Detecting Patient Deterioration Using Artificial Intelligence in a Rapid Response System

Kyung-Jae Cho, MS[1]; Oyeon Kwon, MS[1]; Joon-myoung Kwon, MD, MS[2]; Yeha Lee, PhD[1]; Hyunho Park, MD[1]; Ki-Hyun Jeon, MD, MS[3]; Kyung-Hee Kim, MD, PhD[3]; Jinsik Park, MD, PhD[3]; Byung-Hee Oh, MD PhD[3]

Objectives: As the performance of a conventional track and trigger system in a rapid response system has been unsatisfactory, we developed and implemented an artificial intelligence for predicting in-hospital cardiac arrest, denoted the deep learning-based early warning system. The purpose of this study was to compare the performance of an artificial intelligence-based early warning system with that of conventional methods in a real hospital situation.

Design: Retrospective cohort study.

Setting: This study was conducted at a hospital in which deep learning-based early warning system was implemented.

Patients: We reviewed the records of adult patients who were admitted to the general ward of our hospital from April 2018 to March 2019.

Interventions: The study population included 8,039 adult patients. A total 83 events of deterioration occurred during the study period. The outcome was events of deterioration, defined as cardiac arrest and unexpected ICU admission. We defined a true alarm as an alarm occurring within 0.5–24 hours before a deteriorating event.

Measurements and Main Results: We used the area under the receiver operating characteristic curve, area under the precision-recall curve, number needed to examine, and mean alarm count per day as comparative measures. The deep learning-based early warning system (area under the receiver operating characteristic curve, 0.865; area under the precision-recall curve, 0.066) outperformed the modified early warning score (area under the receiver operating characteristic curve, 0.682; area under the precision-recall curve, 0.010) and reduced the number needed to examine and mean alarm count per day by 69.2% and 59.6%, respectively. At the same specificity, deep learning-based early warning system had up to 257% higher sensitivity than conventional methods.

Conclusions: The developed artificial intelligence based on deep-learning, deep learning-based early warning system, accurately predicted deterioration of patients in a general ward and outperformed conventional methods. This study showed the potential and effectiveness of artificial intelligence in an rapid response system, which can be applied together with electronic health records. This will be a useful method to identify patients with deterioration and help with precise decision-making in daily practice. (*Crit Care Med* 2020; 48:e285–e289)

Key Words: artificial intelligence; cardiology; critical care; deep learning

In-hospital cardiac arrest is a major healthcare burden and rapid response systems (RRSs) are used worldwide to identify deteriorating hospitalized patients and to prevent cardiac arrest (1). Most patients with cardiac arrest show signs of deterioration. However, 209,000 cardiac arrests occur and the survival to discharge rate was only less than 20% in the United States each year (2). One challenge with RRS is the failure to detect the deteriorating signs of patient; thus, several track and trigger systems (TTSs) have been developed (3). However, conventional methods, such as the single parameter TTS (SPTTS) and modified early warning score (MEWS), have been disappointing owing to their limited ability to work together with electronic health records (EHRs) (4).

We previously developed and validated an artificial intelligence (AI) for predicting in-hospital cardiac arrest, denoted the deep learning-based early warning system (DEWS) (5). After fine-tuning and setup, we implemented DEWS with EHR to monitor the risk of deterioration among patients in general wards; we have actively used DEWS in our RRS since April 2018. The purpose of this study was to compare the performance of our developed AI with that of conventional methods. To our best knowledge, this is the first study to apply a deep learning-based AI algorithm in an RRS, verified in an external validation study in an actual hospital setting.

[1]VUNO, Seoul, Korea.

[2]Department of Critical care and Emergency Medicine, Mediplex Sejong Hospital, Incheon, Korea.

[3]Division of Cardiology, Cardiovascular Center, Mediplex Sejong Hospital, Incheon, Korea.

DOI: 10.1097/CCM.0000000000004236

MATERIALS AND METHODS

In this retrospective cohort study, we reviewed records of adult patients (age ≥ 18 yr) admitted to the general ward of our hospital from April 2018 to March 2019. We excluded patients who died with no attempt at resuscitation. The institutional review board of Mediplex Sejong Hospital (2018-055) approved the study and waived the requirement for informed consent based on minimal patient harm.

