Journal of the American Medical Informatics Association, 28(11), 2021, 2423–2432
doi: 10.1093/jamia/ocab118
Advance Access Publication Date: 17 August 2021
Research and Applications

Research and Applications

Developing machine learning models to personalize care levels among emergency room patients for hospital admission

Minh Nguyen ,[1] Conor K. Corbin,[1,*] Tiffany Eulalio,[1,*] Nicolai P. Ostberg,[1,2] Gautam Machiraju,[1] Ben J. Marafino,[1] Michael Baiocchi,[3] Christian Rose ,[4] and Jonathan H. Chen[5]

[1]Department of Biomedical Data Science, Stanford University, School of Medicine, Stanford, California, USA, [2]New York University Grossman School of Medicine, New York, New York, USA, [3]Department of Epidemiology and Population Health, Stanford University, School of Medicine, Stanford, California, USA, [4]Department of Emergency Medicine, Stanford University, School of Medicine, Stanford, California, USA and [5]Stanford Center for Biomedical Informatics Research; Division of Hospital Medicine, Department of Medicine, Stanford University, School of Medicine, Stanford, California, USA

*Authors with equal contributions.

Corresponding Author: Jonathan H. Chen, MD, PhD, Stanford University School of Medicine, 1265 Welch Rd, Medical School Office Building X213, Stanford, CA 94305, USA (jonc101@stanford.edu)

Received 1 March 2021; Revised 6 May 2021; Editorial Decision 25 May 2021; Accepted 26 May 2021

ABSTRACT

Objective: To develop prediction models for intensive care unit (ICU) vs non-ICU level-of-care need within 24 hours of inpatient admission for emergency department (ED) patients using electronic health record data.

Materials and Methods: Using records of 41 654 ED visits to a tertiary academic center from 2015 to 2019, we tested 4 algorithms—feed-forward neural networks, regularized regression, random forests, and gradient-boosted trees—to predict ICU vs non-ICU level-of-care within 24 hours and at the 24th hour following admission. Simple-feature models included patient demographics, Emergency Severity Index (ESI), and vital sign summary. Complex-feature models added all vital signs, lab results, and counts of diagnosis, imaging, procedures, medications, and lab orders.

Results: The best-performing model, a gradient-boosted tree using a full feature set, achieved an AUROC of 0.88 (95%CI: 0.87–0.89) and AUPRC of 0.65 (95%CI: 0.63–0.68) for predicting ICU care need within 24 hours of admission. The logistic regression model using ESI achieved an AUROC of 0.67 (95%CI: 0.65–0.70) and AUPRC of 0.37 (95%CI: 0.35–0.40). Using a discrimination threshold, such as 0.6, the positive predictive value, negative predictive value, sensitivity, and specificity were 85%, 89%, 30%, and 99%, respectively. Vital signs were the most important predictors.

Discussion and Conclusions: Undertriaging admitted ED patients who subsequently require ICU care is common and associated with poorer outcomes. Machine learning models using readily available electronic health record data predict subsequent need for ICU admission with good discrimination, substantially better than the benchmarking ESI system. The results could be used in a multitiered clinical decision-support system to improve ED triage.

Key words: triage, machine learning, electronic health records, medical informatics, clinical decision support, emergency medicine

INTRODUCTION

For hospitalized patients, unplanned elevation in level of care is associated with adverse outcomes such as increased morbidity and mortality.[1,2] Overestimation of level of care is associated with suboptimal use of resources,[3] which is significant during times of high critical care stress, like the COVID-19 pandemic. Most systems predicting clinical deterioration, commonly used to aid decisions around changes in the level of care, are designed for patients already hospitalized. However, the initial level of care is determined in the emergency department (ED) for most hospitalized patients, beginning the subsequent anchoring to this determination.

Each year, roughly 2 million ED visits in the United States result in intensive care unit (ICU) admissions.[4] The number of ED visits over the past 20 years has increased twice as fast as population growth.[5–8] In the ED setting, after a physician confirms a patient needs inpatient care, they triage the patient and admit them to the ICU or general wards. These triage decisions largely depend on human judgment in a high-stakes, evolving ED environment, where the patient's clinical course is most variable. Undertriaging occurs when a patient is assigned to a level of care lower than which they require, delaying eventual ICU admission and increasing mortality risk.[9,10] ICU admission delay contributes up to 44.8% of mortality risk, with each hour of delay associated with a 1.5% increase in the risk of ICU mortality.[11] A national survey indicates that 14% of unplanned ICU transfers are from non-ICUs.[12] Furthermore, most of the 80% of preventable unplanned transfers seem to result from inappropriate admission triage.[13]

Undertriaged patients also directly affect safe nurse-to-patient ratios, potentially overloading admitting nurses and nursing units. Increasing ICU capacity in response may not fully solve this resource allocation problem due to the unintentional creation of additional ICU demand.[14] Multiple factors affect accurate triaging of patients to particular levels of care. Vital signs and diagnosis make up a majority of the clinical reasoning, though gestalt and recall biases play a significant role, in addition to consultant opinion and available resources. As a result, triage decisions often are highly variable, based on limited communication and information.[15–17]

The most common triage tool in EDs across the US is the Emergency Severity Index (ESI),[16,18] which is routinely used to prioritize time-sensitive care among ED patients. It is often considered a correlate for the likelihood of hospital admission and level of care. However, this index relies heavily on clinical judgment, lacks accuracy and interrater reliability,[10,16,17,19] and is not designed to determine the care level for subsequent inpatient admissions. In addition, while the majority of hospitals report having written triaging guidelines in place for ICU transfer,[20] several studies indicate that decision-making remains rooted in qualitative clinical judgement and is not based on these written guidelines.[21,22] Naturally, this intuition-based and subjective decision-making raises concerns for unintentional biases.[23] The difficulty of ED triaging, coupled with inherent biases in decision-making, highlights the need and opportunity for computer-aided clinical decision support to manage difficult triage decisions that otherwise place undue pressure on decision-makers with potentially dire consequences for both under- or overtriaging.

