



# MONITOR: a multi-domain machine learning approach to predicting in-hospital mortality

Christina C. Guerrier<sup>1^</sup>, Salvatore J. D'Acunto<sup>1</sup>, Guillaume P. L. Labilloy<sup>1^</sup>, Rhemar A. Esma<sup>2</sup>, Heather A. Kendall<sup>2,3^</sup>, Daniel A. Norez<sup>1</sup>, Jennifer N. Fishe<sup>1^</sup>

<sup>1</sup>Center for Data Solutions, College of Medicine, University of Florida, Jacksonville, FL, USA; <sup>2</sup>Quality Management, UF Health-Jacksonville, Jacksonville, FL, USA; <sup>3</sup>Quality Management, Brooks Rehabilitation, Jacksonville, FL, USA

*Contributions:* (I) Conception and design: CC Guerrier, RA Esma, HA Kendall, GPL Labilloy, JN Fishe; (II) Administrative support: CC Guerrier, JN Fishe; (III) Provision of study materials or patients: CC Guerrier, JN Fishe; (IV) Collection and assembly of data: SJ D'Acunto, CC Guerrier, JN Fishe; (V) Data analysis and interpretation: SJ D'Acunto, CC Guerrier, GPL Labilloy, DA Norez, JN Fishe; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* Christina C. Guerrier, MBA. Center for Data Solutions, College of Medicine, University of Florida, Jacksonville, FL 32209, USA. Email: christina.guerrier@jax.ufl.edu.

**Background:** Current machine-learning (ML) models have been developed to predict mortality for specific diseases, procedures, and setting at a given time; however, the risk for in-hospital mortality changes throughout a patient's hospital stay. A model that could predict in-hospital mortality throughout a patient's stay regardless of disease or procedure could improve clinical outcomes.

**Methods:** We conducted a prognostic study where cohorts were created from electronic health records (EHR) with encounters between January 1, 2014 and January 30, 2020 at tertiary academic hospital and community hospital. The initial dataset contained 228,405 patients. EHR of 176,526 patients remained in the study after adjusting for age (18 or older), length of stay (LOS) (between 0 and 365 days), and encounter dates within study period. Training and testing cohorts, stratified by length-of-stay and in-hospital mortality, were created with an 80/20 split.

**Results:** The study included 176,526 patients {mean [interquartile range (IQR)] age of 52.2 [34–68] years; 55.3% female, 63.7% white, 92.7% non-Hispanic} who were admitted for 5.6 [2–6] days. The in-hospital mortality rate for the training and testing cohorts was 3.0%. The CatBoost classifier model, trained with a combination of undersampling and oversampling, demonstrated a F2 score of 0.510 [95% confidence intervals (CI): 0.496–0.516]. The F2 score is highest for patients with a one-day LOS (0.811; 95% CI: 0.776–0.843). Even though the F2 score is lower for patients who stayed more than a day, the F2 score generally increases each day until the day of discharge or mortality.

**Conclusions:** This study investigated an ML model that predicted risk of in-hospital mortality regardless of patient demographics and level of care setting. The model accounted for changes in patient condition throughout the LOS. An implementation study should be conducted to determine how this model can be integrated into clinical workflow to support decision making.

**Keywords:** Machine-learning (ML); in-hospital mortality; inpatient mortality; prediction; risk stratification

Received: 21 September 2021; Accepted: 21 January 2022; Published: 30 March 2022.

doi: 10.21037/jmai-21-28

View this article at: <https://dx.doi.org/10.21037/jmai-21-28>

<sup>^</sup> ORCID: Christina C. Guerrier, 0000-0002-8198-2047; Guillaume P. L. Labilloy, 0000-0002-9440-4211; Heather A. Kendall, 0000-0002-4671-8363; Jennifer N. Fishe, 0000-0001-9037-8143.

## Introduction

A multitude of tasks, actions, and medical providers interact with a patient's unique clinical characteristics throughout a hospital stay to result in an outcome. Unfortunately, for more than 700,000 patients annually in the United States (US) that outcome is death (1). Although there are validated clinical scores to predict in-hospital mortality (2-4), they require manual data entry—a time and labor-intensive process and potentially a source of error(s). Additionally, those scoring tools only allow for an analysis of data at a particular moment in time as it relates to a patient's condition (5). Using a machine learning (ML) approach with electronic health record (EHR) data can account for the complexity and number of fluctuating inputs to better predict and reduce in-hospital mortality.

While many open-source ML algorithms for predicting in-hospital mortality exist, to date, they were developed using the Medical Information Mart for Intensive Care III (MIMIC-III) publicly available dataset (6-12). Data in MIMIC-III comprise patients admitted to the intensive care unit (ICU) at a single hospital (13), thus restricting the use of the open-source models to the intensive care setting and limiting the ability to apply them to the general care floor setting. The current literature indicated that many ML approaches focused on mortality due to specific conditions (14-16) and after specific procedures (17-19). With nonspecific diseases being the fifth underlying cause of death for inpatient hospital mortality (20), ideally a model would predict mortality beyond a particular condition or procedure. Additionally, certain patients admitted to a non-ICU can be at risk for rapid deterioration (21,22), and therefore merit further attention. Besides reviewing the patient's mortality risk at triage or admission (23,24), an ideal model would continuously evaluate the mortality risk through the patient's admission, given that more extended hospital stays are more prevalent in patients who died in the hospital (1). Indeed, ICU mortality prediction is more accurate with a continuous model than static scoring or prediction (25).

Given the limitations of static validated scoring systems and existing open-source models, the objective of this study was to develop a machine-learning (ML) model that predicts in-hospital mortality throughout a patient's hospital stay, regardless of the patient's specific cause or location of admission. We present the following article in accordance with the TRIPOD reporting checklist (available at <https://jmai.amegroups.com/article/view/10.21037/jmai-21-28/rc>).

## Methods

### *Setting, participants & outcome*

University of Florida (UF) Health system has two campuses in Northeast Florida. UF Health Jacksonville (UFHJ) is a teaching hospital and level 1-trauma center and UF Health North is a community hospital with an outpatient medical complex. The hospitals combined have 695 licensed beds with approximately 33,000 admissions annually. UF's Institutional Review Board approved the study, and an honest broker at the UF Integrated Data Repository (IDR) extracted the data. This study adheres to the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines for prediction model development. We utilized EHR for patients having an inpatient hospital visit from January 1, 2014 to January 30, 2020. Records were excluded for patients who were less than 18 years old at the time of encounter. Records for inpatient visits were also excluded if the length of stay (LOS) was less than one day or greater than 365 days. For patients with multiple inpatient visits, only the most current visit was kept.

The primary outcome of interest was in-hospital mortality. A patient was determined to have the outcome 'died in hospital' if they had a recorded death date equal to the discharge date of their inpatient visit. Patients were split into training and testing sets using an 80/20 split of the data stratified on a combination of outcome and total hospital LOS. *Figure 1* details the cohort selection and exclusion criteria. The training data utilized the total data available for the patients up until the day before discharge or death. For the patients in the test data, we created observations and predictions for each day the patient was in the hospital using data available up until that day.

### *Ethical statement*

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by University of Florida's Institutional Review Board (IRB201903477) and individual consent for this retrospective analysis was waived.

### *Software*

The model was developed with Python Programming Language (RRID:SCR\_008394). Packages used in data cleaning and manipulation were Pandas

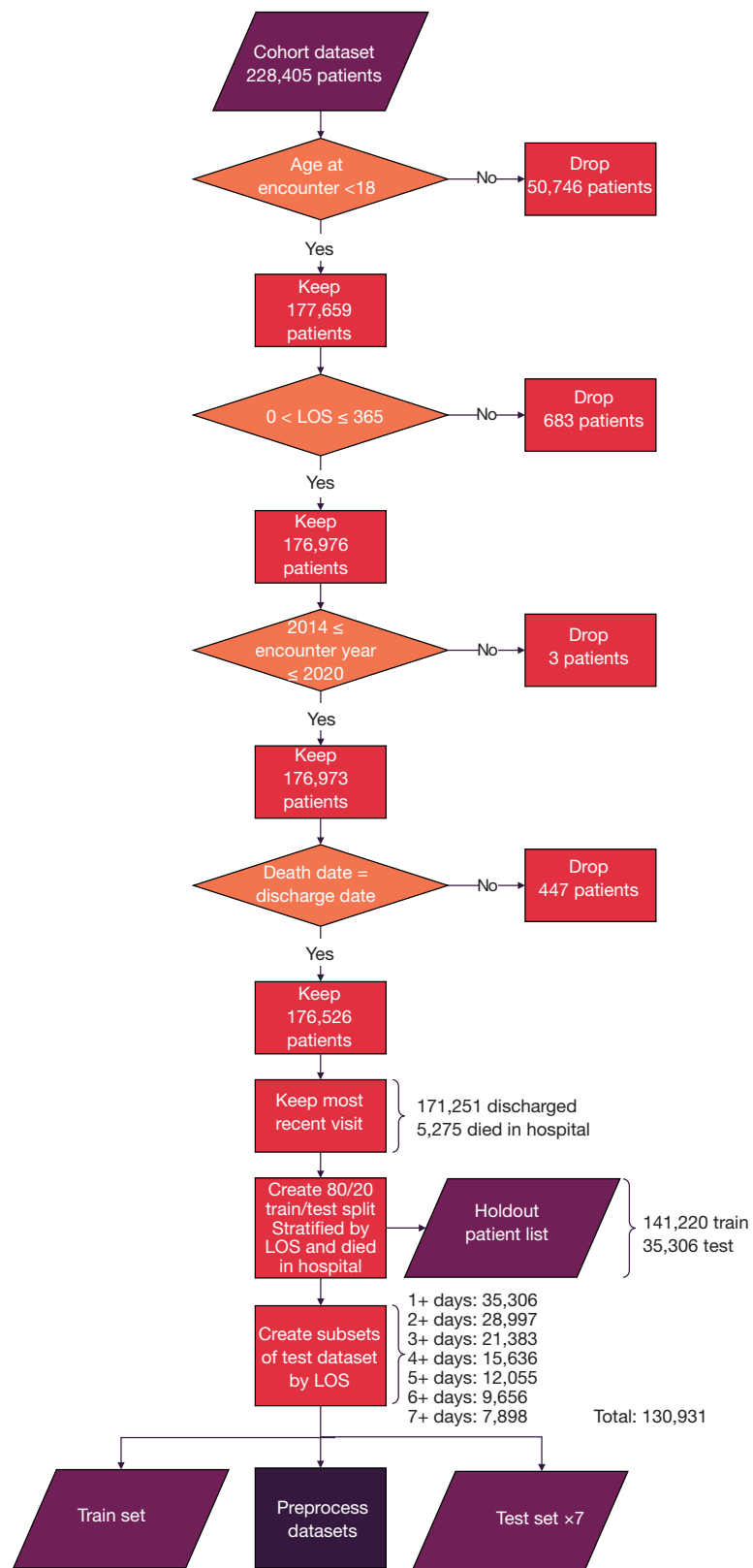


Figure 1 Cohort selection and exclusion criteria. LOS, length of stay.

