ORIGINAL CLINICAL REPORT

OPEN

Developing and Validating a Prediction Model For Death or Critical Illness in Hospitalized Adults, an Opportunity for Human-Computer Collaboration

Amol A. Verma, MD[1,2,3]
Chloe Pou-Prom, MSc[1]
Liam G. McCoy, MD[2]
Joshua Murray, MSc[1]
Bret Nestor, MEng[4,5]
Shirley Bell, RN[1]
Ophyr Mourad, MD[1,2]
Michael Fralick, MD[2,6]
Jan Friedrich, MD[1,2]
Marzyeh Ghassemi, PhD[5,7]
Muhammad Mamdani, PharmD[1,2,3,5,8]

OBJECTIVES: Hospital early warning systems that use machine learning (ML) to predict clinical deterioration are increasingly being used to aid clinical decision-making. However, it is not known how ML predictions complement physician and nurse judgment. Our objective was to train and validate a ML model to predict patient deterioration and compare model predictions with real-world physician and nurse predictions.

DESIGN: Retrospective and prospective cohort study.

SETTING: Academic tertiary care hospital.

PATIENTS: Adult general internal medicine hospitalizations.

MEASUREMENTS AND MAIN RESULTS: We developed and validated a neural network model to predict in-hospital death and ICU admission in 23,528 hospitalizations between April 2011 and April 2019. We then compared model predictions with 3,374 prospectively collected predictions from nurses, residents, and attending physicians about their own patients in 960 hospitalizations between April 30, and August 28, 2019. ML model predictions achieved clinician-level accuracy for predicting ICU admission or death (ML median F1 score 0.32 [interquartile range (IQR) 0.30-0.34], AUC 0.77 [IQ 0.76-0.78]; clinicians median F1-score 0.33 [IQR 0.30–0.35], AUC 0.64 [IQR 0.63–0.66]). ML predictions were more accurate than clinicians for ICU admission. Of all ICU admissions and deaths, 36% occurred in hospitalizations where the model and clinicians disagreed. Combining human and model predictions detected 49% of clinical deterioration events, improving sensitivity by 16% compared with clinicians alone and 24% compared with the model alone while maintaining a positive predictive value of 33%, thus keeping false alarms at a clinically acceptable level.

CONCLUSIONS: ML models can complement clinician judgment to predict clinical deterioration in hospital. These findings demonstrate important opportunities for human-computer collaboration to improve prognostication and personalized medicine in hospital.

KEY WORDS: artificial intelligence; clinical prediction; machine learning; mortality; prognosis

Unrecognized clinical deterioration is the most common cause of unplanned ICU transfer in hospital (1). Although physicians and nurses are able to predict death, critical illness, and recovery in hospital (e.g., area under the receiver operating characteristic curve [AUC] values ranging from 0.70 to 0.85) (2–5), early warning systems are designed to systematically predict a patient's likelihood of clinical deterioration with the goal of protocolizing care escalation and expediting early intervention to prevent deterioration (6–9).

DOI: 10.1097/CCE.0000000000000897

KEY POINTS

Question: How does a machine learning model compare with physician and nurse predictions about the risk of patient deterioration in hospital?

Findings: A machine learning model was developed in 23,528 hospitalizations and compared prospectively to 3,374 physician and nurse predictions in 960 hospitalizations. Combining human and model predictions detected 49% of clinical deterioration events, improving sensitivity by 16% compared with clinicians alone and 24% compared with the model alone while maintaining false alarms at a clinically acceptable level.

Meaning: Machine learning models can complement clinician judgment to improve the prediction of clinical deterioration in hospital.

Simple early warning systems generate a points-based score from a small number of inputs (e.g., patient vital signs and mental status) (10, 11) and have been widely implemented with varying effectiveness in reducing mortality (6). They are generally less accurate than physicians and nurses at predicting in-hospital mortality (4, 12–17). More advanced early warning systems include a broader range of inputs and more sophisticated modeling (9, 18–20). Implementation of one such system across 19 hospitals in California reduced mortality for high risk medical-surgical patients (21). However, most implementations of early warning systems have not significantly improved patient outcomes (22). Early warning systems based on machine learning (ML) are relatively novel, and few have been validated or implemented in real-world clinical settings (5, 23–27). A ML sepsis warning system was associated with reduced mortality only when providers confirmed their agreement with the alert within 3 hours (38% of cases) (27, 28). This highlights the importance of understanding how human judgment and computer predictions interact. Yet, little is understood about how ML early warning systems might supplement physician or nurse judgment about patient risk.

Our objective was to develop and validate a ML-based model to predict ICU admission or in-hospital death among general internal medicine (GIM) inpatients at an academic hospital. We compared ML model predictions with a commonly used points-based deterioration score (National Early Warning Score, NEWS) (10) and real-world physician and nurse predictions. Finally, we investigated the accuracy of predictions in cases of agreement and disagreement between clinicians and the ML model. Taken together, these experiments seek to illuminate opportunities for human-computer collaboration in predicting clinical deterioration in hospital.