While guidelines recommend that the role of the skilled physician and their multidisciplinary care team never be replaced with algorithms,[9] the lack of data-driven or quantitative decision-making tools gives rise to solutions that can aid physicians at the point of care. Rather than replacing the decision-making of the ED triage team, research and implementation efforts are focused on augmenting triage capabilities such as training guidelines[24] and machine learning prediction of diagnosis, mortality, readmission, and length of stay.[25] While predicting hospital admission is also a common application of machine learning in emergency medicine,[26–33] prediction of ICU care need for admitted adult ED patients is a burgeoning area of research.[17,19,34–37] The few studies that have been published on this topic either use national survey data,[19,34,35] include all mortality to the composite critical outcome,[17,19,34,36] or only subset adult patients in a narrow range of illness severity.[36]

Electronic health record (EHR) data offer historical information for clinical insights and allow decision-support algorithms to be updated over time.[38] Our study used EHR data for a large patient population with a full range of illness severity and expanded feature sets, focusing on admission level of care as the outcome. Using data available prior to the point of triage, we applied machine learning to develop a predictive model for clinical decision support of ED triage. The model can be applied at the beginning of the patient's hospital visit when a diagnosis may be unclear, and clinicians must rely heavily on patient presentation. The output of the model is the probability of ICU admission within 24 hours of the original inpatient admission, which can be utilized as a priority score. This score could serve as a data-driven tool to aid timely decision-making, facilitating resource utilization and improving safety, care quality, and efficiency.

OBJECTIVE

Our objective is to develop a prognostic model to predict the level of care (ICU vs non-ICU) needed within 24 hours of inpatient admission for ED patients using EHR data. Specifically, we aim to predict a patient's highest level of care within 24 hours and level of care *at* 24 hours following inpatient admission, using only data available prior to the time when the admission order was written.

MATERIALS AND METHODS

Data source and cohort

The data consist of deidentified EHR data for patient encounters from a tertiary academic hospital and Level I Trauma Center between 2015 and 2019. All adult patients 18 years or older admitted to the hospital as inpatients from the ED were included. The unit of analysis was an inpatient admission. We excluded patients who were not full code, such as do-not-resuscitate and do-not-intubate status, because level of care assignments for these patients may not correspond to their clinical presentation. Figure 1 visualizes the patient process through the ED and our predictive model pipeline.

Outcomes

The primary outcome was a patient's highest level of care within 24 hours following inpatient admission. A positive label was assigned for patients who were admitted to ICUs directly from the ED, and those initially admitted to non-ICUs (acute care units and intermediate care units) but subsequently transferred to ICUs within 24 hours. The secondary outcome was a patient's level of care at the 24th hour since inpatient admission, considering both directions of patient transfer in and out of the ICUs. Admission was defined as the time when an initial inpatient admission order was written. We

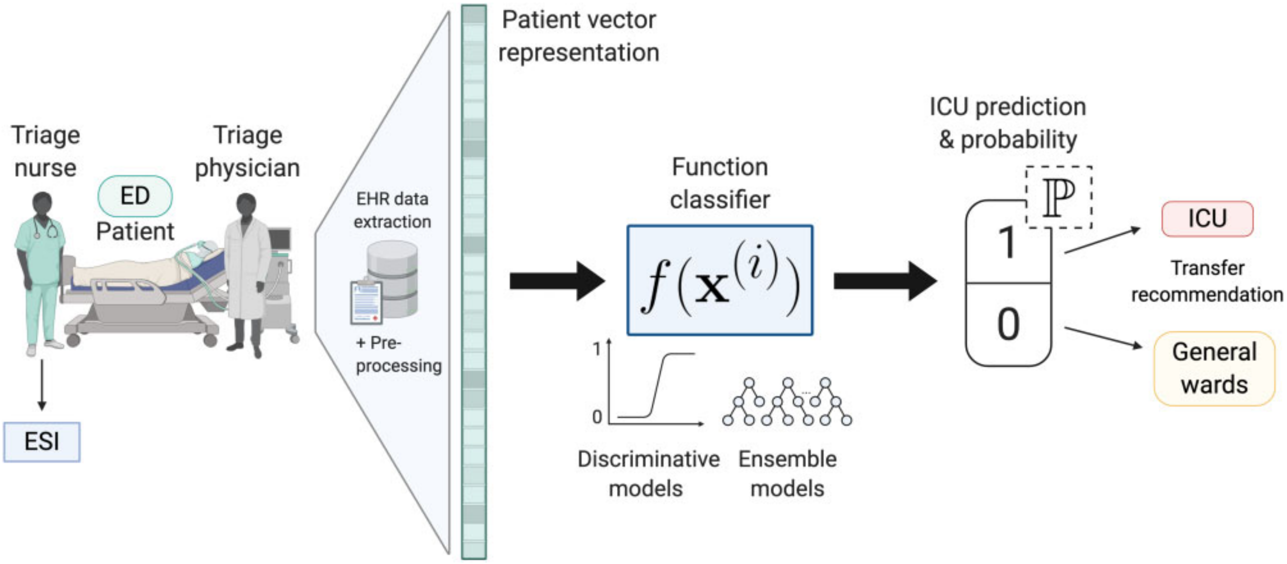

Figure 1. Schematic of patient flow through the ED and our predictive model pipeline. Created with BioRender.com.

chose 24 hours as the cutoff time, consistent with previous studies.[36,40]

Features

We extracted these features prior to admission order time: (1) patient demographics, such as Emergency Severity Index (ESI), vital signs, and laboratory test results, from the current visit; (2) count of specific microbiology orders, counts of specific imaging, procedure, medication, and lab orders within 1 prior year; and (3) count of all patient-specific historical diagnoses prior to the current visits. The ESI system was started by the US Agency for Healthcare Research and Quality and was subsequently acquired by the Emergency Nurses Association in 2019. Primarily used by triage ED nurses, the ESI is a triage score that ranges from 1 (most urgent) to 5 (least urgent) on the basis of patient acuity and resource needs.[18] Data included age, sex, race, language spoken, medical insurance provider (public or private), height, and weight. Vital signs included temperature, pulse, systolic and diastolic blood pressure, respiratory rate, and oxygen saturation. Clinical lab results included glucose, sodium, potassium, magnesium, albumin, creatinine, blood urea nitrogen, CO2, anion gap, aspartate aminotransferase, alanine transaminase, total bilirubin, platelet count, hemoglobin, white blood cell count, and absolute neutrophil count. Tables 1 and 2 include a summary of selected features. Other features are provided in Supplemental Table S1.

Statistical methods

We split data by time to account for feature drift,[41] resulting in the training dataset including all patient admissions from 2015 to 2017, whereas the validation dataset included all 2018 admissions, and the test dataset included all 2019 admissions. While prior research has opted to include only the first admission per patient to preserve model generalizability,[30,42] we relaxed this constraint by assuming that, during deployment, our model would likely encounter patients whose prior admissions were used for model training. However, to estimate generalizability to new patients, we also evaluated model performance on patients not seen during training. Models were trained under 2 regimes corresponding to different feature sets. In Regime I, we constructed a model that used relevant structured features available in EHR data. This served as an upper bound on model performance, assuming that most structured data can be easily queried to feed the model pipeline when deployed. In practice, the amount of data that could be pulled to conduct real time analysis often is limited. Therefore, in Regime II, we trained a simpler model that used only patient demographics and vital signs from the current admission. All features were recorded prior to admission order time.