(RRID:SCR\_018214), NumPy (RRID:SCR\_008633), and FuzzyWuzzy (RRID:SCR\_021699). The dataset was split with scikit-learn (RRID:SCR\_002577). Imputation of missing data was done with LightGBM (RRID:SCR\_021697) and PyCaret (RRID:SCR\_021695). The imbalanced dataset was resampled with imblearn (RRID:SCR\_021698). The ML model was developed with Catboost (RRID:SCR\_021694), and PyCaret (RRID:SCR\_021695). BorutaShap (RRID:SCR\_021696) was used for feature selection. Graphs and plots were created with PyCaret (RRID:SCR\_021695), Matplotlib (RRID:SCR\_008624), and Plotly (RRID:SCR\_013991).

### *Data variables*

Data utilized included demographics, emergency department and inpatient medications, laboratory (lab) data, procedures, and other orders. We also utilized tobacco use history available from the EHR. For each patient, the last available record for demographic data and tobacco use history were used. Inpatient consults were mapped to a priori-determined relevant consult groups according to clinical domain expert authors (JNF & CCG). All international classification of diseases (ICD) 9/10 records were recorded after the conclusion of an inpatient visit; therefore, only ICD information available prior to the start date of each patient's current inpatient visit is used, acting as a proxy for medical history. All ICD9 codes were mapped to ICD10 codes using the crosswalk from the National Bureau of Economic Research (26). Those codes were mapped to the clinical classifications software refined (CCSR) groupings using the crosswalk from the Agency for Healthcare Research and Quality (27). Consult and ICD features indicating end-of-life were dropped.

Medication order descriptions were cleaned and matched to a flattened version (author GPLL) of the anatomical therapeutic chemical (ATC) classification (28) using fuzzy string-matching and proofread by manual review (authors JNF, CCG, & SJD). Each ATC classification was grouped into its highest available level up to Level 4. Lab names and units were cleaned and concatenated. The variations were mapped to a priori-determined relevant labs according to authors with clinical experience (JNF & CCG). Conservative thresholds were set for outliers at the upper/lower quartiles  $\pm 5 \times$  the training data's interquartile range (IQR). Any value outside of that range was replaced with the threshold. Vital sign measurements [blood pressure (BP), Braden Score, heart rate, Glasgow Coma Score (GCS),

pain scale, respiratory measurements, and temperature] were bounded to their appropriate clinically defined or physiologically possible ranges. Any values outside of these ranges were removed. For all labs and vitals, the first, last, minimum, maximum, mean, and standard deviations (SD) for each patient were incorporated. For labs, minimum and maximum values were replaced with an indicator of high or low using hospital lab outlier thresholds. All categorical variables were one-hot encoded.

For data from consults, demographics, ICD codes, medications, and tobacco history, all patients with missing data were assigned a value of 'unknown' for each respective category. All 'unknown' features were dropped from the dataset to reduce multicollinearity except for lab data and vital signs. For lab data, we imputed mean, first, and last values with normal physiologic values. The threshold values used in vital sign measurements and lab data are listed in [Tables S1,S2](#) in the supplementary appendix online. SD was imputed with a value of zero. Missing vital signs were imputed during model training using mean values or an iterative imputation procedure.

### *Feature selection*

The training and test datasets contained a combined 1,479 features. The BorutaShap wrapper method (29-31), using the CatBoost classifier with SHAP values for feature importance, was utilized for feature selection. Both accepted and tentative features were kept for a reduced feature set of 60 attributes.

### *Machine learning algorithm*

CatBoost is a machine learning algorithm that utilizes gradient boosting on decision trees for both regression and classification. Gradient boosted decision trees have recently been used with great success across various disciplines, and CatBoost is a preferable model choice for large datasets with heterogenous and categorical data (32,33). A total of 15 different CatBoost models were trained using the default hyperparameters in PyCaret.

### *Statistical analysis*

Models were tested with both mean and iterative imputation for missing data. Various resampling methods were explored for class imbalance such as random undersampling, random oversampling, SMOTE (34), random undersampling

followed by random oversampling, and random undersampling followed by SMOTEEEN (35). The F2 measure was chosen as the metric for model selection based on its relative weighting of recall and precision. Unlike the F1 measure, the F2 measure, Eq. [1], emphasizes recall more heavily which is vital in medical applications where the cost of a false negative (predicting survival when in fact the patient dies) outweighs the cost of a false positive (misclassifying a patient as more likely to die when they actually survive).

$$F_2 = 5 \times \frac{\text{precision} \times \text{recall}}{4 \times \text{precision} + \text{recall}} \quad [1]$$

We also reviewed accuracy, recall, precision, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and F1 score. Accuracy was measured as the proportion of all patient outcomes, both survival and death, predicted correctly. Recall was measured as the proportion of all patients accurately predicted to die out of all the patients who did die. Precision was measured as the proportion of all patients accurately predicted to die out of all the patients who were prediction to die. The precision-recall curve plots precision as a function of recall for various classification thresholds. The AUPRC is an overall measure of the tradeoff of precision and recall for the model. The ROC curve is a plot of true positive rate against false positive rate for various classification thresholds. The AUROC is an alternative measure of diagnostic accuracy of a model, but is typically more useful when there is not large class imbalance. The F1 score, like the F2 score, is an F-measure. It is defined as the harmonic mean of precision and recall.

The model with the highest F2 score was then tuned to select optimal hyperparameters via a grid search. The tuned model was then calibrated using an isotonic method (36). Baseline models were created using Braden Score and GCS. For each baseline model, the test data set was used to find the threshold of the score that led to the greatest F2 score for classifying patients. The tuned and calibrated model and the baseline models were used to predict mortality on each subset of the test data stratified by the current LOS. Predictions were made using an empirical bootstrapping method using 1,000 iterations with resampling of the test data for each subset to obtain estimates and confidence intervals (CI) for the metrics. We also evaluated the model for patient subsets of total LOS. We then tested the model against sub-populations to uncover potential predictive

biases against gender, race, and age group.

## Results

Our dataset comprised 176,526 patients, of which 63.7% were white, and 27.3% were black. Slightly over half (55.3%) of patients were female and 44.6% were male (Table 1). A total of 5,275 (3.0%) patients died in the hospital. Most patients (77.6%) were discharged or died by day 6. From day 7 and after, the percentage of patients with a total LOS for any specific number of days became low (0.0–4.0%); therefore, patients with an LOS of seven or more days were reclassified as having an LOS of 7+. We analyzed patients' mortality rate against their LOS and noted a high mortality rate for patients staying one day. The mortality rate was the lowest on days 2 and 3 (1.432% and 1.437% vs. 2.005–6.027% for remaining days). From day 3, the mortality rate increased as LOS increased. By day 4, 66.5% were discharged alive. Over half (54.9%) of patients who did not survive died by day 6. Figure 2 illustrates the mortality distribution.

The combination of various imputation and resampling methods produced 15 models (Table 2). The CatBoost classifier using mean imputation and an under-sampling strategy of 0.25 followed by oversampling to achieve class balance performed best with an F2 score of 0.506. That model was tuned and calibrated, and the final model, MONITOR, achieved an F2 score of 0.510, a recall of 0.644, and an AUROC of 0.905 (Figure 3 and Table 3). The feature importance analysis showed that the Braden Score and GCS had the highest predictive power, followed by the last heart rate, the last mean arterial pressure, age, and LOS. Tables S3,S4 in the supplementary appendix lists the data features by the cohorts.

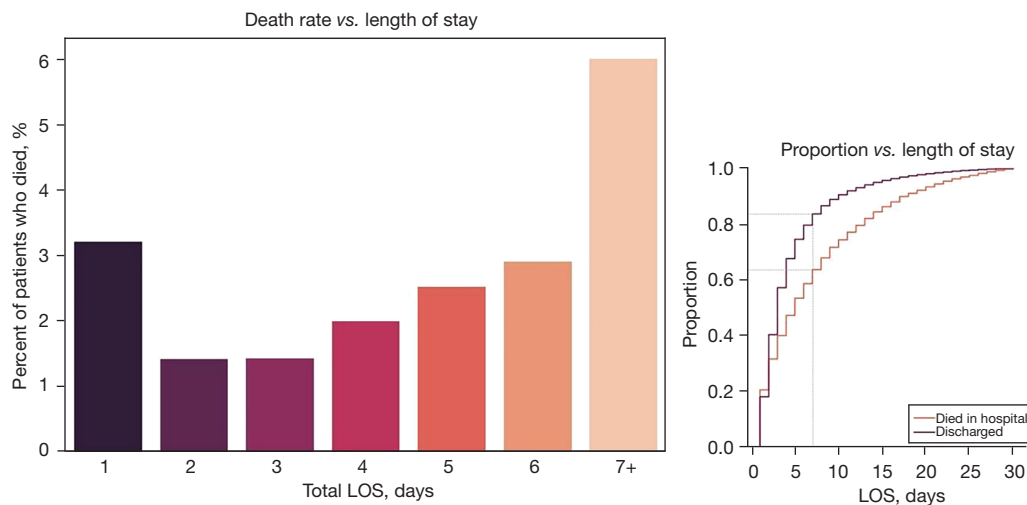
We established the performance metrics for the GCS Score, Braden Score, and MONITOR models for each day of stay. MONITOR systematically outperformed the score-based models for F2 score (0.466–0.543), recall (0.729–0.759), AUROC (0.889–0.920), and AUPRC (0.362–0.391). Figure 4 illustrate the comparison between the three models. The detailed performance metrics between the three models can be found in Table S5 of the supplementary appendix online.

