OPEN

Incorporating Laboratory Values Into a Machine Learning Model Improves In-Hospital Mortality Predictions After Rapid Response Team Call

Peter M. Reardon, MD[1,2]; Enea Parimbelli, PhD[3]; Szymon Wilk, PhD[3,4]; Wojtek Michalowski, PhD[3]; Kyle Murphy, MD[1]; Jennifer Shen, BSc[1]; Brent Herritt , MD[1]; Benjamin Gershkovich, MD[1]; Peter Tanuseputro, MD, MHSc[5,6,7,8]; Kwadwo Kyeremanteng, MD, MHA[1,5,6,9]

Objectives: Machine learning models have been used to predict mortality among patients requiring rapid response team activation. The goal of our study was to assess the impact of adding laboratory values into the model.

Design: A gradient boosted decision tree model was derived and internally validated to predict a primary outcome of in-hospital mortality. The base model was then augmented with laboratory values.

Setting: Two tertiary care hospitals within The Ottawa Hospital network.

Patients: Inpatients over the age of 18 years who experienced a rapid response team activation between January 1, 2015, and May 31, 2016.

Interventions: None.

Measurements and Main Results: A total of 2,061 rapid response team activations occurred during the study period. The in-hospital mortality rate was 29.4%. Patients who died were older (median age, 72 vs 68 yr; $p < 0.001$), had a longer length of stay (length of stay) prior to rapid response team activation (4 vs 2 d; $p < 0.001$), and more often had respiratory distress (31% vs 22%; $p < 0.001$). Our base model without laboratory values performed with an area under the receiver operating curve of 0.71 (95% CI, 0.71–0.72). When the base model was augmented with laboratory values, the area under the receiver operating curve improved to 0.77 (95% CI, 0.77–0.78). Important mortality predictors in the base model were age, estimated ratio of $\mathrm{Pa}{\scriptstyle\mathrm{O}}_2$ to $\mathrm{F}{\scriptstyle\mathrm{IO}}_2$ (calculated using oxygen saturation and estimated $\mathrm{F}{\scriptstyle\mathrm{IO}}_2$), length of stay prior to rapid response team activation, and systolic blood pressure.

Conclusions: Machine learning models can identify rapid response team patients at a high risk of mortality and potentially supplement clinical decision making. Incorporating laboratory values into model development significantly improved predictive performance in this study.

Key Words: critical care; machine learning model; mortality; rapid response team; resuscitation

Accurate prognostication in the rapid response team (RRT) setting can be difficult. The RRT faces high-acuity encounters with little previous knowledge of the patient. Yet, there are a limited number of clinical decision tools to help prognosticate patients in this setting.

Early Warning Scores are designed to identify inpatients at high risk of deterioration through changes in their clinical status or vital

[1] Division of Critical Care, Department of Medicine, University of Ottawa, Ottawa, ON, Canada.

[2] Department of Emergency Medicine, University of Ottawa, Ottawa, ON, Canada.

[3] Telfer School of Management, University of Ottawa, Ottawa, ON, Canada.

[4] Institute of Computing Science, Poznan University of Technology, Poznan, Poland.

[5] Division of Palliative Care, Department of Medicine, University of Ottawa, Ottawa, ON, Canada.

[6] Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada.

[7] Institute for Clinical and Evaluative Sciences, Ottawa, ON, Canada.

[8] Bruyere Research Institute, Ottawa, ON, Canada.

[9] Institut du Savoir Montfort, Montfort Hospital, Ottawa, 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).

The authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: prear074@uottawa.ca

Crit Care Expl 2019; 1:e0023

DOI: 10.1097/CCE.0000000000000023

signs (1). Ideally, early activation of the RRT and subsequent intervention can mitigate negative outcomes. However, performance can be variable when implementing Early Warning Scores into new or different healthcare systems (2). Machine learning models (MLMs) offer an alternative method for developing complex prediction models, and studies have shown promising results in this setting (3, 4).

