Title

Development and Validation of a Deep Learning Model for Prediction of Adult Physiological Deterioration

Permalink

https://escholarship.org/uc/item/3hs1x2dr

Journal

Critical Care Explorations, 6(9)

ISSN

2639-8028

Authors

Shashikumar, Supreeth P
Le, Joshua Pei
Yung, Nathan
et al.

Publication Date

2024

DOI

10.1097/cce.0000000000001151

Peer reviewed

OPEN

Development and Validation of a Deep Learning Model for Prediction of Adult Physiological Deterioration

Supreeth P. Shashikumar, PhD[1]
Joshua Pei Le, BS[2]
Nathan Yung, MD[3]
James Ford, MD[4]
Karandeep Singh, MD[1,5]
Atul Malhotra, MD[6]
Shamim Nemati, PhD[1,7]
Gabriel Wardi, MD, MPH[6,7]

BACKGROUND: Prediction-based strategies for physiologic deterioration offer the potential for earlier clinical interventions that improve patient outcomes. Current strategies are limited because they operate on inconsistent definitions of deterioration, attempt to dichotomize a dynamic and progressive phenomenon, and offer poor performance.

OBJECTIVE: Can a deep learning deterioration prediction model (Deep Learning Enhanced Triage and Emergency Response for Inpatient Optimization [DETERIO]) based on a consensus definition of deterioration (the Adult Inpatient Decompensation Event [AIDE] criteria) and that approaches deterioration as a state "value-estimation" problem outperform a commercially available deterioration score?

DERIVATION COHORT: The derivation cohort contained retrospective patient data collected from both inpatient services (inpatient) and emergency departments (EDs) of two hospitals within the University of California San Diego Health System. There were 330,729 total patients; 71,735 were inpatient and 258,994 were ED. Of these data, 20% were randomly sampled as a retrospective "testing set."

VALIDATION COHORT: The validation cohort contained temporal patient data. There were 65,898 total patients; 13,750 were inpatient and 52,148 were ED.

PREDICTION MODEL: DETERIO was developed and validated on these data, using the AIDE criteria to generate a composite score. DETERIO's architecture builds upon previous work. DETERIO's prediction performance up to 12 hours before T0 was compared against Epic Deterioration Index (EDI).

RESULTS: In the retrospective testing set, DETERIO's area under the receiver operating characteristic curve (AUC) was 0.797 and 0.874 for inpatient and ED subsets, respectively. In the temporal validation cohort, the corresponding AUC were 0.775 and 0.856, respectively. DETERIO outperformed EDI in the inpatient validation cohort (AUC, 0.775 vs. 0.721; $p < 0.01$) while maintaining superior sensitivity and a comparable rate of false alarms (sensitivity, 45.50% vs. 30.00%; positive predictive value, 20.50% vs. 16.11%).

CONCLUSIONS: DETERIO demonstrates promise in the viability of a state value-estimation approach for predicting adult physiologic deterioration. It may outperform EDI while offering additional clinical utility in triage and clinician interaction with prediction confidence and explanations. Additional studies are needed to assess generalizability and real-world clinical impact.

KEYWORDS: critical care; deep learning; emergency medicine; machine learning; prediction algorithms

Physiologic deterioration is a dynamic phenomenon that is associated with substantial morbidity and mortality, and several interventions have sought to improve patient outcomes through early identification and treatment (1–8). The rapid response team (RRT) is an example of one such intervention that has

DOI: 10.1097/CCE.0000000000001151

KEY POINTS

Question: Can a Deep Learning Enhanced Triage and Emergency Response for Inpatient Optimization (DETERIO) model, which implements a consensus definition of deterioration (the Adult Inpatient Decompensation Event criteria) and that approaches deterioration as a state "value-estimation" problem outperform a commercially available deterioration score?

Findings: DETERIO outperformed Epic Deterioration Index in the inpatient validation cohort (area under the receiver operating characteristic curve, 0.775 vs. 0.721; $p < 0.01$) while maintaining superior sensitivity and a comparable rate of false alarms (sensitivity, 45.50% vs. 30.00%; positive predictive value, 20.50% vs. 16.11%).

Meanings: DETERIO demonstrates promise in the viability of a state value-estimation approach for adult physiologic deterioration prediction models and may outperform commercially available models.

shown promise in significantly reducing in-hospital mortality after patients suffer acute decompensation or deterioration episodes (9–14). However, the RRT represents a reactive, rather than proactive, intervention. More recent strategies, such as the Early Warning Score (15, 16) and machine learning (ML)-based prediction models (17, 18), have thus aimed to predict deterioration.

Early prediction of deterioration gives the clinical team time to intervene before the patient suffers a poor outcome. ML-based solutions are particularly suited to this task due to their ability to leverage the data-rich environment of hospital settings (19). Lilly et al (20) demonstrated this exact concept for hemodynamic instability and respiratory failure. Similarly, Escobar et al (21) showed how a remotely monitored deterioration risk score that triggered a structured patient workup significantly decreased patient mortality, length of hospital stay, and need for ICU transfer. However, such studies have relied on different outcomes to determine deterioration, including in-hospital mortality and ICU transfer (17, 18, 21, 22). While these outcomes relate to deterioration, they are prone to institution-dependent subjectivity, which consequently limits study power, model generalizability, and performance comparisons (23, 24). This is especially the case when applied as the primary outcome to defining deterioration. To our knowledge, a recently developed consensus definition, the Adult Inpatient Decompensation Event (AIDE) criteria (24), has not yet been used as the primary outcome in such a prediction model.