Figure 5 shows how F2 score varies by day when patients are grouped by total LOS. The F2 score is highest for patients on day 1 or with a total LOS of one day (0.811) and lowest for patients with a total LOS of seven or more days (0.380). For patients with staying seven or more days, the

**Table 1** Characteristics of patients in training and test data sets

Characteristic	Total (n=176,526)	Train (80%)			Test (20%)		
		Discharged (n=137,000)	Died in hospital (n=4,220)	Total (n=141,220)	Discharged (n=34,251)	Died in hospital (n=1,055)	Total (n=35,306)
LOS, mean [IQR], days	5.6 [2–6]	5.4 [2–6]	10.5 [2–13]	5.6 [2–6]	5.4 [2–6]	10.5 [2–13]	5.6 [2–6]
Age, mean [IQR], years	52.2 [34–68]	51.8 [34–67]	64.4 [56–76]	52.2 [34–68]	51.7 [34–67]	64.8 [56–77]	52.0 [34–67]
Sex, n (%)							
Female	97,665 (55.3)	76,350 (55.7)	1,807 (42.8)	78,157 (55.3)	19,028 (55.6)	480 (45.5)	19,508 (55.3)
Male	78,800 (44.6)	60,600 (44.2)	2,413 (57.2)	63,013 (44.6)	15,214 (44.4)	573 (54.3)	15,787 (44.7)
Unknown	61 (0.0)	50 (0.0)	0 (0.0)	50 (0.0)	9 (0.0)	2 (0.2)	11 (0.0)
Race, n (%)							
White	112,524 (63.7)	87,258 (63.7)	2,754 (65.3)	90,012 (63.4)	21,836 (63.8)	676 (64.1)	22,512 (63.8)
Black	48,131 (27.3)	37,393 (27.3)	1,153 (27.3)	38,546 (27.3)	9,289 (27.1)	296 (28.1)	9,585 (27.1)
Asian	1,961 (1.1)	1,519 (1.1)	28 (1.0)	1,547 (1.1)	411 (1.2)	3 (0.3)	414 (1.2)
Other	11,576 (6.6)	9,076 (6.6)	151 (3.6)	9,227 (6.5)	2,307 (6.7)	42 (4.0)	2,349 (6.7)
Unknown	2,334 (1.3)	1,754 (1.3)	134 (1.3)	1,888 (1.3)	408 (1.2)	38 (3.6)	446 (1.3)
Ethnicity, n (%)							
Not Hispanic	163,713 (92.7)	127,050 (92.7)	3,948 (94.6)	130,998 (92.8)	31,733 (92.6)	982 (93.1)	32,715 (92.7)
Hispanic	10,138 (5.7)	7,949 (5.8)	121 (2.9)	8,070 (5.7)	2,040 (6.0)	28 (2.7)	2,068 (5.9)
Unknown	2,675 (1.5)	2,001 (1.5)	151 (1.5)	2,152 (1.5)	478 (1.4)	45 (4.3)	523 (1.5)

LOS, length of stay; no., number; IQR, interquartile range.



**Figure 2** Distribution and proportion of mortality rate against LOS. LOS, length of stay.

**Table 2** Performance metrics for various models after imputation and resampling

Imputation	Resampling	Accuracy	AUROC	Recall	Precision	AUPRC	F1	F2
Mean	Undersampling.25_ Oversampling	0.8979	0.9060	0.7258	0.2283	0.3731	0.3473	0.5055
Mean	Undersampling.50_ Oversampling	0.8821	0.9085	0.7625	0.2074	0.3783	0.3262	0.4967
Mean	Undersampling.25_ SMOTE_EEN	0.8793	0.9053	0.7607	0.2030	0.3460	0.3205	0.4910
Mean	Oversampling	0.9312	0.9007	0.5878	0.2918	0.3613	0.3900	0.4887
Mean	Undersampling.75_ Oversampling	0.8729	0.9085	0.7762	0.1967	0.3756	0.3138	0.4883
Mean	Undersampling	0.9310	0.9013	0.5872	0.2910	0.3626	0.3892	0.4879
Iterative	Undersampling.25_ SMOTE_EEN	0.8880	0.9005	0.7233	0.2103	0.3456	0.3259	0.4861
Mean	Undersampling.75_ SMOTE_EEN	0.8812	0.8994	0.7374	0.2021	0.3308	0.3173	0.4821
Mean	Undersampling.50_ SMOTE_EEN	0.8729	0.9017	0.7631	0.1946	0.3378	0.3101	0.4817
Iterative	Undersampling	0.8651	0.9076	0.7815	0.1876	0.3682	0.3025	0.4785
Iterative	Undersampling.50_ SMOTE_EEN	0.8816	0.8939	0.7268	0.2010	0.3281	0.3149	0.4772
Iterative	Oversampling	0.9325	0.8984	0.5623	0.2916	0.3521	0.3840	0.4743
Iterative	Undersampling.75_ SMOTE_EEN	0.8838	0.8886	0.7037	0.2003	0.3291	0.3119	0.4683
Mean	SMOTE	0.9571	0.9029	0.3585	0.4160	0.3607	0.3851	0.3687
Iterative	SMOTE	0.9602	0.9061	0.3354	0.4564	0.3747	0.3867	0.3542

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.

F2 score increases to 0.543 for predictions made on day 7. Similar trends in increasing scores are observed for other subgroups for all of the reported metrics (*Figure 5*).

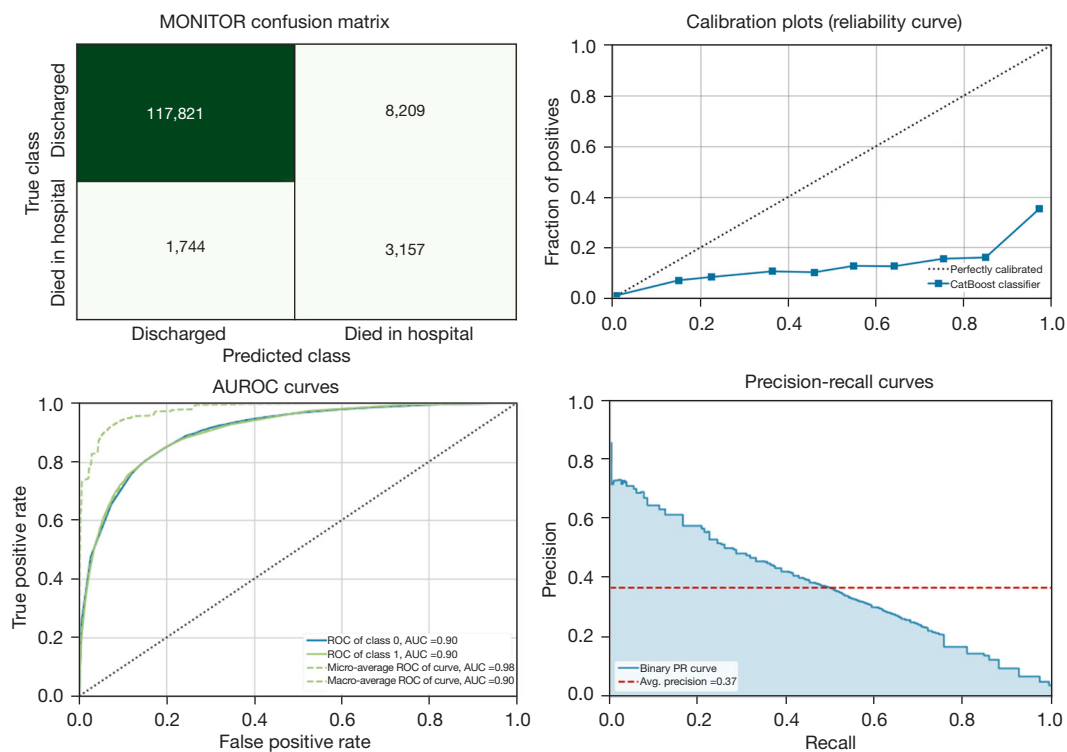
The subgroup analysis revealed no statistical differences between age, race, and sex; with the exception in sex on days 5 and 6 (*Figure 6*). The F2 score for males on day 5 was 0.490 (0.454–0.528) whereas 0.579 (0.540–0.618) for females. The F2 score for males on day 6 was 0.515 (0.475–0.553) whereas 0.577 (0.536–0.617) for females.

## Discussion

The adoption of EHRs has increased the availability of data, facilitating the uptake of ML approaches in the healthcare setting (37). Enhancing EHRs with modern data visualizations and multi-domain machine learning

can provide a 360-degree view of the patient, provider, hospital, and healthcare system to better understand and act upon mortality risk throughout the patient's hospital stay. However, many ML approaches to date for predicting mortality are limited to a specific setting (e.g., the ICU), conditions (e.g., sepsis), or procedures (e.g., PCI) at a singular moment in time. Because clinicians evaluate a patient's conditions throughout his/her hospital stay, the same time-varying approach should be applied when evaluating mortality risk. Therefore, we developed an ML model to predict mortality risk throughout a patient's hospital stay.

