off scores and achieved at most 228.3% and 63.2% higher sensitivity than MEWS in the internal validation and external validation cohorts, respectively. The predictive performance of each hospital is shown in supplemental Fig. 4, and DEWS outperformed MEWS in each of five hospitals.

Key Question 2. Alarming performance

We compared DEWS and MEWS by the MACPD at the same sensitivity level. As shown in Fig. 2, DEWS achieved a lower MACPD than MEWS. This result indicates that DEWS can detect the same number of deteriorating patients with a much lower false alarm rate than MEWS. For example, at MEWS cut-offs of 3, DEWS produced 62.5% and 44.2% 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 h 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 h before the event, respectively. This finding indicates that DEWS can not only predict more IHCA patients within 24 h but can also detect more patients in advance and thus save time for the medical team to effectively manage patients at risk.

Model calibration

We assessed the calibration of DEWS on the entire cohort. As shown in supplemental Fig. 5, DEWS was reasonably calibrated where the curve approaches close to the diagonal. Quantitatively, the average absolute error between the outcome and the estimated probabilities was 0.046, indicating that the prediction scores and the absolute risk are close to perfect concordance.

Inspection of model features

In supplemental Fig. 6, the overall importance of the predictor variables of DEWS shows HR as the most important feature. The second most important feature was RR, but in the case of other features, it was found that the importance was relatively low. Additionally, the feature importance of DEWS according to the order of consecutive time steps shows a rapid increase in the SHAP value at the most recent time point.

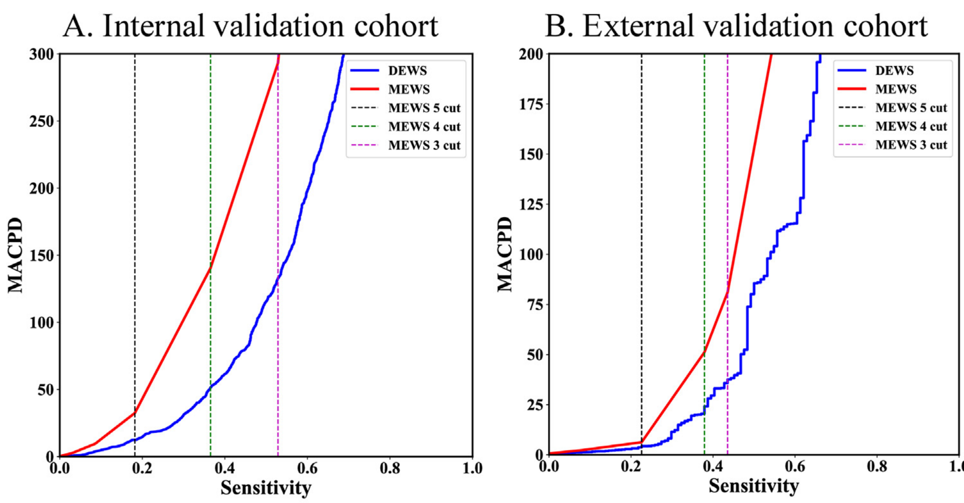
Discussion

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

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 $\geq$ 3	0.484	0.932	0.0011	1	0.002		104	391
DEWS $\geq$ 53.1	0.548	0.932	0.0029	1	0.006	0.0011	103	342
MEWS $\geq$ 4	0.419	0.953	0.0032	1	0.006		71	308
DEWS $\geq$ 60.5	0.484	0.953	0.0037	1	0.007	0.0015	71	269
MEWS $\geq$ 5	0.234	0.992	0.0106	1	0.007		12	94
DEWS $\geq$ 87.5	0.306	0.992	0.0136	1	0.026	0.0032	12	73
External validation cohort								
MEWS $\geq$ 3	0.551	0.908	0.0033	1	0.007		335	302
DEWS $\geq$ 69.9	0.700	0.908	0.0042	1	0.008	0.0014	334	236
MEWS $\geq$ 4	0.386	0.958	0.0050	1	0.010		154	191
DEWS $\geq$ 83.2	0.560	0.958	0.0073	1	0.032	0.0024	154	137
MEWS $\geq$ 5	0.230	0.989	0.0117	1	0.022		39	85
DEWS $\geq$ 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.



	Sensitivity	MACPD	Alarm rate
Internal validation			
MEWS >=3	0.484	104	100%
DEWS >= 62.5		65	62.50%
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%

Fig. 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.

Various studies have attempted to predict mortality in critically ill patients (i.e., those in ICUs) using machine learning (ML).[31–35] 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 an adequate 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 in general, algorithm-based prediction achieved better performances than conventional prognostic systems.[36,37]

However, only a few studies have focused on deteriorating patients admitted to general wards. In 2016, Churpek et al.'s study[38] showed that a ML (i.e., random forest) algorithm (AUROC 0.80) predicted clinical deterioration more accurately than MEWS (AUROC 0.70) in general ward patients. Both ML and DL methods analyse data through self-learning to solve the task or problem. ML requires feature engineering, whereas DL does not; rather, it learns 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 most ML methods.[39] Alvin Rajkomar et al. demonstrated the effectiveness of DL models in a wide variety of predictive problems and settings.[40] However, this study did not focus on general ward patients and sudden CA but rather on the entire length of stay, including both 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 multicentre cohort.