MATERIALS AND METHODS

Design and Setting

This was a retrospective and prospective study conducted among GIM patients at St. Michael's Hospital, an academic health center in Toronto, ON, Canada. GIM patients are cared for by four clinical teaching unit (CTU) teams and one nonteaching "day admission" team. Every team is supervised by an attending physician. The CTU teams have one senior medical resident (in their second or third year of postgraduate medical training) and several first-year residents and medical students. CTU teams typically care for 15–25 patients at any given time. Nurses on the GIM ward were trained as registered nurses and typically cared for four patients on each 12-hour shift during the study period. There was no active early warning system or other similar system (e.g., sepsis risk prediction) at our hospital during the study period.

This study was reported in alignment with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement (29). This study was conducted in accordance with ethical standards consistent with the Helsinki Declaration of 1975 and approved by the St. Michael's Hospital Research Ethics Board (REB no.:19-008, April 04, 2019, Title: "Comparing clinical predictions with an automated early warning system tool for detecting clinical deterioration").

Study Outcomes

The primary outcome of interest was a composite of in-hospital death or ICU transfer. Secondary outcomes included in-hospital death or ICU transfer separately.

Model Development: Outcomes

Because in-hospital death or ICU transfer is relatively rare, we trained models on a composite outcome that also included transfer to the GIM step-up unit (a four-bed monitored unit) in addition to in-hospital death

and ICU transfer. To train the models, we censored patient visits at the first occurrence of any component of the training outcome.

Model Development: Data Sources

For the training and validation data, we extracted patient visit data from the hospital's enterprise data warehouse. We included all complete inpatient admissions to the GIM service between April 2011 and April 2019, when data from electronic medical records were available. We split the data into training and validation datasets by date: April 2011 to November 2018 for training and December 2018 to April 2019 for validation. To avoid biasing model development with outlier cases, we used the latter 40 days of a visit for any visits whose length of stay was greater than 40 days.

Inputs to the model consisted of static and time-varying features: patient demographics (age, sex, housing status), ICU admission prior to GIM admission, laboratory test results, and vital signs (for all inputs, please see **Appendix Table 6**, http://links.lww.com/CCX/B174). We selected these inputs based on a combination of most frequent counts in the training data and clinical judgment.

Model Development: Data Preprocessing

Considering Each Visit as a Time Series. For time-varying features, we binned the data into 6-hour intervals. The time-series was censored at hospital discharge or the first occurrence of the composite training outcome. When there were multiple input measures within the same interval, we took the average value. We created a "window since start" variable to track the number of 6-hour intervals that elapsed since the patient was in hospital. To avoid "signal leakage" (generating predictions based on factors that are directly related to the outcome, such as predicting death based on withdrawal of ventilator support, which would lack clinical utility [30]), we removed the 6-hour block before outcomes in the training data.

Numeric Measures and Missing Data. To minimize the impact of extreme values, we trimmed all numeric labs and vitals values less than the first percentile and greater than the 99th percentile (as determined from the training data) and then normalized the values using the first and 99th percentile. An indicator variable flagged whether or not the feature was measured in the 6-hour window. We also added a "time since last-measured" variable for each feature. To address missingness, we imputed data using last observation carried forward followed by population mean imputation if no prior measurement was available. We calculated the population mean using all observations in the training data. The mean and median values and missingness for all model inputs are reported in **Appendix Table 7** (http://links.lww.com/CCX/B174).

Model Development

We trained a gated recurrent unit neural network model on the training data and with early stopping on the validation data. In order to determine the alerting threshold, we performed 10-fold cross-validation on the training dataset to find the probability threshold that would give a 30% encounter-level positive predictive value. This was selected after engagement with physicians and nurses who felt that there should be no more than two false alarms for every one true alarm, to minimize alarm fatigue and excessive workload (31).

Prospective Clinical Validation

Between April 30 and August 28, 2019, a research assistant attended the GIM ward on weekdays and interviewed attending physicians, senior medical residents, and nurses once daily about their patients using a standardized questionnaire. The clinicians were asked to review every patient on their roster and identify, "Which of the patients that you are currently caring for are likely to die or require ICU care at any point during this hospital visit?" The time of each clinical prediction was documented.

Early warning scores are not designed to function based on a single prediction, but to alert clinicians whenever a patient's risk crosses a predefined threshold. Given that it was not feasible to collect real-time clinical predictions every 6 hours, we designed this study to approximate the real-world use of an early warning system. We considered all ML and NEWS model predictions in the 48 hours prior to the clinical prediction and selected the highest model risk prediction to capture all patients who would receive an "alert" in an early warning system. Alerts are generally what drives clinical action in an early warning system, and focusing on highest predicted risk has been used

to assess model performance in other studies (32). A 48-hour lookback period to include model predictions was selected because clinicians reported that this is a useful time window to predict patient deterioration (25). Both clinical and model predictions were assessed for outcomes occurring at any time during the rest of the hospitalization.

Based on published estimates of clinician accuracy in predicting clinical deterioration (2, 12) and a historical rate of death or ICU transfer in 7% of GIM admissions at our hospital, we calculated that a sample of 948 admissions would be needed to identify a 20% difference in sensitivity between clinician and model predictions, with 80% power.