Feature representation in regime I

We used the following feature types: (1) demographics and ESI; (2) lab results from the current ED visit; (3) vital signs from the current ED visit; (4) counts of historical diagnosis codes from prior admissions; (5) counts of imaging, procedure, medication, lab, and microbiology orders within the prior year. To accommodate sparsity and the wide intra- and inter-feature variances, continuous features (lab results and vital signs) were also turned into counts. This was done

Table 1. Summary of selected categorical variables

Variables		Count	Proportion
Gender			
	Female	19 961	47.9%
	Male	21 693	52.1%
Race			
	Asian	6284	15.1%
	Black	2933	7.0%
	Native American	182	0.4%
	Pacific Islander	861	2.1%
	White	21 402	51.4%
	Other	9627	23.1%
	Unknown	365	0.9%
Insurance			
	*Public**	21 263	51.0%
	Private	20 391	49.0%
Language			
	English	35 045	84.1%
	Non-English	6609	15.9%

Note: * includes publicly insured and uninsured patients.

Table 2. Summary of selected numerical variables

Variables	Mean	Standard Deviation
Age	58.2	18.5
Weight (kg)	77.2	23.1
Height (cm)	168.2	11.1
ESI	2.66	0.51
Medication count[a]	127.3	245.9
Imaging count[a]	23.8	32.6
Diagnosis count[b]	74.1	79.1
Procedure count[a]	5.7	10.9
Lab order count[a]	168.1	315.4
Microbiology order count[a]	7.4	2.9

Notes: [a]orders within 1 year prior to admission time, including both current and past visits.

[b]historical diagnosis from all prior visits, excluding current visits.

Actual values of lab results and vital signs: values for the current visits prior to admission time.

by quantizing feature distributions into decile bins, then assigning values to bins, and finally counting the frequencies of bin membership. This method naturally handles missingness by yielding count vectors of zeros over all bins if a particular numerical feature is not available.[43] The full-feature set included all 99 667 features.

Feature representation in regime II

The following features were used: (1) demographics and ESI; (2) first and last available measurements of vital signs from the current ED visit; (3) summary statistics (count, min, max, mean, median, standard deviation, median absolute deviation, interquartile range, difference between first and last values) of all available vital signs.

Supplemental Figure S1 shows our data processing pipeline for both feature regimes.

There were only small percentages of missing values for ESI (4.0%), height (3.9%), and weight (0.85%). We applied the method of multivariate imputation by chained equations (MICE) to impute these missing data. This method is flexible and often used to account for the uncertainty in the imputations by imputing the missing values many times with predicted values to create different "complete" datasets.[44] Indicators for missingness were also added as features.

We applied 4 machine learning algorithms to predict ICU care needs: (1) elastic net regularized logistic regression, (2) random forests, (3) gradient-boosted trees, and (4) 4-layer perceptron feedforward neural networks. These most used algorithms have been shown to perform well on large EHR data and are well understood.[45] For each algorithm type, a grid search was used during model training to tune algorithm-specific hyperparameters (Supplemental Table S2). The best tuned hyperparameters for each algorithm were selected using prediction results from the validation data set. The 4 final models were then evaluated and compared on the holdout test set to select the single best-performing model.

We used 2 primary metrics for model performance evaluation: Area Under the Receiver Operating Characteristic (AUROC) curve and Area Under the Precision-Recall Curve (AUPRC). The ROC curve summarizes the trade-off between true positive rate (sensitivity) and true negative rate (specificity) for a predictive model using different discrimination thresholds. However, AUROC can be misleading, especially with highly imbalanced data where the positive class is rare.[46] Hence, AUPRC can aid in interpreting model performance characteristics, as it is less sensitive to the true negative rate. The precision-recall curve distinguishes the trade-off between the true positive rate (recall or sensitivity) and the positive predictive value (precision). We calculated 95% confidence intervals (CIs) for the AUROC and AUPRC estimates based on 2000 bootstrap replicates.

Because we split the data based on time, the same patients could appear multiple times in the data if they had multiple admissions. We evaluated the model performance on all patient admissions and also on patients present only in the test set. Finally, given the large number of detailed features, we performed ablation studies to assess the relative feature importance by removing 1 feature type from the feature set at a time. Under Regime I, the following feature types were removed 1 at a time: (1) demographics, including height, weight, and ESI; (2) vital signs; (3) medications; (4) imaging; (5) diagnosis; (6) procedure; (7) lab orders; and (8) lab results. Under Regime II, we only removed the summary statistics of all vital signs for the ablation study.

Finally, because ESI is the triage score assigned when a patient is first assessed in the ED by triage nurses, we hypothesized that it can be informative in predicting the outcomes. ESI scores are used to guide treatment priority in the ED,[47] not for inpatient admissions. We carried out a univariate logistic regression with ESI as the sole predictor as a benchmark method and ultimately included it in the main model as well.

The study was conducted and is reported in accordance with the TRIPOD statement.[39] Our study was approved by the institutional review board of the Stanford University School of Medicine.

RESULTS

Our cohort consisted of 41 654 distinct patient admissions, of which 70% were from unique patients. In the holdout test set of 10 096 admissions, 75% of patients were not in the training and validation sets. On average, it took 4.9 hours (standard deviation of 6.6 hours) from the time patients were assigned to the ED service until they were admitted as inpatients to the hospital. We identified 5568 patient admissions, representing 13.4% of the cohort, who were admitted to the ICUs at some point within 24 hours following inpatient admissions. Twenty-three non-ICU patients died and 845 initially admitted to non-ICUs were subsequently transferred to ICUs within 24 hours. When we considered both directions of patient transfer in and out of the ICUs, there were 3892 patient visits, about 9.3% of the cohort, in the ICUs at the 24th hour after admission. Supplemental Figures S2 and S3 show the care level trajectories and cumulative number of ICU transfers over time.

Model performance

The AUROC and AUPRC from the 4 competing algorithm types, using both the full-feature (Regime I) and simple-feature (Regime II) on the test set, are displayed in Tables 3 and 4 for comparison. All algorithm types appeared to perform comparably well. Gradient-boosted trees consistently performed better or just as well as other algorithm types. Hence, the following reported results are from the best gradient-boosted tree model using different feature sets and with 2 different outcome labels.

