Title: Multicenter Development and Prospective Validation of eCARTv5: A Gradient Boosted Machine Learning Early Warning Score

Authors: Matthew M. Churpek, MD, MPH, PhD, ATSF[1,2]; Kyle A. Carey, MPH[3]; Ashley Snyder, MPH[4]; Christopher J Winslow, MD[5]; Emily Gilbert, MD[6]; Nirav S Shah, MD, MPH[5]; Brian W. Patterson, MD, MPH[2,7]; Majid Afshar, MD, MSCR[1,2]; Alan Weiss, MD, MBA[8]; Devendra N. Amin, MD[8]; Deborah J. Rhodes, MD[9]; Dana P. Edelson, MD, MS[3,4]

Affiliations:
[1]Department of Medicine, University of Wisconsin-Madison, Madison, WI
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI
[3]Department of Medicine, University of Chicago, Chicago, IL
[4]AgileMD, San Francisco, CA
[5]Department of Medicine, Endeavor Health, Evanston, IL
[6]Department of Medicine, Loyola University Medical Center, Chicago, IL
[7]Department of Emergency Medicine, University of Wisconsin-Madison, Madison WI
[8]BayCare, Clearwater, FL
[9]Department of Medicine, Yale University, New Haven, CT

Corresponding author:
Matthew M Churpek, MD, MPH, PhD
University of Wisconsin School of Medicine and Public Health
600 Highland Ave, Madison, WI 53726
Email: mchurpek@medicine.wisc.edu

Disclosures: Drs. Churpek and Edelson are inventors on a patent for patient risk evaluation (US11410777) and receive royalties from this intellectual property from the University of Chicago. Dr. Edelson is employed by and has an equity stake in AgileMD, which markets and distributes eCART.

Funding source: This work was supported by funding from the National Institutes of Health (PI: MMC; R01HL157262) and Biomedical Advanced Research and Development Authority (BARDA) as part of its Division of Research Innovation and Ventures (DRIVe) under contract number 75A50121C00043 (PI: DPE).

Contributions: MMC takes full responsibility for the content of the manuscript. MMC and DPE conceptualized the study. KAC conducted statistical analysis of the data. MMC wrote the first draft of the manuscript and revised subsequent versions. All authors contributed to the interpretation of data, reviewed and edited the initial drafts, and approved the final manuscript.

Keywords: Early Warning Score; Clinical Deterioration; Machine Learning; Rapid Response Systems; Artificial Intelligence

Word Count: 3,310

ABSTRACT

Rationale: Early detection of clinical deterioration using early warning scores may improve outcomes. However, most implemented scores were developed using logistic regression, only underwent retrospective internal validation, and were not tested in important patient subgroups.

Objectives: To develop a gradient boosted machine model (eCARTv5) for identifying clinical deterioration and then validate externally, test prospectively, and evaluate across patient subgroups.

Methods: All adult patients hospitalized on the wards in seven hospitals from 2008-2022 were used to develop eCARTv5, with demographics, vital signs, clinician documentation, and laboratory values utilized to predict intensive care unit transfer or death in the next 24 hours. The model was externally validated retrospectively in 21 hospitals from 2009-2023 and prospectively in 10 hospitals from February to May 2023. eCARTv5 was compared to the Modified Early Warning Score (MEWS) and the National Early Warning Score (NEWS) using the area under the receiver operating characteristic curve (AUROC).

Measurements and Main Results: The development cohort included 901,491 admissions, the retrospective validation cohort included 1,769,461 admissions, and the prospective validation cohort included 46,330 admissions. In retrospective validation, eCART had the highest AUROC (0.835; 95%CI 0.834, 0.835), followed by NEWS (0.766 (95%CI 0.766, 0.767)), and MEWS (0.704 (95%CI 0.703, 0.704)). eCART's performance remained high (AUROC ≥0.80) across a range of patient demographics, clinical conditions, and during prospective validation.

Conclusions: We developed eCARTv5, which accurately identifies early clinical deterioration in hospitalized ward patients. Our model performed better than the NEWS and MEWS retrospectively, prospectively, and across a range of subgroups.

Introduction

Clinical deterioration occurs in up to 5% of hospitalized patients, and early detection and treatment has been associated with improved patient outcomes (1-9). These events are often heralded by physiologic abnormalities, such as deranged vital signs and laboratory values, in the hours to days before the event, which has led to the development of early warning scores aimed at identifying high-risk patients before deterioration (10, 11). Early warning scores have evolved over time from aggregated weighted scores, such as the Modified Early Warning Score (MEWS) (12), which can be calculated by hand, to those based on logistic regression and other traditional statistical frameworks (1, 2), which can be summed with a calculator or spreadsheet, and more recently to advanced machine learning models, such as gradient boosted machines (GBM), which have been shown to be more accurate across multiple tasks in large datasets (13-16). Computational resources and electronic record use have grown in parallel to the complexity of these scores, which provides the capability to implement more advanced scores in real-time for patient care.

