Multicenter validation of a deep-learning–based pediatric early-warning system for prediction of deterioration events

Yunseob Shin¹, Kyung-Jae Cho¹, Yeha Lee¹, Yu Hyeon Choi², Jae Hwa Jung³, Soo Yeon Kim³, Yeo Hyang Kim⁴, Young A Kim⁵, Joongbum Cho⁶, Seong Jong Park⁷, Won Kyoung Jhang⁷

¹VUNO Inc., Seoul; ²Department of Pediatrics, Seoul National University Children’s Hospital, Seoul; ³Department of Pediatrics, Severance Children’s Hospital, Yonsei University College of Medicine, Seoul; ⁴Department of Pediatrics, Kyungpook National University Children’s Hospital, School of Medicine, Kyungpook National University, Daegu; ⁵Department of Pediatrics, Pusan National University Children’s Hospital, Yangsan; ⁶Department of Critical Care Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul; ⁷Department of Pediatrics, Asan Medical Center Children’s Hospital, University of Ulsan College of Medicine, Seoul, Korea

Background: Early recognition of deterioration events is crucial to improve clinical outcomes. For this purpose, we developed a deep-learning–based pediatric early-warning system (pDEWS) and aimed to validate its clinical performance.

Methods: This is a retrospective multicenter cohort study including five tertiary-care academic children’s hospitals. All pediatric patients younger than 19 years admitted to the general ward from January 2019 to December 2019 were included. Using patient electronic medical records, we evaluated the clinical performance of the pDEWS for identifying deterioration events defined as in-hospital cardiac arrest (IHCA) and unexpected ward-to-pediatric intensive care unit transfer (UIT) within 24 hours before event occurrence. We also compared pDEWS performance to those of the modified pediatric early-warning score (PEWS) and prediction models using logistic regression (LR) and random forest (RF).

Results: The study population consisted of 28,758 patients with 34 cases of IHCA and 291 cases of UIT. pDEWS showed better performance for predicting deterioration events with a larger area under the receiver operating characteristic curve, fewer false alarms, a lower mean alarm count per day, and a smaller number of cases needed to examine than the modified PEWS, LR, or RF models regardless of site, event occurrence time, age group, or sex.

Conclusions: The pDEWS outperformed modified PEWS, LR, and RF models for early and accurate prediction of deterioration events regardless of clinical situation. This study demonstrated the potential of pDEWS as an efficient screening tool for efferrnt operation of rapid response teams.

Key Words: cardiac arrest; critical care; deep learning; early warning score; pediatrics

INTRODUCTION

Many healthcare centers worldwide continue to develop and introduce various early-warning scoring systems to identify patients in critical condition in advance of onset to perform...
prompt intervention to improve patient safety and clinical outcome [1-5]. For afferent limbs, such systems widely range from simple and easy bedside calculations to more sophisticated complex scoring systems that include laboratory test results, a combination of patient clinical characteristics and medical histories, and therapeutic interventions [6-9]. Furthermore, due to recent revolutionary progress in artificial intelligence (AI) and machine learning, these algorithms can be implemented for more precise and earlier prediction of critical events [10-13]. However, most research has focused on adult populations and has rarely been externally validated or widely implemented in real clinical practice.

Previously, we developed a deep-learning-based early-warning system (DEWS) for predicting in-hospital cardiac arrest (IHCA) in an adult population [14] and demonstrated its excellent clinical performance. After fine-tuning and setting up the DEWS, we implemented electronic medical records (EMR) to monitor the risk of deterioration among adult patients in general wards, presenting better performance that conventional methods. The DEWS was successfully implemented in rapid response systems (RRTs) [15] and was validated by a multicenter study including adult patients [16]. Subsequently, further upgrades for learning and additional training using pediatric data led to development of a deep-learning-based pediatric early-warning system (pDEWS) that can predict pediatric IHCA and unexpected general ward-to-pediatric intensive care unit (PICU) transfer (UIT), which were validated in a single-center study [17]. In this study, we aimed to validate the clinical performance of pDEWS externally for predicting deterioration events in a larger multicenter cohort and compared it to several conventional predicting models.