Recently, Shappell et al (3) developed a novel gradient boosted decision tree MLM to predict in-hospital mortality after RRT activation. Their model was developed from data including 282,710 RRT activations across 274 hospitals (3). The model had an area under the receiver operating characteristic curve (AUC) of 0.78 and outperformed the National Early Warning Score, which had an AUC of 0.66.

However, laboratory values were not included in the analysis conducted by Shappell et al (3) and the authors hypothesized that these data would likely improve model performance. In our study, we sought to evaluate whether adding laboratory values to a similar MLM would improve predictive performance.

MATERIALS AND METHODS

Ethics approval for this study was obtained from The Ottawa Health Science Network Research Ethics Board.

Study Setting and Subjects

This study was performed at two academic hospitals within The Ottawa Hospital network (TOH). TOH contains 1,163 beds and sees over 50,000 general inpatient admissions each year. Each hospital site has a combined medical-surgical ICU with 28 beds.

Included patients were over the age of 18 years and had an RRT activation between January 1, 2015, and May 31, 2016. We excluded patients missing disposition data and activations for nonacute issues such as IV access needs.

The RRT at our center comprised of an attending intensivist, a critical care nurse, and a respiratory therapist. On weekends and overnight, the intensivist is substituted by a resident physician with the intensivist available on call. Activation criteria include abnormalities in vital signs or signs of clinical deterioration (**Appendix 1**, Supplemental Digital Content 1, http://links.lww.com/CCX/A61).

Data Collection

Patient data regarding the RRT assessment as well as previous admissions were retrieved from the Ottawa Hospital Data Warehouse. These included demographics, number of emergency department, inpatient, and ICU admissions over the preceding year, limits of care, length of stay (LOS), total number of RRT calls during admission, reason for RRT call, vital signs, laboratory values, and disposition data. Included variables are detailed in **Appendix 2** (Supplemental Digital Content 2, http://links.lww.com/CCX/A62). We used only the vital signs and laboratory values that were closest to the RRT call. We made this decision after our analysis revealed that trending (or temporal) data could not be used because of uneven granularity of the results and lengths of the time series. If patients experienced more than one activation of the RRT during their admission, data from the first activation were included for model development.

Statistical Analysis

Similar to Shappell et al (3), we developed a gradient boosted decision tree model. The base model was derived without laboratory values, and then we independently developed an augmented

TABLE 1. Baseline Characteristics of Patients Seen by the Rapid Response Team

Variables	Survived (*n* = 1,456)	Died (*n* = 605)	*p*
Age, median (IQR)	68 (56–79)	72 (63–83)	< 0.001
Male, *n* (%)	759 (52)	355(58)	0.007
Admission source, *n* (%)			0.07
Home	1,155 (79)	457 (76)	
Acute care facility transfer	135 (9)	67 (11)	
Long-term care facility transfer	68 (5)	39 (6)	
Unknown	98 (7)	42 (7)	
Emergency department visits in past year, median (IQR)	1 (0–2)	1 (0–3)	< 0.001
Hospital admissions in past year, median (IQR)	0 (0–1)	1 (0–2)	< 0.001
ICU admissions in past year, median (IQR)	0 (0–0)	0 (0–0)	0.236
Limits of care, *n* (%)			< 0.001
Full care	1,066 (73)	303 (50)	
No ICU-level care	198 (14)	193 (32)	
Do not resuscitate	146 (10)	88 (15)	
Other/unknown	46 (3)	21 (3)	

IQR = interquartile range.