While machine learning models come with the promise of decreased false alarms and increased detection rates over both traditional statistical models and the original aggregated weighted scores, they can suffer from overfitting and poor calibration when trained on insufficiently sized and representative datasets (17, 18). Further, the use of these models in clinical practice raises concerns about model fairness and bias, and little is known regarding how they perform across a range of important patient subgroups. For example, a model may perform well overall in a population, but may underperform in specific subgroups, which can lead to the diversion of resources away

from disadvantaged, vulnerable patients. Evaluating models across these subgroups is critical to ensuring the fairness of these tools in practice so that they have the potential to benefit all patients, including marginalized groups. Finally, most of the work in this area has been done retrospectively (11), and it is not known whether these complex models will perform similarly in production environments, with prospective calculations executed in real-time.

Therefore, we aimed to develop and externally validate a GBM model for identifying clinical deterioration in a large, geographically diverse set of hospitals. After retrospective validation, which included extensive subgroup analyses, we then silently tested the model's performance prospectively. These results serve as the foundation for a submission to the Food and Drug Administration for a first-in-class advanced machine learning analytic for inpatient clinical deterioration detection.

Methods

Study Overview and Population

In this observational cohort study, a machine learning model to predict impending clinical deterioration was developed and validated in hospitalized adult (age ≥18 years) medical-surgical ward patients. Patients who were only admitted to the intensive care unit (ICU), labor and delivery, or emergency department and were never transferred to a medical-surgical (non-ICU) unit during their hospital encounter were excluded. The model, eCARTv5 (eCART), was developed in a dataset of seven hospitals from three health systems in Illinois (D1-D3), spanning the years 2008-2022. The model was then externally validated in two phases: (1) a retrospective cohort of admissions to 21 hospitals from three health systems (R1-R3) in Florida, Wisconsin, Connecticut, and

Rhode Island, encompassing years 2009-2023; and (2) a 16-week prospective cohort of consecutive admissions to the 10 Florida hospitals where eCART ran in production with scores hidden from clinicians from February to May in 2023 (P3). eCART was compared to the MEWS (12)and the National Early Warning Score (NEWS) (19), which are commonly used and cited tools for clinical deterioration. The study was funded by the Biomedical Advanced Research and Development Authority (BARDA) as part of its Division of Research Innovation and Ventures (DRIVe) under contract number 75A50121C00043 and the National Institutes of Health (R01HL157262). The study was approved for each health system by the following Institutional Review Boards (IRB): University of Chicago Biological Sciences Division IRB (#18-0447), Loyola University Chicago Health Sciences Division IRB (#215437), NorthShore University HealthSystem Research Institute IRB (#EH16-210T), University of Wisconsin-Madison Minimal Risk Research IRB (#2019-1258), BayCare Health System IRB (#2022.014-B.MPH & #2022.015-B.MPH) and Yale Human Research Protection Program IRBs (#2000035317). Each of the IRBs waived study-specific informed consent. Procedures were followed in accordance with the ethical standards of the responsible institutional committee on human experimentation and with the Helsinki Declaration of 1975. Please see the supplemental digital content for more information.

Outcome

The study outcome was clinical deterioration, defined as death or ICU transfer from the medical-surgical wards within 24 hours of a score (1, 13, 19). Death was determined using the discharge disposition from the admission, discharge, transfer (ADT) data feed

in the EHR, with the time of death being the last recorded vital sign. ICU transfer was defined as a direct ward to ICU transfer and was determined using the transfer disposition from the ADT data feed in the EHR, with the time of transfer being the last vital sign on the ward.

Predictor Variables

A total of 97 features were included as predictor variables in the eCART machine learning model. These variables included patient characteristics (e.g., age, body mass index (BMI), vital signs, laboratory values, time of day, time since admission, and nursing/respiratory therapist documentation (e.g., the amount of delivered oxygen, Braden scale), as well as vital sign and laboratory value trends (20). A full list of model predictor variables is found in Appendix Table E1 in the Online Supplement.

Non-physiologic flowsheet data were considered to be input errors and treated as missing, as per prior publications (see Appendix Table E2 in the Online Supplement). If no physiologic range data for a particular variable were available at a specific time, then the most recent prior value, if available, was pulled forward. If no prior values were available, the variable was left as missing. Given the dynamic nature of blood gas and lactate values and the tendency of providers to only order them on actively deteriorating patients, those values were only pulled forward for 24 hours, after which they were treated as missing in the model.

Model Development

A gradient boosted machine (GBM) model was developed with the aforementioned features to predict clinical deterioration in all adult patients hospitalized on the wards in the training data. A standard statistical framework, known as discrete-time survival analysis, was used during model development (1, 13, 21). This framework utilizes non-overlapping time segments and is analogous to life tables, whereby the patient's risk of an event in the future is conditional on the fact that they have survived to that time point. Time was discretized into 8-hour blocks, and the data at the beginning of each time block were used to predict whether an outcome occurred within eight hours of the beginning of that block. This approach allowed the inclusion of time-varying predictor variables, removed the bias of sicker patients receiving more frequent measurements, and provided results analogous to the Cox survival model (21). Because tree-based models can perform poorly in highly imbalanced data (i.e., when the outcome of interest is uncommon), down-sampling of the training dataset to obtain a 50% outcome prevalence was performed prior to model fitting (13). Model hyperparameters were tuned in the training cohort using five-fold cross-validation to maximize the area under the receiver operating characteristic curve (AUROC). No variable selection was performed, as earlier research has shown that this does not improve the accuracy (and may degrade performance) of tree-based machine learning algorithms (22). Variable importance was calculated using the relative influence of each variable based on the improvement of each split averaged across all trees (17). Additional modeling details can be found in the Online Supplement.