Table 5 summarizes the results on the holdout test set for all patient visits in 2019, including ablation studies using the best tuned gradient-boosted tree models under both feature regimes and for both outcomes. For the primary outcome, the full-feature model achieved an AUROC of 0.88 (95% CI: 0.87–0.89) and an AUPRC

Table 3. Area under the receiving characteristic curve (AUROC) with 95% confidence intervals, to compare and select the best model

Models	Primary outcome (highest care level within 24 hours)		Secondary outcome (care level at the 24th hour)	
	Full feature	Simple feature	Full feature	Simple feature
Gradient Boosting*	*0.88 (0.87–0.89)*	*0.82 (0.80–0.83)*	*0.86 (0.85–0.87)*	*0.81 (0.79–0.82)*
Random Forest	0.86 (0.85–0.87)	0.78 (0.77–0.80)	0.84 (0.82–0.85)	0.78 (0.76–0.79)
Logistic Regression (elastic net)	0.84 (0.82–0.85)	0.79 (0.77–0. 80)	0.82 (0.80–0.83)	0.78 (0.77–0.80)
Feed-Forward Neural Networks	0.85 (0.83–0.86)	0.77 (0.76–0.79)	0.82 (0.81–0.84)	0.78 (0.76–0.79)

Notes: The baseline of AUROC is fixed at 0.5, which is equal to random guessing.
*Best models with highest AUROC .

Table 4. Area under the precision-recall curve (AUPRC) with 95% confidence intervals, to compare and select the best model

Models	Primary outcome		Secondary outcome	
	(highest care level within 24 hours)		(care level at the 24th hour)	
	Baseline AUPRC = 0.13		Baseline AUPRC = 0.09	
	Full feature	Simple feature	Full feature	Simple feature
Gradient Boosting*	*0.65 (0.63–0.68)*	*0.52 (0.50–0.56)*	*0.50 (0.47–0.54)*	*0.41 (0.39–0.45)*
Random Forest	0.59 (0.57–0.62)	0.49 (0.47–0.52)	0.46 (0.42–0.49)	0.39 (0.36–0.42)
Logistic Regression (elastic net)	0.52 (0.49–0.55)	0.46 (0.44–0.49)	0.39 (0.37–0.43)	0.36 (0.34–0.40)
Feed-Forward Neural Networks	0.56 (0.53–0.59)	0.45 (0.43–0.49)	0.42 (0.39–0.45)	0.38 (0.36–0.42)

Notes: The baseline of AUPRC is equal to the fractions of positive cases for each outcome.
*Best models with highest AUPRC.

Table 5. Evaluation results from the best model with ablation studies for both outcomes

GBM models		Primary outcome		Secondary outcome	
		(highest care level within 24 hours)		(care level at the 24th hour)	
		AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
Regime I	Full-feature[a]	0.88 (0.87–0.89)	0.65 (0.63–0.68)	0.86 (0.85–0.87)	0.50 (0.47–0.53)
	(-) Demographics & ESI	0.88 (0.87–0.89)	0.64 (0.62–0.67)	0.86 (0.84–0.87)	0.49 (0.47–0.53)
	(-) Vital signs[b]	*0.85 (0.84–0.86)*	*0.60 (0.58–0.63)*	*0.83 (0.82–0.84)*	*0.45 (0.43–0.49)*
	(-) Meds	0.88 (0.87–0.89)	0.63 (0.61–0.66)	0.85 (0.84–0.86)	0.48 (0.45–0.52)
	(-) Imaging	0.88 (0.87–0.89)	0.64 (0.62–0.67)	0.85 (0.84–0.86)	0.49 (0.46–0.53)
	(-) Diagnosis Codes	0.88 (0.87–0.89)	0.65 (0.63–0.67)	0.86 (0.85–0.87)	0.49 (0.47–0.53)
	(-) Procedures	0.88 (0.87–0.89)	0.65 (0.63–0.68)	0.86 (0.84–0.87)	0.50 (0.47–0.54)
	(-) Lab orders	0.87 (0.86–0.88)	0.63 (0.61–0.66)	0.87 (0.86–0.88)	0.48 (0.46–0.52)
	(-) Lab results	0.88 (0.87–0.89)	0.64 (0.63–0.67)	0.88 (0.87–0.89)	0.48 (0.46–0.52)
Regime II	Simple-feature[a]	0.82 (0.80–0.83)	0.52 (0.50–0.56)	0.81 (0.79–0.82)	0.41 (0.38–0.45)
	(-) Vitals summary[b]	0.75 (0.73–0.76)	0.41 (0.39–0.44)	0.74 (0.73–0.76)	0.32 (0.29–0.35)
ESI-only logistic regression		0.67 (0.65 - 0.70)	0.37 (0.35–0.40)	0.67 (0.65–0.70)	0.28 (0.26–0.31)
		(intercept: 1.022; coefficient: −1.143)		*(intercept: 0.774; coefficient: −1.209)*	

Notes: [a]indicates models without removing any feature types.
[b]Models with vital signs as the feature type removed had the most reduction in AUROC and AUPRC. Vital signs are the most important predictors.

of 0.65 (95% CI: 0.63–0.68). The full-feature model outperformed the simple-feature model, which achieved an AUROC of 0.82 (95% CI: 0.80–0.83) and an AUPRC of 0.52 (95% CI: 0.50–0.56). Using a threshold such as 0.6, our best model had an 85% positive predictive value (precision), 89% negative predictive value, 30% true positive rate (recall or sensitivity), and 99% specificity (true negative rate). For the secondary outcome, the full-feature model achieved an AUROC of 0.86 (95% CI: 0.85–0.87) and an AUPRC of 0.50 (95% CI: 0.47–0.54). It outperformed the simple-feature model, which achieved an AUROC of 0.81 (95% CI: 0.79–0.82) and an AUPRC of 0.41 (95% CI: 0.39–0.45).