In this study, we propose a deep learning deterioration prediction model Deep Learning Enhanced Triage and Emergency Response for Inpatient Optimization (DETERIO) that is built on a state value-estimation approach and implements a consensus definition of deterioration. Rather than directly predicting deterioration as a single event in time, temporal difference (TD) learning is used to predict the value of each state in the temporal trajectory of a patient, where the value of each state is determined in terms of improvements or deteriorations that the patient may experience in the near future (25). For example, a hypotensive event in certain patient scenarios may be inconsequential (e.g., small drop in blood pressure during sleep) but may represent an early manifestation of circulatory shock in other instances (26, 27). By incentivizing the algorithm to predict the patient trajectory (e.g., hypotension leading to circulatory shock), the model will be able to better capture the relationship between patient phenotypes and long-term outcomes. Using TD learning to predict the future events of shock and ICU admission "in addition to" the immediate event of hypotension, the algorithm can account for the variability and complexity of patient responses. This approach can lead to more personalized and effective interventions and treatments. In terms of physiologic deterioration, temporal integration of a series of adverse events enables DETERIO to learn the value of each patient state (or a "deterioration score") via back-propagation in time. We generate this score from the consensus AIDE criteria, placing decreased emphasis on in-hospital mortality and ICU transfer, which have been argued to be workflow-dependent and noisy surrogates for deterioration (24). We hypothesize that our approach centered around the concept of value-estimation would have good predictive ability and outperform a commonly used commercially available deterioration score. We also seek to compare our novel predictive model to a commercially available deterioration index available in Epic.

MATERIALS AND METHODS

Study Design and Patient Cohorts

We conducted a retrospective cohort study using de-identified electronic health record data of all adult patients (≥ 18 yr) who were admitted to an inpatient service or presented to the emergency department (ED) between January 1, 2016, and October 31, 2022, at two hospitals within the University of California San Diego Health System. This study was completed in accordance with Strengthening the Reporting of Observational Studies in Epidemiology guidelines and other relevant guidance (28, 29) (**Appendix STROBE**, http://links.lww.com/CCX/B397), the ethical standards of the University of California San Diego on human experimentation, and the Helsinki Declaration of 1975. Institutional review board (IRB) approved protocol No. 201476 with waiver of consent ("Enhanced Metadata Design, Architecture, and Learning [MeDAL] for Development of Generalizable Deep Learning-Based Predictive Analytics from Electronic Health Records") was initially approved on August 13, 2020, with a latest approval date of February 14, 2024.

We included all adult patients (age 18 yr old or older) who presented to our EDs or were admitted directly to our inpatient services. Patients were excluded if: 1) their care unit length of stay was less than 2 hours, 2) physiologic deterioration occurred before hour 2 of care unit admission, 3) there was no measurement of heart rate or blood pressure or laboratories before the prediction start time, or 4) they were receiving comfort measures only. Patients in a procedure suite (e.g., catheterization laboratory or perioperative area) or obstetrics units were also excluded (**Appendix A**, http://links.lww.com/CCX/B397, details these exact services). For prediction purposes, patients were followed throughout their stay until either: 1) the time of their first episode of physiologic deterioration or 2) the time of transfer out of a given care unit. To allow for adequate data collection, prediction began 2 hours after the start of an inpatient service or ED stay. If a patient initially presented to the ED and was later admitted to an inpatient service, they were included in both cohorts with separate predictions occurring 2 hours after the start of each unit stay. These predictions were made and updated at every hour based on the newest clinical data.

The overall dataset was split up as follows: 1) a development cohort consisting of encounters with hospital admission dates between January 1, 2016, and September 30, 2021 and 2) a temporal validation cohort consisting of encounters with hospital admission dates between October 1, 2021, and October 31, 2022 (**Fig. 1**). The development cohort was further randomized into a training cohort (consisting of 80% encounters from the development cohort) and testing cohort (consisting of the remaining 20% encounters). The development cohort was used for model training and internal testing while the temporal validation cohort was solely used for model testing purposes.

In this study, we used a consensus definition for physiologic deterioration (AIDE) (24). To define the

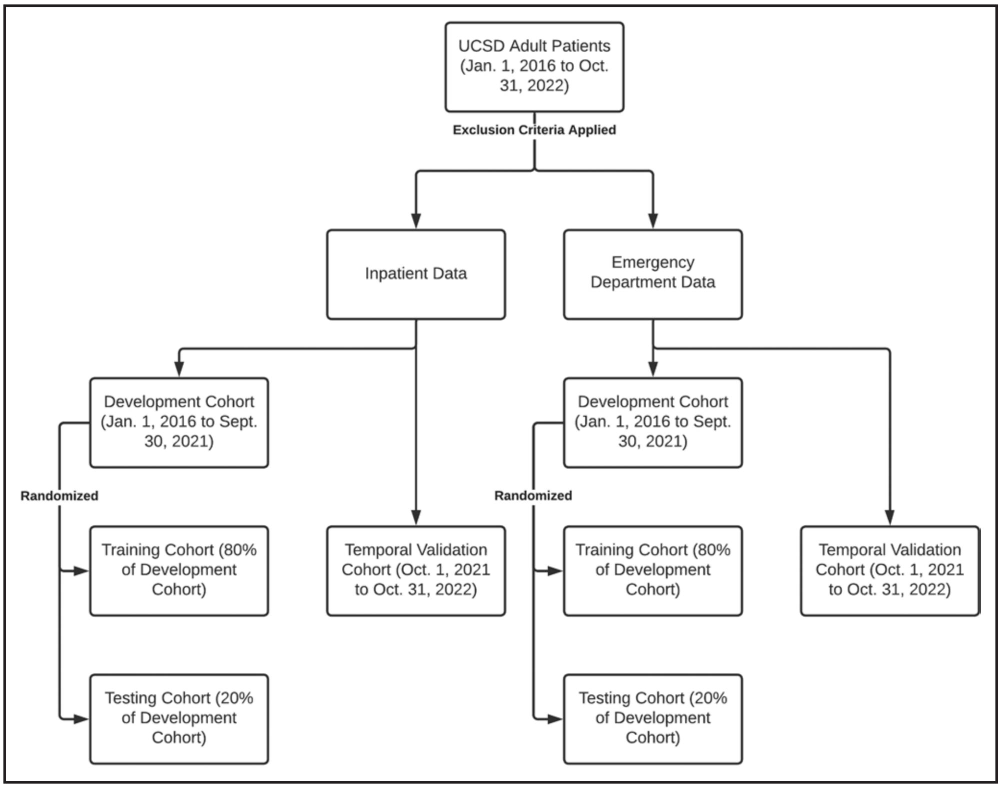
UCSD Adult Patients (Jan. 1, 2016 to Oct. 31, 2022)
Exclusion Criteria Applied
Inpatient Data
Emergency Department Data
Development Cohort (Jan. 1, 2016 to Sept. 30, 2021)
Randomized
Training Cohort (80% of Development Cohort)
Testing Cohort (20% of Development Cohort)
Temporal Validation Cohort (Oct. 1, 2021 to Oct. 31, 2022)
Development Cohort (Jan. 1, 2016 to Sept. 30, 2021)
Randomized
Training Cohort (80% of Development Cohort)
Testing Cohort (20% of Development Cohort)
Temporal Validation Cohort (Oct. 1, 2021 to Oct. 31, 2022)

