MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220160

Using Data-Driven Machine Learning to Predict Unplanned ICU Transfers with Critical Deterioration from Electronic Health Records

Lingyun Shi, MS[a,b], Naveen Muthu, MD[b,c], Gerald P. Shaeffer, MS[b], Yujie Sun, MS[a,b], Victor M Ruiz Herrera, PhD[a,b] , and Fuchiang R. Tsui, PhD[a,b, †]

[a] Tsui Laboratory, Children's Hospital of Philadelphia (CHOP)
[b] Department of Biomedical Informatics, CHOP, [c] University of Pennsylvania Perelman School of Medicine

Abstract

Objective: *We aimed to develop a data-driven machine learning model for predicting critical deterioration events from routinely collected EHR data in hospitalized children.*

Materials: *This retrospective cohort study included all pediatric inpatients hospitalized on a medical or surgical ward between 2014-2018 at a quaternary children's hospital.*

Methods: *We developed a large data-driven approach and evaluated three machine learning models to predict pediatric critical deterioration events. We evaluated the models using a nested, stratified 10-fold cross-validation. The evaluation metrics included C-statistic, sensitivity, and positive predictive value. We also compared the machine learning models with patients identified as high-risk Watchers by bedside clinicians.*

Results: *The study included 57,233 inpatient admissions from 34,976 unique patients. 3,943 variables were identified from the EHR data. The XGBoost model performed best (C-statistic=0.951, CI: 0.946 ~ 0.956).*

Conclusions: *Our data-driven machine learning models accurately predicted patient deterioration. Future sociotechnical analysis will inform deployment within the clinical setting.*

Keywords:

Machine Learning, Data-Driven Science, Clinical Deterioration

Introduction

Background

Critical deterioration events (CDEs) are clinical declines in pediatric inpatients defined by a transfer to an intensive care unit (ICU) with initiation of vasopressors or positive pressure ventilation within 12 hours of transfer, a measure associated with increased morbidity and mortality[1]. CDEs are associated with higher medical costs and higher rates of morbidity and mortality. Each event adds nearly $100,000 to the cost of hospitalization. A review of events at pediatric institutions suggests that more than 40% of events may be preventable by providing advance warning of deterioration to clinical and nursing staff[2–4].

In the past two decades, several pediatric early warning systems (PEWSs) have been developed and integrated into the inpatient clinical processes to predict clinical deterioration using vital signs and other clinical data[5–8]. The Children's Hospital of Philadelphia (CHOP) has made a substantial investment in improving the recognition and response to clinical deterioration by implementing an inpatient Watcher program [9]. The Watcher program allows clinicians to proactively identify and document high-risk patients who may get sicker based on clinical findings and judgment. Watcher designations are binary judgments by the clinical team.

PEWS aim to improve human processes such as Watcher programs. Traditionally, they are typically based on a small number of clinical variables, commonly including vital signs, aggregated into a single risk score by some expert-defined method. While such systems show promise, the sole cluster-randomized trial studying the use of a pediatric early warning scoring system demonstrated a reduction in deterioration events without a reduction in mortality [5]. There is, therefore, substantial room for improvement of pediatric early warning systems.

Recent studies have explored the use of machine learning for improvement of pediatric early warning systems in adults and children [10,11] at clinical sites in the US and abroad [12,13], with state of the art machine-learning systems consistently outperforming traditional EWS systems, including the Modified Early Warning System [14], Parshuram's Bedside PEWS[15] and Monaghan's PEWS [12], and the National Early Warning System[10]. The current best ML-based systems for pediatric patients reach C-statistics from 0.79 to 0.91 with prediction horizons from 6 to 12 hours, while traditional PEWS only achieve C-statistics around 0.7-0.8 [12,13,16–20]. Previous work has been limited in the machine learning approach due to inadequate feature engineering, short prediction horizon in predicting CDEs, and/or no comparison with human judgments. For example, many studies use only static measures of clinical variables, rather than some measure of a trend in recent values (e.g., [17]). It is possible that two patients with the same clinical state ought to have different risk scores if one has recently been improving and the other deteriorating. Many ML-based PEWS use vital signs exclusively, ignoring other data including laboratory-test results or nurse assessments, leaving out possibly predictive information (e.g., [21]). Others (e.g.,[12]) use purely linear machine learning models, which miss potential important non-linear interactions between clinical variables (e.g., a 'shock index' computed as heart rate divided by systolic blood pressure), while some do not predict deterioration more than an hour out, which is of limited clinical utility as it does not give clinicians much time to intervene (e.g.,[17]). Some studies don't compare ML-based PEWS to some standard of care – whether clinician predictions or a conventional PEWS – leaving doubt as to whether machine learning actually entails some improvement over standard of care (e.g.,[22]).