Previously published models that predicted in-hospital mortality at the time of admission achieved an AUROC of 0.84–0.94 (24,38–40). In addition to using data from the entire LOS to train our model, we also



**Figure 3** Confusion matrix, calibration plot, AUROC, and PRC for MONITOR. AUROC, area under the receiver operating characteristic curve; PRC, precision-recall curve.

updated our predictions each day as new time-series data became available. Models trained on time-series data have demonstrated increased predictive ability (25). For patients that had longer LOS, it was more difficult to ascertain the discharge outcome early during the patient's stay. However, the model's performance increased as more data became available. Another main difference between our model and prior studies is that we selected the F2 measure to rank the different models instead of AUROC. Real-world datasets can exhibit imbalanced class distributions, especially in mortality where there is a binary minority class. Using the F-measure instead of accuracy can be more useful when dealing with class imbalance (41). Furthermore, the F2 measure balances precision and recall with more attention towards minimizing false negatives, considering the grave consequences of failing to identify an at-risk patient. For the same reason, other studies also optimized the F-measure in developing their models (42,43).

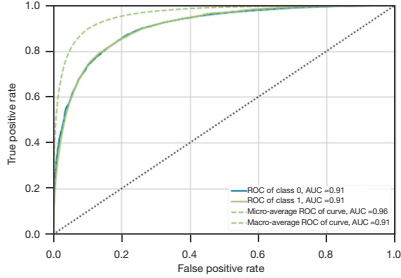
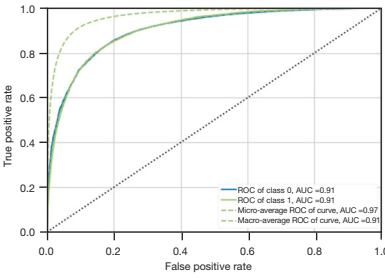
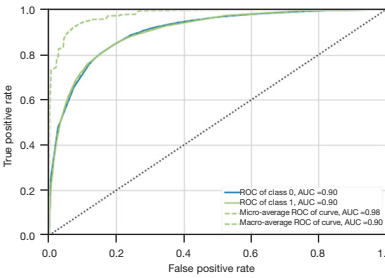
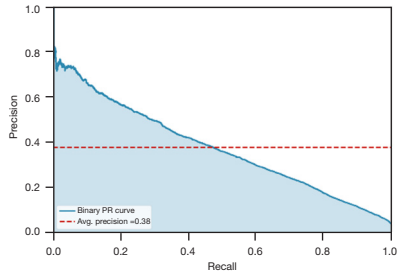
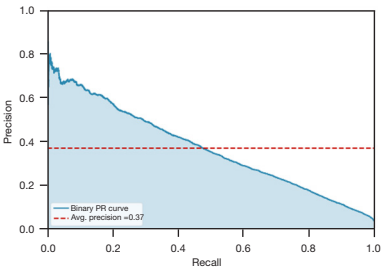
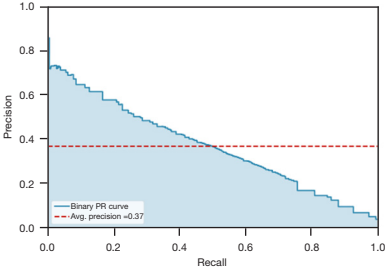
Healthcare professionals have been using medical calculations and scores as part of clinical decision tools for decades. Both the Braden Score and GCS, among other scoring systems, are integral parts of clinical practice, albeit

not intended originally as mortality risk prediction scores. The Braden Scale is used to assess pressure-injury risk in various healthcare settings. A GCS assessment is done to determine a patients' level of consciousness. However, both the Braden and GCS have been expanded to predict mortality in different studies (44-48). Even though the Braden and GCS scores were the highest ranked individual features in our model, our results suggested a more holistic ML approach is better as it has the ability to capture data spanning various clinical functions, including sequential time series temporal trends.

Our study has several limitations to consider. Lab and vital sign measurements were only included in the data if they occurred after an inpatient designation status for a patient. Data collected before patient handoff from other departments (e.g., emergency department) might be insightful but were discarded with our methodology. Additionally, we omitted several other possible predictors, including missingness indicators and features from clinical notes that other studies have shown to be helpful. Our choice of statistics for time-series data (first, last, maximum, minimum, mean, and SD) was also limited, and more



**Table 3** AUROC, PRC, and performance metrics for CatBoost, tuned CatBoost, and calibrated tuned CatBoost (MONITOR) models

Model	CatBoost model	Tuned CatBoost model	Calibrated tuned CatBoost (MONITOR) model
AUROC	 <p>0.9072</p>	 <p>0.9065</p>	 <p>0.9047</p>
PRC	 <p>0.3751</p>	 <p>0.3702</p>	 <p>0.3663</p>
Accuracy	0.8975	0.9091	0.9240
Recall	0.7278	0.6890	0.6442
Precision	0.2279	0.2454	0.2778
F1	0.3471	0.3619	0.3881
F2	0.5059	0.5061	0.5097

AUROC, area under the receiver operating characteristic curve; PRC, precision-recall curve.

sophisticated techniques may have led to better model performance. Our model was also impacted by incomplete patient records and missing data. Alternatively, other models such as recurrent neural networks or LSTMs may be utilized to capture signals from time series data that may retain some more clinically meaningful variations. The output of these models can be used as features in the final Catboost model. While different imputation types were utilized, more complete patient records would provide better results. Additionally, we assigned the survival outcome to each patient in the training dataset associated with the occurrence of a discharge event. However, in practice and in the LOS-stratified test set, patients may experience several events, including discharge, mortality, a continuation of stay, or departmental transfer. Although

we believe that the use of F2 measure over AUROC as the guiding metric is more indicative of the real impact of misclassifications in the hospital setting, we were unable to validate our decision without a complete cost-benefit analysis. The tradeoff between false positives and false negatives may depend on the current conditions and constraints of each hospital and need to be evaluated on an individual basis.

### Conclusions

This study investigated an ML model that can be utilized for both ICU and non-ICU inpatient mortality. Though limited to a single institution, it performs equally well regardless of patient demographics. Additionally, it accounts

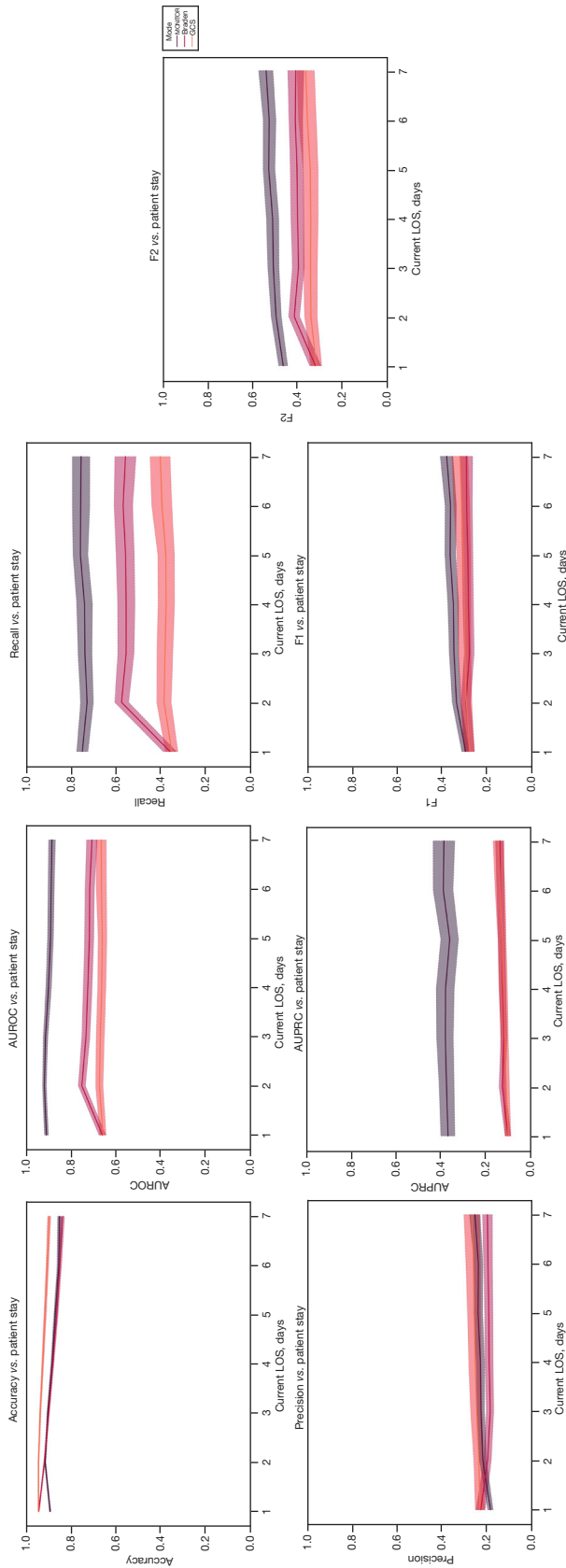


Figure 4 Comparison of performance metrics between MONITOR, Braden Score, and GCS. LOS, length of stay; GCS, Glasgow Coma Score.