Figure 1. Formation of development and temporal validation cohorts for inpatient and emergency department (ED) data. This flowchart represents how the aggregate data were allocated to form development and temporal validation cohorts for the inpatient and ED data. These data were collected from two hospitals within the University of California San Diego (UCSD) Health System.

long-term value of a given patient state, we defined a composite deterioration score by summing over the points associated with four temporally separate adverse events: need for vasopressors or inotropes (three points), severe hypoxemia/invasive mechanical ventilation (three points), ICU transfer (one point), and in-hospital mortality (one point) (**Appendix B**, http://links.lww.com/CCX/B397). Importantly, the cardiovascular and respiratory adverse events were derived directly from the AIDE criteria, and in-hospital mortality and ICU transfer were only used as labels to differentiate between acute and chronic events during model training; they were "not" involved in temporal prediction. Each component had a binary score, but the overall composite deterioration score (or the patient "state value") could take on values in the range 0–8. For model evaluation purposes, a composite deterioration score of greater than or equal to 3 was defined as the positive class, and a composite deterioration score of less than 3 was defined as the control class. As a result, a positive class could only be triggered if at least one AIDE criterion was fulfilled. Crucially, this approach does not characterize a modification of the AIDE criteria because we selected a binary classification cutoff for model evaluation here. The implementation of a risk model allows for the generation of actionable steps in an otherwise negative class (e.g., score of 2), such as the activation of smart order sets or notification of "control tower" team members who can provide a second set of eyes on the patient. Characteristics of patients who fulfilled at least one AIDE criterion have been tabulated in **Appendix C** (http://links.lww.com/CCX/B397).

Model Features

Data from both sites were automatically abstracted into a clinical data repository (Epic Clarity; Epic Systems, Verona, WI). In addition to using the clinical variables from a previously published predictive model for sepsis (30), we included additional features into DETERIO (129 total; **Appendix D**, http://links.lww.com/CCX/B397). Variables consisted of 50 vital signs and laboratory measurements, six demographic features, ten medication features, and 62 comorbidities. Vital signs and laboratory variables were organized into 1 hour nonoverlapping time series bins to accommodate for different sampling frequencies of available data. All the variables with sampling frequencies greater than once per hour were uniformly resampled into 1 hour time bins by calculating the median values. Variables were updated hourly when new data became available; otherwise, the old values were carried forward (sample-and-hold interpolation) for 24 hours. Mean imputation was used to replace all remaining missing values, mainly at the start of each record. We report missing data on an hourly (not encounter) basis in relevant tables (**Appendix E**, http://links.lww.com/CCX/B397). In addition to the 129 clinical variables, we calculated 100 features derived from the 50 vital signs and laboratory measurements. Specifically, for each vital sign and laboratory measurement, local trends (slope of change), baseline values (mean value measured over the previous 72 hr), and the time since the variable was last measured were derived. The Epic Deterioration Index (EDI) was available only for admitted patients, and we did not impute values. At our institution, the EDI is updated every 20 minutes.

Model Development, Evaluation, and Statistical Analyses

The model architecture of DETERIO was similar to a previously published model for early prediction of sepsis called COnformal Multidimensional Prediction Of SEpsis Risk (COMPOSER) (30). DETERIO was a three-layer feedforward neural network of size 100, 80, and 64. DETERIO was trained using a TD-learning approach that leveraged the value iteration algorithm (25) to predict the future value of a patient state, starting from hour 2 of admission up to the time T0 (physiologic deterioration episode or transfer out of a given care unit). The patient state in this setting was a 64 dimensional vector, which was mapped to a single state value using a fully connected neural network layer. Model performance was evaluated by thresholding on the predicted state value. For additional details, see **Appendix G** (http://links.lww.com/CCX/B397). In the case of control patients, the model was trained to predict up to the end-of-stay in a care unit or 14 days, whichever occurred first. Similar to COMPOSER, DETERIO had a conformal prediction module whose functionality was to detect out-of-distribution samples and thus defined the "conditions for use" of the model (23, 30). DETERIO was developed using TensorFlow, Version 2.14 (TensorFlow, Mountain View, CA). The parameters of DETERIO were initialized randomly and optimized using the training dataset from the development cohort with L1–L2 regularization and dropout to avoid overfitting. Additionally, DETERIO was made interpretable by calculating the relevance score of each input variable for every predicted risk score (**Appendix F**, http://links.lww.com/CCX/B397).

The decision threshold was chosen corresponding to 50% sensitivity at the encounter level. A predicted risk score beyond this threshold meant that DETERIO predicted that the patient would undergo physiologic deterioration within the prediction window (up to 12 hr before T0). A predicted risk score less than the decision threshold meant that DETERIO did not predict physiologic deterioration within the prediction window. Note that prior studies have reported sensitivities in the range 20–50% for prediction of clinical deterioration (31, 32).

For all continuous variables, we have reported the median and interquartile range. For binary variables, we have reported percentages. The area under the receiver operating characteristic curve (AUC) has been reported at the hourly window level. Specificity, sensitivity, and positive predictive value (PPV) at a fixed decision threshold have been reported at the encounter level. The four atomic elements—number of true positives, false positives, true negatives, and false negatives—required to compute specificity, sensitivity, and PPV at the encounter level have been described in **Appendix H** (http://links.lww.com/CCX/B397). We have additionally reported the number of false alarms per patient hour (FAPH), which can be used to calculate the expected number of false alarms per unit of time in a typical care unit. FAPH was calculated by dividing the total number of false alarms by the total number of data points (the sum of hourly time points across all patients) in a given cohort. Comparisons between models were achieved with the DeLong test at α = 0.01. AUC was calculated under an end-user clinical response policy wherein the model was silenced for 6 hours after an alarm was fired.