MATERIALS AND METHODS

Study Design
This was a retrospective multicenter observational cohort study of five tertiary-care academic children’s hospitals. The requirement for informed consent was waived due to the retrospective nature of the study. External validation of the clinical performance of previously developed pDEWS for identifying deterioration events defined as either UIT or IHCA within 24 hours before event occurrence was performed [17]. This study was approved by the Institutional Review Board of each participating hospital (Seoul National University Children’s Hospital: 2003-229-1115; Severance Hospital: 4-2019-1304; Kyungpook National University Children’s Hospital: 2020-02-002; Pusan National University Yangsan Hospital: 05-2020-005; Samsung Medical Center: 2020-03-148-0020, respectively).

Deep-Learning-Based Pediatric Early-Warning System
The pDEWS architecture includes an embedding layer, three bi-directional long short-term memory (LSTM) layers for modeling the sequential characteristics of EMR data as an encoder, and three fully connected (FC) layers as a classifier. Before the LSTM encoder, we embedded the input data, consisting of respiratory rate (RR), heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), body temperature (BT), age, and a time feature through the FC embedding layer. To reflect the vital sign trend for each patient, 20 consecutive series of vital signs were used as inputs to the LSTM layer [18]. We used the last time step LSTM output to pass to the FC layer. Batch normalization and dropout were used on each FC layer in the classifier to regularize and stabilize the pDEWS model [19]. By adding a softmax layer at the end, the pDEWS model output a score between 0 and 1. We optimized the parameters of the pDEWS model by minimizing the cross-entropy loss function with the Adam optimizer [20]. The hyperparameters were tuned with the best performance from 10% of the derivation data. To resolve the class imbalance problem, we over-sampled the event data allowing duplication during the training process. We trained our model for 1000 epochs and selected the model with the highest area under the receiver operating characteristic curve (AUROC) score in the validation data. We also applied a transfer learning technique [21] to encourage our pDEWS model to obtain additional knowledge from other data by initializing the pDEWS model’s connection weights from the DEWS model, which was developed to predict adult patient IHCA.

Materials
Pediatric patients (<19 years old) admitted to the general wards of five university affiliate tertiary care medical centers in the Republic of Korea over a 12-month period between January 2019 and December 2019 were included. From EMRs and

KEY MESSAGES
- We developed a deep-learning-based pediatric early-warning system (pDEWS) and performed a multicenter validation.
- pDEWS showed excellent performance in predicting clinical deterioration events regardless of clinical setting.
the medical database, we collected patient data including age, sex, event occurrence, exact time and location of event occurrences, and length of hospital stay and extracted five basic vital signs—RR, HR, SBP, DBP, and BT—during hospitalization for pDEWS and other early-warning system calculation.

We excluded patients with data recorded <30 minutes after admission, no vital signs at 24 hours prior to the deterioration event, incorrect demographics, and do-not-resuscitate orders. Patient information was anonymized and de-identified prior to analysis. Outlier values outside the normal range of each vital sign or non-numeric values were excluded from the initially collected data and treated as missing values (Supplementary Table 1). Missing values were replaced with the most recent previous values. Based on these data, we also calculated the modified pediatric early-warning score (PEWS) to include only vital sign parameters (HR, RR, SBP, oxygen saturation, and temperature).

Outcome Measures
The primary outcome of interest was deterioration event, defined as a composite of IHCA and UIT. UIT was defined as “PICU admission due to acutely deteriorating clinical conditions,” excluding routine scheduled post-surgical treatment or PICU admission for scheduled procedures. Secondary outcomes were numbers of each type of deterioration event, IHCA and UIT. We also performed subgroup analyses by hospital, age groups, event occurrence time, and sex.