model using all data, including laboratory values. For missing data, imputation was performed using the closest value to the RRT call or a median/mode value if no prior observations were available. Calibration curves (established on a hold-out sample with 50% of randomly selected patients) for the base model and the augmented model are shown in **Supplemental Figure 1** (Supplemental Digital Content 3, http://links.lww.com/CCX/A63; **legend**, Supplemental Digital Content 5, http://links.lww.com/CCX/A65). Mean-predicted value represents the output of the predictive models, interpreted as a probability of belonging to the positive outcome class (i.e., in-hospital mortality), whereas fraction of positives is the corresponding proportion of cases with positive outcome class observed in the dataset. The model was considered to be well calibrated when these two values were linearly correlated (or dependent). We evaluated both models using embedded evaluation that included a 10-fold cross-validation repeated 10 times to evaluate models' performance in predicting a primary outcome of in-hospital mortality. We conducted hyperparameter optimization for three main model parameters: depth of a decision tree, number of decision trees, and a learning rate. This parameters' optimization was conducted using a nested three-fold cross-validation to avoid model overfitting and obtain reliable results (5). We constructed average receiver operating characteristic curves over all cross-validation iterations and used them to establish model performance with an AUC value, reported with 95% CIs. We also evaluated differences between the

TABLE 2. **Call Characteristics of Rapid Response Team Activations**

Variables	Survived (*n* = 1,456)	Died (*n* = 605)	*p*
Days since admission, median (IQR)	2 (1–6)	4 (1–9)	< 0.001
Most recent vital signs			
Systolic blood pressure, mm Hg, mean (SD)	126 (31)	117 (29)	< 0.001
Diastolic blood pressure, mm Hg, mean (SD)	72 (16)	68 (16)	< 0.001
Heart rate, beats/min, mean (SD)	101 (31)	99 (29)	0.35
Temperature, °C, mean (SD)	36.9 (0.7)	36.7 (0.7)	< 0.001
Oxygen saturation, %, median (IQR)	95 (93–97)	94 (92–97)	< 0.001
Most recent blood work			
WBC count, × 10^9/L, median (IQR)	10.1 (7.1–13.9)	11.5 (7.8–17)	< 0.001
Hemoglobin, g/L, mean (SD)	108.3 (23.2)	105.4 (24.1)	0.01
Platelets, × 10^9/L, mean (SD)	230.7 (129.0)	212.5 (136.9)	0.004
Potassium, mmol/L, mean (SD)	4.03 (0.6)	4.34 (0.8)	< 0.001
Creatinine, μmol/L, median (IQR)	126 (135)	157 (153)	< 0.001
Urea, mmol/L, median (IQR)	9.0 (6.8)	13.6 (10.0)	< 0.001
Lactate, mmol/L, median (IQR)	2.6 (1.7)	3.4 (3.0)	< 0.001
Albumin, g/L, mean (SD)	27.4 (6.6)	24.8 (6.3)	< 0.001
International normalized ratio, median (IQR)	1.2 (1.1–1.3)	1.3 (1.1–1.5)	< 0.001
Reason for call, *n* (%)			< 0.001
Respiratory distress	324 (22)	186 (31)	
Tachycardia/bradycardia/arrhythmia	326 (22)	66 (11)	
Altered level of consciousness	225 (16)	122 (20)	
Hypotension	212 (15)	89 (15)	
Hypertension	36 (2)	5 (1)	
Airway concern	49 (3)	25 (4)	
Seizure	18 (1)	4 (1)	
Worried about patient	156 (11)	64 (11)	
Other/unknown	110 (8)	44 (7)	
Transferred to ICU post rapid response team, *n* (%)	394 (27)	228 (38)	< 0.001

IQR = interquartile range.