Objective

To our knowledge, there is no single study of machine learning-based PEWS that avoids all of the above pitfalls. Therefore, our objective was to develop and evaluate data-driven machine learning models for predicting pediatric deterioration events in hospitalized children. We used routinely collected EHR data for prediction with the goal of providing more accurate recognition of pediatric deterioration than existing models. We used static and dynamic variables derived from clinical measures, including patient demographics, vital signs, laboratory test results, and nurse assessments to predict deterioration 24 hours before the ICU admission. We developed both linear and non-linear models and compared these ML models' performance to clinical team judgment (patients identified by the team during care as high-risk Watchers).

Methods

The Institutional Review Board at Children's Hospital of Philadelphia (CHOP) approved this study as exempt research (IRB 20-017837). This study follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [23].

Study Population

In this hospital-wide retrospective cohort study, we collected five years of EHR data at the Children's Hospital of Philadelphia, a quaternary children's hospital. The institution has 559 inpatient beds with over 29,000 annual admissions to the hospital. This includes 87 beds in the pediatric and cardiac ICUs, which see over 400 admissions annually. We included all inpatient admissions between January 2014 and December 2018 to medical or surgical wards, excluding ICUs. Admissions with less than a 48-hour length of stay (LOS) were excluded from the study. We applied the LOS criterion to ensure enough data for machine learning modeling. No additional exclusion criteria were applied. We collected 5 EHR data types: registration data (admission, discharge, and transfer), patient demographics (age, race, sex, insurance), laboratory test results, vital signs, and nurse assessments.

Outcome and prediction window

The primary outcome was admission to a medical or surgical ward with a critical deterioration event (a ward-to-ICU transfer with the initiation of vasopressors or positive pressure ventilation within 12 hours of the transfer). All non-case admissions were defined as control admissions, including admissions with ward-to-ICU transfers that did not qualify as critical deterioration events. The case prediction time was 24-hours ahead of the ICU transfer time. The control prediction times were randomly picked from 24 hours after hospital admission to 2 hours ahead of disposition/discharge from the ward. The 24 hours after admission were reserved as baseline data to ensure the availability of required EHR data by machine learning models, and the 2 hours before disposition/discharge were untouched to prevent early leakage of patient disposition signals. Figure 1 describes the cohort selection process.

Table 1 described the cohort patient characteristics from sex, race, and insurance.

Table 1 *– Demographic characteristics of cohort admissions*

Variables	**Cases (n=2,069)**	**Controls (n=55,164)**
Sex		
Female (49.3%)	915 (44.2%)	27,288 (49.5%)
Male (50.7%)	1,154 (55.8%)	27,874 (50.3%)
Race		
White (48.9%)	847 (40.9%)	27,164 (49.2%)
African American (30.6%)	656 (31.7%)	16,868 (30.6%)
Other (15.6%)	452 (21.8%)	8,452 (15.3%)
Asian (3.3%)	69 (3.3%)	1,800 (3.3%)
Indian (1.2%)	32 (1.5%)	652 (1.2%)
Unknown (0.2%)	13 (0.6%)	228 (0.4%)
Insurance		
***Medicaid** (44.1%)*	1,041 (50.3%)	24,205 (43.9%)
Medicare (0.4%)	11 (0.5%)	231 (0.4%)
Commercial (50.3%)	905 (43.7%)	27,891 (50.6%)
Self-pay (5.2%)	112 (5.4%)	2,837 (5.1%)