Additionally, when using only new patients on the test set that were not seen during model building, both models achieved almost identical AUROC and AUPRC values regardless of the outcomes or feature sets used (Supplemental Table S3). Under Regime II, using patient summary data on vital signs improved model performance significantly compared to using only the first and last available vital sign measurements. Similarly, under Regime I, vital signs were the

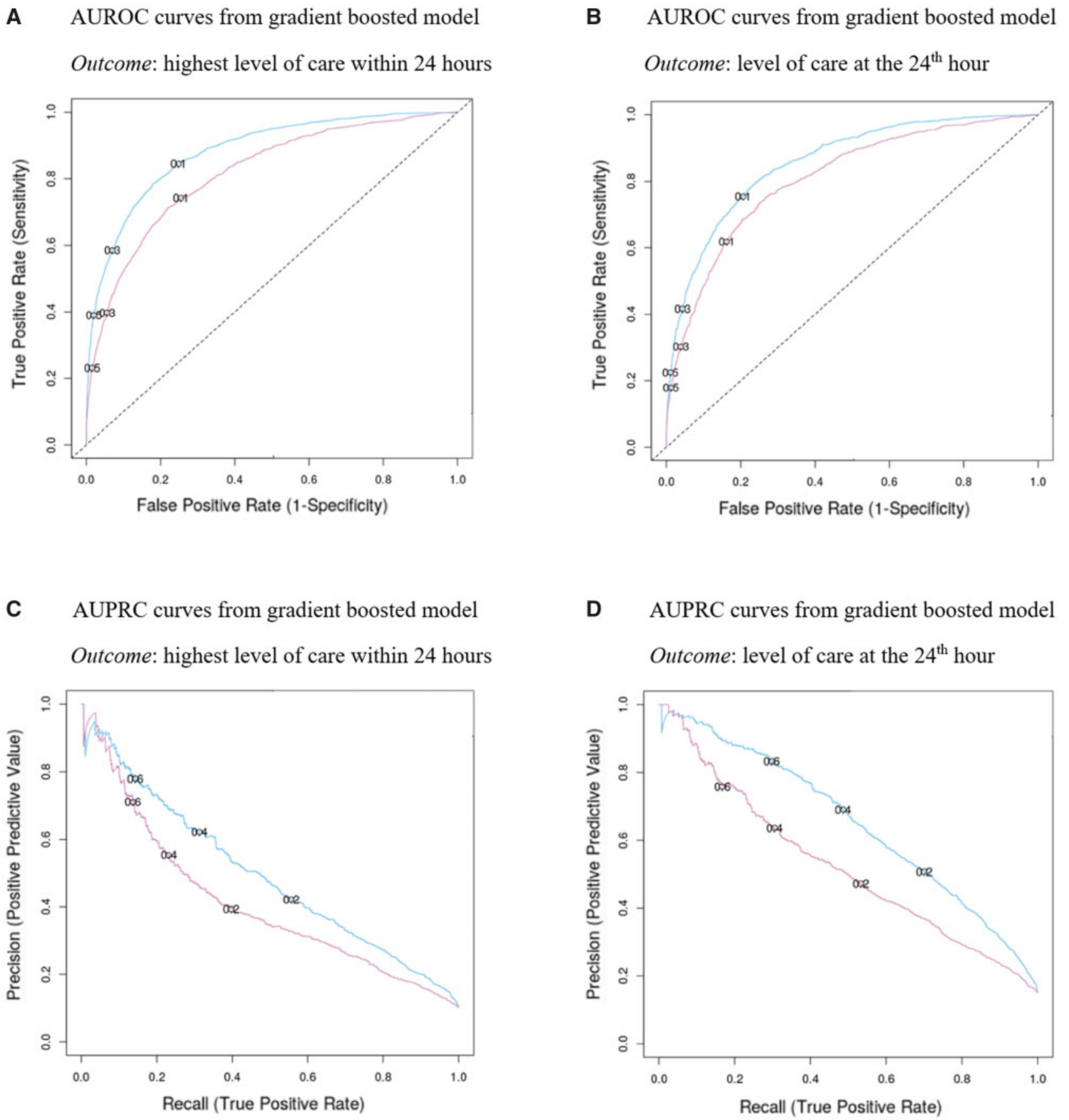

Figure 2. Evaluation results from the best model and ablation studies.

only stand-alone feature type that significantly impacted the model predictive performance. Although the predictive power of medications, imaging orders, and labs individually is insignificant, together they appeared to noticeably improve model performance compared to other feature types.

Figure 2 shows the ROC and PR curves from evaluation results on the test set, using the best tuned gradient-boosted tree in both feature regimes and for both outcomes. Figure 3 shows the calibration plots on test data, showing how consistent the predicted probabilities are with observed ICU admission percentage. Finally, our data also showed that ICU vs non-ICU admissions followed similar distributions of ESI. The logistic regression model with ESI as the only predictor suggested a negative association between the ESI and the predicted probability of ICU admission (Figure 4). With AUROC of 0.67 (95% CI: 0.65–0.70) for both outcomes and AUPRC of 0.37 (95% CI: 0.35–0.40) and 0.28 (95% CI: 0.26–0.31) for the primary and secondary outcomes, respectively, ESI is far less predictive of ICU triage than any of our data-driven models.

Analysis of prediction errors:

We inspected the most erroneous predicted values (less than 0.2) from the best models under both Regimes I and II for patients in the test set, using the primary outcome. Among patients whose first in-