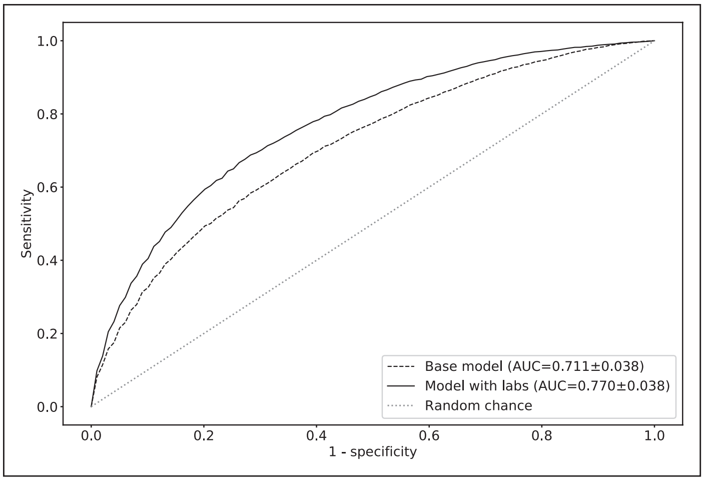

Figure 1. Receiver operating curve demonstrating performance of the base model versus the model augmented with laboratory values, compared with random chance. AUC = area under the receiver operating curve.

patients who survived to hospital discharge and those who died. Chi-square tests were used for categorical variables and Student *t* test/Wilcoxon signed rank for continuous variables. *p* values less than 0.05 were considered statistically significant. Data analysis was performed using the Anaconda environment and the gradient boosted decision tree implementation offered by XGBoost (Anaconda Inc, Austin, TX) (6).

RESULTS

There were 72,167 inpatient admissions during the study period, and 2,118 patients (2.9%) experienced an RRT activation. After applying exclusion criteria, 2,061 patients remained for the analysis (**Appendix 3**, Supplemental Digital Content 4, http://links.lww.com/CCX/A64). In-hospital mortality was 29.4%. The patients who died were older (median age, 72 vs 68 yr; $p < 0.001$) (**Table 1**). Patients who died also had a longer LOS prior to RRT assessment (4 vs 2 d; $p < 0.001$) (**Table 2**). There were differences in laboratory values for the patients who died including higher creatinine levels (median, 157 vs 126 µmol/L; $p < 0.001$) and higher lactate (median, 3.4 vs 2.6 mmol/L; $p < 0.001$).

Model performance is illustrated in **Figure 1**. The base MLM demonstrated an AUC of 0.71 (95% CI, 0.71–0.72). The top 20 most important variables in the base model, graded on a scale of 0–100, are presented in **Figure 2**. Age, estimated ratio of Pao_2 to Fio_2 (P:F ratio; calculated using oxygen saturation and estimated Fio_2), LOS prior to RRT, systolic blood pressure, and pulse were among the most important. The MLM developed with laboratory values demonstrated a statistically significant improved performance ($p < 0.001$), with an AUC of 0.77 (95% CI, 0.77–0.78). Age, platelet count, temperature, creatinine, and neutrophils were the most important variables (**Fig. 3**). Laboratory values accounted for half of the top 20 most important variables.

DISCUSSION

This was a retrospective study examining the effect of augmenting a MLM, designed to predict in-hospital mortality, with laboratory values. When laboratory values were included in the dataset used to build the model, the MLM performance significantly improved. Laboratory values accounted for 50% of the most important variables in the augmented model.

In our base model, age and vital signs ranked highly, which is consistent with prior MLMs (3, 4, 7) and with Early Warning Scores (1). LOS was also consistently important across both models as in the study by Shappell et al (3). However, compared with oxygen saturation, the estimated P:F ratio was a more important variable, reflecting the improved quantification of the degree of respiratory insufficiency when the amount of supplemental oxygen is accounted for. This may be important to consider in future development of MLMs.

Improved accuracy has been previously demonstrated when incorporating laboratory values into Early Warning Scores (4). Utility has also been demonstrated in previous MLMs (7, 8). In our study, the introduction of laboratory values, particularly platelet count, creatinine, neutrophils, and albumin, resulted in an improved AUC. Laboratory metrics are prominently featured in most commonly used ICU mortality risk scores, such as the Sequential Organ Failure Assessment and Acute Physiology and Chronic