The outcome was events of deterioration, defined as cardiac arrest (lack of a palpable pulse with attempted resuscitation) and unexpected admission to the ICU. The hospital RRS simultaneously implements DEWS and conventional methods. We defined a true alarm as an alarm occurring within 0.5–24 hours before a deteriorating event.

The DEWS (score between 0 and 100) is composed of four vital signs: systolic blood pressure (SBP), heart rate (HR), respiratory rate (RR), and body temperature (BT); it is calculated every time the vital sign values are inputted into the EHR. When patients with DEWS over the cutoff (60) are updated to the RRS dashboard, an alarm is sent to the rapid response team. Once the patient's name appears, it remains on the dashboard and no additional alarm is generated based on the same criteria for 4 hours. Therefore, the number of true alarms may be more than the number of true events.

The conventional methods were MEWS and SPTTS. The MEWS (0–14) comprises SBP, HR, RR, BT, and mental status and is calculated automatically. Patients with MEWS over 5 are updated to the dashboard (6). The SPTTS criteria comprises vital signs and laboratory tests. If a patient meets any of the SPTTS criteria, the patient's name is updated to the dashboard. Even if a patient does not meet the DEWS, MEWS, and SPTTS criteria, our RRS monitors patient's 24 hours after transfer from the ICU to a general ward, major operation, and angiointervention.

The DEWS architecture includes three bidirectional recurrent neural network (RNN) layers with a long short-term memory unit, four fully connected layers with a rectified linear unit, and the Softmax layer at the end to output a score between 0 and 1; the risk score is obtained by multiplying the Softmax output score by 100. Before passing the RNN output to the fully connected layer, we use the output of the last step only. An Adam optimizer is used to train the DEWS with default parameters and binary cross-entropy as a loss function.

The unit of performance testing was not the number of patients but rather the number of alarms; repeated alarms could exhaust RSS resources, and the alarm count is important to implementing and maintaining the RRS. We used the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity, specificity, positive predictive value (PPV), negative predictive value, net reclassification index, number needed to examine (NNE), and mean alarm count per day (MACD) as comparative measures. All conventional methods were aggregated as the criteria of MEWS and SPTTS using the "OR" condition.

As a secondary experiment, we hypothesized that physicians could select same vital signs as the most contributing factor to a DEWS alarm. We randomly selected 600 alarms based on DEWS and compiled the data of each vital sign from 3 days ago to the time of each alarm. Four board-certified practicing intensive care physicians with over 10 years of critical care experience were asked to select the greatest contributing factor for triggering an alarm from the four vital signs used in the DEWS.

RESULTS

After excluding 17 patients who died with no attempt at resuscitation, the study population included 8,039 adult patients. A total 83 events of deterioration (11 cardiac arrest, 72 unexpected ICU admission) occurred during the study period. Baseline characteristics of study population are shown in **Table 1**.

The ROC curve of DEWS exceeded those of the other TTS methods (**Fig. 1*A***). DEWS (AUROC: 0.865 and AUPRC: 0.066) outperformed MEWS in detecting a deteriorating event. Using each cutoff point, DEWS outperformed MEWS in most metrics. Furthermore, sensitivity and 1-specificity points of the method combining DEWS and conventional methods were under the DEWS ROC curve. DEWS alone appeared to perform better than using DEWS with other conventional methods. Additionally, the cumulative percentage of event time points showed that DEWS detected 122% and 68.7% more deteriorating events than MEWS and all conventional methods, respectively, 15 hours before the event (**Fig. 1*B***).

Compared with MEWS (≥ 5), DEWS could reduce the NNE by 69.2% (**Fig. 1*C***). With DEWS, physician must examine 20.4 fewer patients than with MEWS to detect one deteriorating patient. Although the PPV of DEWS was relatively low, the algorithm was not used for diagnosis but rather for screening deteriorating patients, and the RRS could exclude false alarms. And patients who were not admitted to the ICU after treatment in a general ward were counted as false positives. As shown in **Figure 1*D***, we confirmed the sensitivity of each TTS when the number of alarms was acceptable for RRS implementation. When the MACD was 15.8, the sensitivity of the MEWS was only 22.9%; however, the sensitivity of DEWS was acceptable at 43.5%. DEWS detects a greater number of deteriorating patients than existing methods at the same time point; thus, more patients could be detected early with DEWS (Fig. 1*B*).