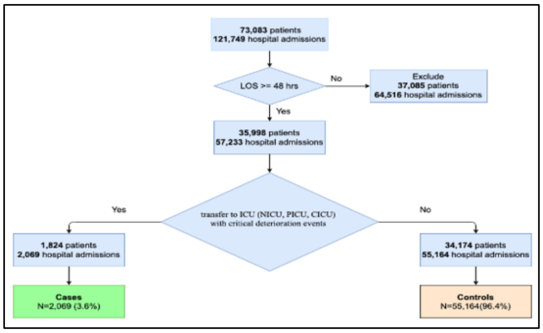

73,083 patients
121,749 hospital admissions
LOS >= 48 hrs
No
Exclude
37,085 patients
64,516 hospital admissions
Yes
35,998 patients
57,233 hospital admissions
transfer to ICU (NICU, PICU, CICU) with critical deterioration events
Yes
No
1,824 patients
2,069 hospital admissions
34,174 patients
55,164 hospital admissions
Cases
N=2,069 (3.6%)
Controls
N=55,164(96.4%)

Figure 1 *– Cohort selection flow and final case-control size; NICU, neonatal intensive care unit; PICU, pediatric intensive care unit; CICU, cardiac intensive care unit.*

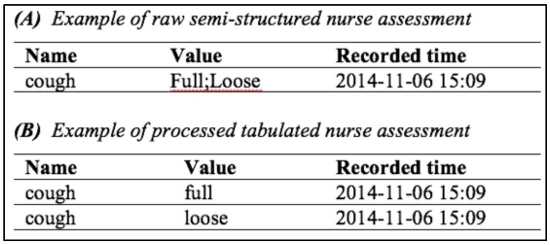
Feature Engineering

Model features were built from 72 raw EHR elements, including demographic characteristics, laboratory test results, vital signs, and nurse assessments that occurred between the hospital admission and the prediction time.

For the laboratory data, we constructed 14 time-series features for each numeric lab test. We first conducted preprocessing steps, including deleting labs with invalid units, scaling values to ensure that all labs with the same name were measured using the same unit. For example, we harmonized all potassium values to mmol/L unit, e.g., all mmol/mL values were divided by 1000. Then, we aggregated the most granular time-series data points into hourly averaged values. Finally, we constructed 14 time-series features based on the hourly values, including : (1) the first value, (2) the last value, (3) the maximum value, (4) the minimum value, (5) the difference between the last two values, (6) the difference between the last two values divided by the last value, (7) the difference between the last value and the maximum value, (8) the difference between the last value and the minimum value, (9) the difference between the last value and the maximum value, divided by the maximum value, (10) the difference between the last value and the minimum value, divided by the minimum value, (11) the difference between the first value and the last value, (12) the difference between the first value and the last value, divided by the first value, (13) the slope of the last two values, and (14) the linear regression slope from all values. Other than the time-series features, we also calculated the count and percentage of different result flags (e.g., normal, abnormal, high, and low).

We constructed the same set of time-series features for vital signs such as systolic/diastolic blood pressures and temperature.

Nurse assessments were semi-structured data. Figure 2.A shows an example of original assessment of cough with two characteristics: full and loose.



(A) Example of raw semi-structured nurse assessment

Name	Value	Recorded time
cough	Full;Loose	2014-11-06 15:09

(B) Example of processed tabulated nurse assessment

Name	Value	Recorded time
cough	full	2014-11-06 15:09
cough	loose	2014-11-06 15:09

***Figure 2** – Example of processing nurse assessments.*

To process nurse assessments, we lowercased the value, split the value by semicolon (;), and stacked the split value. For example, the cough assessment in Figure 2.A was processed into what is shown in Figure 2.B. Then, we selected the most recent value(s) for each assessment and used one-hot encoding to generate assessment features.

Finally, we concatenated the demographic, laboratory test, vital signs, and nurse assessment features and formed the final dataset for machine learning.

Missing Data and Data Imputation

We created a "missing" label for missing values of categorical variables. For numeric variables, the missing data were imputed using the mean value from the training dataset.