We followed encounters for up to 30 days after the end of the data collection period to capture outcomes.

Statistical Analysis

We report the demographic and clinical characteristics of the model training and validation datasets and the prospective clinical validation cohort. We report performance of the ML model using area under the receiver operating characteristic curve (AUC) and F1-score. The F1-score is the harmonic mean of sensitivity and positive predictive value (PPV) and is a good measure of model performance when there is class imbalance (i.e., the outcome is rare) (33). Because an important clinical aim is to minimize false alarms, we report the model's sensitivity, specificity, PPV, and negative predictive value at a prespecified threshold PPV of 30% in the training data (31). We considered predictions above this threshold to be "high risk." We calculated 95% CIs by bootstrapping across 500 random samples with replacement, within each dataset.

To report model performance, we censored patient visits at the first occurrence of any outcome event for the composite primary outcome but not for individual secondary outcomes. We compared clinical predictions with ML model predictions and NEWS model predictions (10, 11) in the prospective cohort. As recommended (10), we considered a NEWS score of more than 6 to be "high risk." We report the performance of all types of clinicians grouped together and of each group separately. Because multiple predictions were sometimes made for single patient admissions, we randomly sampled one clinical prediction per encounter to compare with the highest model-predicted risk in the prior 48 hours. We repeated this sampling 500 times and report the median and interquartile range (IQR) of results across the bootstrapped samples.

Finally, we report the accuracy of predictions when clinicians and the ML model agreed or did not agree, and when clinician and ML model predictions were combined. To combine predictions, we considered "low risk" to be the cases where both clinicians and the ML model agreed on low risk and "high risk" to be when either clinicians or the ML model predicted so.

The ML model was trained using Python 3.7 (Python Software Foundation) and the Pytorch library (PyTorch Foundation) (34). Analyses were performed using R 3.6 (R Foundation for Statistical Computing, Vienna, Austria) and the tidyverse package (35).

RESULTS

There were 22,361 hospitalizations in the model training cohort, 1,167 in the historical validation cohort, and 960 in the testing (prospective clinical) cohort. Across the three cohorts, the median age was approximately 67 years, and approximately 40% of patients were female (**Table 1**). Baseline characteristics across the three cohorts were generally comparable. Death or ICU admission occurred in 1,800 hospitalizations (8.0%) in the training cohort, 83 hospitalizations (7.1%) in the validation cohort, and 61 hospitalizations (6.4%) in the testing cohort (Table 1).

Model Development and Historical Validation

Model performance across the training, validation, and testing cohorts is presented in **Table 2** and **Appendix Tables 1** and **2** (http://links.lww.com/CCX/B174). In the historical validation cohort, the model had an AUC of 0.75 (95% CI 0.68–0.81) and F1-score of 0.49 (95% CI 0.39–0.58) for predicting ICU or death. At a PPV of 0.43 (95% CI 0.33–0.53), the sensitivity was 0.57 (95% CI 0.46–0.68), specificity was 0.95 (95% CI 0.94–0.96), and negative predictive value (NPV) was 0.97 (95% CI 0.96–0.98). In the testing cohort, the model had an AUC of 0.81 (95% CI 0.73–0.87) and F1-score of 0.43 (95% CI 0.34–0.53). At a PPV of 0.34 (95% CI 0.25–0.44), the sensitivity was 0.60 (95% CI 0.48–0.73), specificity was 0.93 (95% CI

TABLE 1.
Patient Characteristics and Outcomes in Retrospective and Prospective Cohorts

Characteristic	Training	Validation	Testing
Number of admissions	22,361	1,167	960
Number of unique patients	14 567	991	847
Age, median (IQR)	67.0 (53.1–80.1)	66.6 (52.9–79.7)	65.7 (52.9–79.4)
Sex female, *n* (%)	9,554 (42.7)	463 (39.7)	383 (39.9)
Charlson Comorbidity Score, *n* (%)			
0	5,608 (25.1)	269 (23.1)	219 (22.8)
1	5,118 (22.9)	239 (20.5)	205 (21.4)
2+	11,635 (52.0)	659 (56.5)	536 (55.8)
Temperature, median (IQR)[a]	36.4 (35.9–36.8)	36.6 (36–37.3)	36.4 (36–36.8)
Heart rate, median (IQR)[a]	84 (72–98)	92 (78–109)	81 (71–94)
Systolic BP, median (IQR)[a]	129 (114–146)	130 (113–150)	129 (115–145)
Diastolic BP, median (IQR)[a]	73 (65–82)	76 (66–86)	73 (65–81)
O_2 saturation, median (IQR)[a]	97 (95–98)	97 (95–99)	97 (95–98)
Respiratory rate, median (IQR)[a]	19 (18–20)	18 (18–20)	18 (18–20)
Hospital length of stay (d), median (IQR)	4.9 (2.7–9.2)	4.7 (2.7–9.0)	6.5 (3.8–11.8)
ICU admission, *n* (%)	906 (4.1)	43 (3.7)	31 (3.2)
Step-up unit admission, *n* (%)	768 (3.4)	20 (1.7)	28 (2.9)
Death, *n* (%)	1134 (5.1)	49 (4.2)	37 (3.9)
Composite death or ICU	1800 (8.1)	83 (7.1)	61 (6.4)

BP = blood pressure, IQR = interquartile range.
[a]Vital sign measurement contributing to first machine learning model prediction.
Temperature in degrees Celsius.