In the secondary experiment, interobserver agreement among physicians in selecting the greatest contributing factor in triggering an alarm was 90.7% (kappa = 0.705; 95% CI, 0.672–0.738). The proportion of physician agreement was 38.1%, 26.7%, 13.4%, and 20.8% for SBP, HR, RR, and BT, respectively.

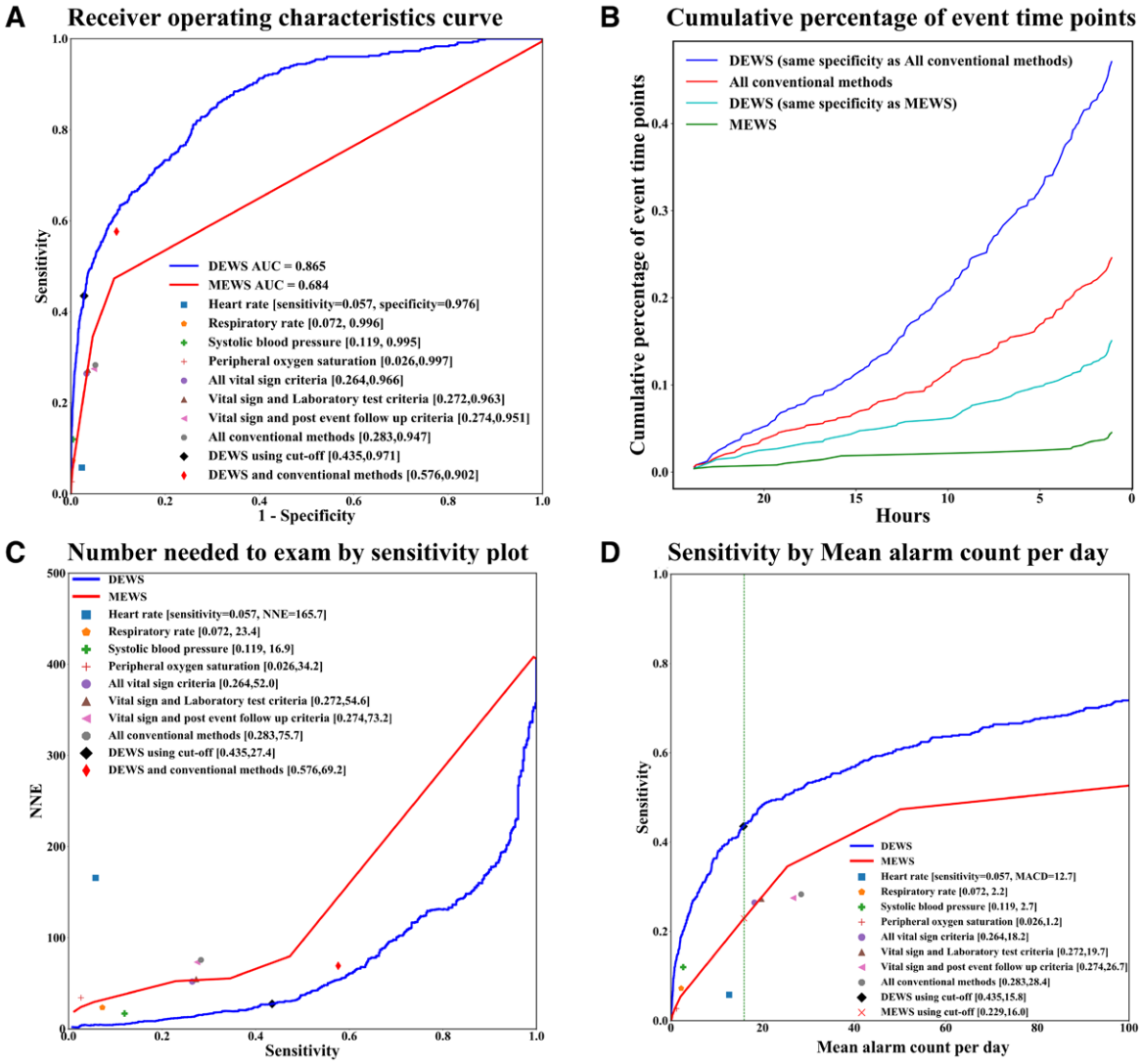
DISCUSSION

In this study, the DEWS outperformed the other methods investigated. The single-parameter TTS was too simple to reflect intervariable relationships. Several aggregate-weighted TTSs based on logistic regression and several predictive algorithms based on machine learning have previously been developed (7, 8). However, logistic regression and machine