Retrospective and Prospective Score Calculation

For model validation, data pre-processing was performed in the same manner as during model development with the exception that in the validation cohorts, both retrospective and prospective, data were not blocked and no down-sampling was performed. Specifically, each time a new observation was recorded in the EHR (i.e., a new data point becomes available), predicted probabilities from the eCART model, as well as MEWS and NEWS scores were calculated. The transformed eCART model output probabilities were then scaled to eCART scores ranging from 0-100 for ease of interpretation. For the retrospective validation, these scores were calculated on a static multicenter dataset stored on secured, laboratory servers. In the prospective validation, the model features, outputted scores, and outcomes were collected in real-time utilizing Health Level-Seven Version 2 (HL-7 V2) messaging standard interfaces, a clinical data standard to protocolize how data are shared and exchanged in EHR operations, and stored on cloud-based computing and storage resources hosted at Amazon Web Services. The model scores were not available to clinicians during this silent validation.

Statistical Analysis

Descriptive statistics were used to characterize patient demographics across the separate development, retrospective validation, and prospective validation cohorts. Model performance was calculated by assessing the ability of the scores at each observation time to predict clinical deterioration in the following 24 hours. Discrimination was measured using the AUROC and then compared using the method of DeLong (23). Subgroup analyses were performed in the retrospective validation cohort across patient

demographics (age, sex, race) and clinical conditions (surgical, obstetric, sepsis, COVID-19, congestive heart failure (CHF), and chronic obstructive pulmonary disease (COPD)). Definitions of these subgroups can be found in the Supplementary Methods in the Online Supplement. Sensitivity, specificity, and positive and negative predictive values were calculated for each threshold, with confidence intervals calculated using the Clopper-Pearson method. Performance at a moderate-risk and high-risk threshold for each score (eCART of ≥93 and ≥97; NEWS ≥5 and ≥7; MEWS ≥3 and ≥4) was also compared. Model calibration was assessed in the prospective validation by comparing observed to expected deterioration rates across eCART score values. Analyses were performed using Stata version 16.1 (StataCorps; College Station, Texas) and R version 4.2.1 (The R Foundation for Statistical Computing, Vienna, Austria).

Results

The training dataset included 901,491 adult inpatient admissions with ward stays at seven hospitals from three health systems while the retrospective validation cohort (R1-R3) included 1,769,461 adult admissions to 21 hospitals from three health systems. The prospective validation cohort (P3) included 46,330 consecutive adult admissions to 10 hospitals. There was considerable variation in demographics across the three cohorts (Table 1). When compared to the two validation cohorts, the training cohort had a higher proportion of black patients (32% in the training cohort vs. 14% and 16% in the retrospective and prospective cohorts) and a lower proportion of most of the Elixhauser comorbidities, except for malignancies. Meanwhile, the prospective validation cohort differed from the retrospective validation cohort in having a higher median age (66 vs 62

years), a shorter length of stay (60 vs 72 hours), a lower proportion of encounters with surgical procedures, and a higher prevalence of most Elixhauser comorbidities. There was also considerable variation in missing variables across the health systems (Table E3 in the Online Supplement). Most notably, R3 had higher rates of missing Braden scores and respiratory rate trends, suggesting a lower frequency of respiratory rate documentation, which was even more pronounced in the prospective cohort (P3). R1 had higher rates of missing hematology labs, particularly white blood cell differential distributions. R2 had higher missing rates for mental status.

The most important variables in the final eCART model were maximum respiratory rate in the prior 24 hours, delivered FiO2, minimum systolic blood pressure in the prior 24 hours, and heart rate (Figure 1). Partial plots illustrating the relationship between values of these variables and risk of deterioration are shown in Figure E1.

In the retrospective validation dataset, a total of 132,873,833 eCART, MEWS and NEWS scores were calculated. The AUROC for eCART was 0.835 (0.834, 0.835) for the full retrospective cohort (Table 2). eCART consistently outperformed NEWS (AUROC 0.766 (0.766, 0.767)), which consistently outperformed MEWS (AUROC 0.704 (0.703, 0.704)). eCART's sensitivity in the retrospective cohort at the moderate-risk threshold (≥93) was 51.8% with a positive predictive value (PPV) of 9.0%. At the high-risk threshold (≥97), PPV increased to 14.2% with a decrease in sensitivity to 38.6% (Appendix Table E4 in the Online Supplement). In contrast, MEWS had sensitivities of 38.9% and 22.6% at the moderate (≥3) and high-risk (≥4) thresholds, with corresponding PPVs of 5.7% and 10.4%, respectively, while NEWS had sensitivities of 49.7% and 28.0% at the moderate (≥5) and high-risk (≥7) thresholds, with

corresponding PPVs of 5.4% and 9.9%, respectively (Appendix Tables E5 and E6 in the Online Supplement). The precision-recall curve, Figure 2, plots PPV as a function of sensitivity and demonstrates a consistently more favorable tradeoff between sensitivity and PPV for eCART compared to both NEWS and MEWS. At the moderate-risk threshold, NEWS provided the longest lead time prior to the deterioration event (17 hours (IQR 1, 73)), followed by eCART (16 hours (IQR 1, 68)), and then MEWS (13 hours (IQR 1, 66)). At the high-risk threshold, eCART alerted a median of 5 (IQR 0, 43) hours in advance of clinical deterioration, significantly earlier than NEWS (3 hours (IQR 0, 40) and MEWS (2 hours (IQR 0, 30)) ($p<0.01$ for all comparisons).