RESULTS

Patient Characteristics

After applying the exclusion criteria, a total of 71,735 (258,994) and 13,750 (52,148) inpatient (ED) encounters in the development and validation cohorts were included, respectively. Patient characteristics of the development and temporal validation cohorts of inpatient and ED data have been tabulated in **Tables 1** and **2**. Additionally, comparison of patient characteristics

TABLE 1.
Demographics of Inpatient Data

Demographic	Development Data	Temporal Validation Data
Number of encounters	71,735	13,750
Age (yr), median (IQR)	59.39 (46.24–70.44)	61.18 (46.86–72.13)
Male gender (%)	55.32% (*n* = 39,684)	54.98% (*n* = 7,560)
White (%)	52.63% (*n* = 37,754)	49.90% (*n* = 6,861)
African American (%)	10.91% (*n* = 7,826)	9.90% (*n* = 1,361)
Asian (%)	6.22% (*n* = 4,462)	6.34% (*n* = 872)
Stay (hr), median (IQR)	91.96 (52.28–162.47)	106.64 (68.18–189.10)
Charlson Comorbidity Index, median (IQR)	1 (0–3)	1 (0–3)
Sequential Organ Failure Assessment score, median (IQR)	1 (1–2)	1 (1–3)
AIDE cardiovascular–pressors/inotropes (%)	1.06% (*n* = 760)	0.65% (*n* = 89)
AIDE respiratory (%)	4.54% (*n* = 3,257)	4.81% (*n* = 661)
Mortality (%)	1.21% (*n* = 868)	1.29% (*n* = 177)
Transfer to ICU (%)	5.46% (*n* = 3,917)	4.86% (*n* = 668)

AIDE = Adult Inpatient Decompensation Event, IQR = interquartile range.

TABLE 2.
Demographics of Emergency Department Data

Demographic	Development Data	Temporal Validation Data
Number of encounters	258,994	52,148
Age (yr), median (IQR)	54.5 (38.50–66.40)	56.14 (38.94–68.65)
Male gender (%)	52.62% (*n* = 136,283)	51.82% (*n* = 27,023)
White (%)	52.08% (*n* = 134,884)	49.37% (*n* = 25,745)
African American (%)	12.30% (*n* = 31,856)	11.00% (*n* = 5,736)
Asian (%)	5.50% (*n* = 14,245)	5.76% (*n* = 3,004)
Stay (hr), median (IQR)	8.11 (4.81–49.90)	9.41 (5.15–68.86)
Charlson Comorbidity Index, median (IQR)	0 (0–2)	0 (0–2)
Sequential Organ Failure Assessment score, median (IQR)	0 (0–1)	0 (0–1)
AIDE cardiovascular–pressors/inotropes (%)	0.76% (*n* = 1,968)	0.42% (*n* = 219)
AIDE respiratory (%)	2.53% (*n* = 6,553)	3.13% (*n* = 1,632)
Mortality (%)	0.66% (*n* = 1,709)	0.71% (*n* = 370)
Transfer to ICU (%)	5.17% (*n* = 13,390)	5.37% (*n* = 2,800)

AIDE = Adult Inpatient Decompensation Event, IQR = interquartile range.

between patients with composite deterioration score greater than or equal to 3 (positive class) and patients with composite deterioration score less than 3 (control class) are tabulated in Appendix C (http://links.lww.com/CCX/B397).

Development and Temporal Validation Cohorts

DETERIO's performance (AUC/PPV/sensitivity/specificity/FAPH) on inpatient and ED test cohorts were 0.797/22.20%/46.40%/92.50%/0.00337 and 0.874/24.30%/49.60%/96.60%/0.00529, respectively. This compares to 0.775/20.50%/45.50%/92.01%/0.00358 and 0.856/20.80%/46.30%/95.90%/0.00458 of inpatient and ED temporal validation cohorts, respectively (**Table 3**). The AUCs for testing and temporal validation cohorts are shown in **Figure 2**. Heat maps of the top 15 clinical variables contributing to the increase in risk score for physiologic deterioration up to 12 hours before T0 have been shown in **Figure 3**. It was also observed that the samples rejected by the conformal prediction module had higher data missingness compared with the samples accepted by the conformal prediction module (**Appendix J**, http://links.lww.com/CCX/B397).

When compared with the commercially available EDI, DETERIO exhibited significantly better performance (AUC 0.775 vs. 0.721; $p < 0.01$) on the temporal validation inpatient cohort. Additionally, DETERIO achieved higher sensitivity and PPV in comparison to EDI at a comparable rate of false positive alarms (sensitivity 45.50% vs. 30.00%, PPV 20.50% vs. 16.11%, specificity 92.01% vs. 92.93%, FAPH 0.00458 vs. 0.00399) (Table 3). A similar comparison with the ED temporal validation cohort was not possible because EDI is only available in the inpatient setting.

DISCUSSION

In this study, we developed and validated DETERIO, a deep learning-based prediction model for adult physiologic deterioration using a large cohort of adult patients at several hospitals within the University of California San Diego Health system. Our model builds upon previous work that defines model-specific "conditions for use" using conformal prediction and identifies the most important available clinical variables for deterioration (23, 30). Additionally, it implements a consensus definition of physiologic deterioration (the AIDE criteria [24]) to create a composite deterioration score for prediction. By modeling the long-term value of a patient state, the proposed model weighs the contribution of various adverse events to predict the likelihood of patient deterioration. It demonstrated significantly better performance on the temporal inpatient data when

TABLE 3.
Model Performance Summary on Inpatient and Emergency Department Cohorts

Cohort Type	Area Under the Receiver Operating Characteristic Curve[a]	Sensitivity[b]	Specificity[b]	Positive Predictive Value[b]
Inpatient cohorts				
DETERIO: Development cohort (training)	0.823	52.10%	91.90%	23.40%
DETERIO: Development cohort (test)	0.797	46.40%	92.50%	22.20%
DETERIO: Temporal validation cohort	0.775	45.50%	92.01%	20.50%
Epic Deterioration Index: Temporal validation cohort	0.721	30.00%	92.93%	16.11%
Emergency department cohorts				
DETERIO: Development cohort (training)	0.887	50.60%	96.60%	24.10%
DETERIO: Development cohort (test)	0.874	49.60%	96.60%	24.30%
DETERIO: Temporal validation cohort	0.856	46.30%	95.90%	20.80%

DETERIO = Deep Learning Enhanced Triage and Emergency Response for Inpatient Optimization.