TABLE 1. Baseline Characteristics

Characteristics	Patients With No Events	Patients With a Deteriorating Event[a]
Baseline information		
Number of patients	7,956	83
Number of vital sign data set	444,213	33,592
Male, *n* (%)	3,759 (47.3)	45 (54.2)
Age, yr, mean ± SD	62.4 ± 16.8	73.1 ± 13.9
Length of stay, median (interquartile range)	3 (1–7)	21 (12–35)
Initial vital signs and laboratory results, mean ± SD[b]		
Systolic blood pressure, mm Hg	128.1 ± 17.7	121.9 ± 26.7
Diastolic blood pressure, mm Hg	77.4 ± 12.3	72.2 ± 15.6
Heart rate, beats/min	77.0 ± 14.5	83.5 ± 18.7
Respiratory rate, breaths/min	17.9 ± 1.8	18.9 ± 3.8
Temperature	36.8 ± 0.4	36.9 ± 0.5
Peripheral oxygen saturation, %	98.0 ± 1.6	97.0 ± 3.2
Venous serum P_{CO_2}, mm Hg	41.2 ± 9.2	34.7 ± 12.1
Potassium, mmol/L	4.1 ± 0.5	4.2 ± 0.6
Lactate, mmol/L	1.9 ± 1.6	2.4 ± 2.5
Last vital signs and laboratory results, mean ± SD[c]		
Systolic blood pressure, mm Hg	120.0 ± 15.2	116.1 ± 19.6
Diastolic blood pressure, mm Hg	73.4 ± 11.2	68.2 ± 13.3
Heart rate, beats/min	68.3 ± 11.7	75.7 ± 17.0
Respiratory rate, breaths/min	17.4 ± 1.2	17.9 ± 3.3
Temperature	36.7 ± 0.3	36.7 ± 0.4
Peripheral oxygen saturation, %	97.4 ± 2.3	97.0 ± 3.2
Venous serum P_{CO_2}, mm Hg	40.9 ± 8.7	37.5 ± 14.1
Potassium, mmol/L	4.1 ± 0.5	4.3 ± 0.5
Lactate, mmol/L	1.8 ± 1.8	2.8 ± 2.9
Total vital signs and laboratory results, mean ± SD		
Systolic blood pressure, mm Hg	123.4 ± 17.5	115.4 ± 19.8
Diastolic blood pressure, mm Hg	73.8 ± 11.9	67.5 ± 13.4
Heart rate, beats/min	74.9 ± 14.8	85.9 ± 17.8
Respiratory rate, breaths/min	17.7 ± 2.4	19.5 ± 4.8
Temperature	36.9 ± 0.4	36.9 ± 0.5
Peripheral oxygen saturation, %	97.9 ± 2.3	98.3 ± 2.6
Venous serum P_{CO_2}, mm Hg	41.9 ± 10.8	40.4 ± 11.1
Potassium, mmol/L	4.0 ± 0.9	4.0 ± 0.7
Lactate, mmol/L	2.2 ± 2.4	2.1 ± 2.7

[a]Patients with a deteriorating event indicates those patients who experience cardiac arrest or are admitted to the ICU.
[b]Initial vital signs indicate vital signs that were checked at the time of admission.
[c]The last results mean the results before discharge in the nonevent group and the last results before the deteriorating event in the event group.



E Comparison of DEWS and other conventional methods with the same specificity

Track and Trigger system	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	NRI (95% CI)	True Alarm	False Alarm
MEWS ≥ 3	0.345 (0.309–0.380)	0.953 (0.953–0.954)	0.018 (0.016–0.020)	0.998 (0.998–0.998)		167	9095
DEWS ≥ 58.09	0.506 (0.470–0.547)	0.953 (0.953–0.954)	0.026 (0.023–0.029)	0.999 (0.999–0.999)	0.008 (0.005–0.012)	245	9090
MEWS ≥ 4	0.229 (0.198–0.260)	0.9708 (0.970–0.971)	0.019 (0.016–0.022)	0.998 (0.998–0.998)		111	5716
DEWS ≥ 61.8	0.442 (0.405–0.484)	0.9708 (0.970–0.971)	0.036 (0.032–0.040)	0.999 (0.998–0.999)	0.018 (0.012–0.024)	214	5714
MEWS ≥ 5	0.053 (0.039–0.070)	0.996 (0.996–0.996)	0.034 (0.024–0.045)	0.997 (0.997–0.998)		26	739
DEWS ≥ 80.1	0.188 (0.158–0.217)	0.996 (0.996–0.996)	0.110 (0.093–0.129)	0.998 (0.998–0.998)	0.085 (0.048–0.111)	91	733
All conventional methods	0.283 (0.252–0.321)	0.947 (0.947–0.949)	0.013 (0.011–0.015)	0.998 (0.998–0.998)		137	10215
DEWS ≥ 57.25	0.52 (0.483–0.561)	0.947 (0.947–0.949)	0.024 (0.022–0.027)	0.999 (0.999–0.999)	0.011 (0.006–0.014)	252	10201