Statistical Analysis
For validation, we performed extensive statistical analysis using scikit-learn (Scikit-learn 0.23.1; community-driven project sponsored by BCG GAMMA) and pandas (Pandas 1.0.5; community-driven project sponsored by NumFOCUS). We evaluated deterioration prediction performance by calculating the AUROC and the area under the precision-recall curve (AUPRC) \([22,23]\). AUROC is one of the most generally used metrics and shows the area of sensitivity versus the false-positive rate. Compared with AUROC, AUPRC describes class imbalance data by measuring the area under the plot of precision versus sensitivity. Additionally, we calculated F-1 score \([2x(\text{precision}\times\text{recall})/(\text{precision+recall})]\), the net reclassification index (NRI), positive predictive value \([\text{PPV}=\text{true positive}/(\text{true positive+false positive})]\), negative predictive value \([\text{NPV}=\text{true negative}/(\text{true negative+false negative})]\), mean alarm count per day (MACPD) per 1,000 beds, and patient number needed to examine (NNE) \([23,24]\). The NRI is used to compare improvement in prediction performance. To compare the clinical prediction of deterioration events within 24 hours prior to occurrence, we trained random forest (RF) models with various hyperparameter sets and logistic regression (LR) models with an L2 regularization penalty and 1e-4 tolerance for stopping criteria to obtain comparable performance to that of our pDEWS model. Then, we evaluated the clinical performance of pDEWS by comparing to modified PEWS, RF, and LR. In addition to predictive performance, we evaluated the alarm rate with comparison to MACPD at a set sensitivity level and the cumulative prediction percentage of deterioration events at the same time point within 24 hours of the event.

Additionally, we calibrated pDEWS to reflect the real probability of deterioration events because a predictive model should infer proper output probabilities without being extreme. We visualized pDEWS model calibration performance by comparing it with the ideal calibration line. We also performed feature importance analysis to interpret which characteristics of vital signs influence pDEWS model decision and calculated the importance of each feature and time-step by applying Shapley Additive Explanations (SHAP) values \([25]\).

RESULTS
Study Population
Among the 29,035 patients admitted to five hospitals over a 12-month duration, 277 were excluded (Figure 1). Among the remaining 28,758 patients, 16,167 were male (56.2%). The median hospital stay was 3.35 days (2.25–6.13 days) (Table 1). A total of 996,874 vital sign sets was evaluated to validate the pDEWS. There were 34 cases of IHCA, 291 cases of UIT, and 325 related vital sign sets.

![Figure 1. A flowchart for patient inclusion and exclusion.](https://www.accjournal.org)
Table 1. Baseline characteristics of the study population

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Admissions</td>
<td>28,758</td>
</tr>
<tr>
<td>Vital sign data set</td>
<td>996,874</td>
</tr>
<tr>
<td>Admission with unexpected PICU transfer</td>
<td>291</td>
</tr>
<tr>
<td>Vital sign data set</td>
<td>6,050</td>
</tr>
<tr>
<td>Admission with in-hospital cardiac arrest</td>
<td>34</td>
</tr>
<tr>
<td>Vital sign data set</td>
<td>371</td>
</tr>
<tr>
<td>Male</td>
<td>16,167 (56.2)</td>
</tr>
<tr>
<td>Age (yr)</td>
<td>6.28±5.24</td>
</tr>
<tr>
<td>Length of stay (day)</td>
<td>3.35 (2.25–6.13)</td>
</tr>
</tbody>
</table>

**Initial vital sign**

- Systolic blood pressure (mm Hg): 104.51±5.24
- Diastolic blood pressure (mm Hg): 62.18±10.38
- Heart rate (/min): 109.66±25.66
- Respiratory rate (/min): 24.97±6.87
- Body temperature (°C): 36.81±0.60
- SpO2: 98.42±2.20

**Vital sign within 24 hours before outcome**

- Systolic blood pressure (mm Hg): 103.51±5.24
- Diastolic blood pressure (mm Hg): 61.75±13.38
- Heart rate (/min): 129.00±30.74
- Respiratory rate (/min): 32.21±14.42
- Body temperature (°C): 37.12±0.75
- SpO2: 98.36±2.20

**Total vital sign**

- Systolic blood pressure (mm Hg): 105.00±5.24
- Diastolic blood pressure (mm Hg): 62.54±11.26
- Heart rate (/min): 109.05±25.39
- Respiratory rate (/min): 25.41±7.33
- Body temperature (°C): 36.82±0.59
- SpO2: 98.28±2.72

Values are presented as number (%), mean±standard deviation, or median (interquartile range).