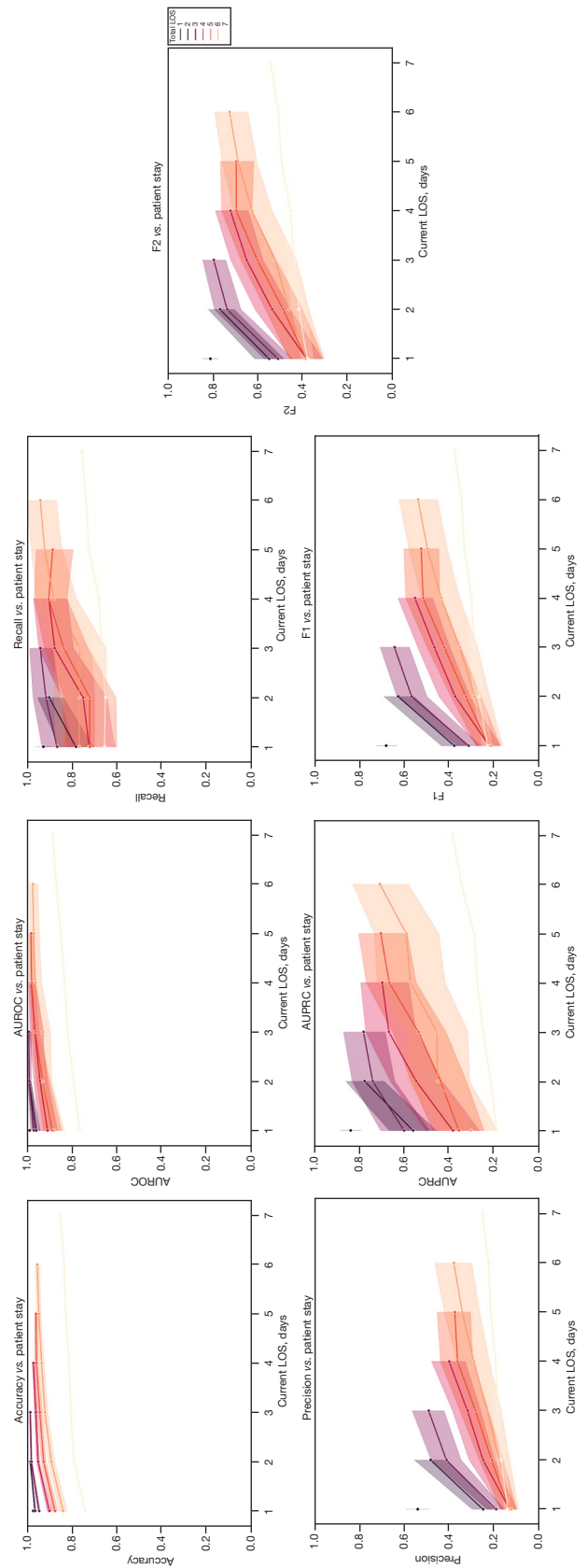


Figure 5 Performance metrics with CI for patients with different LOS. LOS, length of stay; CI, confidence interval.

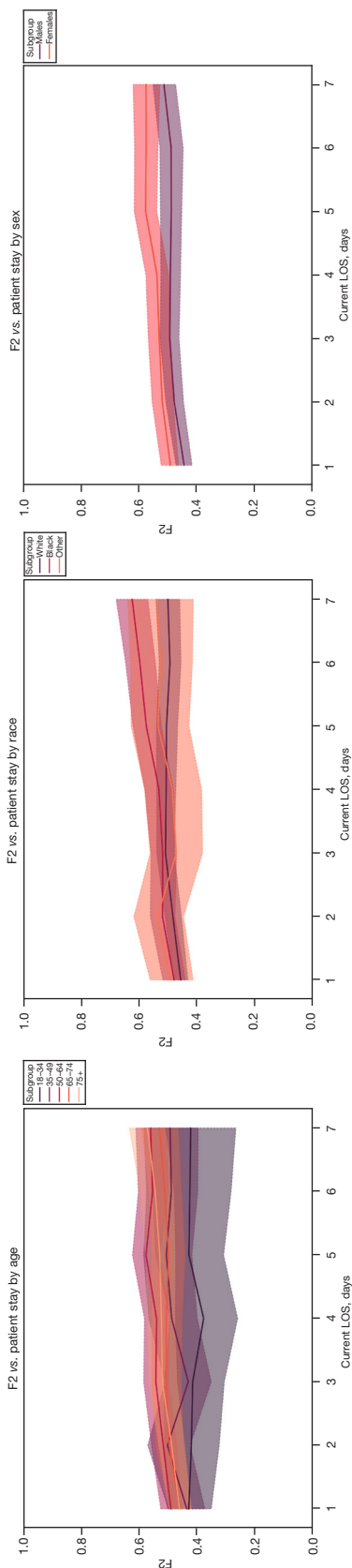


Figure 6 F2 scores for subgroup analysis. LOS, length of stay.

for the change in patient condition as the patient continues to stay in the hospital. Moving forward, the model needs to be validated internally with prospective data as well as externally to determine how the model performs with other patient cohorts. Further implementation research is required to understand how we can integrate this model with clinical workflow in order for this model to be impactful in an operational setting.

**Acknowledgments**

We acknowledge the University of Florida (UF) Integrated Data Repository (IDR) and the UF Health Office of the Chief Data Officer for providing the analytic data set for this project.

Funding: None.

**Footnote**

*Reporting Checklist:* The authors have completed the TRIPOD reporting checklist. Available at <https://jmai.amegroups.com/article/view/10.21037/jmai-21-28/rc>

*Data Sharing Statement:* Available at <https://jmai.amegroups.com/article/view/10.21037/jmai-21-28/dss>

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at <https://jmai.amegroups.com/article/view/10.21037/jmai-21-28/coif>). The authors have no conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by University of Florida’s Institutional Review Board (IRB201903477) and individual consent for this retrospective analysis was waived.

*Open Access Statement:* This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license).

See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

## References

1. Hall MJ, Levant S, DeFrances CJ. Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010. NCHS Data Brief 2013. Available online: <https://www.cdc.gov/nchs/data/databriefs/db118.pdf>
2. Chang SH, Hsieh CH, Weng YM, et al. Performance Assessment of the Mortality in Emergency Department Sepsis Score, Modified Early Warning Score, Rapid Emergency Medicine Score, and Rapid Acute Physiology Score in Predicting Survival Outcomes of Adult Renal Abscess Patients in the Emergency Department. *Biomed Res Int* 2018;2018:6983568.
3. Brink A, Alsma J, Verdonshot RJCG, et al. Predicting mortality in patients with suspected sepsis at the Emergency Department; A retrospective cohort study comparing qSOFA, SIRS and National Early Warning Score. *PLoS One* 2019;14:e0211133.
4. Ferreira FL, Bota DP, Bross A, et al. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* 2001;286:1754-8.
5. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLoS Med* 2018;15:e1002701.
6. El-Manzalawy Y, Abbas M, Hoaglund I, et al. OASIS+: leveraging machine learning to improve the prognostic accuracy of OASIS severity score for predicting in-hospital mortality. *BMC Med Inform Decis Mak* 2021;21:156.
7. Sadeghi R, Banerjee T, Romine W. Early Hospital Mortality Prediction using Vital Signals. *Smart Health (Amst)* 2018;9-10:265-74.
8. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. *AMIA Annu Symp Proc* 2018;2017:994-1003.
9. Zhang D, Yin C, Zeng J, et al. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak* 2020;20:280.
10. Harutyunyan H, Khachatrian H, Kale DC, et al. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019;6:96.
11. Purushotham S, Meng C, Che Z, et al. Benchmarking deep learning models on large healthcare datasets. *J Biomed Inform* 2018;83:112-34.
12. Pirracchio R, Petersen ML, Carone M, et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med* 2015;3:42-52.
13. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
14. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J* 2017;38:500-7.
15. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Acad Emerg Med* 2016;23:269-78.
16. Moll M, Qiao D, Regan EA, et al. Machine Learning and Prediction of All-Cause Mortality in COPD. *Chest* 2020;158:952-64.
17. Hernandez-Suarez DF, Kim Y, Villablanca P, et al. Machine Learning Prediction Models for In-Hospital Mortality After Transcatheter Aortic Valve Replacement. *JACC Cardiovasc Interv* 2019;12:1328-38.
18. Al'Aref SJ, Singh G, van Rosendaal AR, et al. Determinants of In-Hospital Mortality After Percutaneous Coronary Intervention: A Machine Learning Approach. *J Am Heart Assoc* 2019;8:e011160.
19. Li Y, Chen M, Lv H, et al. A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery. *Injury* 2021;52:1487-93.
20. Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying cause of death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Available online: <http://wonder.cdc.gov/ucd-icd10.html>
21. Saliba P, Hornero A, Cuervo G, et al. Mortality risk factors among non-ICU patients with nosocomial vascular catheter-related bloodstream infections: a prospective cohort study. *J Hosp Infect* 2018;99:48-54.
22. Nemer DM, Wilner BR, Burkle A, et al. Clinical Characteristics and Outcomes of Non-ICU Hospitalization for COVID-19 in a Nonpicenter, Centrally Monitored Healthcare System. *J Hosp Med* 2021;16:7-14.
23. Raita Y, Goto T, Faridi MK, et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23:64.
24. Brajer N, Cozzi B, Gao M, et al. Prospective and External Evaluation of a Machine Learning Model to Predict In-