Performance across subgroups in the retrospective validation (Table 3) demonstrated that eCART (AUROCs 0.810-0.909) consistently outperformed NEWS (AUROCs 0.745-0.793), which outperformed MEWS (AUROCs 0.672-0.726). Among the different age groups, the AUROC for eCART was highest in 18-33 year old patients (0.861) and lowest in 65-78 year old patients (0.822). In subgroup analysis by race, the eCART AUROC was highest in the Native Hawaiian/Other Pacific Islander cohort (0.862) and lowest in the American Indian or Alaska Native cohort (0.814). eCART performance was slightly higher for female compared to male patients (0.844 vs 0.824) and was exceptionally high in obstetric encounters (0.909). Among the clinical conditions, performance was highest in patients with COVID-19 across all scores and lowest in heart failure. Across all subgroups studied, eCART retained high discrimination (AUROC ≥0.81).

In the prospective analysis, there were 4,778,200 scores calculated, and 1,579 encounters had clinical deterioration within 24 hours following an observation.

Performance in the prospective cohort (P3 in Table 2) was similar to the retrospective results in this health system (R3), and eCART (AUROC 0.800 (0.798, 0.801)) outperformed NEWS (AUROC 0.736 (0.734, 0.738)) and MEWS (AUROC 0.681 (0.678, 0.683)). Model calibration is shown in Figure 3, demonstrating close agreement between the observed and expected deterioration rates.

Discussion

In a large retrospective validation of nearly two million inpatient encounters with over 130 million calculated scores from three geographically distinct health systems in the United States, eCART outperformed MEWS and NEWS for predicting impending clinical deterioration. Results were robust across age, sex, and race, with eCART performing better than NEWS and MEWS in all subgroups. Further, eCART had consistently high discrimination in the predetermined clinical conditions of heart failure, COPD, sepsis, and COVID-19, as well as in surgical and obstetric patients. This illustrates an important strength of developing an all-cause deterioration model because it can enhance early identification and potentially improve outcomes across a wide range of patients as opposed to more narrowly developed models for specific conditions (e.g., sepsis). Clinical performance was confirmed in a prospective study of nearly 50,000 admissions with over four million scores in 10 hospitals. These results, which constitute the largest validation of an early warning score to date, provide confidence that eCART's performance is strong and generalizable. Prospective implementation of eCART would lead to increased detection and decreased false alarm rates compared to MEWS and NEWS, which could improve patient outcomes and decrease alarm fatigue.

The original version of eCART, which our group developed in 2014, included 269,999 admissions from three health systems in Illinois and used discrete-time logistic regression with splines to predict ICU transfer, cardiac arrest, or death (1). The AAM score developed by Kipnis and colleagues used a similar approach to predict unplanned ICU transfer in the Kaiser Permanente Northern California health system (2). Prospective implementation studies of both of these scores found an association with decreased mortality (8, 9). However, our group and others have found that more advanced machine learning methods, such as GBM, can outperform logistic regression for predicting clinical deterioration (13, 14). Further, the inclusion of trends has also been shown to improve model discrimination (20). Therefore, in this new version of eCART, we utilized both GBM and trends to optimize performance. GBM can automatically learn interactions between variables, fit non-linear relationships, and handle missing data, which makes it ideal for developing clinical models in EHR data. The comparatively high performance of the GBM version of eCART is consistent with prior early warning scores developed using machine learning, including the Hospital-wide Alerting Via Electronic Noticeboard (HAVEN), which is a GBM model developed and then validated in four hospitals in the United Kingdom that outperformed other scoring systems based on logistic regression (14). The increased discrimination of these more advanced models allows for the same (or higher) detection rates while limiting false positive alerts that can lead to alarm fatigue.

While previously published studies have demonstrated that more advanced models can outperform standard tools, such as MEWS and NEWS, across the entire medical-surgical cohort (2, 13, 14), little is known regarding comparative performance

across important patient subgroups. Therefore, in this study, we performed extensive subgroup analysis that included age, sex, race, and medical conditions. We found that eCART had high discrimination across all subgroups and consistently outperformed both MEWS and NEWS. The highest results were in the post-partum cohort, followed by age 18-33, Native Hawaiian and Other Pacific Islanders, and patients with COVID-19. Across all 19 tested subgroups, eCART maintained an AUC of ≥0.81. To our knowledge, this is the largest and most comprehensive subgroup analysis performed on early warning scores to date and demonstrates the excellent performance of eCART across these cohorts. This type of analysis is critical to ensure that these scores can be safely used across a wide range of hospitalized patients.