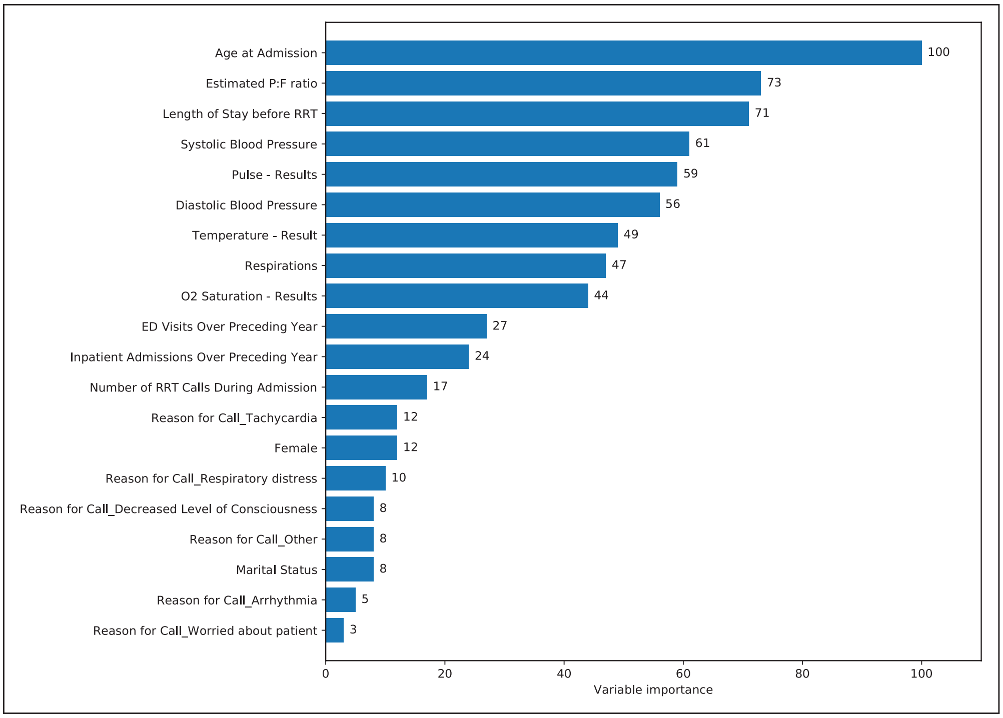

Figure 2. Importance of the predictor variables in the base machine learning model, scaled to a maximum of 100. ED = emergency department, O_2 = oxygen, P:F ratio = estimated Pao_2 to Fio_2 ratio, RRT = rapid response team.