TABLE 2.
Performance of Machine Learning Model on the Composite Outcome of Death or ICU Transfer

	Death or ICU		
Measure	Training	Validation	Testing
Area under the receiver operating characteristic curve	0.81 (0.80–0.82)	0.75 (0.68–0.81)	0.81 (0.73–0.87)
F1-score	0.44 (0.42–0.46)	0.49 (0.39–0.58)	0.43 (0.34–0.53)
Positive predictive value	0.36 (0.34–0.38)	0.43 (0.33–0.53)	0.34 (0.25–0.44)
Sensitivity	0.56 (0.53–0.58)	0.57 (0.46–0.68)	0.60 (0.48–0.73)
Specificity	0.92 (0.92–0.93)	0.95 (0.94–0.96)	0.93 (0.91–0.95)
Negative predictive value	0.96 (0.96–0.97)	0.97 (0.96–0.98)	0.98 (0.96–0.99)

F1-score = harmonic mean of sensitivity and positive predictive value.
We calculated 95% CIs through bootstrapping. Results were calculated on 500 random samples, with replacement, within each dataset.

0.91–0.95), and NPV was 0.98 (95% CI 0.96–0.99). In the testing cohort, the median time from the first "high risk" model prediction to the first occurrence of ICU admission or death was 117 hours (IQR 70–244 hr).

Prospective Clinical Comparison

The prospective clinical study included 3,374 real-time clinical predictions for 960 hospitalizations representing 847 unique patients. This included 1,264 predictions from 23 attending physicians, 1,097 predictions by 20 resident physicians, and 1,013 predictions by 129 nurses.

Across 500 bootstrapped samples, clinicians (nurses or physicians) predicted ICU or death with a median AUC of 0.64 (IQR 0.63–0.66) and F1-score of 0.33 (IQR 0.30–0.35) (**Fig. 1**) (**Appendix Figs**. **1** and **2** and **Appendix Table 3**, http://links.lww.com/CCX/B174). In comparison, the ML model predicted ICU or death with a median AUC of 0.77 (IQR 0.76-0.78) and F1-score of 0.32 (IQR 0.30-0.34). The NEWS model was less accurate, with median AUC of 0.60 (IQR 0.59–0.61) and F1-score of 0.12 (IQR 0.10–0.14). Across bootstrapped samples, the median F1-scores for nurses, residents, and attending physicians in predicting ICU or death were 0.27 (IQR 0.25–0.30), 0.38 (IQR 0.36–0.41), and 0.37 (IQR 0.34–0.39), respectively (Appendix Fig. 2, http://links.lww.com/CCX/B174; Appendix Table 3, http://links.lww.com/CCX/B174).

The ML model was consistently more accurate than all types of clinicians for predicting ICU admission (Fig. 1) (Appendix Figs. 1 and 2 and **Appendix Tables 3-5**, http://links.lww.com/CCX/B174).

Agreement and Disagreement Between ML Model and Clinicians

Across bootstrapped samples, clinicians and the ML model agreed on 92.0% (IQR 91.6–92.3%) of predictions overall (**Table 3**). Among all high risk predictions (median $n = 88$), clinicians and the model agreed on only 11 predictions (12.5%) (Table 3). Death or ICU admission occurred in 60.0% of patients (IQR