Although numerous predictive models have been developed using retrospective data, few have been implemented prospectively. An important step towards implementation is building the informatics infrastructure to calculate the model prospectively and to assess performance during a silent implementation. Therefore, in addition to performing a large, extensive retrospective validation, we also tested eCART in a silent prospective study in 10 hospitals. We found that our model had similar discrimination to the retrospective evaluation in the same health system and had excellent calibration. These results are encouraging, given that data quality and timing can differ between prospective and retrospective data (e.g., back-dating of vital sign documentation) and increase confidence that eCART will continue to perform well during implementation studies. Furthermore, we demonstrate the feasibility of testing these models in environments for clinical operations and integrated with HL7 for production use.

Our study has several limitations. First, although GBM models are flexible and can be highly accurate, they are complex and difficult to interpret. Therefore, explainable machine learning approaches are needed to provide insights to clinicians regarding the variables that are driving an individual patient's risk of deterioration. In addition, there are myriad other machine learning approaches available to develop prediction models, and numerous possible comparator scores that have been published. We chose GBM due to its excellent discrimination and calibration from prior publications, and NEWS and MEWS due to their widespread use across the country and around the world. Finally, it is also important to note that high model discrimination may not translate to improved patient outcomes, so prospective implementation of eCART is required to study its impact on patient care.

In conclusion, we developed and validated a new GBM model, called eCARTv5, which accurately identifies early clinical deterioration. Our model was validated retrospectively in a geographically diverse set of health systems and performed better than the NEWS and MEWS overall and across a range of subgroups. Prospective validation of eCART found similar performance, and these results served as the foundation for an FDA submission. Future implementation of our score could identify more high-risk patients at a lower false alarm rate than commonly used tools.

References

1. Churpek MM, Yuen TC, Winslow C, Robicsek AA, Meltzer DO, Gibbons RD, Edelson DP. Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014; 190: 649-655.
2. Kipnis P, Turk BJ, Wulf DA, LaGuardia JC, Liu V, Churpek MM, Romero-Brufau S, Escobar GJ. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *Journal of biomedical informatics* 2016; 64: 10-19.
3. Churpek MM, Wendlandt B, Zadravecz FJ, Adhikari R, Winslow C, Edelson DP. Association between intensive care unit transfer delay and hospital mortality: A multicenter investigation. *J Hosp Med* 2016; 11: 757-762.
4. Bellomo R, Ackerman M, Bailey M, Beale R, Clancy G, Danesh V, Hvarfner A, Jimenez E, Konrad D, Lecardo M, Pattee KS, Ritchie J, Sherman K, Tangkau P, Vital Signs to Identify T, Assess Level of Care Study I. A controlled trial of electronic automated advisory vital signs monitoring in general hospital wards. *Crit Care Med* 2012; 40: 2349-2361.
5. Hodgetts TJ, Kenward G, Vlackonikolis I, Payne S, Castle N, Crouch R, Ineson N, Shaikh L. Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resuscitation* 2002; 54: 115-123.
6. 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.
7. Gupta S, Green C, Subramaniam A, Zhen LD, Low E, Tiruvoipati R. The impact of delayed rapid response call activation on patient outcomes. *Journal of critical care* 2017; 41: 86-90.
8. Winslow CJ, Edelson DP, Churpek MM, Taneja M, Shah NS, Datta A, Wang CH, Ravichandran U, McNulty P, Kharasch M, Halasyamani LK. The Impact of a Machine Learning Early Warning Score on Hospital Mortality: A Multicenter Clinical Intervention Trial. *Crit Care Med* 2022; 50: 1339-1347.
9. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *The New England journal of medicine* 2020; 383: 1951-1960.
10. Churpek MM, Yuen TC, Edelson DP. Risk stratification of hospitalized patients on the wards. *Chest* 2013; 143: 1758-1765.
11. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, Collins GS. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020; 369: m1501.
12. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM* 2001; 94: 521-526.
13. 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: 368-374.

14. Pimentel MAF, Redfern OC, Malycha J, Meredith P, Prytherch D, Briggs J, Young JD, Clifton DA, Tarassenko L, Watkinson PJ. Detecting Deteriorating Patients in the Hospital: Development and Validation of a Novel Scoring System. *Am J Respir Crit Care Med* 2021; 204: 44-52.
15. Koyner JL, Carey KA, Edelson DP, Churpek MM. The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. *Crit Care Med* 2018; 46: 1070-1077.
16. Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, Haga A, Shinzawa M, Yamakawa M, Fukui S, Moriyama T. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Sci Rep* 2022; 12: 15889.
17. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. New York, NY: Springer; 2009.
18. Sanchez-Pinto LN, Luo Y, Churpek MM. Big Data and Data Science in Critical Care. *Chest* 2018; 154: 1239-1248.
19. 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: 465-470.
20. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016; 102: 1-5.
21. Willett JB, Singer JD. Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *J Consult Clin Psychol* 1993; 61: 952-965.
22. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. *Int J Med Inform* 2018; 116: 10-17.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44: 837-845.

Table 1. Patient characteristics by cohort.