Machine Learning Modeling

We built and evaluated three types of machine learning models: (1) generalized linear model (GLM), (2) gradient boost model (XGBoost), and (3) deep neural network (DNN). GLM is a type of regression model which works very well with a limited number of predictors. Once the model is trained, the computation for predicting new samples is extremely fast and can be easily distributed. It is also a very interpretable model. Thus, GLM has been a popular choice in real clinical applications. XGBoost is an implementation of gradient boost model optimized on computational speed and model performance. It has recently been used to obtain the state of the art performance in various biomedical prediction tasks using tabular data. DNN is a type of highly non-linear model, which has recently dominated the applied machine learning tasks for image, video, and free text. We also included DNN in this study as it can potentially capture the non-linear interactions between clinical variables.

Model Performance Evaluation

We performed nested cross-validation (CV) for optimizing hyperparameters and evaluating model performance. Nested CV is a more rigorous protocol that overcomes the overfitting pitfall of non-nested CV [24]. Figure 3 depicts this nested modeling strategy. First, the dataset was randomly split into 10 outer folds. One fold was reserved within each outer round as a testing dataset, and the remaining 9 folds were combined and used as a training dataset. We then performed a grid search on the training dataset through inner 5-fold cross-validation to find the best hyperparameters. Table 2 lists the hyperparameters of the grid search for each model type. For GLM and DNN, we used a cartesian grid search strategy to exhaust all combinations of the hyperparameters. A random grid search strategy was used for XGBoost because there were too many hyperparameter combinations to run an exhaustive search. The random grid search was stopped after a maximum of 50 models trained or after the area under the receiver operating characteristics (AUROC) did not increase more than 0.001 for five consecutive runs. We trained a model from the training dataset with the obtained best hyperparameters and tested it on the reserved testing dataset for this outer fold. Finally, we computed the evaluation metrics from the stacked test model predictions from 10 outer folds.

We evaluated the machine learning model performance with four metrics: AUROC (C-statistic), sensitivity, specificity, and positive predictive value (PPV). To assess the feasibility of implementing our systems in a real clinical setting, we compare the number of variables among different models. We also compared the machine learning models with the clinician's judgment – the Watcher score – by comparing the sensitivity, specificity, and PPV in the ROC curve and the precision-recall curve.

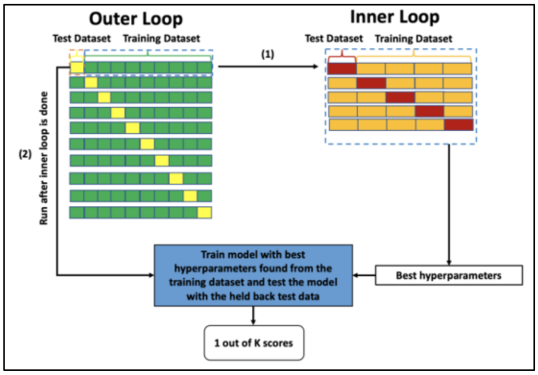

Outer Loop
Test Dataset
Training Dataset
Inner Loop
Test Dataset
Training Dataset
(1)
(2)
Run after inner loop is done
Train model with best hyperparameters found from the training dataset and test the model with the held back test data
Best hyperparameters
1 out of K scores

***Figure 3** – 10-fold nested cross-validation strategy; The red box test dataset was used as validation dataset for hyperparameter search for the inner loop CV; The yellow box test dataset was hold-out test dataset for the outer loop CV; CV, cross-validation.*

***Table 2** – Hyperparameters of model grid search*

Model	Hyperparameter
GLM	alpha (distribution between l1-LASSO and l2-Ridge regression penalties)
	lambda (regularization strength)
XGBoost	learning rate (the weighting of new trees added to the model)
	number of trees
	maximum tree depth
	booster function (gbtree, gblinear, dart)
	alpha (distribution between l1-LASSO and l2-Ridge regression penalties for regression tree)
	Lambda (regularization strength for regression tree)
DNN	number and size of hidden layers
	activation function (Tanh, Rectifier, Maxout)
	dropout ratio

Results

Cohort Description

We identified 57,233 hospital admissions from 35,998 patients, among which 3.6% were case admissions and 96.4% were control admissions.

In the raw EHR data, there were 72 variables comprising 26 laboratory tests, 5 vital signs, and 41 nurse assessments. After feature engineering and one-hot encoding, we obtained a total of 3,943 features (predictors) for model training.