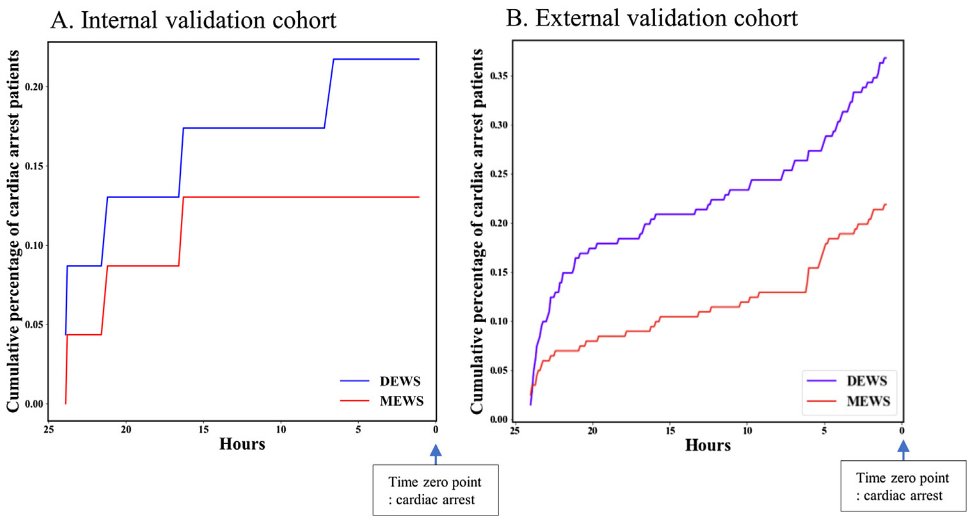


Fig. 3 – Comparison of the cumulative percentages of cardiac arrest patients. DEWS indicates the deep learning-based early warning score, and MEWS indicates the modified early warning score.

One of the critical strengths of DEWS is that it consists of a limited number of predictor variables. In this validation study, DEWS used only five basic vital signs: SBP, DBP, HR, RR and BT along with age and recorded time of vital signs. The two previous AI-based studies[38,40] 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 but essential vital signs that are almost always checked in admitted patients and with low missing rates. Therefore, DEWS can be applied worldwide without any difficulties in technical implementation.

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 more than 300 beds; one is a cardiovascular-specific hospital, and the other is a community general hospital. The hospitals in the external validation set have more than 900 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 on 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. We consider only the first CA for each patient admission, although second and third CAs are also important for the patient's prognosis. Nonetheless, the first CA has the highest priority because the care level the patient receives after CA will be maximal. Additionally, since this study was performed in a retrospective manner, a well-designed prospective clinical trial is necessary to apply DEWS in clinical practices as an alternative to other triggering score systems in RRS.

Conclusion

We compared DEWS and MEWS in multiple centres via extensive experiments. The results showed that DEWS not only predicts IHCA 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 h to 0.5 h before the event than MEWS. These findings demonstrate the potential of DEWS as an effective screening tool in RRSs that can be efficiently applied to identify high-risk patients.

Conflict of interest

All authors have reported that no potential conflicts of interest exist with any companies/organizations whose products or services may be discussed in this article.

Funding

None.

Authors' contribution

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.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.resuscitation.2021.04.013.

REFERENCES

1. Bellomo R. Rapid-response teams. . p. 139–46 (Table 2).
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:e592–9.
3. 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.
4. Lee BY, Hong S-B. Rapid response systems in Korea. Acute Crit Care 2019;34:108.
5. Duckitt RW, 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.
6. Paterson R, MacLeod DC, Thetford D, et al. Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. Clin Med (Northfield Il) 2006;6:281.
7. 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.
8. 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.
9. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM 2001;94:521–6.
10. Smith GB, Prytherch DR. Widely used track and trigger scores: are they ready for automation in practice? Resuscitation 2014;85:e157.
11. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI. Review and performance evaluation of aggregate weighted 'track and trigger'systems. Resuscitation 2008;77:170–9.
12. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI, Higgins B. A review, and performance evaluation, of single-parameter "track and trigger" systems. Resuscitation 2008;79:11–21.
13. Kwon J, 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:e008678.
14. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020;26:364–73.
15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80.
16. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
17. Kingma DP, Ba J, Adam:. A method for stochastic optimization. 2014 arXiv Prepr arXiv14126980.
18. 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.
19. Weng CG, Poon J. A new evaluation measure for imbalanced datasets. Proceedings of the 7th Australasian data mining conference – vol. 87 2008;27–32.
20. Leening MJG, Vedder MM, Witteman JCM, 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.
21. 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.
22. 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 – vol. 4. IntechOpen; 2019.
23. 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.
24. 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.
25. Boniatti MM, Azzolini N, Viana MV, et al. Delayed medical emergency team calls and associated outcomes. Crit Care Med 2014;42:26–30.
26. 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.
27. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on machine learning 2005;625–32.
28. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Informatics Assoc 2020;27:621–33.
29. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. . p. 4765–74.
30. Subbe C, Davies RG, 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.
31. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019;6:1–18.
32. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. Machine learning for healthcare conference 2017;361–76.
33. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: customized mortality prediction. J Pers Med 2012;2:138–48.
34. Gupta P, Malhotra P, Vig L, Shroff GM. Using features from pre-trained TimeNET for clinical predictions. KHD@ IJCAI. . p. 38–44.
35. 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.
36. 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.
37. 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 Heal Sci 2017;6:1–7.
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. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
40. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18.