Patient Characteristics	Derivation Cohort	Retrospective Validation Cohort
Hospitals, N	7	21
Encounters, N	901,491	1,769,461
Admission age, years, median (IQR)	61 (44, 74)	62 (45, 75)
Female sex	512,198 (56.8%)	993,297 (56.1%)
Race: American Indian or Alaska Native	1,445 (0.2%)	6,468 (0.4%)
Race: Asian/Mideast Indian	23,334 (2.6%)	26,681 (1.5%)
Race: Black/African American	286,901 (31.8%)	252,982 (14.3%)
Race: Pacific Islander/ Hawaiian Native	666 (0.1%)	2,496 (0.1%)
Race: White/Caucasian	496,154 (55.0%)	1,384,075 (78.2%)
Race: Other	92,991 (10.3%)	96,759 (5.5%)
Surgical	312,153 (34.6%)	539,875 (30.5%)
Obstetric	65,594 (7.3%)	154,759 (8.7%)
Sepsis	240,651 (26.7%)	639,802 (36.2%)
COVID-19	4,365 (0.5%)	49,834 (2.8%)
Congestive heart failure	131,644 (14.6%)	306,140 (17.3%)
Chronic pulmonary disease	150,043 (16.6%)	443,263 (25.1%)
Length of stay, hours, median (IQR)	70 (37, 124)	72 (43, 130)
Ward to ICU transfer	32,320 (3.6%)	57,789 (3.3%)
Mortality	10,568 (1.2%)	26,319 (1.5%)

Abbreviations: ICU = Intensive Care Unit; COVID-19 = Coronavirus Disease 2019; AIDS/HIV = Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome

Table 2. Area under the receiver operating characteristic curve (AUROC) of the risk scores for predicting the outcome of clinical deterioration within 24 hours in the external retrospective validation cohort.

Cohort	Encounters, n	Observations, n	eCART AUROC (95% CI)	NEWS AUROC (95% CI)	MEWS AUROC (95% CI)
Retrospective (All)	1,769,461	132,873,833	0.835 (0.834, 0.835)	0.766 (0.766, 0.767)	0.704 (0.703, 0.704)
•Retrospective (R1)	246,949	19,262,093	0.862 (0.861, 0.862)	0.775 (0.774, 0.777)	0.730 (0.729, 0.732)
•Retrospective (R2)	592,504	37,930,348	0.872 (0.872, 0.873)	0.808 (0.807, 0.809)	0.749 (0.748, 0.749)
•Retrospective (R3)	930,008	75,681,392	0.807 (0.807, 0.808)	0.744 (0.743, 0.744)	0.674 (0.674, 0.675)
Prospective (P3)	46,330	4,778,200	0.800 (0.798, 0.801)	0.736 (0.734, 0.738)	0.681 (0.678, 0.683)

Abbreviations: eCART = electronic Cardiac Arrest Risk Triage score; NEWS = National Early Warning Score; MEWS = Modified Early Warning Score

Table 3. Subgroup analysis results showing the area under the receiver operating characteristic curve (AUROC) values for predicting clinical deterioration in the full retrospective cohort by risk score and subgroup.

Category	Subgroup	Encounters, n	eCART AUROC	NEWS AUROC	MEWS AUROC
All	–	1,769,461	0.835	0.766	0.704
Age	18-33	232,353	0.861	0.770	0.726
	34-48	271,904	0.845	0.758	0.709
	49-64	475,033	0.827	0.755	0.702
	65-78	458,470	0.822	0.758	0.700
	≥79	331,701	0.832	0.776	0.716
Sex	Male	776,164	0.824	0.761	0.699
	Female	993,297	0.844	0.775	0.710
Race	American Indian or Alaska Native	6,468	0.814	0.746	0.672
	Asian/Mideast Indian	26,681	0.847	0.779	0.722
	Black/African-American	252,982	0.831	0.769	0.707
	Native Hawaiian/Other Pacific Islander	2,496	0.862	0.772	0.709
	White	1,384,075	0.834	0.764	0.702
	Other/Unknown	96,759	0.859	0.785	0.724
Procedure	Surgical	539,875	0.817	0.745	0.690
	Obstetric	154,759	0.909	0.758	0.691
Clinical condition	Sepsis	639,802	0.836	0.774	0.714
	COVID-19	49,834	0.858	0.793	0.710
	CHF	306,140	0.810	0.750	0.694
	COPD	443,263	0.824	0.757	0.698

Abbreviations: eCART = electronic Cardiac Arrest Risk Triage score; NEWS = National Early Warning Score; MEWS = Modified Early Warning Score; COVID-19 = Coronavirus Disease 2019; CHF = Congestive Heart Failure; COPD = Chronic Obstructive Pulmonary Disease

Figure Legends:

Figure 1. Variable importance plot illustrating the top 25 most important variables in the eCART model.

[Figure 1]

Abbreviations: eCART = electronic Cardiac Arrest Risk Triage score; FiO2 = Fraction of Inspired Oxygen; BUN = Blood Urea Nitrogen; aPTT = Activated Partial Thromboplastin Clotting Time; AVPU = Alert, responds to Voice, responds to Pain, Unresponsive

Figure 2. Precision-recall curves of the risk scores in the full retrospective dataset (n=132,873,833 observations). Sensitivity is plotted along the X-axis and positive predictive value is plotted along the Y-axis for eCART, NEWS and MEWS. The markers on the lines correspond to a MEWS of 3 and 4, NEWS of 5 and 7 and eCART of 93 and 97, representing commonly used moderate (higher sensitivity) and high-risk (higher PPV) thresholds for each score.