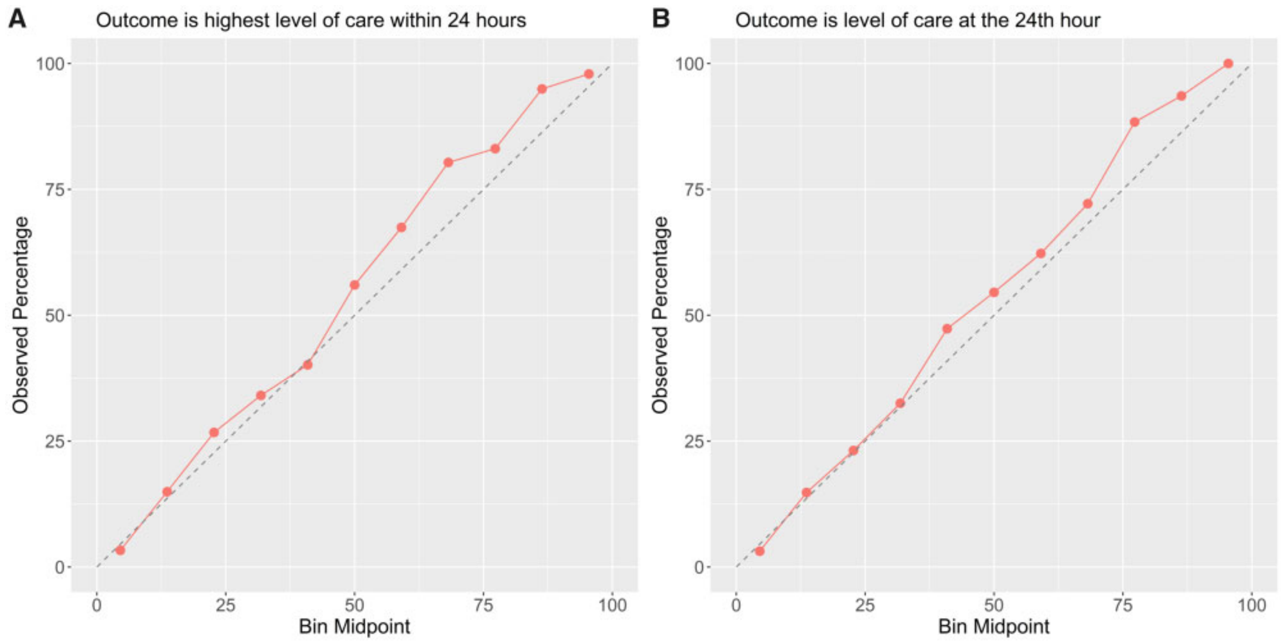

Figure 3. Calibration plots for class probabilities predicted by the best models for both outcomes.

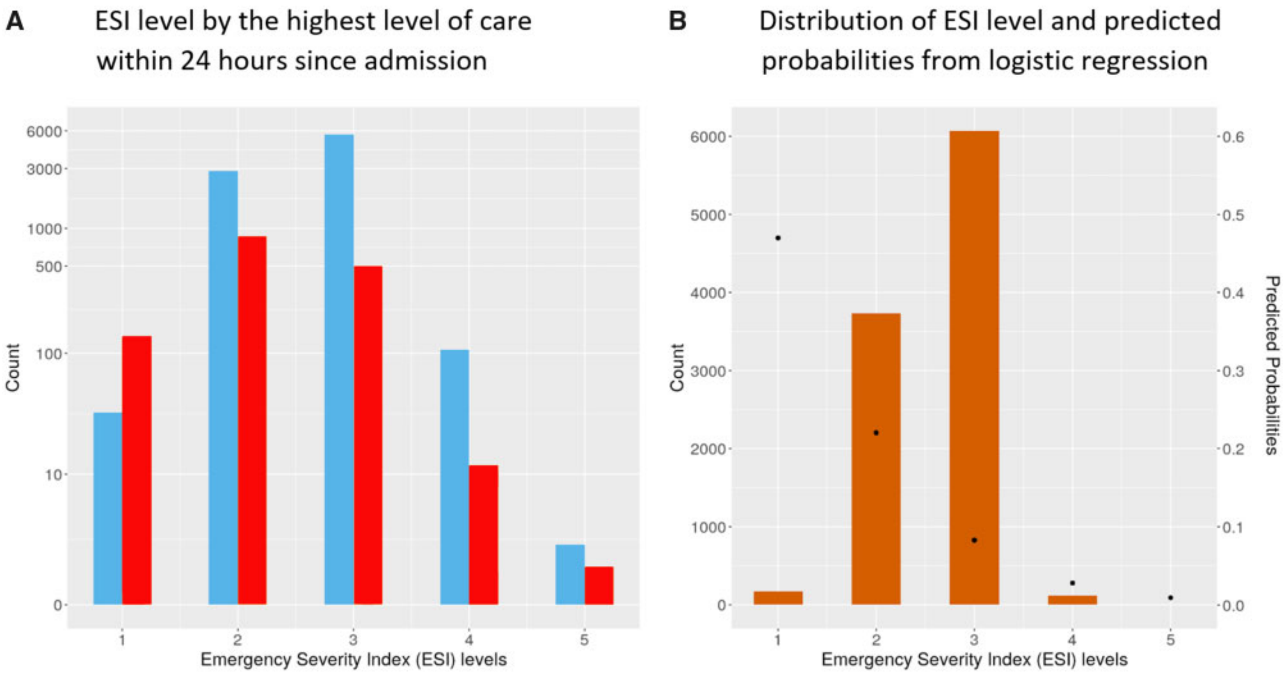

Figure 4. ESI level by highest level of care within 24 hours since admission and the average predicted probabilities for each ESI level.

patient admitting physicians were not critical care services, the most erroneous predicted values were for patients from neurosurgery, pulmonary hypertension, and trauma services. Patients from these services accounted for 10% to 18% of the errors, compared to the next most erroneous service, general surgery, of 5%. Furthermore, we reviewed patients' level of care trajectories within the first 24 hours following inpatient admission. Among patients with positive labels, those whose levels of care were escalated, instead of downgraded, were 1.5 times more likely to be incorrectly predicted. These trajectories include: acute to critical, acute to intermediate to critical, and intermediate to critical. Among the negative labels, 36% of the erroneous predictions were from patients who were admitted to intermediate ICUs.

DISCUSSION

Appropriate designation of level of care from the ED is of critical importance. Our model appeared to accurately discriminate patients requiring ICU care from those who did not, which could help identify higher risk patients and improve decision-making. The gradient-boosted tree model, using the full-feature set, performed best out of all algorithm-feature pairs. When comparing the 2 outcomes, prediction results were better for the primary outcome than the secondary outcome (AUROC of 0.86 vs 0.81 and AUPRC of 0.65 vs 0.52). Comparing results between Regime I vs Regime II, as well as within the ablation studies, vital signs were the most important predictors overall.

Analysis of cases of prediction error indicated greater uncertainty amongst patients initially admitted to certain specialty services. Some specialty services have admission protocols which supersede clinical decision-making based on specific disease severity measures, which may have affected this result. For example, patients with otherwise stable gastrointestinal bleeding may be admitted to the ICU within 24 hours from admission mostly to perform endoscopy. Notably, although a lower ESI was associated with worse outcomes, its ability to predict the assigned level of care was much more limited, highlighting the need for a separate clinical index or decision support tool to aid physicians in triaging inpatient admissions after initial assessment, evaluation, and stabilization.