Primary Outcome

The pDEWS yielded an AUROC of 0.892 (95% CI, 0.888–0.895) for predicting deterioration events, which was larger than those of modified PEWS, LR, and RF models (Figure 2). The pDEWS AUPRC for predicting critical events was 0.093 (95% CI, 0.089–0.098), which was larger than the modified PEWS (0.029; 95% CI, 0.028–0.031), LR (0.045; 95% CI, 0.042–0.049), and RF (0.042; 95% CI, 0.040–0.044) models. We evaluated sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), PPV, NPV, F-score, NNE, and MACPD for each cutoff value for predicting critical events (Table 2). Given that the cutoff value of the pDEWS was 90, it showed an acceptable F-1 score, corresponding to the most acceptable PPV and NPV for clinical integration, MACPD, and NNE.

In a paired comparison to the modified PEWS, the LR, and RF models at the same specificity, pDEWS showed superior performance with the highest sensitivity, PLR, PPV, and F-1 score and the lowest NLR and NNE (Supplementary Table 2). The pDEWS provided much lower MACPD and NNE for these deterioration events under the same sensitivity than did the modified PEWS, LR, and RF models (Figure 3). It markedly reduced false alarms in detecting these deterioration events by 56%, 37%, and 66%, respectively, at the cutoff value of the modified PEWS≥5 compared with modified PEWS, LR, and RF models (Figure 3). The cumulative percentage of deteriorating patients for these deterioration events was larger in the pDEWS than the modified PEWS or the LR or RF prediction models at the same cutoff level (Figure 4). The pDEWS showed markedly larger values than other methods for the true alarm count at 12 hours before the occurrence and for the total period.

Secondary Outcome

For prediction of each type of deterioration event by pDEWS, the AUROC for IHCA was 0.865 (95% CI, 0.847–0.883) and for UIT was 0.897 (95% CI, 0.893–0.901) (Figure 5). For AUPRC,
Table 2. Performance of the pDEWS for prediction of deterioration events at different cutoff levels

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PLR</th>
<th>NLR</th>
<th>PPV</th>
<th>NPV</th>
<th>F-1 score</th>
<th>MACPD</th>
<th>NNE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.980</td>
<td>0.293</td>
<td>1.386</td>
<td>0.068</td>
<td>1.000</td>
<td>0.018</td>
<td>1.936</td>
<td>112.310</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.951</td>
<td>0.498</td>
<td>1.893</td>
<td>0.099</td>
<td>0.999</td>
<td>0.024</td>
<td>1.379</td>
<td>82.483</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.917</td>
<td>0.638</td>
<td>2.536</td>
<td>0.130</td>
<td>0.999</td>
<td>0.032</td>
<td>0.997</td>
<td>61.834</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.879</td>
<td>0.730</td>
<td>3.260</td>
<td>0.165</td>
<td>0.999</td>
<td>0.040</td>
<td>0.999</td>
<td>74.7</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.839</td>
<td>0.794</td>
<td>4.066</td>
<td>0.203</td>
<td>0.999</td>
<td>0.050</td>
<td>0.574</td>
<td>38.940</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.795</td>
<td>0.838</td>
<td>4.919</td>
<td>0.245</td>
<td>0.998</td>
<td>0.059</td>
<td>0.452</td>
<td>32.361</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.751</td>
<td>0.871</td>
<td>5.827</td>
<td>0.286</td>
<td>0.998</td>
<td>0.069</td>
<td>0.362</td>
<td>27.473</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.712</td>
<td>0.895</td>
<td>6.804</td>
<td>0.322</td>
<td>0.998</td>
<td>0.080</td>
<td>0.296</td>
<td>23.672</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.673</td>
<td>0.914</td>
<td>7.859</td>
<td>0.357</td>
<td>0.997</td>
<td>0.090</td>
<td>0.244</td>
<td>20.627</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.639</td>
<td>0.930</td>
<td>9.115</td>
<td>0.388</td>
<td>0.997</td>
<td>0.103</td>
<td>0.201</td>
<td>17.922</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.598</td>
<td>0.943</td>
<td>10.534</td>
<td>0.426</td>
<td>0.997</td>
<td>0.116</td>
<td>0.164</td>
<td>15.643</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.558</td>
<td>0.954</td>
<td>12.224</td>
<td>0.463</td>
<td>0.997</td>
<td>0.130</td>
<td>0.133</td>
<td>13.619</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>0.510</td>
<td>0.964</td>
<td>14.205</td>
<td>0.508</td>
<td>0.997</td>
<td>0.145</td>
<td>0.106</td>
<td>11.859</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>0.452</td>
<td>0.973</td>
<td>16.544</td>
<td>0.563</td>
<td>0.996</td>
<td>0.160</td>
<td>0.082</td>
<td>10.324</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>0.387</td>
<td>0.981</td>
<td>20.120</td>
<td>0.625</td>
<td>0.996</td>
<td>0.178</td>
<td>0.058</td>
<td>8.667</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>0.306</td>
<td>0.988</td>
<td>25.382</td>
<td>0.702</td>
<td>0.995</td>
<td>0.193</td>
<td>0.038</td>
<td>7.077</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>0.196</td>
<td>0.994</td>
<td>30.909</td>
<td>0.809</td>
<td>0.995</td>
<td>0.180</td>
<td>0.020</td>
<td>5.990</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.105</td>
<td>0.997</td>
<td>40.856</td>
<td>0.897</td>
<td>0.994</td>
<td>0.140</td>
<td>0.010</td>
<td>4.775</td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>0.031</td>
<td>1.000</td>
<td>66.013</td>
<td>0.969</td>
<td>0.994</td>
<td>0.056</td>
<td>0.001</td>
<td>3.337</td>
<td></td>
</tr>
</tbody>
</table>