- Hospital Mortality of Adults at Time of Admission. *JAMA Netw Open* 2020;3:e1920733.
25. Shickel B, Loftus TJ, Adhikari L, et al. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. *Sci Rep* 2019;9:1879.
  26. National Bureau of Economic Research (NBER). ICD-9-CM to and from ICD-10-CM And ICD-10-PCS crosswalk or general equivalence mappings. Cambridge, MA, USA: NBER, 2016. Available online: <https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings>
  27. Agency for Healthcare Research and Quality (AHRQ). HCUP tools and software. Rockville, MD, USA: AHRQ, 2021. Available online: [https://www.hcup-us.ahrq.gov/tools\\_software.jsp](https://www.hcup-us.ahrq.gov/tools_software.jsp)
  28. World Health Organization Collaborating Center (WHOCC). ATC/DDD index 2021. Oslo: WHOCC, 2020. Available online: [https://www.whocc.no/atc\\_ddd\\_index/](https://www.whocc.no/atc_ddd_index/)
  29. Cohen A. Python Package Index (PyPI). Fuzzywuzzy 0.18.0. 2011. Available online: <https://pypi.org/project/fuzzywuzzy/>
  30. Chierigato M, Frangiamore F, Morassi M, et al. A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. *arXiv* 2021. Available online: <https://arxiv.org/pdf/2105.06141.pdf>
  31. Kleiman MJ, Barenholtz E, Galvin JE, et al. Screening for Early-Stage Alzheimer's Disease Using Optimized Feature Sets and Machine Learning. *J Alzheimers Dis* 2021;81:355-66.
  32. Andersson P, Johnsson J, Björnsson O, et al. Predicting neurological outcome after out-of-hospital cardiac arrest with cumulative information; development and internal validation of an artificial neural network algorithm. *Crit Care* 2021;25:83.
  33. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2018:6639-49. Available online: <https://dl.acm.org/doi/abs/10.5555/3327757.3327770>
  34. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data* 2020;7:94.
  35. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321-57.
  36. Batista GE, Bazzan AL, Monard MC. Balancing training data for automated annotation of keywords: A case study. *WOB* 2003;10. Available online: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.2192&rep=rep1&type=pdf>
  37. Niculescu-Mizil A, Caruana RA. Obtaining calibrated probabilities from boosting. *arXiv* 2012. Available online: <https://arxiv.org/ftp/arxiv/papers/1207/1207.1403.pdf>
  38. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)* 2014;33:1123-31.
  39. Seki T, Kawazoe Y, Ohe K. Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. *PLoS One* 2021;16:e0246640.
  40. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;1:18.
  41. Soffer S, Klang E, Barash Y, et al. Predicting In-Hospital Mortality at Admission to the Medical Ward: A Big-Data Machine Learning Model. *Am J Med* 2021;134:227-234.e4.
  42. Meng Liu, Chang Xu, Yong Luo, et al. Cost-Sensitive Feature Selection by Optimizing F-Measures. *IEEE Trans Image Process* 2018;27:1323-35.
  43. Shah V, Turkbey B, Mani H, et al. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. *Med Phys* 2012;39:4093-103.
  44. Khajehali N, Khajehali Z, Tarokh MJ. The prediction of mortality influential variables in an intensive care unit: a case study. *Pers Ubiquitous Comput* 2021. [Epub ahead of print]. doi: 10.1007/s00779-021-01540-5.
  45. Bandle B, Ward K, Min SJ, et al. Can Braden Score Predict Outcomes for Hospitalized Heart Failure Patients? *J Am Geriatr Soc* 2017;65:1328-32.
  46. Jentzer JC, Anavekar NS, Brenes-Salazar JA, et al. Admission Braden Skin Score Independently Predicts Mortality in Cardiac Intensive Care Patients. *Mayo Clin Proc* 2019;94:1994-2003.
  47. Yousefifard M, Shahsavarinia K, Faridaalee G, et al. Comparison of Glasgow Coma Scale with physiologic scoring scales in prediction of in-hospital outcome of

trauma patients: a diagnostic accuracy study. *Front Emerg Med* 2020;4:e89.

48. Weingarten S, Bolus R, Riedinger MS, et al. The principle

of parsimony: Glasgow Coma Scale score predicts mortality as well as the APACHE II score for stroke patients. *Stroke* 1990;21:1280-2.

doi: 10.21037/jmai-21-28

**Cite this article as:** Guerrier CC, D'Acunto SJ, Labilloy GPL, Esma RA, Kendall HA, Norez DA, Fische JN. MONITOR: a multi-domain machine learning approach to predicting in-hospital mortality. *J Med Artif Intell* 2022;5:3.

**Table S1** Thresholds used in vital sign measurements

Measurement	Low cutoff	High cutoff	Unit
Braden Score	6	23	Score
Mean arterial pressure	0	200	mmHg
GCS	3	15	Score
Pain scale	0	10	Score
SpO <sub>2</sub>	0	100	%
Respiration rate	0	200	BPM
FiO <sub>2</sub>	0	100	%
BMI	0	500	kg/m <sup>2</sup>
Weight	0	700	lbs
Height	20	110	in
Heart rate	0	500	BPM
Temperature	0	120	°F

GCS, Glasgow Coma Score.

**Table S2** Thresholds used in laboratory data

Name	Sex	Low_value	High_value	Normal_value
abg_lactic_acid	B	0.7	2.7	1.7
Albumin	B	3.8	4.9	4.35
alkaline_phosphatase	F	35	104	69.5
alkaline_phosphatase	M	40	129	84.5
alt_(sgpt)	B	10	42	26
anion_gap	B	4	16	10
arterial_o2	B	80	100	90
arterial_pco2	B	35	45	40
arterial_po2	B	80	100	90
arterial-venous_ph	B	7.35	7.45	7.4
ast_(sgot)	B	14	33	23.5
atypical_lymphocyte	B	0	12	0
banded_neutrophils	B	34	73	53.5
Bicarbonate	B	21	27	24
bilirubin_direct	B	0	0.2	0.1
bilirubin_indirect	B	0.2	0.9	0.55
bilirubin_total	B	0.2	1	0.6
blood_urean_nitrogen	B	6	22	14
bun-creatinine_ratio	B	7.5	34	20.75
Calcium	B	8.6	10	9.3
calcium_ionized	B	4.64	5.28	4.96
carbon_dioxide	B	22	30	26
Chloride	B	101	110	105.5
creatine_kinase	B	22	195	108.5
Creatinine	F	0.4	0.9	0.65
Creatinine	M	0.8	1.2	1
EGFR	B	59	500	60
fibrinogen_level	B	186	461	323.5
Globulin	B	2.3	3.5	2.9
Glucose	B	71	99	85
Hematocrit	M	40	54	47
Hematocrit	F	37	47	42
Hemoglobin	M	14	18	16
Hemoglobin	F	12	16	14
Inr	B	0.8	1.1	0.95
Ketones_ua	B	0	20	20
Lipase	B	0	60	30
Lymphocytes	B	24	44	34
Magnesium	B	1.8	2.6	2.2
Neutrophils	B	0	5	2.5
nt_pro_bnp	B	0	2,000	400
partial_thromboplastin_time	B	25	37	31
Phosphorus	B	2.5	4.5	3.5
plasma_lactic_acid	B	0.7	2.7	1.7
platelet_count	B	140	440	290
Potassium	B	3.3	4.6	3.95
pro_bnp	B	0	125	62.5
Procalcitonin	B	0	0.15	0.075
Prottime	B	9.4	12.5	10.95
rbc_count	M	4.5	6.3	5.4
rbc_count	F	4.2	5.4	4.8
Sodium	B	136	145	140.5
total_protein	B	6.5	8.3	7.4
venous_o2	B	37	43	40
venous_pco2	B	44	46	45
venous_po2	B	37	43	40
wbc_count	B	4.5	11	7.75
arterial_o2_content	B	16	20	18
venous_o2_content	B	12	15	13.5

B, both sexes; F, female; M, male.



Table S3 Data features (continuous variables)