To implement our predictive models, a potential approach would be to categorize the predicted probabilities into 3 zones as proposed by O'Brien: "red," "yellow," and "green."[48] This approach corresponds to 2 probability thresholds: setting 1 at 0.6 achieves 85% precision overall. Patients with values above this threshold are considered to be in the "red" zone and have the highest risk of ICU admission. In addition, setting a threshold at 0.2 would yield a sensitivity of 75% and a precision of 47%. Patients with predicted scores between 0.2 to 0.6 could be considered in the "yellow" zone, capturing potential ICU admissions that may require further evaluation. The "green" zone consists of patients with predicted probabilities under 0.2, representing those who are most likely safe to be admitted to the general wards. These thresholds can be customized along the precision-recall curve with respect to a setting's risk tolerance and resource constraints. Such a multitiered alert system can support admitting physician decision processes, while reducing alarm fatigue. Since vital signs were the most significant contributor, the simple-feature model can potentially be extended to inform providers about patient acuity throughout the hospital course. However, as with many similar scoring tools, such a system could inadvertently lead to anchoring bias or increased testing or spending on care to attempt to address high-risk patients. Care must be given to make sure that the tool is used appropriately.

A key limitation to the generalizability of this study is due to changes in standards of care and guidelines for admission. Protocols and thresholds for level of care assignments vary depending on hospital capacity, unit level nursing structure, patient acuity, and case mix. For example, patients requiring ICU care at higher-level hospitals might be more ill than those at smaller, lower-level hospitals. Larger hospitals also have more mid-level and specialized units where guidelines and admission criteria could exhibit more variation in these settings. Additionally, our cohort demographics, such as race, might not be representative of the general population. For optimal performance, models should be retrained for different populations to reflect potentially different distributions and optimized hyperparameters that may not translate readily across sites and time. Determining the appropriate thresholds for a multitiered alert system requires assessment of the complete risk distribution since individual patient risks are relative to their specific patient population. Our results illustrate the potential of this process, while our open-source code base allows for the reproduction and tailoring of models to specific applied areas.

Finally, another limitation lies in the observational nature of our study. The EHR data used involves practical data issues including missingness, entry errors, and inconsistent coding practices. In addition, due to the deidentified nature of the data, we could not access information regarding instantaneous ICU bed availability to better characterize the undertriaging problem. Bed availability, considered a resource constraint, can be informative in this analysis, as ED patients may be undertriaged as a matter of necessity if ICU beds are not available. Future studies incorporating bed availability could be fruitful, as it is a strong instrumental variable that can be used to estimate the causal effect of an ICU admission for ED patients.

CONCLUSION

Improper triaging is influenced by many factors, such as time- and information-limited ED environment, heterogeneous decision-making among physicians, and a lack of quantitative clinical decision-support tools. Ultimately, these triage decisions for inpatient admissions influence patient outcomes and the dynamics of hospital resources. Our proposed prediction models demonstrated good discrimination for ICU vs non-ICU level of care within 24 hours and at the 24th hour after initial admission and substantially better than the ESI system. Outputs of such models could be used as priority scores to aid triage decision-making and improving the efficiency and quality of emergency hospital care.

FUNDING

This research was supported in part by: NIH/National Library of Medicine via Award R56LM013365, Gordon and Betty Moore Foundation through Grant GBMF8040, and Stanford Clinical Excellence Research Center (CERC). GM is supported by the NLM training grant 5T32LM012409. MN, TE, and BJM are supported by the NLM training grant 5T15LM007033. CC is supported by the National Science Foundation Graduate Research Fellowship grant DGE-1656518.

AUTHOR CONTRIBUTIONS

MN, MB, and JHC conceived the study. MN, CKC, and TE queried and processed the data. MN, CKC, TE, NPO, and GM performed statistical analyses and implemented the algorithm. MN drafted the initial manuscript. All authors contributed to the study and analysis design, critically revised, and reviewed the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY STATEMENT

The electronic health records data underlying this article were provided by STAnford medicine Research data Repository (STARR) (https://med.stanford.edu/starr-tools.html). The data can be accessed for research purposes after Institutional Review Board approval via the Stanford Research Informatics Center.

ACKNOWLEDGMENTS

We thank Dr. Russ Altman, Dr. Erika Strandberg, Laurynas Kalesinskas, and Dr. Benjamin Goldstein for their valuable feedback during the initial phase of this work. We thank the anonymous reviewers for their valuable comments that helped us improve the manuscript.

This rsearch used data or services provided by STARR, STAnford medicine Research data Repository, a clinical data warehouse containing live Epic data from Stanford Health Care (SHC), the University Healthcare Alliance (UHA), and Packard Children's Health Alliance (PCHA) clinics and other auxiliary data from hospital applications, such as radiology PACS. The STARR platform is developed and operated by the Stanford Medicine Research IT team and is made possible by the Stanford School of Medicine Research Office.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or Stanford Health Care.

CONFLICT OF INTEREST STATEMENT

JHC is a co-founder of Reaction Explorer LLC that develops and licenses organic chemistry education software and has received paid consulting or speaker fees from the National Institute of Drug Abuse Clinical Trials Network, Tuolc Inc., Roche Inc., and Younker Hyde MacFarlane PLLC.

CR is a co-founder of MyVitae, a physician credential management platform, and has received consulting fees from Caption Inc. to evaluate the utility of point-of-care-ultrasound AI.

CKC has received consulting fees from Fountain Therapeutics Inc.