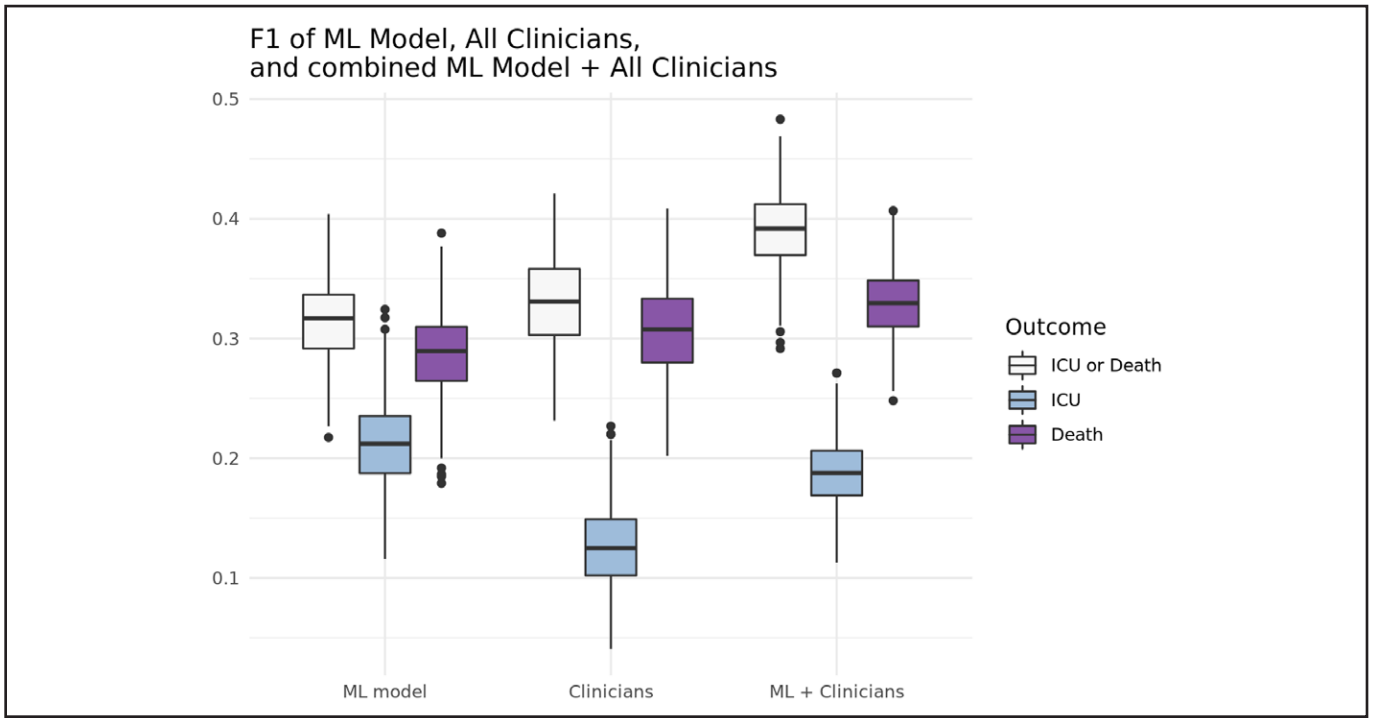

Figure 1. Accuracy of clinicians and the machine learning (ML) model in predicting ICU and death across 500 bootstrapped samples. Boxplots depict the distribution (*horizontal bar* denotes median, box is 25/75 percentile) of the F1-score of the ML model and clinicians across 500 bootstrapped samples. Different bootstrapped samples were used for each clinician type, resulting in slightly different estimates of ML model performance in each comparison. The ML model demonstrates superior F1-scores in predicting ICU admission, but the model was not more accurate than clinicians in predicting death. Combining model and clinician predictions leads to improvements in F1-scores, which is particularly important given that the value of identifying true positive cases outweighs identification of true negatives in an early warning system.

TABLE 3.
Patient Characteristics and Outcomes When Clinicians and the Model Agree and Disagree

Measure	Agree High Risk	Agree Low Risk	Disagree–Model Predicts High Risk	Disagree–Physician Predicts High Risk
Number of predictions	11 (9–13)	872 (868–875)	23 (21–25)	54 (51–57)
ICU or death, % (IQR)	60.0 (53.9–66.7)	3.7 (3.5–3.9)	34.8 (30.0–38.5)	26.5 (23.6–28.9)
ICU, % (IQR)	16.7 (10.0–23.1)	2.3 (2.2–2.4)	20.0 (16.7–23.8)	7.8 (6.0–9.6)
Death, % (IQR)	46.7 (40.0–55.6)	1.9 (1.7–2.1)	20.8 (17.4–23.8)	19.6 (17.4–21.6)
Age, median (IQR)	66.9 (62.4–71.8)	65.7 (65.6–65.9)	70.3 (69.6–70.9)	63.5 (63.1–63.9)
Sex female, % (IQR)	37.5 (30.0–45.5)	40.3 (40.1–40.6)	30.4 (26.9–34.8)	36.4 (33.7–39.6)
Charlson Comorbidity Score, % (IQR) 0	0 (0–0)	24.6 (24.4–24.7)	0 (0–0)	9.1 (7.0–10.9)
1	10 (6.7–16.7)	21.8 (21.7–22.0)	15.4 (12.5–18.2)	18.3 (16.1–20.8)
2+	88.9 (81.8–92.3)	53.6 (53.4–53.8)	84.6 (81.8–87.5)	72.9 (70.1–75.5)
Temperature, median (IQR)[a]	36.6 (36.4–36.7)	36.4 (36.4–36.4)	36.4 (36.4–36.6)	36.5 (36.5–36.5)
Heart rate, median (IQR)[a]	97 (93–99.5)	80 (80–80)	96 (96–100)	86.5 (84–88)
Systolic BP, median (IQR)[a]	124 (123–126)	129 (129–129.5)	122 (119–123)	122 (119.5–124)
Diastolic BP, median (IQR)[a]	76 (73.5–77.5)	73 (73–73)	67.5 (65–70)	70.5 (70–71.5)
Oxygen saturation, median (IQR)[a]	95 (95–95.62)	97 (97–97)	95 (94–95)	96 (96–96)
Respiratory rate, median (IQR)[a]	20 (20–20)	18 (18–18)	20 (20–20)	20 (20–20)
Hospital length of stay (d), median (IQR)	27.5 (18.3–31.9)	6.2 (6.1–6.2)	18.6 (12.9–23.8)	8.7 (8.2–9.5)

BP = blood pressure, IQR = interquartile range.
[a]Vital sign measurement contributing to first machine learning model prediction.
Temperature in degrees Celsius.