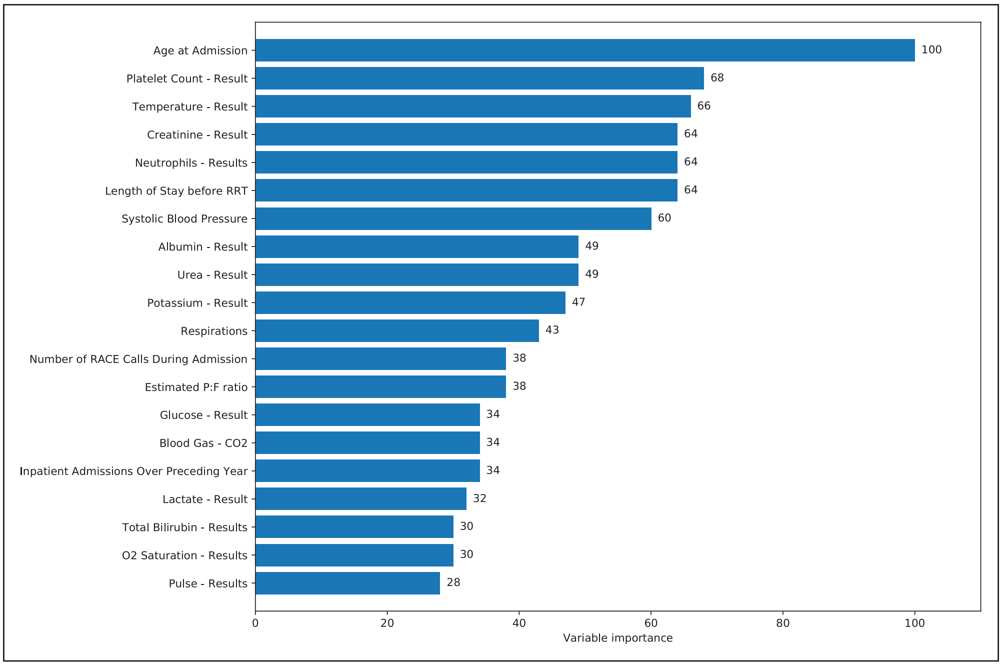

Figure 3. Importance of the predictor variables in the machine learning model augmented with laboratory values, scaled to a maximum of 100. Co_2 = carbon dioxide, O_2 = oxygen, P:F ratio = estimated Pao_2 to Fio_2 ratio; RRT = rapid response team.

Health Evaluation (APACHE) IV scores (9, 10). It is important to note, however, that the interactions between variables in the MLM are complex and nonlinear, and so the independent association with mortality of each variable in isolation is hard to quantify from the variable importance alone.

When compared with traditional predictive scoring systems such as APACHE IV, the MLM may have several advantages. For example, APACHE IV is a complex system of 129 different variables (10). The MLM allows for capturing nonlinear interactions between different variables and will also provide prediction for incomplete data. APACHE IV was also developed using an ICU population and requires input of the worst measure of a variable within the initial 24 hours of admission, potentially limiting use at the time of RRT assessment.

Although our base model did not achieve the same AUC as reported by Shappell et al (3), there were some key differences to mention. Specifically, there were a number of important variables included in their model that are not routinely collected at TOH, such as illness category, or sedation or postanesthetic care unit within 24 hours. Nevertheless, our augmented model with laboratory values performed similarly, and the statistically significant difference between our models supports the Shappell et al (3) hypothesis that laboratory values should improve model performance.

This study has several strengths. It further supports the use of a MLM to help predict in-hospital mortality in the RRT setting. It also demonstrates the importance of laboratory values, which are not typically included in Early Warning Scores. However, our study is limited by its sample size. We also could not consider clinical history such as medical comorbidities, which may improve performance. It should also be noted that although we included the patient's limits of care at the time of RRT assessment, goals of care can change throughout an admission, and limitations can also portend mortality and may have affected our results. Our study was conducted within a single hospital network and therefore generalizability is limited. Finally, although in-hospital mortality was the primary outcome of the study, future studies might consider including measures of function, such as the capability to carry out activities of daily living. Models predicting the risk of deterioration in functional status may benefit goals of care discussions with patients and caregivers.

CONCLUSIONS

Our study presents further evidence that MLMs can support in-hospital mortality prediction in the RRT setting and potentially supplement clinical decision making. Incorporating laboratory values into model development significantly improved performance in this case.

REFERENCES

1. Royal College of Physicians: National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS. Update Report of a Working Party. 2017. Available at: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2. Accessed June 28, 2019
2. 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
3. Shappell C, Snyder A, Edelson DP, et al: Predictors of in-hospital mortality after rapid response team calls in a 274 hospital nationwide sample. *Crit Care Med* 2018; 46:1041–1048
4. 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
5. Cawley GC, Talbot NL: On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010; 11:2079–2107
6. Chen T, Guestrin C: XGBoost. *In:* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. 2016. Available at: https://dl.acm.org/citation.cfm?id=2939785. Accessed June 15, 2018
7. Churpek MM, Yuen TC, Winslow C, et al: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*. 2016; 44:368–374
8. Escobar GJ, LaGuardia JC, Turk BJ, et al: Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record. *J Hosp Med* 2012; 7:388–395
9. Vincent J-L, de Mendonca A, Cantraine F, et al: Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. *Crit Care Med* 1998; 26:1793–1800
10. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310