pDEWS: deep-machine-learning-based pediatric early warning system; PLR: positive likelihood ratio; NLR: negative likelihood ratio; PPV: positive predictive value; NPV: negative predictive value; MACPD: mean alarm count per day; NNE: number needed to examine.

Figure 3. Comparison of (A) mean alarm count per day (MACPD) at the same sensitivity and (B) sensitivity at the same number needed to examine (NNE) for deterioration events. pDEWS: deep-learning-based pediatric early-warning system; PEWS: pediatric early-warning score.
Figure 4. Cumulative percentages of deteriorating patients. The cutoffs of the models for each figure were set at threshold points with the same specificity as (A) pediatric early-warning score (PEWS) ≥4, (B) PEWS ≥5, and (C) PEWS ≥6. pDEWS: deep-learning-based pediatric early-warning system.

Figure 5. Areas under the receiver operating characteristic curves (AUROC) for the prediction of (A) in-hospital cardiac arrest (IHCA) and (B) unexpected ward-to-pediatric intensive care unit transfer (UIT). CI: confidence interval; pDEWS: deep-learning-based pediatric early-warning system; PEWS: pediatric early-warning score.

IHCA was 0.006 (95% CI, 0.005–0.008), and UIT was 0.100 (95% CI, 0.096–0.106).

Subgroup Analysis
The five participating hospitals had different characteristics in that hospital B had a higher proportion of UIT (1.8%), and hospital A had a higher proportion of IHCA than the other hospitals. In a comparison of AUROCs for predicting deterioration events by individual hospital, the pDEWS yielded a larger AUROC than other prediction models (Figure 6). For comparing patients by age (<3 months, 3 months to <1 year, 1 to <4 years, 4 to <12 years, and 12 to <19 years), the pDEWS AUROC for predicting deterioration events increased with increasing age group (Figure 7A). It also outperformed other models regardless of age group. pDEWS AUROCs were similar between day and night, weekday and weekend, and male and female com-
Figure 6. Areas under the receiver operating characteristic curves (AUROC) for prediction of deterioration events by hospital: (A) hospital A, (B) hospital B, (C) hospital C, (D) hospital D, and (E) hospital E. CI: confidence interval; pDEWS: deep-learning-based pediatric early-warning system; PEWS: pediatric early-warning score.