REFERENCES

1. 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 (3): 224–30.
2. Sutherland ME, Yarmis SJ, Lemkin DL, *et al.* National early warning score is modestly predictive of care escalation after emergency department-to-floor admission. *J Emerg Med* 2020; 58 (6): 882–91.
3. Orsini J, Blaak C, Yeh A, *et al.* Triage of patients consulted for ICU admission during times of ICU-bed shortage. *J Clin Med Res* 2014; 6 (6): 463–8.
4. Rui P, Kang K. National hospital ambulatory medical care survey: 2017 emergency department summary tables. National Center for Health Statistics; 2017: 37. https://www.cdc.gov/nchs/data/nhamcs/web_tables/2017_ed_web_tables-508.pdf. Accessed January 5, 2021.
5. Niska R, Bhuiya F, Xu J. National hospital ambulatory medical care survey: 2007 emergency department summary. *Natl Health Stat Rep* 2010; (26): 1–31.
6. Tang N, Stein J, Hsia RY, *et al.* Trends and characteristics of US emergency department visits, 1997-2007. *JAMA* 2010; 304 (6): 664–70.
7. Greenwood-Ericksen MB, Kocher K. Trends in emergency department use by rural and urban populations in the United States. *JAMA Netw Open* 2019; 2 (4): e191919.doi:10.1001/jamanetworkopen.2019.1919
8. Agency for Healthcare Research and Quality. Trends in Emergency Department Visits, 2006–2014 #227. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb227-Emergency-Department-Visit-Trends.jsp?utm_source=ahrq&utm_medium=en1&utm_term=&utm_content=1&utm_campaign=ahrq_en10_24_2017. Accessed January 5, 2021.
9. Blanch L, Abillama FF, Amin P, *et al.* Council of the World Federation of Societies of Intensive and Critical Care Medicine. Triage decisions for ICU admission: report from the Task Force of the World Federation of Societies of Intensive and Critical Care Medicine. *J Crit Care* 2016; 36: 301–5.
10. Hinson JS, Martinez DA, Schmitz PSK, *et al.* Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. *Int J Emerg Med* 2018; 11 (1): 3.
11. Cardoso LT, Grion CM, Matsuo T, *et al.* Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. *Crit Care* 2011; 15 (1): R28.
12. Angus DC, Shorr AF, White A, *et al.* Critical care delivery in the United States: distribution of services and compliance with Leapfrog recommendations. *Crit Care Med* 2006; 34 (4): 1016–24.
13. 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 (2): 68–72.
14. Rice TH, Labelle RJ. Do physicians induce demand for medical services? *J Health Polit Policy Law* 1989; 14 (3): 587–600.
15. Nates JL, Nunnally M, Kleinpell R, *et al.* ICU admission, discharge, and triage guidelines: a framework to enhance clinical operations, development of institutional policies, and further research. *Crit Care Med* 2016; 44 (8): 1553–602.
16. Mistry B, Stewart De Ramirez S, Kelen G, *et al.* Accuracy and reliability of emergency department triage using the emergency severity index: an international multicenter assessment. *Ann Emerg Med* 2018; 71 (5): 581–7.e3.
17. Levin S, Toerper M, Hamrock E, *et al.* Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann Emerg Med* 2018; 71 (5): 565–74.e2.
18. Emergency Severity Index (ESI): A Triage Tool for Emergency Departments. http://www.ahrq.gov/professionals/systems/hospital/esi/index.html. Accessed April 30, 2020.
19. Raita Y, Goto T, Faridi MK, *et al.* Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care Lond Care* 2019; 23 (1): 64.
20. Walter KL, Siegler M, Hall JB. How decisions are made to admit patients to medical intensive care units (MICUs): a survey of MICU directors at academic medical centers across the United States. *Crit Care Med* 2008; 36 (2): 414–20.
21. Azoulay E, Pochard F, Chevret S, *et al.* Compliance with triage to intensive care recommendations. *Crit Care Med* 2001; 29 (11): 2132–6.
22. Graf J, Janssens U. Still a black box: What do we really know about the intensive care unit admission process and its consequences? *Crit Care Med* 2005; 33 (4): 901–3.
23. Truog RD, Brock DW, Cook DJ, *et al.*; Task Force on Values, Ethics, and Rationing in Critical Care (VERICC). Rationing in the intensive care unit. *Crit Care Med* 2006; 34 (4): 958–63. quiz 971.
24. Oredsson S, Jonsson H, Rognes J, *et al.* A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scand J Trauma Resusc Emerg Med* 2011; 19: 43.
25. Shafaf N, Malek H. Applications of machine learning approaches in emergency medicine; a review article. *Arch Acad Emerg Med* 2019; 7 (1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6732202/.Accessed April 29, 2020.
26. Peck JS, Gaehde SA, Nightingale DJ, *et al.* Generalizability of a simple approach for predicting hospital admission from an emergency department. *Acad Emerg Med Off Med* 2013; 20 (11): 1156–63.
27. Cameron A, Rodgers K, Ireland A, *et al.* A simple tool to predict admission at the time of triage. *Emerg Med J* 2015; 32 (3): 174–9.
28. Zlotnik A, Alfaro MC, Pérez MCP, *et al.* Building a decision support system for inpatient admission prediction with the Manchester triage system and administrative check-in variables. *Comput Inform Nurs* 2016; 34 (5): 224–30.
29. Golmohammadi D. Predicting hospital admissions to reduce emergency department boarding. *Int J Prod Econ* 2016; 182: 535–44.
30. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *Plos ONE* 2018; 13 (7): e0201016.
31. Graham B, Bond R, Quinn M, *et al.* Using data mining to predict hospital admissions from the emergency department. *IEEE Access* 2018; 6: 10458–69.
32. Araz OM, Olson D, Ramirez-Nafarrate A. Predictive analytics for hospital admissions from the emergency department using triage information. *Int J Prod Econ* 2019; 208: 199–207.
33. Parker CA, Liu N, Wu SX, *et al.* Predicting hospital admission at the emergency department triage: a novel prediction model. *Am J Emerg Med* 2019; 37 (8): 1498–504.

34. Dugas AF, Kirsch TD, Toerper M, *et al.* An electronic emergency triage system to improve patient distribution by critical outcomes. *J Emerg Med* 2016; 50 (6): 910–8.

35. Kwon J, Lee Y, Lee Y, *et al.* Validation of deep-learning-based triage and acuity score using a large national dataset. *PLoS ONE* 2018; 13 (10): e0205836.

36. Fernandes M, Mendes R, Vieira SM, *et al.* Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PLoS ONE* 2020; 15 (3): e0229331.

37. Miles J, Turner J, Jacques R, *et al.* Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagn Progn Res* 2020; 4 (1): 16.

38. Ye C, Wang O, Liu M, *et al.* A real-time early warning system for monitoring inpatient mortality risk: prospective study using electronic medical record data. *J Med Internet Res* 2019; 21 (7): e13719.

39. Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594.

40. Leong J, Madhok J, Lighthall GK. Mortality of patients requiring escalation to intensive care within 24 hours of admission in a mixed medical-surgical population. *Clin Med Res* 2020; 18 (2–3): 68–74.

41. Barddal JP, Gomes HM, Enembreck F, *et al.* A survey on feature drift adaptation: definition, benchmark, challenges and future directions. *J Syst Softw* 2017; 127: 278–94.

42. Hong WS, Haimovich AD, Taylor RA. Predicting 72-hour and 9-day return to the emergency department using machine learning. *JAMIA Open* 2019; 2 (3): 346–52.

43. Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; 1 (1): 18.

44. Buuren S. V, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *J Stat Softw* 2011; 45 (1): 1–67.

45. Uddin S, Khan A, Hossain ME, *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* 2019; 19 (1): 281.

46. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 2015; 10 (3): e0118432.

47. Dong SL, Bullard MJ, Meurer DP, *et al.* The effect of training on nurse agreement using an electronic triage system. *CJEM* 2007; 9 (4): 260–6.

48. O'Brien C, Goldstein BA, Shen Y, *et al.* Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. *MDM Policy Pract* 2020; 5 (1): 2381468319899663.