Machine Learning and Clinical Team Watcher Performance

Our data-driven machine learning models achieved excellent predictive performance with C-statistics ranging from 0.929 to 0.951. Figures 4.A and 4.B show ROC and Precision-recall curves, respectively, for all three ML models. These were compared to the performance of bedside teams identifying patients as high-risk Watchers during clinical care (depicted in red). The ROC curves were created from stacked test predictions from 10 folds, and showed that XGBoost outperformed GLM and DNN in predicting CDEs.

The Watcher program achieved a 0.261 positive predictive value (PPV) with 0.197 sensitivity, confirming that most patients with CDEs were not predicted by the clinical team.

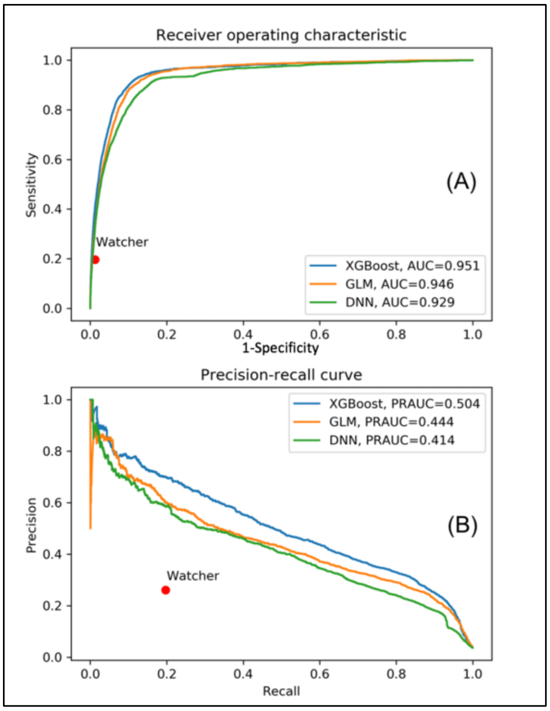


Figure 4 *- Prediction performance comparison between the machine learning models and the Watcher score. Because Watcher score is a binary (0 or 1) flag that clinicians assign to hospital admission, we can not plot the ROC or PRC curves by moving the prediction threshold. Instead, we calculated its sensitivity, specificity, and PPV, and plotted the Watcher performance as a red dot in the figure.*

The GLM model contained 190 features; the XGBoost and DNN models had 587 and 3,943, respectively. As expected, GLM model had the lowest number of features, while DNN had the most. Figure 5 is a stacked bar chart further showing the distribution of model features in each category of demographics, lab tests, vital signs, and nurse assessments. GLM model features were more evenly distributed in all 4 categories, while XGBoost used more laboratory tests. There was no feature selection in DNN; thus DNN used all developed features.

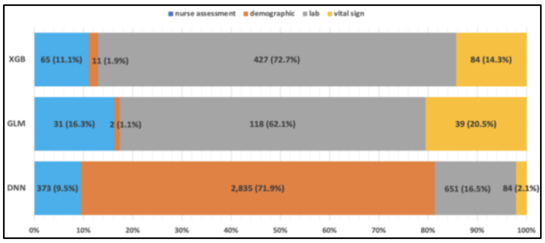


Figure 5 *– The distribution of feature numbers of our 3 machine learning models in the 4 categories: demographics, vital signs, laboratory test, and nurse assessments. DNN used all 3,943 features, and because we had one-hot encoding for every zip code in demographics, the number of demographic features in the DNN model was inflated.*

Discussion

We developed and evaluated three different types of data-driven machine learning models that can accurately predict acute patient deterioration within 12 hours of ICU transfer, represented by subsequent administration of vasopressors or positive pressure ventilation. Our XGBoost model attained better performance (C-statistic = 0.951) than the state of the art model with an earlier prediction horizon (24-hours prior to ICU transfer), which can give clinicians more time to intervene or prepare. The GLM model achieved similar performance (C-statistic = 0.946) with a third as many model features as the XGboost model. We may consider implementing the GLM model in clinical care due to its simplicity with minimal drop in prediction performance. Compared with the existing Watcher program, which captures clinical team concern, our best model achieved 4.5 times higher sensitivity given the same PPV.