Figure 1. Performances of track and trigger system for detecting deterioration events. The single parameter track and trigger system criteria consists of vital signs (systolic blood pressure < 85 mm Hg, heart rate < 50 or > 130, respiratory rate < 8 or > 25, peripheral oxygen saturation < 89%) and laboratory test criteria (pH < 7.3, Pao_2 < 55 mm Hg, $Paco_2$ > 50 mm Hg, lactic acid > 2.0 mmol/L, potassium > 6.5 mEq/dL). *Mean alarm count per day was calculated assuming that the number of hospital beds was 313 beds. AUC = area under the receiver operating characteristic curve, DEWS = deep-learning-based early warning system, MACD = mean alarm count per day, MEWS = modified early warning score, NNE = number needed to examination, NPV = negative predictive value, NRI = net reclassification index, PPV = positive predictive value.

learning are based on fixed assumptions about data behavior and the need to preselect variables, and careful engineering in the development phase may cause information loss and limited performance (9). However, deep learning includes feature learning, a method in which a model can be fed raw data, to automatically identify the features needed for conducting a task. Deep learning is effective for discovering the intricate structures in high-dimensional data without information loss. DEWS can also reflect temporal changes in vital signs and differences between individuals, as previously shown (10).

The hospital in which DEWS was developed differed from the present study hospital; therefore, this study serves as external validation of DEWS. As the AI algorithm was not derived from medical knowledge but rather from observable relationships between data (10), its performance in one hospital is not guaranteed in other hospitals without external validation.

Our study has limitations. First, deep learning known as a "black box." We could not know the exact reason of alarm. However, as only four variables were used in our DEWS, we could surmise the reason of alarm in the secondary experiment. In the future, we will focus on an interpretable deep learning model to tackle this limitation. Second, this study was a retrospective review. Clinical trials are needed to confirm the exact impact of DEWS on an RRS and the team's response, excluding bias owing to patients, physicians, and systems. We plan to conduct a randomized controlled trial comparing DEWS with conventional methods in multiple hospitals and countries in the future.

CONCLUSIONS

We showed that biological signals, such as vital signs, can be effectively analyzed by AI based on deep learning to predict in-hospital cardiac arrest and unexpected ICU admission more accurately than conventional methods. We demonstrated the potential effectiveness of AI in an RRS, which can be applied with EHR as a useful method to identify deterioration in patients and help with precise decision-making in practice.

Drs. Cho, O. Kwon, and J.-m. Kwon contributed equally to this study.

Drs. O. Kwon and H. Park 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: kwonjm@sejongh.co.kr

REFERENCES

1. Salvatierra G, Bindler RC, Corbett C, et al: Rapid response team implementation and in-hospital mortality*. *Crit Care Med* 2014; 42:2001–2006
2. Merchant RM, Yang L, Becker LB, et al; American Heart Association Get With The Guidelines-Resuscitation Investigators: Incidence of treated cardiac arrest in hospitalized patients in the United States. *Crit Care Med* 2011; 39:2401–2406
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. Kwon J-M, 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
6. Subbe CP, Kruger M, Rutherford P, et al: Validation of a modified early warning score in medical admissions. *QJM* 2001; 94:521–526
7. Kang MA, Churpek MM, Zadravecz FJ, et al: Real-time risk prediction on the wards: A feasibility study. *Crit Care Med* 2016; 44:1468–1473
8. 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
9. Breiman L: Statistical modeling: The two cultures. *Stat Sci* 2001; 16:199–231
10. LeCun Y, Bengio Y, Hinton G: Deep learning. *Nature* 2015; 521:436–444