DISCUSSION

In this study, the pDEWS showed excellent performance for predicting deterioration events in a multicenter validation cohort. It showed earlier identification of deterioration events with fewer false alarms, MACPD, and NNE than other early-warning prediction models including pDEWS, RF, and LR models. Previously, the pDEWS was validated in a single-center retrospective study [17]. However, it was performed in a different cohort from the same hospital from which patient data were collected for its development. Even though it was composed of different patients split into development and validation cohorts by admission period, its performance for different groups could not be guaranteed and, thus, had limit-
ed generalizability. Consistent with this, as many PEWS have been developed and introduced, there have been problems associated with considerable variation in the performance in different settings [5,6,26-28]. Therefore, external validation is required to widely implement this method in clinical practice.

Similar to previous single-center validation studies in pediatric populations, pDEWS also showed good performance in the multicenter cohort. Although the five included hospitals had different settings and characteristics, the pDEWS demonstrated excellent and consistent clinical performance in each hospital, suggesting strong advantages of this method using deep learning methods based on only five basic vital signs. One of the significant obstacles for a multicenter study is disparity in EMR quality, which could vary widely across hospitals. However, vital-sign data are essential for all admitted patients and are usually systematically checked and recorded using the same measurement units regardless of institution, which enables the pDEWS model to be applied without specialized staged modifications across hospitals. Consistent with previous reports, this study showed that complex systems that include many parameters are not necessary for improving performance quality.

The pDEWS also showed consistent performance in several subgroup analyses categorized by sex, age, and event-occurrence time. pDEWS success could be due partially to the advantages of AI, which reduce human error. The performance of prediction deterioration events improved with increasing age group, though this might also be related to the characteristics of AI and deep learning methods, where few event numbers in the younger age group could pose a difficulty for deep learning and model training. Nevertheless, the pDEWS yielded the largest AUROC for predicting deterioration events among the

Figure 7. Areas under the receiver operating characteristic curves (AUROC) for prediction of deterioration events by subgroup analysis: (A) age group, (B) event occurring time, and (C) sex. CI: confidence interval; pDEWS: deep-learning-based pediatric early-warning system; PEWS: pediatric early-warning score (Continued to the next page).
Figure 7. Areas under the receiver operating characteristic curves (AUROC) for prediction of deterioration events by subgroup analysis: (A) age group, (B) event occurring time, and (C) sex. CI: confidence interval; pDEWS: deep-learning-based pediatric early-warning system; PEWS: pediatric early-warning score.
prediction models, highlighting its excellent performance.

The primary goal of the early-warning system (EWS) is to reduce critical events by timely recognition and intervention for deteriorating patients. As EWS are increasingly introduced and used in clinical practice, clinical outcomes have improved and in-hospital critical events have decreased significantly [3,29-31]. These results could be related to the timeliness of EWS recognition [32,33]. As compared with the cumulative prediction percentage of deterioration events at the same time point within 24 hours of the event, pDEWS yielded a larger area than those of other prediction models.

On the other hand, regarding RRT implementation, alarm count is a key point of interest for validating EWS. As previously reported, there is the challenge of increased alarm rates, which is related to not only accuracy and efficacy, but also practicality. A false alarm results in unnecessary activation of RRT, which could lead to RRT exhaustion with alarm fatigue and additional workload [34,35]. Consequently, excessive false alarms and alarm fatigue might result in inappropriate responses and desensitization and reduced or missing responses to clinically significant events, putting the patient at substantial risk of decreased safety and poor quality of care [36,37]. However, the pDEWS has an outstanding feature for controlling the alarm count with smaller MACPD and NNE at the same specificity as other prediction models. Thus, in real clinical practice, implementation of the pDEWS to an EMR system could be automatic by manipulating the input vital sign data. From this information, the pDEWS could predict and detect deterioration events early and accurately, generating alarms for RRT activation. As the pDEWS showed acceptable levels of MACPD.