The strengths of our study include a data-driven approach and time-series feature engineering, which expanded 72 raw EHR variables into a total of 3,943 features. The best-performing model contained 587 features, the majority of which (427) came from laboratory tests.

The main limitation of our study is that the models were developed and tested within a single hospital. Using data from additional hospitals may further test the generalizability of our models. Another limitation is that the study employed a large number of variables from EHR data that may limit the potential adoption from other hospitals, especially for those hospitals not using the same EHR system at CHOP, the Epic EHR system. A common data model such as OMOP may be needed to serve as a research data warehouse for addressing the limitation of different EHR systems such as Cerner and AllScripts used in different hospitals. Finally, our study required at least 48 hours of historical EHR data prior to the onset of ICU transfers, which may limit the coverage of the patient population with a short floor stay. Further exploration using a shorter LOS may be considered in the future.

Future Work

We plan to further test the reliability of models using 2019 and 2020 data before deploying the model to a real-time decision support system. The reliability test can verify if the models using the data from 2014 to 2018 can perform similarly. Further model training, calibration, and bias testing across different demographics and model calibration are also critical before deployment. To maximize the benefit for clinical care, we also plan to seamlessly integrate the model into the clinician's workflow by embedding the models into the Epic EHR system using a real-time, event-driven system, based on our previous experience[25,26]. It would be important to carefully select and evaluate the alert thresholds for sending out alerts to reduce clinicians' alert fatigue in practice.

Conclusions

In this sizeable single-hospital retrospective inpatient study, we developed and evaluated predictive models for CDEs and compared the model performance with the bedside care team's Watcher program. Compared with other state of the art models, our best model achieved a better C-statistic with an earlier prediction horizon. Since all our models outperformed the current Watcher program, we believe our data-driven machine learning model could be implemented as part of a real-time decision support system to help identify more at-risk patients and thus better prepare clinicians and potentially reduce preventable mortality and morbidity. We plan to validate our model prospectively and/or at other sites. If this performance holds, we would achieve – to the best of our knowledge – state-of-the-art in ML- or expert-based pediatric early warning systems. Our ML approach with a large EHR data feature space may be valuable to identify patients with rare clinical events like CDEs.

Acknowledgements

This work was supported by the Children's Hospital of Philadelphia (CHOP). The content is solely the responsibility of the authors and does not necessarily represent the official views of CHOP. All authors have no conflicts of interest to disclose.

References

[1] C.P. Bonafide, A.R. Localio, K.E. Roberts, V.M. Nadkarni, C.M. Weirich, and R. Keren, Impact of rapid response system implementation on critical deterioration events in children, *JAMA Pediatr.* **168** (2014) 25–33. doi:10.1001/jamapediatrics.2013.3266.

[2] J. Tibballs, and S. Kinney, Reduction of hospital mortality and of preventable cardiac arrest and death on introduction of a pediatric medical emergency team, *Pediatr. Crit. Care Med.* **10** (2009). doi:10.1097/PCC.0b013e318198b02c.

[3] G. Pearson, Why children Die: The report of a pilot confidential enquiry into child death by CEMACh (Confidential Enquiry into Maternal and Child Health), *Clin. Risk*. **14** (2008) 166–168. doi:10.1258/cr.2008.080042.

[4] C.P. Bonafide, A.R. Localio, L. Song, K.E. Roberts, V.M. Nadkarni, M. Priestley, C.W. Paine, M. Zander, M. Lutts, P.W. Brady, and R. Keren, Cost-benefit analysis of a medical emergency team in a children's hospital, *Pediatrics*. **134** (2014) 235–241. doi:10.1542/peds.2014-0140.