53.9–66.7%) when clinicians and the model agreed patients were high risk and 3.7% of patients (IQR 3.5–3.9%) when clinicians and the model agreed patients were low risk. When only the model predicted patients were high risk, death or ICU admission occurred in 34.8% (IQR 30.0–38.5%). When only clinicians predicted patients were high risk, death or ICU admission occurred in 26.5% (IQR 23.6–28.9%).

Combining clinician and model predictions yielded a median AUC 0.71 (IQR 0.69–0.73), F1-score 0.39 (IQR 0.37–0.41), sensitivity 0.49 (IQR 0.46–0.52), and PPV 0.33 (IQR 0.31–0.34) for ICU or death (Fig. 1) (Appendix Table 3, http://links.lww.com/CCX/B174). This approach improved detection of ICU admissions or death by 16% compared with clinicians alone and 24% compared with the model alone, while keeping the PPV at two false alarms for one true alarm.

DISCUSSION

A ML model performed comparably to clinicians in predicting clinical deterioration in hospital. Overall, the ML model was superior to a simpler model (NEWS) and was better than clinicians in predicting ICU transfers, whereas physicians were better at predicting deaths. The ML model and human predictions were complementary, with each identifying cases that the other missed. Thus, collaboration between humans

and ML models can improve the prediction of clinical deterioration events. We found that combining human and model predictions led to the detection of 49% of deaths or ICU transfers, improving sensitivity by 16% compared with clinicians alone and by 24% compared with the model alone while maintaining false alarms at an acceptable level (PPV of 33%). This is particularly promising, as a central goal of early warning systems is to prevent a "failure to rescue" (1, 36) by increasing the recognition of deteriorating patients.

A recent scoping review identified 18 studies of automated systems for detecting clinical deterioration and found that only four studies reported significant improvements in clinical outcomes (22). The authors highlighted the importance of clinical response protocols and involvement of attending physicians as determinants of successful implementation (22). Yet, we have poor understanding of how early warning systems interact with clinical judgment. Numerous studies have demonstrated that clinicians are modestly accurate in predicting death, critical illness, or recovery (2–5, 37). Simpler risk prediction tools are generally inferior to clinical judgment (4, 12–17, 37), but advances in data science have improved prediction of clinical deterioration (9, 19, 38). A recent systematic review identified 24 studies of ML models to predict clinical deterioration (19). Of these, only one was prospectively validated, and none were compared with real-world clinical predictions. Outside of this systematic review, two small studies showed that ML predictions of patient deterioration were comparable or superior to physician predictions (5, 39). ML models have also been compared with clinical judgment in other contexts (23), and human-computer collaboration is promising across several applications of computer vision (40–42). Comparisons between ML and human performance are often artificial as clinicians are asked to interpret a clinical scenario (43) or photograph (44, 45) rather than actually assess a patient. Such experiments disadvantage clinicians, who typically make decisions through holistic patient assessments that include narrative history, physical examination, and formation of a clinical gestalt (46, 47). A major strength of our study is including real-world clinician predictions about their own patients.

Our study has several limitations. First, this was a single-center analysis of GIM patients, who have relatively high rates of severe illness (48, 49), at an urban teaching hospital. Our results may not generalize to other settings or patient populations. For example, accuracy of clinical judgment or criteria for admission to ICU may vary across centers. Additionally, at our institution, admission and comorbidity diagnoses were not available from the electronic medical record in real-time (they were coded after discharge); therefore, they were not included as model inputs. We would expect their inclusion to improve ML model performance. However, given that our ML model had similar performance to those in the literature (8, 9, 19) and because, in many cases, ML models perform best when trained within a specific context, our article demonstrates how models can be validated locally to inform implementation. Second, clinicians may have had different interpretations of the question about which patients were "likely" to deteriorate. Although our survey aligned with other approaches that rely on a subjective assessment of likelihood (3), asking clinicians to quantify the probability of their prediction or to assess clinical agreement with model predictions in real-time may strengthen future research. Third, we only collected physician and nurse predictions during working hours on weekdays. It is possible that clinician predictions would be less accurate on the weekends, or overnight, when staffing levels may be lower and quality of care may differ (50). After-hours settings may represent an even greater opportunity for ML model predictions to inform clinical decisions. Differences in the apparent accuracy of physician and nurse predictions should be interpreted cautiously as our study was not designed or powered to investigate these differences. Fourth, we were unable to present model predictions to clinicians in real-time to assess how model predictions would shape clinical predictions. This remains an important area for future research in prognostic accuracy, as previous studies have demonstrated that faulty model predictions can mislead clinicians in diagnostic accuracy (42). It remains unknown how regular interaction with ML model predictions might influence clinicians, as this could plausibly improve their accuracy by providing more systematic feedback or reduce their accuracy by fostering dependency on automated support.