[a] Hourly window wise.

[b] Encounter wise.

Decision threshold corresponding to 50% sensitivity threshold on the development cohort training set (3.5 and 3.0 for inpatient and emergency department cohorts, respectively).

Decision threshold for Epic Deterioration Index (60) was chosen based on thresholds used in the literature (32).

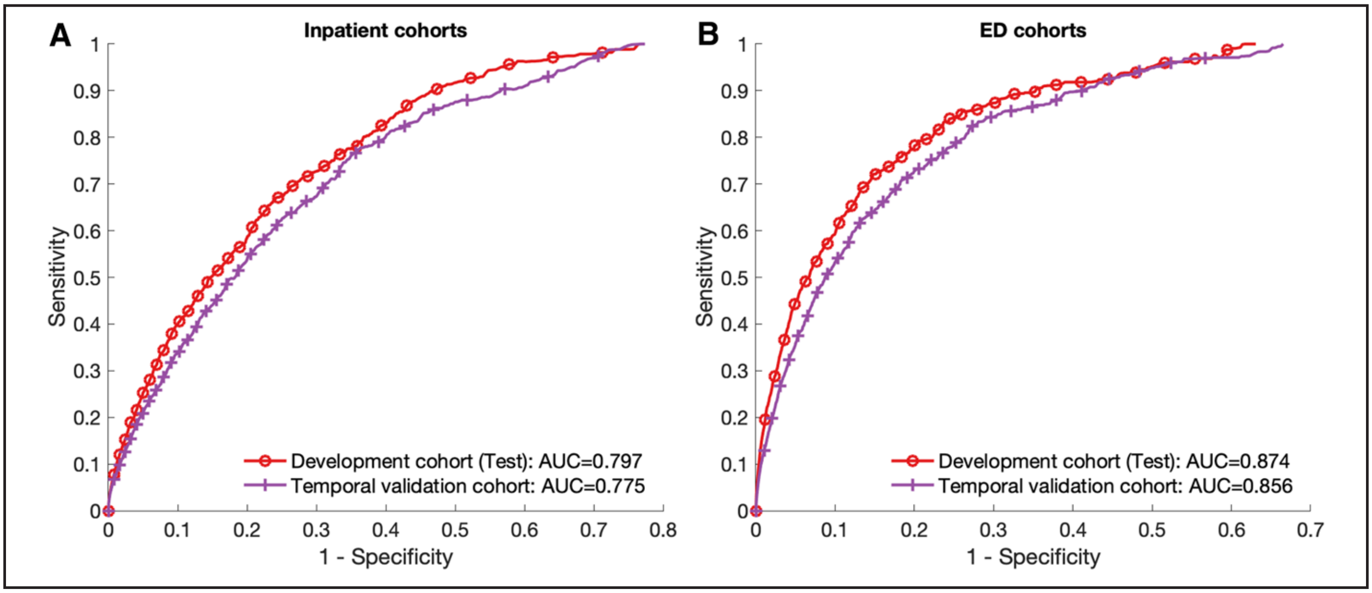

Figure 2. Receiver operating characteristic curves for inpatient and emergency department (ED) cohorts. The *X*-axis represents 1− specificity or the false positive rate. The *Y*-axis represents sensitivity or the true positive rate. Receiver operating characteristic curves were plotted for the development cohort and temporal validation cohort under inpatient (**A**) and ED (**B**) data. Area under the receiver operating characteristic curve (AUC) represents the response of the true positive rate to the false positive rate as the decision threshold is decreased.