**Figure 8.** Deep-learning-based pediatric early-warning system (pDEWS) model calibration analysis.

**Figure 9.** Deep-learning-based pediatric early-warning system (pDEWS) model feature importance analysis. (A) Sequence-wise feature importance and (B) average feature importance. RR: respiratory rate; HR: heart rate; DBP: diastolic blood pressure; SBP: systolic blood pressure; BT: body temperature.
and NNE, it could be helpful for more efficient RRT operations with few false alarms, which could reduce physician workload, enable prompt and effective intervention, and consequently decrease critical event occurrence and improve clinical outcomes. Furthermore, pDEWS has the advantage of being adjustable according to site characteristics by controlling the alarm threshold. Collectively speaking, pDEWS has a promising role for improving clinical practice.

In addition to other supporting outcome data for clinical performance of pDEWS, we evaluated and performed calibration of this prediction model, which showed acceptable findings. The previous deep-learning-based model has been criticized for being a “black box” in terms of decision making with nontransparent, unknown, and non-traceable algorithms [38,39]. However, in this study, we performed feature importance analysis, which showed the importance of each factor by time-step. This could partially explain the underlying process of pDEWS.

This study has several advantages. To our knowledge, this is the first evaluation of the clinical performance of an EWS for predicting deterioration events composed of IHCA and UIT in a pediatric multicenter validation cohort. It demonstrated excellent performance using various statistical approaches. This study also included subgroup analysis to consider various clinical situations, which could be helpful for application in real clinical practice. We performed model calibration and feature importance analysis, which were rarely performed in previous studies, and which demonstrated the high quality of this prediction model and partially explained the deep-learning-based algorithm.

Additionally, this study has several limitations. Because the primary outcome was deterioration events composed of IHCA and UIT, all included institutions were tertiary academic children’s hospitals because of the need for a PICU. Therefore, this study could have a selection bias and its generalizability is limited. Because it is a multicenter study, EMR quality, data collection, and the related missing rate could be different across all hospitals, which could affect the results. There was a smaller number of events in the younger age group, which could affect pDEWS performance.

The pDEWS showed excellent clinical performance for predicting deterioration events, including IHCA and UIT, compared with modified PEWS and other prediction models, like RF or LR. The pDEWS offered earlier prediction with fewer false alarms and higher accuracy, which could be promising if implemented in real clinical practice. It may provide more precise and timely identification of deterioration events, which could be helpful for more efficient operation of RRT with decreased workload and improved clinical outcomes.

**CONFLICT OF INTEREST**

No potential conflict of interest relevant to this article was reported.

**FUNDING**

None.

**ACKNOWLEDGMENTS**

None.

**ORCID**

Yunseob Shin https://orcid.org/0000-0002-1955-1908
Kyung-Jae Cho https://orcid.org/0000-0003-3564-3287
Yeha Lee https://orcid.org/0000-0002-6248-7729
Yu Hyeon Choi https://orcid.org/0000-0002-3057-0886
Jae Hwa Jung https://orcid.org/0000-0001-7443-9073
Soo Yeon Kim https://orcid.org/0000-0003-4965-6193
Yeo Hyang Kim https://orcid.org/0000-0002-1631-574X
Young A Kim https://orcid.org/0000-0002-8332-5200
Joongbum Cho https://orcid.org/0000-0001-5931-7553
Seong Jong Park https://orcid.org/0000-0003-0250-2381
Won Kyoung Jhang https://orcid.org/0000-0003-2309-0494

**AUTHOR CONTRIBUTIONS**

Conceptualization: SYS, CKJ, PSJ, JWK. Data curation: SYS, CKJ, LYH, CYH, JJH, KSY, KYH, KYA, CJB, JWK. Formal analysis: SYS, CKJ, LYH. Methodology: SYS, CKJ, LYH, JWK. Project administration: all authors. Visualization: SYS, CKJ, JWK. Writing—original draft: JWK, SYS, CKJ. Writing—review & editing: JWK, SYS, CKJ.

**SUPPLEMENTARY MATERIALS**

Supplementary materials can be found via https://doi.org/10.4266/acc.2022.00976.
REFERENCES