[5] C.S. Parshuram, K. Dryden-Palmer, C. Farrell, R. Gottesman, M. Gray, J.S. Hutchison, M. Helfaer, E.A. Hunt, A.R. Joffe, J. Lacroix, M.A. Moga, V. Nadkarni, N. Ninis, P.C. Parkin, D. Wensley, A.R. Willan, G.A. Tomlinson, A. Willems, M. Hazim, B. Wenderickx, A. Kotsakis, S. Gander, W. Harris, J. Holland, J. MacLean, D. Boliver, S. Zavalkoff, M. Dagenais, S. Shea, J. Gaudreault, M.A. Dugas, L. Gosselin, C. Proulx-Clerc, L. Bertout, I. Grisoni, J. Duff, J. Pugh, D. Capito, G. Krahn, A. Barclay, F. Auld, L. Robson, E. Carrick, J. Gilleland, L. Saunders, D. Fraser, P. Bechard, C. Martin, L. Spear, K. Tobler, K. Kulbaba, N. Peiris, D.R. Doherty, F. Fjficmi, E. Ladewig, S. Somanadhan, L. Greensmith, C. Breatnach, C. O'Rourke, B. Gesù, C. Cecchetti, O. Gawronski, A. Ruscitto, E.P. Cabillon, M.C.D. Atti, M. Raponi, G. Nuthall, G.D. Williams, C. Sherring, T. Bushell, M. Rea, L. Armriding, G. Olykan, C. Van Der Starre, A. Hogeboom, A. De Oude-Lubbers, N. Hemat, S. Broughton, S. Harris, E. Downing, D. Inwald, R. Sinha, S. Raghunanan, M. Vaidya, L. Reardon, M. Burmester, K. Kailay, L. Haidu, S. Ferri, J. Grillo, N. Shahid, S. Ashley, S. Singh, K. Byrne, A. Kamath, and K. Middaugh, Effect of a pediatric early warning system on all-cause mortality in Hospitalized pediatric patients: The epoch randomized clinical trial, *JAMA - J. Am. Med. Assoc.* **319** (2018) 1002–1012. doi:10.1001/jama.2018.0948.

[6] V. Lambert, A. Matthews, R. MacDonell, and J. Fitzsimons, Paediatric early warning systems for detecting and responding to clinical deterioration in children: A systematic review, *BMJ Open*. **7** (2017). doi:10.1136/bmjopen-2016-014497.

[7] S.M. Chapman, J. Wray, K. Oulton, and M.J. Peters, Systematic review of paediatric track and trigger systems for hospitalised children, *Resuscitation*. **109** (2016) 87–109. doi:10.1016/j.resuscitation.2016.07.230.

[8] S.M. Chapman, and I.K. Maconochie, Early warning scores in paediatrics: An overview, *Arch. Dis. Child.* **104** (2019) 395–399. doi:10.1136/archdischild-2018-314807.

[9] S. Muething, U. Kotagal, M. Ashby, R. Gallagher, D. Hall, M. Goodfriend, C. White, T.M. Bracke, V. Decastro, M. Geiser, J. Simon, K.M. Tucker, J. Olivea, P.H. Conway, and D.S. Wheeler, Improving Situation Awareness to Reduce Unrecognized Clinical Deterioration and Serious Safety Events abstract, *Pediatrics*. **131** (2013).

[10] S. Romero-Brufau, D. Whitford, M.G. Johnson, J. Hickman, B.W. Morlan, T. Therneau, J. Naessens, and J.M. Huddleston, Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS), *J. Am. Med. Informatics Assoc.* **00** (2021) 1–9. doi:10.1093/jamia/ocaa347.

[11] H. Lonsdale, A. Jalali, L. Ahumada, and C. Matava, Machine Learning and Artificial Intelligence in Pediatric Research: Current State, Future Prospects, and Examples in Perioperative and Critical Care, *J. Pediatr.* **221** (2020) S3–S10. doi:10.1016/j.jpeds.2020.02.039.

[12] H. Zhai, P. Brady, Q. Li, T. Lingren, Y. Ni, D.S. Wheeler, and I. Solti, Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children, *Resuscitation*. **85** (2014) 1065–1071. doi:10.1016/j.resuscitation.2014.04.009.

[13] L. Xiang, H. Wang, S. Fan, W. Zhang, H. Lu, B. Dong, S. Liu, Y. Chen, Y. Wang, L. Zhao, and L. Fu, Machine Learning for Early Warning of Septic Shock in Children With Hematological Malignancies Accompanied by Fever or Neutropenia: A Single Center Retrospective Study, *Front. Oncol.* **11** (2021) 1–9. doi:10.3389/fonc.2021.678743.