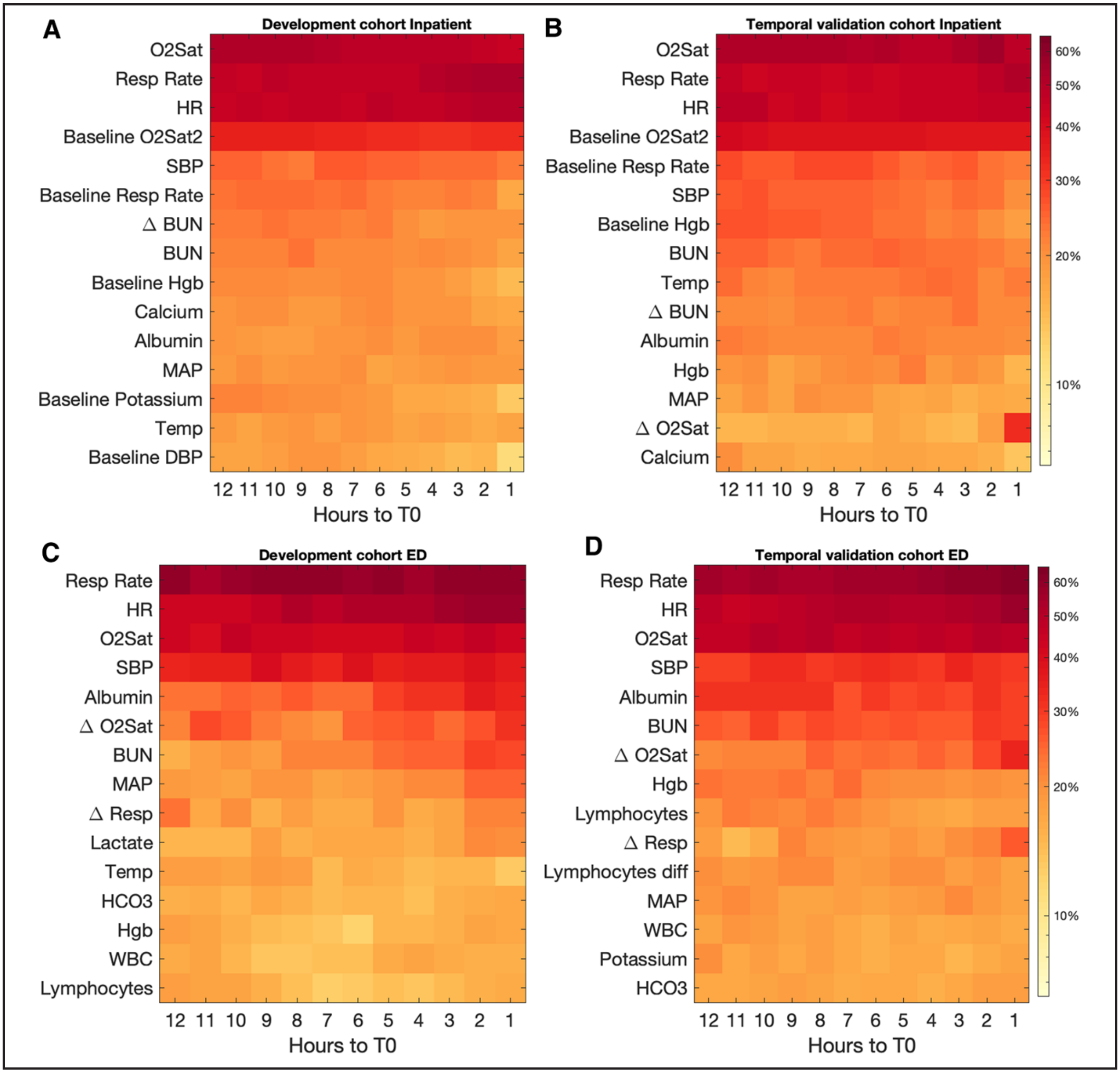

Figure 3. Population-level plot of top contributing factors to the increase in model risk score. The *X*-axis represents hours before T0. The *Y*-axis represents the top factors (sorted by the magnitude of relevance score) across the patient populations at the development cohort—inpatient (**A**), temporal validation cohort—inpatient (**B**), development cohort—emergency department (ED) (**C**), and temporal validation cohort—ED (**D**). Only dynamically changing variables are shown. The *heat map* shows the percentage of patients for whom a given variable was an important contributor to their risk score, up to 12 hr before T0 (Appendix F, http://links.lww.com/CCX/B397). Albumin = serum albumin, Baseline DBP = mean diastolic blood pressure over past 72 hr, Baseline Hgb = mean hemoglobin over past 72 hr, Baseline o_2Sat = mean oxygen saturation over past 72 hr, Baseline Potassium = mean serum potassium over past 72 hr, Baseline Resp Rate = mean respiratory rate over past 72 hr, BUN = blood urea nitrogen, Calcium = serum calcium, Hco_3 = serum bicarbonate, Hgb = serum hemoglobin, HR = heart rate, Lactate = serum lactate, Lymphocytes = serum lymphocytes, Lymphocytes Diff = lymphocyte differential, MAP = mean arterial pressure, o_2Sat = oxygen saturation, Resp Rate = respiratory rate, SBP = systolic blood pressure, Temp = body temperature, ΔBUN = change in blood urea nitrogen since last measurement, Δo_2Sat = change in oxygen saturation since last measurement, ΔResp = change in respiratory rate.

compared with the commercially available EDI (AUC 0.775 vs. 0.721; $p < 0.01$). Furthermore, DETERIO outperformed EDI at correctly identifying patients at-risk for physiologic deterioration while maintaining a comparable rate of false alarms.

Although the commercially available EDI demonstrated relatively acceptable performance at the encounter level in this study (AUC, 0.721), its generalizability remains questionable. A recent prospective validation of the same algorithm on similar inpatient hospital services across the Midwestern United States found an encounter-level AUC of 0.685 (95% CI, 0.671–0.700) (33). This shift in performance can be attributed to several factors: differences in definitions of deterioration, patient characteristics, and institution-dependent practices. Crucially, such factors are not unique to this direct comparison and remain barriers to generalizability at any institution (23). Along similar lines, a rigorous analysis of EDI using regression discontinuity design found that while an EDI-triggered response was associated with a statistically significant decrease in patient care escalations, there was no impact on patient mortality (34). It is important to note that this result was generated in Northern California and may not apply elsewhere due to the previously mentioned generalizability constraints. Furthermore, to our knowledge, Epic Systems has not discussed a systematic approach to model performance monitoring, for instance via periodic temporal validation (35). These reasons should thus highlight the need to test the performance of the EDI at individual health systems before adoption, and periodically thereafter, as the model is potentially subject to performance degradation.

When compared with EDI, DETERIO maintains key advantages and clinical potential outside of pure model performance. In addition to the advantages of utilizing a consensus definition, DETERIO's novel composite score allows it to reflect more accurately the dynamic and complex pathophysiology of deterioration. Prediction of the long-term value of a patient state allows for taking into account various downstream adverse events, which can be interpreted as a form of "label confidence" (36, 37). This approach is particularly useful when a given single event (e.g., activation of code blue) may not reflect the overall prognosis or outcome of the patient. The resulting predicted risk score enables the model to provide severity information instead of a simple binary deterioration flag. Clinically speaking, this information can be used to triage patients in a timely fashion, and consequently improving patient safety (38). DETERIO also contains a conformal prediction module, which has been previously demonstrated to increase sepsis prediction model generalizability and performance during external validation (30). The module facilitates an assessment of DETERIO's capabilities before prediction in a new setting. Based on available clinical data, patient demographics, and other data differences in a validation vs. the training cohort (23), DETERIO quantifies confidence in its predictions. This approach further bolsters clinical utility as it enables clinicians to determine how much "trust" to put into a given prediction. Finally, DETERIO identifies the most important clinical information to a predicted score, an ability that could be leveraged to inform further clinical workup (i.e., actionability); a strong influence from respiratory measurements (e.g., oxygen saturation and respiratory rate) could lead clinicians to implement pulmonary-specific investigations and prophylactic interventions.