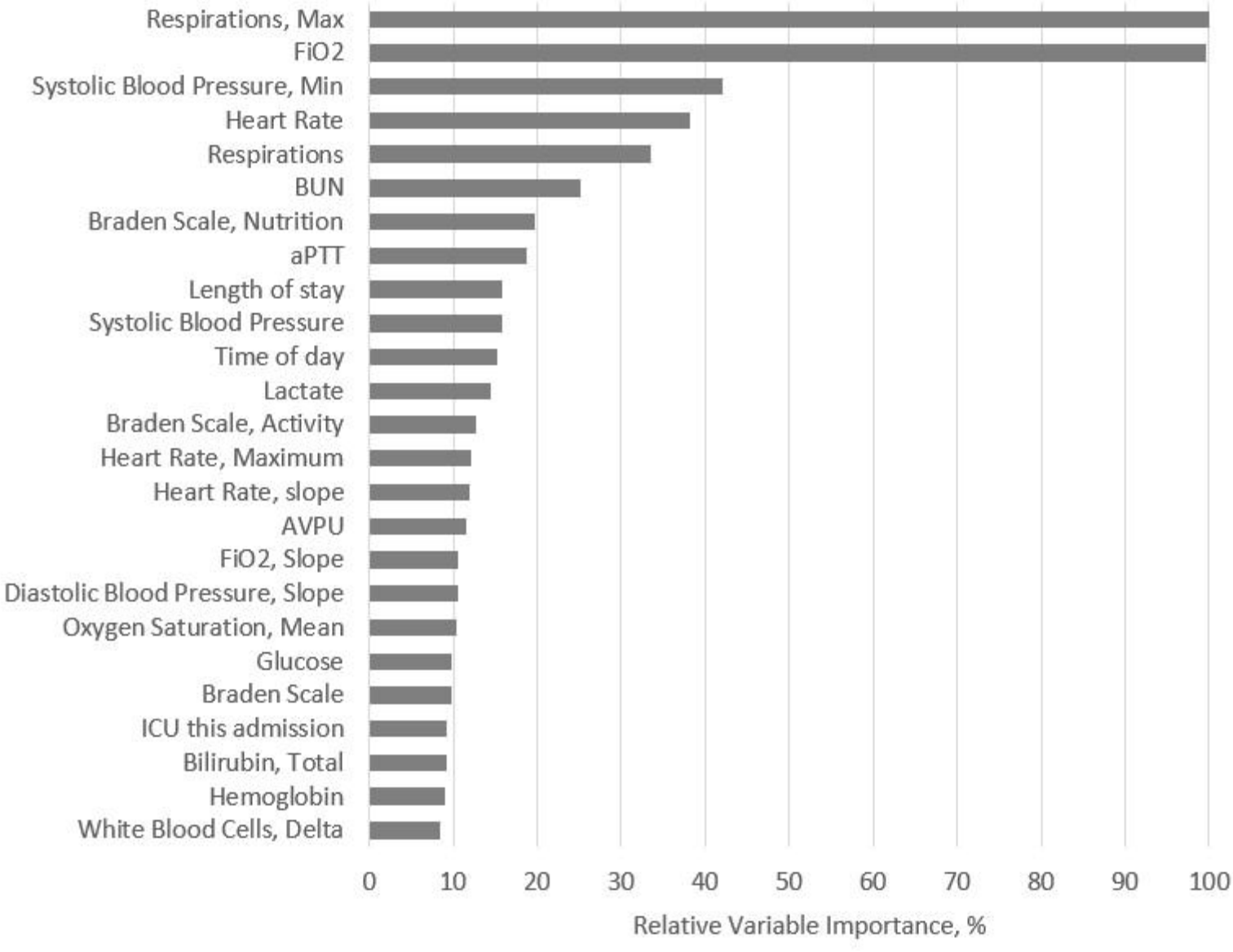
[Figure 2]