[14] A. Kia, P. Timsina, H.N. Joshi, E. Klang, R.R. Gupta, R.M. Freeman, D.L. Reich, M.S. Tomlinson, J.T. Dudley, R. Kohli-Seth, M. Mazumdar, and M.A. Levin, MEWS++: Enhancing the Prediction of Clinical Deterioration in Admitted Patients through a Machine Learning Model, *J. Clin. Med.* **9** (2020) 343. doi:10.3390/jcm9020343.

[15] C.S. Parshuram, H.P. Duncan, A.R. Joffe, C.A. Farrell, J.R. Lacroix, K.L. Middaugh, J.S. Hutchison, D. Wensley, N. Blanchard, J. Beyene, and P.C. Parkin, Multicentre validation of the bedside paediatric early warning system score: A severity of illness score to detect evolving critical illness in hospitalised children, *Crit. Care*. **15** (2011). doi:10.1186/cc10337.

[16] S. Romero-Brufau, D. Whitford, M.G. Johnson, J. Hickman, B.W. Morlan, T. Therneau, J. Naessens, and J.M. Huddleston, Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS), *J. Am. Med. Informatics Assoc.* **28** (2021) 1207–1215. doi:10.1093/jamia/ocaa347.

[17] J. Kwon, K.-H. Jeon, M. Lee, K.-H. Kim, J. Park, and B.-H. Oh, Deep Learning Algorithm to Predict Need for Critical Care in Pediatric Emergency Departments, *Pediatr. Emerg. Care*. **Publish Ah** (2019) 1–7. doi:10.1097/pec.0000000000001858.

[18] S.Y. Kim, S. Kim, J. Cho, Y.S. Kim, I.S. Sol, Y. Sung, I. Cho, M. Park, H. Jang, Y.H. Kim, K.W. Kim, and M.H. Sohn, A deep learning model for real-time mortality prediction in critically ill children, *Crit. Care*. **23** (2019) 1–10. doi:10.1186/s13054-019-2561-z.

[19] S.J. Park, K.-J. Cho, O. Kwon, H. Park, Y. Lee, W.H. Shim, C.R. Park, and W.K. Jhang, Development and validation of a deep-learning-based pediatric early warning system: A single-center study, *Biomed. J.* (2021) 1–14. doi:10.1016/j.bj.2021.01.003.

[20] P.M. Report, Predicting Future Care Requirements Using Machine Learning for Pediatric Intensive and Routine Care Inpatients, (2018) 1–13. doi:10.1097/CCE.0000000000000505.

[21] A. Mayampurath, P. Jani, Y. Dai, R. Gibbons, D. Edelson, and M.M. Churpek, A vital sign-based model to predict clinical deterioration in hospitalized children, *Pediatr. Crit. Care Med.* **21** (2020) 820–826. doi:10.1097/PCC.0000000000002414.

[22] B. Wellner, J. Grand, E. Canzone, M. Coarr, P.W. Brady, J. Simmons, E. Kirkendall, N. Dean, M. Kleinman, and P. Sylvester, Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements, *JMIR Med. Informatics*. **5** (2017) 1–18. doi:10.2196/medinform.8680.

[23] G.S. Collins, J.B. Reitsma, D.G. Altman, and K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement, *BMC Med.* **13** (2015) 1–10. doi:10.1186/s12916-014-0241-z.

[24] G.C. Cawley, and N.L.C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, *J. Mach. Learn. Res.* **11** (2010) 2079–2107.

[25] F.-C. Tsui, J.U. Espino, V.M. Dato, P.H. Gesteland, J. Hutman, and M.M. Wagner, Technical Description of RODS- A Real-time Public Health Surveillance System, *J Am Med Inf. Assoc*. **10** (2003) 399–408. doi:10.1197/jamia.M1345.Covert.

[26] F.C. Tsui, J.U. Espino, Y. Weng, a Choudary, H.D. Su, and M.M. Wagner, Key design elements of a data utility for national biosurveillance: event-driven architecture, caching, and Web service model, *AMIA Annu Symp Proc*. (2005) 739–743. doi:58729 [pii].

Address for correspondence

Fuchiang Tsui, PhD, FAMIA; tsuif@chop.edu; 267-425-1294.