Before development of the AIDE criteria, physiologic deterioration could imply a broad range of clinical trends, syndromes, and conditions. Muralitharan et al (39) demonstrate this finding in a recent systematic review of ML-based prediction models for deterioration. Outcomes of the included studies varied widely: emergencies, cardiorespiratory instability, cardiac decompensation, cardiac arrest, in-hospital mortality, hospitalization, ICU transfer, ICU readmission, development of critical illness, onset of sepsis, mortality due to sepsis, vital sign changes at a set threshold, patient-specific anomalies, and abnormal clinical events were all used to define deterioration (39). While these outcomes all "relate" to deterioration, they are ultimately inconsistent and sometimes emphasize subjectivity or introduce confounders. DETERIO is the first predictive model to incorporate a consensus definition of deterioration. By doing so, it places greater emphasis on objectivity and decreases the potential for confounders. Of the previously listed outcomes used to define deterioration, ICU transfer criteria are institution dependent, and in-hospital mortality may occur due to factors unrelated to their episode of deterioration (24). This notion can limit the external validity and power of models focused on such outcomes; however, local fine-tuning of these models is possible and can significantly improve test characteristics (40). Our composite score relies on the objective cardiovascular and respiratory considerations in the AIDE criteria. Our score also allows for the incorporation of additional decompensation events and their contributions to the overall score, which can be learned and adjusted accordingly.

We acknowledge several limitations. First, this model was developed and validated at a single-academic institution. Although our sample was large and heterogeneous, we are supportive of multicenter studies. Second, although DETERIO predicts based on the consensus AIDE criteria, these criteria were validated to a limited extent (24). Both of these factors may limit generalizability. However, our health system notably has several locations that give a broad range of patients in terms of race, ethnicity, and socioeconomic statuses. Furthermore, DETERIO's conformal prediction module has been previously shown to increase generalizability (30). As such, with appropriate fine-tuning, model performance should not substantially decrease at other centers. Third, a scenario where a negative class is generated despite "patient deterioration" may arise; consider a patient who has been transferred to ICU and ultimately experienced in-hospital mortality but did not experience severe hypoxemia or receive vasopressors, inotropes, or mechanical ventilation. Their composite score would be two, and a negative class would be assigned despite their deteriorated clinical status. We justify this design because we did not want to assign a positive class based purely on subjective criteria, limiting model generalizability. Additionally, this design increases performance (41) and yields benefits over a binary scoring system because the score of two provides DETERIO with consistency during training. Finally, further workflow integration and prospective evaluation are required to assess fully the clinical impact of the proposed model; we hope to implement a rigorous analytical method, such as regression discontinuity design, in pursuit of this goal.

CONCLUSIONS

Our findings suggest that adult physiologic deterioration can be successfully predicted using a novel composite scoring system rooted in a consensus definition of deterioration. DETERIO achieved significantly better performance than the commercially available EDI. Its "label confidence" approach, conformal prediction and interpretability modules may provide additional clinical utility in terms of triage and clinician interaction with prediction confidence and explanations. Prospective studies conducted at external sites are required to further validate these findings.

[1] Department of Biomedical Informatics, University of California San Diego, San Diego, CA.

[2] School of Medicine, University of Limerick, Limerick, Ireland.

[3] Division of Hospital Medicine, University of California San Diego, San Diego, CA.

[4] Department of Emergency Medicine, University of California San Francisco, San Francisco, CA.

[5] Division of Nephrology, Division of Hospital Medicine, University of California San Diego, San Diego, CA.

[6] Division of Pulmonary, Critical Care, Sleep Medicine and Physiology, University of California San Diego, San Diego, CA.

[7] Department of Emergency Medicine, University of California San Diego, San Diego, CA.

*Supplemental digital content is available for this article (**Appendix I**, https://links.lww.com/CCX/B397). Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).*

Dr. Shashikumar, Mr. Le, Dr. Nemati, and Dr. Wardi were involved in the original conception and design of the work. Drs. Shashikumar and Nemati developed the network architectures, conducted the experiments, and analyzed the data. Mr. Le, Dr. Yung, Dr. Ford, C.A.L., Dr. Malhotra, Dr. Singh, and Dr. Wardi provided clinical expertise and contributed to interpretation of the results. All authors contributed to article preparation and critical revisions.

Dr. Wardi has been supported by the National Foundation of Emergency Medicine and the National Institutes of Health (NIH; No. K23GM146092). Dr. Nemati is funded by the NIH (Nos. R01LM013998, R01HL157985, and R35GM143121). Dr. Malhotra is funded by the NIH. Drs. Nemati, Malhotra, and Shashikumar are co-founders of a University of California San Diego (UCSD) start-up, Healcisio, which is focused on commercialization of advanced analytical decision support tools. Healcisio is formed in compliance with the UCSD conflict of interest policies. Dr. Malhotra reports income from Livanova, Eli Lilly, Zoll, and Powell Mansfield outside of this topic area. The remaining authors have disclosed that they do not have any potential conflicts of interest.

Dr. Shashikumar and Mr. Le are joint first authors.

Drs. Nemati and Wardi are joint senior authors.

For information regarding this article, E-mail: gwardi@health.ucsd.edu

The opinions or assertions contained herein are the private ones of the authors and are not to be construed as official or reflective of the views of the National Institutes of Health or any other agency of the U.S. Government.