CONCLUSIONS

ML models can enable human-computer collaboration to improve prediction of deterioration in hospitalized patients. Combining clinical and model predictions

augmented case detection without excessively increasing false alarms. These insights can inform the implementation of ML-based early warning systems to improve early recognition of clinical deterioration and reduce the harms associated with a "failure to rescue" deteriorating patients.

1 *St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada.*

2 *Department of Medicine, University of Toronto, Toronto, ON, Canada.*

3 *Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada.*

4 *Department of Computer Science, University of Toronto, Toronto, ON, Canada.*

5 *Vector Institute, Toronto, ON, Canada.*

6 *Sinai Health System, Toronto, ON, Canada.*

7 *Massachusetts Institute of Technology, Cambridge, MA.*

8 *Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada.*

Supplemental digital content is available for this article. 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).

Drs. Verma, Pou-Prom, Murray, and Mamdani are coinventors of a patient deterioration early warning system similar to the one presented in this article, which was acquired by a start-up company, Signal1.

Supported, in part, by Associated Medical Services (AMS) Healthcare Fellowship in Compassion and Artificial Intelligence (to Dr. Verma) and the Vector Institute Pathfinder Projects. Dr. Verma is supported by the Temerty Professorship for Artificial Intelligence Research and Education in Medicine at the University of Toronto.

Drs. Verma and Pou-Prom are cofirst authors.

Drs. Verma, Pou-Prom, Murray, and Mamdani conceptualized the study. Dr. Pou-Prom was primarily responsible for data analysis. Drs. Verma, Pou-Prom, and McCoy contributed to data collection. Drs. Verma and Pou-Prom drafted the article. All authors made substantial contributions to study design, interpretation, and critically revising the article.

For information regarding this article, E-mail: amol.verma@mail.utoronto.ca

The datasets generated and/or analyzed during the current study are not publicly available because they contain personal health information but are available from the corresponding author on reasonable request and with necessary institutional ethics and privacy approvals.

REFERENCES

1. Van Galen LS, Struik PW, Driesen BEJM, et al: Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: A root cause analysis of unplanned ICU admissions. *PLoS One* 2016; 11:e01613931–e01613914
2. Detsky ME, Harhay MO, Bayard DF, et al: Discriminative accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an ICU admission. *JAMA* 2017; 317:2187–2195
3. Rojas JC, Lyons PG, Jiang T, et al: Accuracy of clinicians' ability to predict the need for intensive care unit readmission. *Ann Am Thorac Soc* 2020; 17:847–853
4. Sinuff T, Adhikari NKJ, Cook DJ, et al: Mortality predictions in the intensive care unit: Comparing physicians with scoring systems*. *Crit Care Med* 2006; 34878–885
5. van Doorn WPTM, Stassen PM, Borggreve HF, et al: A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS One* 2021; 16:e02451571–e02451515
6. Smith MEB, Chiovaro JC, O'Neil M, et al: Early warning system scores for clinical deterioration in hospitalized patients: A systematic review. *Ann Am Thorac Soc* 2014; 11:1454–1465
7. Churpek MM, Snyder A, Han X, et al: Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. *Am J Respir Crit Care Med* 2017; 195:906–911
8. Liu VX, Lu Y, Carey KA, et al: Comparison of early warning scoring systems for hospitalized patients with and without infection at risk for in-hospital mortality and transfer to the intensive care unit. *JAMA Netw open* 2020; 3:e205191
9. Linnen DT, Escobar GJ, Hu X, et al: Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: A systematic review. *J Hosp Med* 2019; 14:161–169
10. Royal College of Physicians: *National Early Warning Score (NEWS)*, 2017. Available at: www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2. Accessed April 1, 2022.
11. McGinley A, Pearse RM: A national early warning score for acutely ill patients. *BMJ* 2012; 345:e5310–e5310
12. Brabrand M, Hallas J, Knudsen T: Nurses and physicians in a medical admission unit can accurately predict mortality of acutely admitted patients: A prospective cohort study. *PLoS One* 2014; 9:e101739
13. Brannen AL 2nd, Godfrey LJ, Goetter WE: Prediction of outcome from critical illness. A comparison of clinical judgment with a prediction rule. *Arch Intern Med* 1989; 149:1083–1086
14. Garrouste-Orgeas M, Montuclard L, Timsit J-F, et al: Triaging patients to the ICU: A pilot study of factors influencing admission decisions and patient outcomes. *Intensive Care Med* 2003; 29:774–781
15. Marks RJ, Simons RS, Blizzard RA, et al: Predicting outcome in intensive therapy units--A comparison of APACHE II with subjective assessments. *Intensive Care Med* 1991; 17:159–163
16. Minne L, Toma T, de Jonge E, et al: Assessing and combining repeated prognosis of physicians and temporal models in the intensive care. *Artif Intell Med* 2013; 57:111–117
17. Scholz N, Bäsler K, Saur P, et al: Outcome prediction in critical care: Physicians' prognoses vs. scoring systems. *Eur J Anaesthesiol* 2004; 21:606–611
18. Fenn A, Davis C, Buckland DM, et al: Development and validation of machine learning models to predict admission from

emergency department to inpatient and intensive care units. *Ann Emerg Med* 2021; 78:290–302