Features	Train—80%									Test—20%								
	Alive (n=137,000—97.0%)			Dead (n=4,220—3.0%)			Total (n=141,220)			Alive (n=34,251—97.0%)			Dead (n=1,055—3.0%)			Total (n=35,306)		
	Missing	Mean	SD	Missing	Mean	SD	Missing	Mean	SD	Missing	Mean	SD	Missing	Mean	SD	Missing	Mean	SD
Features—from cohort and vital sign measurements																		
Age at encounter (y)	0.0%	51.8	19.5	0.0%	64.4	15.5	0.0%	52.2	19.5	0.0%	51.7	19.5	0.0%	64.8	16.0	0.0%	52.0	19.6
Braden Score Last	5.8%	19.4	2.7	7.9%	11.6	3.0	5.8%	19.1	3.0	5.8%	19.3	2.8	7.7%	12.5	3.3	5.8%	19.1	3.0
Braden Score Mean	5.8%	19.1	2.6	7.9%	12.8	2.9	5.8%	18.9	2.8	5.8%	19.1	2.7	7.7%	13.2	3.1	5.8%	18.9	2.9
Braden Score Min	5.8%	17.7	3.3	7.9%	10.6	2.9	5.8%	17.5	3.5	5.8%	17.8	3.2	7.7%	11.5	3.0	5.8%	17.6	3.3
FiO <sub>2</sub> Last (%)	87.7%	34.6	17.3	40.2%	56.7	24.3	86.3%	37.5	19.8	88.5%	35.8	17.0	46.4%	53.8	22.8	87.3%	38.0	18.8
GCS score Last	37.6%	14.7	1.1	21.3%	8.4	4.5	37.1%	14.5	1.8	38.5%	14.7	1.2	22.8%	9.4	4.6	38.0%	14.5	1.8
GCS score Mean	37.6%	14.6	1.2	21.3%	9.9	4.1	37.1%	14.4	1.7	38.5%	14.6	1.3	22.8%	10.1	4.2	38.0%	14.4	1.7
GCS score Min	37.6%	13.7	2.7	21.3%	6.8	4.4	37.1%	13.5	3.1	38.5%	13.8	2.6	22.8%	7.8	4.6	38.0%	13.6	2.9
Heart rate Last (BPM)	0.7%	80.1	14.6	0.7%	94.5	24.0	0.7%	80.5	15.2	0.8%	80.4	14.7	1.0%	92.4	21.9	0.8%	80.7	15.1
Heart rate Max (BPM)	0.7%	114.9	32.7	0.7%	145.9	42.8	0.7%	115.9	33.5	0.8%	113.7	31.0	1.0%	138.3	41.2	0.8%	114.4	31.7
Heart rate Mean (BPM)	0.7%	81.3	13.2	0.7%	92.2	16.8	0.7%	81.6	13.5	0.8%	81.5	13.3	1.0%	91.6	16.9	0.8%	81.8	13.5
Heart rate Min (BPM)	0.7%	57.8	19.6	0.7%	51.3	27.0	0.7%	57.6	19.9	0.8%	58.8	18.9	1.0%	56.4	25.6	0.8%	58.8	19.2
Mean arterial pressure Last (mmHg)	0.4%	87.6	13.5	0.5%	74.5	19.1	0.4%	87.2	13.9	0.4%	87.7	13.7	0.9%	78.2	17.8	0.5%	87.4	13.9
Mean arterial pressure Mean (mmHg)	0.4%	87.5	11.1	0.5%	80.5	12.2	0.4%	87.3	11.2	0.4%	87.5	11.2	0.9%	81.0	11.9	0.5%	87.3	11.3
Mean arterial pressure SD (mmHg)	1.3%	11.6	4.3	0.8%	14.1	5.2	1.3%	11.6	4.3	1.3%	11.5	4.4	1.5%	14.2	5.9	1.3%	11.6	4.4
Pain Last	2.7%	2.6	3.1	33.3%	2.2	3.3	3.6%	2.6	3.1	3.3%	2.6	3.1	35.1%	2.3	3.3	4.2%	2.6	3.1
Pain Mean	2.7%	3.1	2.5	33.3%	2.3	2.5	3.6%	3.1	2.5	3.3%	3.1	2.5	35.1%	2.5	2.5	4.2%	3.1	2.5
Pain Min	2.7%	0.5	1.5	33.3%	0.5	1.9	3.6%	0.5	1.6	3.3%	0.5	1.5	35.1%	0.5	1.8	4.2%	0.5	1.5
Respiration rate Last (BPM)	25.6%	17.4	2.6	25.4%	21.1	7.4	25.6%	17.5	3.0	25.5%	17.4	2.7	25.5%	20.5	7.3	25.5%	17.5	3.0
Respiration rate Max (BPM)	25.6%	26.0	9.8	25.4%	37.6	14.2	25.6%	26.3	10.2	25.5%	25.5	9.2	25.5%	34.4	13.3	25.5%	25.8	9.4
Respiration rate Mean (BPM)	25.6%	16.8	2.9	25.4%	20.0	4.3	25.6%	16.9	3.0	25.5%	16.8	2.9	25.5%	19.4	4.4	25.5%	16.8	3.0
Respiration rate Min (BPM)	25.6%	10.6	5.4	25.4%	9.4	5.2	25.6%	10.6	5.4	25.5%	10.8	5.3	25.5%	10.2	5.2	25.5%	10.8	5.3
Respiration rate SD (BPM)	26.4%	3.0	1.9	25.6%	4.9	2.0	26.4%	3.0	1.9	26.2%	2.9	1.8	25.8%	4.6	2.5	26.2%	3.0	1.8
SpO <sub>2</sub> Last (%)	26.6%	96.8	3.0	25.5%	93.6	9.8	26.6%	96.7	3.4	26.5%	96.8	2.8	25.7%	95.0	7.6	26.4%	96.8	3.1
SpO <sub>2</sub> Mean (%)	26.6%	97.4	1.8	25.5%	96.3	3.4	26.6%	97.4	1.9	26.5%	97.4	1.9	25.7%	96.4	3.4	26.4%	97.4	2.0
SpO <sub>2</sub> Min (%)	26.6%	89.1	10.4	25.5%	75.5	19.3	26.6%	88.7	11.0	26.5%	89.6	9.5	25.7%	79.6	17.1	26.4%	89.3	10.0
SpO <sub>2</sub> SD (%)	27.4%	2.0	1.4	25.8%	3.7	2.9	27.4%	2.0	1.5	27.2%	2.0	1.3	25.9%	3.5	2.9	27.2%	2.0	1.4
Temperature Last (°F)	16.2%	98.2	1.0	13.4%	98.2	2.3	16.1%	98.2	1.1	16.7%	98.3	0.8	12.6%	98.2	2.2	16.6%	98.3	0.9
Temperature SD (°F)	18.7%	1.1	2.2	15.4%	1.6	2.0	18.6%	1.1	2.2	19.3%	1.0	2.1	14.9%	1.4	2.0	19.1%	1.0	2.1
Features—from laboratory data																		
Complete blood count																		
Lymphocytes SD (%)	0.0%	27.7	10.4	0.0%	19.0	14.6	0.0%	27.4	10.6	0.0%	27.5	10.5	0.0%	19.3	14.4	0.0%	27.3	10.8
Neutrophils Last (%)	0.0%	41.9	33.9	0.0%	60.0	34.8	0.0%	42.5	34.1	0.0%	41.7	34.3	0.0%	58.4	35.4	0.0%	42.2	34.5
Neutrophils SD (%)	0.0%	41.9	33.9	0.0%	60.0	34.8	0.0%	42.5	34.1	0.0%	41.7	34.3	0.0%	58.4	35.4	0.0%	42.2	34.5
PLT count Last (thousand/ $\mu$ L)	0.0%	248.3	96.8	0.0%	194.3	120.7	0.0%	246.6	98.0	0.0%	241.2	88.1	0.0%	188.0	112.7	0.0%	239.6	89.4
PLT count SD (thousand/ $\mu$ L)	0.0%	248.3	96.8	0.0%	194.3	120.7	0.0%	246.6	98.0	0.0%	241.2	88.1	0.0%	188.0	112.7	0.0%	239.6	89.4
RBC count Last (million/ $\mu$ L)	0.0%	4.2	0.9	0.0%	3.6	1.0	0.0%	4.2	0.9	0.0%	4.2	0.9	0.0%	3.6	1.0	0.0%	4.2	0.9
WBC count Last (million/ $\mu$ L)	0.0%	9.0	3.8	0.0%	13.6	8.4	0.0%	9.2	4.1	0.0%	9.1	3.9	0.0%	13.0	8.2	0.0%	9.2	4.2
WBC count SD (million/ $\mu$ L)	0.0%	9.0	3.8	0.0%	13.6	8.4	0.0%	9.2	4.1	0.0%	9.1	3.9	0.0%	13.0	8.2	0.0%	9.2	4.2
Chemistry results																		
Albumin Last (g/dL)	0.0%	4.0	0.6	0.0%	3.1	0.9	0.0%	3.9	0.6	0.0%	4.0	0.6	0.0%	3.2	0.9	0.0%	4.0	0.6
Albumin Mean (g/dL)	0.0%	4.0	0.6	0.0%	3.2	0.8	0.0%	4.0	0.6	0.0%	4.0	0.6	0.0%	3.3	0.9	0.0%	4.0	0.6
Anion gap Last (mEq/L)	0.0%	10.9	2.8	0.0%	13.8	6.0	0.0%	11.0	3.0	0.0%	10.9	2.8	0.0%	13.3	5.5	0.0%	11.0	2.9
AST (SGOT) (units/L of serum)	0.0%	28.4	25.5	0.0%	67.3	71.0	0.0%	29.5	28.7	0.0%	28.6	26.7	0.0%	61.2	66.1	0.0%	29.6	29.2
Bilirubin total Last ( $\mu$ mol/L)	0.0%	0.6	0.5	0.0%	1.2	1.3	0.0%	0.6	0.5	0.0%	0.6	0.5	0.0%	1.2	1.3	0.0%	0.6	0.5
BUN Last (mg/dL)	0.0%	15.9	11.0	0.0%	36.5	27.3	0.0%	16.5	12.3	0.0%	15.8	10.8	0.0%	32.4	23.1	0.0%	16.3	11.7
BUN Mean (mg/dL)	0.0%	16.4	10.9	0.0%	32.3	21.8	0.0%	16.9	11.7	0.0%	16.3	10.9	0.0%	30.5	21.0	0.0%	16.7	11.6
BUN SD (mg/dL)	0.0%	15.9	11.0	0.0%	36.5	27.3	0.0%	16.5	12.3	0.0%	15.8	10.8	0.0%	32.4	23.1	0.0%	16.3	11.7
EGFR Mean (mL/min/1.73 m <sup>2</sup> )	0.0%	62.0	20.0	0.0%	51.3	24.8	0.0%	61.7	20.2	0.0%	61.9	19.8	0.0%	51.5	25.6	0.0%	61.6	20.1
Glucose SD (mmol/L)	0.0%	114.6	45.2	0.0%	138.3	60.0	0.0%	115.3	45.9	0.0%	115.2	46.2	0.0%	139.9	60.6	0.0%	116.0	46.9
Sodium Last (mEq/L)	0.0%	139.0	3.2	0.0%	141.3	6.8	0.0%	139.0	3.4	0.0%	139.0	3.3	0.0%	140.8	6.7	0.0%	139.1	3.5
Arterial blood gas																		
Arterial pCO <sub>2</sub> (mmHg)	0.0%	40.2	3.6	0.0%	40.7	9.9	0.0%	40.2	4.0	0.0%	40.2	3.6	0.0%	39.9	8.5	0.0%	40.2	3.8
Arterial pO <sub>2</sub> (mmHg)	0.0%	96.0	28.8	0.0%	123.0	54.9	0.0%	96.8	30.3	0.0%	95.6	28.3	0.0%	120.4	55.0	0.0%	96.3	29.7
Coagulation																		
PTT Mean (s)	0.0%	31.7	6.9	0.0%	38.1	15.0	0.0%	31.9	7.3	0.0%	31.6	6.7	0.0%	37.7	15.6	0.0%	31.8	7.2
Protime Last (s)	0.0%	12.4	3.0	0.0%	17.5	7.7	0.0%	12.5	3.4	0.0%	12.3	3.0	0.0%	17.0	7.2	0.0%	12.5	3.3
Protime Mean (s)	0.0%	12.4	3.0	0.0%	17.2	6.5	0.0%	12.5	3.2	0.0%	12.4	3.0	0.0%	16.9	6.4	0.0%	12.5	3.2
Protime SD (s)	0.0%	12.4	3.0	0.0%	17.5	7.7	0.0%	12.5	3.4	0.0%	12.3	3.0	0.0%	17.0	7.2	0.0%	12.5	3.3

SD, standard deviation; GCS, Glasgow Coma Score; Min, minimum; Max, maximum; PLT, platelet; RBC, red blood cell; WBC, white blood cell; BUN, blood urea nitrogen; PTT, partial thromboplastin time.