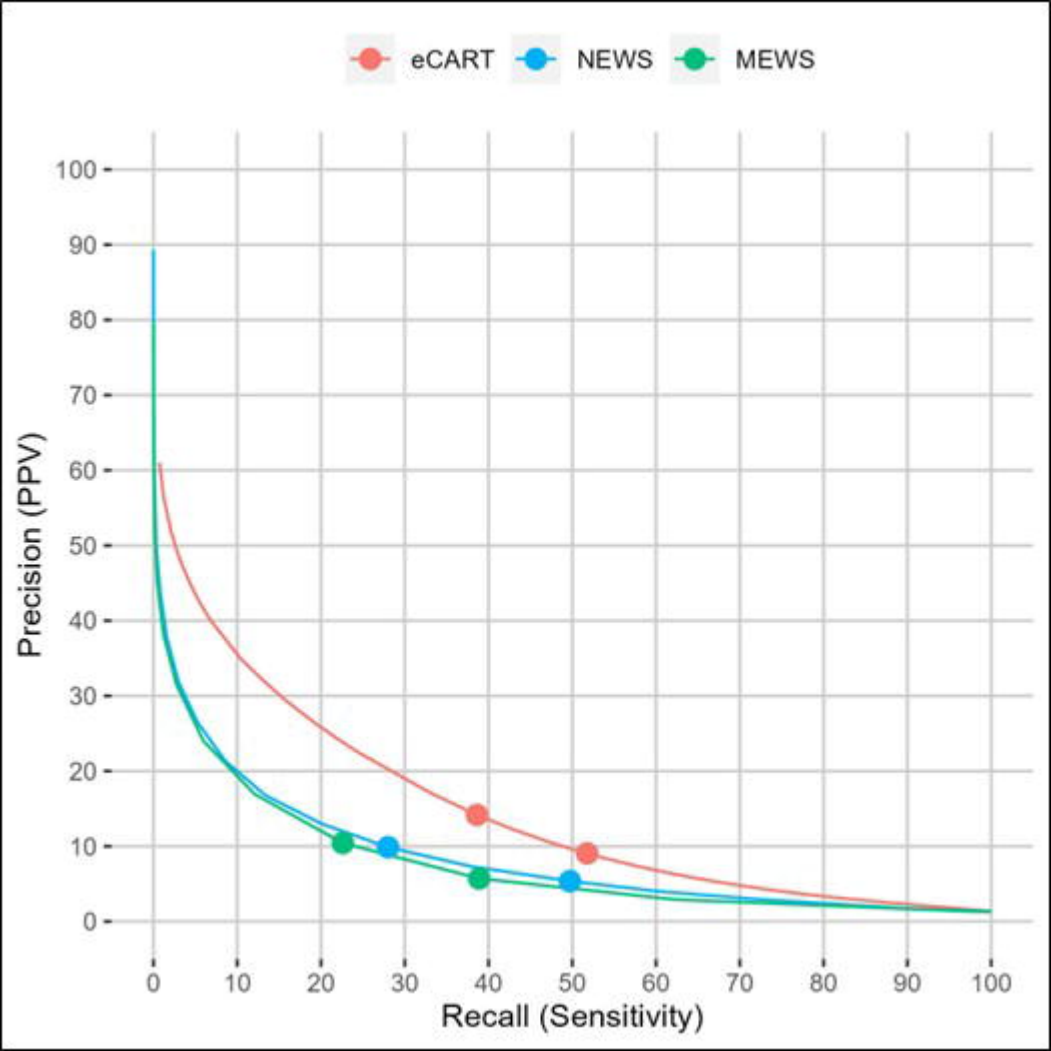
Abbreviations: eCART = electronic Cardiac Arrest Risk Triage score; NEWS = National Early Warning Score; MEWS = Modified Early Warning Score; PPV = Positive Predictive Value

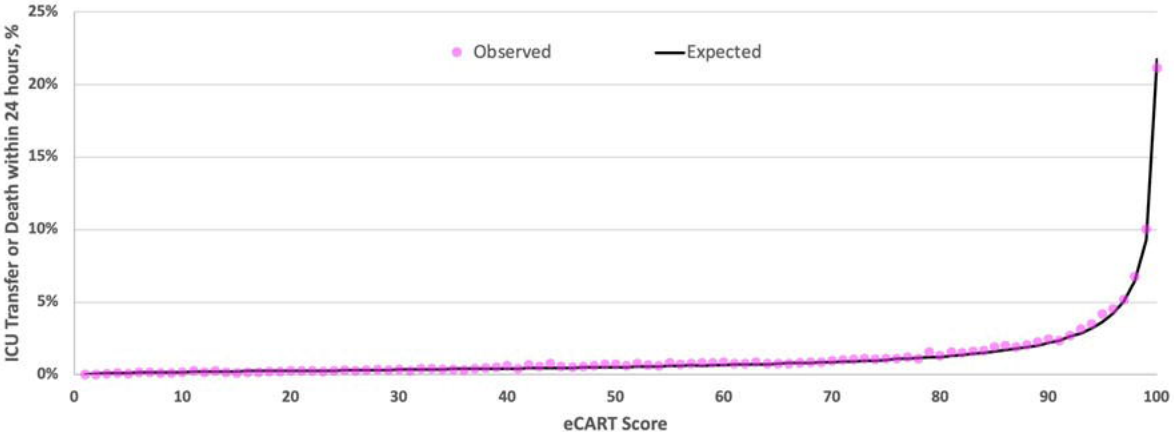
Figure 3. Calibration plot in the prospective validation cohort illustrating the observed and expected outcome rates across values of the eCART score.

[Figure 3]

Abbreviations: eCART = electronic Cardiac Arrest Risk Triage score; ICU = Intensive Care Unit


Respirations, Max
FiO2
Systolic Blood Pressure, Min
Heart Rate
Respirations
BUN
Braden Scale, Nutrition
aPTT
Length of stay
Systolic Blood Pressure
Time of day
Lactate
Braden Scale, Activity
Heart Rate, Maximum
Heart Rate, slope
AVPU
FiO2, Slope
Diastolic Blood Pressure, Slope
Oxygen Saturation, Mean
Glucose
Braden Scale
ICU this admission
Bilirubin, Total
Hemoglobin
White Blood Cells, Delta
0
10
20
30
40
50
60
70
80
90
100
Relative Variable Importance, %


eCART
NEWS
MEWS
Precision (PPV)
Recall (Sensitivity)
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100


ICU Transfer or Death within 24 hours, %
25%
20%
15%
10%
5%
0%
Observed
Expected
0
10
20
30
40
50
60
70
80
90
100
eCART Score