19. 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
20. Cho K-J, Kwon O, Kwon J-M, et al: Detecting patient deterioration using artificial intelligence in a rapid response system. *Crit Care Med* 2020; 48:e285–e289
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. Blythe R, Parsons R, White NM, et al: A scoping review of real-time automated clinical deterioration alerts and evidence of impacts on hospitalised patient outcomes. *BMJ Qual Saf* 2022; 31:725–734
23. Topol EJ: High-performance medicine: The convergence of human and artificial intelligence. *Nat Med* 2019; 25:44–56
24. Ben-Israel D, Jacobs WB, Casha S, et al: The impact of machine learning on patient care: A systematic review. *Artif Intell Med* 2020; 103:101785
25. Verma AA, Murray J, Greiner R, et al: Implementing machine learning in medicine. *Can Med Assoc J* 2021; 193:E1351–E1357
26. Sendak MP, Ratliff W, Sarro D, et al: Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Med informatics*. 2020; 8:e15182
27. Adams R, Henry KE, Sridharan A, et al: Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med* 2022; 28:1455–1460
28. Henry KE, Adams R, Parent C, et al: Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat Med* 2022; 28:1447–1454
29. Collins GS, Reitsma JB, Altman DG, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med* 2015; 13:1–10
30. Ghassemi M, Naumann T, Schulam P, et al: A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci* 2020; 2020:191–200
31. Nestor B, McCoy LG, Verma AA, et al. Preparing a clinical support model for silent mode in general internal medicine. *Proc Mach Learn Res* 2020; 126:950–972
32. Churpek MM, Yuen TC, Winslow C, et al: Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014; 190:649–655
33. Van Rijsbergen CJ: Foundation of evalution. *J Doc* 1974; 30:365–373
34. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*. Vol. 32. Wallach H, Larochelle H, Beygelzimer A (Eds). Curran Associates, Inc., 2019. Available at: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
35. Wickham H, Averick M, Bryan J, et al: Welcome to the Tidyverse. *J Open Source Softw* 2019; 4:1686
36. Burke JR, Downey C, Almoudaris AM: Failure to rescue deteriorating patients: A systematic review of root causes and improvement strategies. *J Patient Saf* 2020; 18:e140–e155
37. Copeland-Fields L, Griffin T, Jenkins T, et al: Comparison of outcome predictions made by physicians, by nurses, and by using the mortality prediction model. *Am J Crit Care* 2001; 10:313–319
38. Churpek MM, Yuen TC, Winslow C, et al: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44:368–374
39. Arnold J, Davis A, Fischhoff B, et al: Comparing the predictive ability of a commercial artificial intelligence early warning system with physician judgement for clinical deterioration in hospitalised general internal medicine patients: A prospective observational study. *BMJ Open* 2019; 9:e0321871–e0321877
40. Nishikawa RM, Bae KT: Importance of better human-computer interaction in the era of deep learning: Mammography computer-aided diagnosis as a use case. *J Am Coll Radiol* 2018; 15:49–52
41. Zhang Z, Chen P, McGough M, et al: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat Mach Intell* 2019; 1:236–245
42. Tschandl P, Rinner C, Apalla Z, et al: Human–computer collaboration for skin cancer recognition. *Nat Med* 2020; 26:1229–1234
43. Saposnik G, Cote R, Mamdani M, et al: JURaSSiC: Accuracy of clinician vs risk score prediction of ischemic stroke outcomes. *Neurology* 2013; 81:448–455
44. Kanagasingam Y, Xiao D, Vignarajan J, et al: Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care. *JAMA Netw Open* 2018; 1:e182665–e182665
45. Esteva A, Kuprel B, Novoa RA, et al: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542:115–118
46. Soto-Mota A, Marfil-Garza BA, Castiello-de Obeso S, et al. Prospective predictive performance comparison between clinical gestalt and validated COVID-19 mortality scores. *J Investig Med* 2022; 70:415–420
47. Dale AP, Marchello C, Ebell MH: Clinical gestalt to diagnose pneumonia, sinusitis, and pharyngitis: A meta-analysis. *Br J Gen Pract* 2019; 69:e444–e453
48. Verma AA, Guo Y, Kwan JL, et al: Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: The general medicine inpatient initiative (GEMINI) retrospective cohort study. *C Open*. 2017; 5:E842–E849
49. Verma AA, Guo Y, Kwan JL, et al. Prevalence and costs of discharge diagnoses in inpatient general internal medicine: A multi-center cross-sectional study. *J Gen Intern Med* 2018; 33:1899–1904
50. Kostis WJ, Demissie K, Marcella SW, et al; Myocardial Infarction Data Acquisition System (MIDAS 10) Study Group: Weekend versus weekday admission and mortality from myocardial infarction. *N Engl J Med* 2007; 356:1099–1109