REFERENCES

1. Bapoje SR, Gaudiani JL, Narayanan V, et al: Unplanned transfers to a medical intensive care unit: Causes and relationship to preventable errors in care. *J Hosp Med* 2011; 6:68–72

2. Churpek MM, Wendlandt B, Zadravecz FJ, et al: Association between intensive care unit transfer delay and hospital mortality: A multicenter investigation. *J Hosp Med* 2016; 11:757–762
3. Delgado MK, Liu V, Pines JM, et al: Risk factors for unplanned transfer to intensive care within 24 hours of admission from the emergency department in an integrated healthcare system. *J Hosp Med* 2013; 8:13–19
4. Escobar GJ, Greene JD, Gardner MN, et al: Intra-hospital transfers to a higher level of care: Contribution to total hospital and intensive care unit (ICU) mortality and length of stay (LOS). *J Hosp Med* 2011; 6:74–80
5. Kause J, Smith G, Prytherch D, et al; Intensive Care Society (UK): A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom—the ACADEMIA study. *Resuscitation* 2004; 62:275–282
6. Liu V, Kipnis P, Rizk NW, et al: Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. *J Hosp Med* 2012; 7:224–230
7. Smith AF, Wood J: Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey. *Resuscitation* 1998; 37:133–137
8. Jones D, Mitchell I, Hillman K, et al: Defining clinical deterioration. *Resuscitation* 2013; 84:1029–1034
9. Solomon RS, Corwin GS, Barclay DC, et al: Effectiveness of rapid response teams on rates of in-hospital cardiopulmonary arrest and mortality: A systematic review and meta-analysis. *J Hosp Med* 2016; 11:438–445
10. Teuma Custo R, Trapani J: The impact of rapid response systems on mortality and cardiac arrests—a literature review. *Intensive Crit Care Nurs* 2020; 59:102848
11. Ko BS, Lim TH, Oh J, et al: The effectiveness of a focused rapid response team on reducing the incidence of cardiac arrest in the general ward. *Medicine (Baltim)* 2020; 99:e19032
12. Yang E, Lee H, Lee S-M, et al: Effectiveness of a daytime rapid response system in hospitalized surgical ward patients. *Acute Crit Care* 2020; 35:77–86
13. Liaw S, Tee A, Carpio G, et al: Review of systems for recognising and responding to clinical deterioration in Singapore hospitals: A nationwide cross-sectional study. *Singapore Med J* 2020; 61:184–189
14. Maharaj R, Raffaele I, Wendon J: Rapid response systems: A systematic review and meta-analysis. *Crit Care* 2015; 19:254
15. Gerry S, Bonnici T, Birks J, et al: Early warning scores for detecting deterioration in adult hospital patients: Systematic review and critical appraisal of methodology. *BMJ (Clin Res Ed)* 2020; 369:m1501
16. Bedoya AD, Clement ME, Phelan M, et al: Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Crit Care Med* 2019; 47:49–55
17. Kipnis P, Turk BJ, Wulf DA, et al: Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016; 64:10–19
18. Steitz BD, McCoy AB, Reese TJ, et al: Development and validation of a machine learning algorithm using clinical pages to predict imminent clinical deterioration. *J Gen Intern Med* 2023; 39:27–35
19. Krumholz HM: Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)* 2014; 33:1163–1170
20. Lilly CM, Kirk D, Pessach IM, et al: Application of machine learning models to biomedical and information system signals from critically ill adults. *Chest* 2023; 165:1139–1148
21. Escobar GJ, Liu VX, Schuler A, et al: Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020; 383:1951–1960
22. Churpek MM, Yuen TC, Edelson DP: Predicting clinical deterioration in the hospital: The impact of outcome selection. *Resuscitation* 2013; 84:564–568
23. Le JP, Shashikumar SP, Malhotra A, et al: Making the improbable possible: Generalizing models designed for a syndrome-based, heterogeneous patient landscape. *Crit Care Clin* 2023; 39:751–768
24. Mitchell OJL, Dewan M, Wolfe HA, et al: Defining physiological decompensation: An expert consensus and retrospective outcome validation. *Crit Care Explor* 2022; 4:e0677
25. Sutton RS: Learning to predict by the methods of temporal differences. *Mach Learn* 1988; 3:9–44
26. Park CH, Jhee JH, Chun K-H, et al: Nocturnal systolic blood pressure dipping and progression of chronic kidney disease. *Hypertens Res* 2024; 47:215–224
27. El Jamal N, Brooks TG, Cohen J, et al: Prognostic utility of rhythmic components in 24-h ambulatory blood pressure monitoring for the risk stratification of chronic kidney disease patients with cardiovascular co-morbidity. *J Hum Hypertens* 2024; 38:420–429
28. von Elm E, Altman DG, Egger M, et al; STROBE Initiative: The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *J Clin Epidemiol* 2008; 61:344–349
29. Leisman DE, Harhay MO, Lederer DJ, et al: Development and reporting of prediction models: Guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020; 48:623–633
30. Shashikumar SP, Wardi G, Malhotra A, et al: Artificial intelligence sepsis prediction algorithm learns to say “I don’t know.” *NPJ Digit Med* 2021; 4:1–9
31. Li RC, Smith M, Lu J, et al: Using AI to empower collaborative team workflows: Two implementations for advance care planning and care escalation. *NEJM Catal* 2022; 3:CAT.21.0457
32. Faians A: CLEWICU—Instructions for Use
33. Byrd TF IV, Southwell B, Ravishankar A, et al: Validation of a proprietary deterioration index model and performance in hospitalized adults. *JAMA Netw Open* 2023; 6:e2324176
34. Gallo RJ, Shieh L, Smith M, et al: Effectiveness of an artificial intelligence-enabled intervention for detecting clinical deterioration. *JAMA Intern Med* 2024; 184:557–562
35. Lu JH, Callahan A, Patel BS, et al: Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: A systematic review. *JAMA Netw Open* 2022; 5:e2227779
36. Wang M, Yu H-T, Min F: Noise label learning through label confidence statistical inference. *Knowl Based Syst* 2021; 227:107234

37. Northcutt C, Jiang L, Chuang I: Confident learning: Estimating uncertainty in dataset labels. *J Artif Intell Res* 2021; 70:1373–1411

38. Ruskin KJ, Hueske-Kraus D: Alarm fatigue: Impacts on patient safety. *Curr Opin Anaesthesiol* 2015; 28:685–690

39. Muralitharan S, Nelson W, Di S, et al: Machine learning–based early warning systems for clinical deterioration: Systematic scoping review. *J Med Internet Res* 2021; 23:e25187

40. Wardi G, Carlile M, Holder A, et al: Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann Emerg Med* 2021; 77:395–406

41. Vrudhula A, Hughes JW, Yuan N, et al: The impact of task set-up in algorithm design: Regression versus classification. *NEJM AI* 2024; 1:AIcs2300176