**Table S4** Data features (categorical variables)

Features	Train—80%						Test—20%					
	Alive (n=137,000—97.0%)		Dead (n=4,220—3.0%)		Total (n=141,220)		Alive (n=34,251—97.0%)		Dead (n=1,055—3.0%)		Total (n=35,306)	
	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion
ATC R03AA—alpha and beta adrenoreceptor agonists												
0	132,663	96.8%	3,166	75.0%	135,829	96.2%	33,357	97.4%	866	82.1%	34,223	96.9%
1	4,337	3.2%	1,054	25.0%	5,391	3.8%	894	2.6%	189	17.9%	1,083	3.1%
Demographics—sex (M)												
0	76,400	55.8%	1,807	42.8%	78,207	55.4%	19,037	55.6%	482	45.7%	19,519	55.3%
1	60,600	44.2%	2,413	57.2%	63,013	44.6%	15,214	44.4%	573	54.3%	15,787	44.7%
ICD/CCSR—other lower respiratory disease												
0	122,239	89.2%	3,187	75.5%	125,426	88.8%	30,612	89.4%	811	76.9%	31,423	89.0%
1	14,761	10.8%	1,033	24.5%	15,794	11.2%	3,639	10.6%	244	23.1%	3,883	11.0%
Lab—arterial pCO <sub>2</sub> low												
0	129,216	94.3%	2,169	51.4%	131,385	93.0%	32,560	95.1%	599	56.8%	33,159	93.9%
1	7,784	5.7%	2,051	48.6%	9,835	7.0%	1,691	4.9%	456	43.2%	2,147	6.1%
LOS												
1	24,423	17.8%	814	19.3%	25,237	17.9%	6,106	17.8%	203	19.2%	6,309	17.9%
2	30,020	21.9%	436	10.3%	30,456	21.6%	7,505	21.9%	109	10.3%	7,614	21.6%
3	22,656	16.5%	330	7.8%	22,986	16.3%	5,664	16.5%	83	7.9%	5,747	16.3%
4	14,034	10.2%	287	6.8%	14,321	10.1%	3,509	10.2%	72	6.8%	3,581	10.1%
5	9,352	6.8%	243	5.8%	9,595	6.8%	2,338	6.8%	61	5.8%	2,399	6.8%
6	6,828	5.0%	206	4.9%	7,034	5.0%	1,707	5.0%	51	4.8%	1,758	5.0%
7+	29,687	21.7%	1,904	45.1%	31,591	22.4%	7,422	21.7%	476	45.1%	7,898	22.4%
Tobacco status—never smoker												
0	76,329	55.7%	3,013	71.4%	79,342	56.2%	19,030	55.6%	753	71.4%	19,783	56.0%
1	60,671	44.3%	1,207	28.6%	61,878	43.8%	15,221	44.4%	302	28.6%	15,523	44.0%

ICD, international classification of diseases; CCSR, clinical classifications software refined; Lab, laboratory; LOS, length of stay.

**Table S5** Performance metrics with CI between MONITOR, Braden, and GCS

Model	LOS	Accuracy	AUROC	Recall	Precision	AUPRC	F1	F2
MONITOR	1	0.894 (95% CI: 0.891–0.897)	0.911 (95% CI: 0.902–0.919)	0.750 (95% CI: 0.725–0.775)	0.185 (95% CI: 0.174–0.198)	0.370 (95% CI: 0.339–0.402)	0.297 (95% CI: 0.281–0.314)	0.466 (95% CI: 0.446–0.486)
	2	0.916 (95% CI: 0.913–0.919)	0.920 (95% CI: 0.911–0.928)	0.729 (95% CI: 0.701–0.758)	0.220 (95% CI: 0.204–0.235)	0.376 (95% CI: 0.342–0.409)	0.338 (95% CI: 0.317–0.356)	0.498 (95% CI: 0.475–0.520)
	3	0.904 (95% CI: 0.900–0.908)	0.915 (95% CI: 0.904–0.924)	0.738 (95% CI: 0.707–0.769)	0.228 (95% CI: 0.212–0.245)	0.381 (95% CI: 0.346–0.421)	0.349 (95% CI: 0.328–0.370)	0.510 (95% CI: 0.485–0.535)
	4	0.885 (95% CI: 0.880–0.891)	0.901 (95% CI: 0.888–0.913)	0.740 (95% CI: 0.705–0.775)	0.231 (95% CI: 0.212–0.249)	0.381 (95% CI: 0.342–0.422)	0.352 (95% CI: 0.328–0.376)	0.513 (95% CI: 0.486–0.541)
	5	0.871 (95% CI: 0.865–0.877)	0.892 (95% CI: 0.879–0.905)	0.759 (95% CI: 0.725–0.791)	0.241 (95% CI: 0.221–0.261)	0.362 (95% CI: 0.323–0.402)	0.366 (95% CI: 0.341–0.390)	0.531 (95% CI: 0.503–0.556)
	6	0.856 (95% CI: 0.849–0.863)	0.890 (95% CI: 0.876–0.903)	0.757 (95% CI: 0.717–0.795)	0.240 (95% CI: 0.220–0.259)	0.391 (95% CI: 0.347–0.436)	0.364 (95% CI: 0.339–0.389)	0.529 (95% CI: 0.499–0.555)
	7	0.852 (95% CI: 0.845–0.860)	0.886 (95% CI: 0.870–0.900)	0.756 (95% CI: 0.717–0.795)	0.255 (95% CI: 0.232–0.279)	0.388 (95% CI: 0.340–0.436)	0.381 (95% CI: 0.354–0.410)	0.543 (95% CI: 0.512–0.575)
Braden Score	1	0.945 (95% CI: 0.942–0.947)	0.661 (95% CI: 0.646–0.676)	0.358 (95% CI: 0.330–0.389)	0.229 (95% CI: 0.210–0.250)	0.102 (95% CI: 0.089–0.115)	0.280 (95% CI: 0.257–0.302)	0.322 (95% CI: 0.297–0.347)
	2	0.919 (95% CI: 0.916–0.922)	0.752 (95% CI: 0.737–0.768)	0.575 (95% CI: 0.543–0.605)	0.199 (95% CI: 0.182–0.215)	0.127 (95% CI: 0.114–0.141)	0.295 (95% CI: 0.275–0.317)	0.417 (95% CI: 0.393–0.442)
	3	0.901 (95% CI: 0.897–0.905)	0.734 (95% CI: 0.716–0.752)	0.554 (95% CI: 0.519–0.591)	0.187 (95% CI: 0.171–0.204)	0.119 (95% CI: 0.106–0.134)	0.280 (95% CI: 0.259–0.302)	0.398 (95% CI: 0.372–0.425)
	4	0.882 (95% CI: 0.877–0.888)	0.726 (95% CI: 0.706–0.746)	0.554 (95% CI: 0.515–0.592)	0.191 (95% CI: 0.174–0.210)	0.125 (95% CI: 0.111–0.141)	0.284 (95% CI: 0.262–0.309)	0.402 (95% CI: 0.374–0.432)
	5	0.864 (95% CI: 0.859–0.870)	0.718 (95% CI: 0.699–0.740)	0.557 (95% CI: 0.518–0.599)	0.193 (95% CI: 0.175–0.212)	0.129 (95% CI: 0.115–0.145)	0.287 (95% CI: 0.264–0.311)	0.404 (95% CI: 0.377–0.434)
	6	0.850 (95% CI: 0.843–0.857)	0.717 (95% CI: 0.696–0.738)	0.568 (95% CI: 0.525–0.608)	0.197 (95% CI: 0.177–0.216)	0.135 (95% CI: 0.119–0.151)	0.292 (95% CI: 0.267–0.316)	0.412 (95% CI: 0.379–0.440)
	7	0.838 (95% CI: 0.830–0.847)	0.707 (95% CI: 0.684–0.731)	0.557 (95% CI: 0.510–0.605)	0.199 (95% CI: 0.177–0.222)	0.138 (95% CI: 0.120–0.157)	0.293 (95% CI: 0.265–0.322)	0.410 (95% CI: 0.374–0.446)
GCS Score	1	0.946 (95% CI: 0.944–0.948)	0.657 (95% CI: 0.644–0.671)	0.351 (95% CI: 0.323–0.378)	0.232 (95% CI: 0.212–0.253)	0.101 (95% CI: 0.089–0.113)	0.279 (95% CI: 0.258–0.301)	0.318 (95% CI: 0.294–0.342)
	2	0.946 (95% CI: 0.943–0.948)	0.674 (95% CI: 0.658–0.690)	0.385 (95% CI: 0.353–0.417)	0.238 (95% CI: 0.215–0.260)	0.110 (95% CI: 0.096–0.125)	0.294 (95% CI: 0.269–0.319)	0.342 (95% CI: 0.314–0.370)
	3	0.938 (95% CI: 0.935–0.941)	0.670 (95% CI: 0.652–0.688)	0.381 (95% CI: 0.346–0.417)	0.247 (95% CI: 0.224–0.274)	0.116 (95% CI: 0.101–0.133)	0.300 (95% CI: 0.273–0.328)	0.344 (95% CI: 0.313–0.376)
	4	0.927 (95% CI: 0.923–0.931)	0.663 (95% CI: 0.645–0.682)	0.376 (95% CI: 0.338–0.411)	0.254 (95% CI: 0.227–0.283)	0.122 (95% CI: 0.105–0.141)	0.303 (95% CI: 0.273–0.333)	0.343 (95% CI: 0.310–0.375)
	5	0.917 (95% CI: 0.911–0.921)	0.661 (95% CI: 0.641–0.679)	0.377 (95% CI: 0.339–0.412)	0.259 (95% CI: 0.230–0.288)	0.128 (95% CI: 0.110–0.146)	0.307 (95% CI: 0.275–0.336)	0.345 (95% CI: 0.310–0.377)
	6	0.906 (95% CI: 0.901–0.912)	0.665 (95% CI: 0.642–0.688)	0.394 (95% CI: 0.350–0.440)	0.262 (95% CI: 0.230–0.295)	0.137 (95% CI: 0.115–0.159)	0.315 (95% CI: 0.279–0.349)	0.358 (95% CI: 0.318–0.397)
	7	0.898 (95% CI: 0.891–0.905)	0.666 (95% CI: 0.643–0.688)	0.402 (95% CI: 0.358–0.447)	0.269 (95% CI: 0.235–0.304)	0.144 (95% CI: 0.123–0.168)	0.322 (95% CI: 0.287–0.358)	0.365 (95% CI: 0.327–0.404)

CI, confidence intervals; GCS, Glasgow Coma Score; LOS, length of stay; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.