Health Outcome Predictive Modelling in Intensive Care Units

Chengqian Xian[1*], Camila P. E. de Souza[1†] and Felipe F. Rodrigues[1,2†]

[1*]Department of Statistical and Actuarial Sciences, University of Western Ontario, 1151 Richmond Street, London, N6A 5B7, Ontario, Canada.
[2]School of Management, Economics, and Mathematics, King's University College at Western University, 266 Epworth Avenue, London, N6A 2M3, Ontario, Canada.

*Corresponding author(s). E-mail(s): cxian3@uwo.ca;
Contributing authors: camila.souza@uwo.ca; frodrig7@uwo.ca;
†These authors contributed equally to this work.

Abstract

The literature in Intensive Care Units (ICUs) data analysis focuses on predictions of length-of-stay (LOS) and mortality based on patient acuity scores such as Acute Physiology and Chronic Health Evaluation (APACHE), Sequential Organ Failure Assessment (SOFA), to name a few. Unlike ICUs in other areas around the world, ICUs in Ontario, Canada, collect two primary intensive care scoring scales, a therapeutic acuity score called the "Multiple Organs Dysfunctional Score" (MODS) and a nursing workload score called the "Nine Equivalents Nursing Manpower Use Score" (NEMS). The dataset analyzed in this study contains patients' NEMS and MODS scores measured upon patient admission into the ICU and other characteristics commonly found in the literature. Data were collected between January 1st, 2015 and May 31st, 2021, at two teaching hospital ICUs in Ontario, Canada. In this work, we developed logistic regression, random forests (RF) and neural network (NN) models for mortality (discharged or deceased) and LOS (short or long stay) predictions. Considering the effect of mortality outcome on LOS, we also combined mortality and LOS to create a new categorical health outcome called LMClass (short stay & discharged, short stay & deceased, or long stay without specifying mortality

outcomes), and then applied multinomial regression and RF for its prediction. Five repetitions corresponding to five random starting points have been done in RF and NN for model optimization, and 5-fold cross-validation (CV) was also carried out for model stability investigation. Results show that logistic regression is the optimal model in mortality prediction with the highest area under the curve (AUC) of 0.795 and also in LMClass prediction with the highest accuracy of 0.630. In contrast, in LOS prediction, RF outperforms the other methods with the highest AUC of 0.689. This study also demonstrates that MODS and NEMS, as well as their components measured upon patient arrival, significantly contribute to health outcome prediction in ICUs.

Keywords: Intensive Care Units, Health outcome, Predictive modelling, MODS, NEMS

1 Introduction

The Intensive Care Unit (ICU) is a unique hospital department, providing the highest level of treatment for a hospital's highest acuity patients. It is an intrinsically technological environment where each patient generates thousands of data points per day, and data-driven management applied to ICU allows not only an evaluation of ICU performance but has other implications, including better planning of scarce resources and transition of care and discharge [1, 2].

ICU scoring system is an essential tool to describe the severity of patients' illnesses to improve clinical decision-making and predict patients' health outcomes [3–7]. Existing scoring systems to assess patient illness severity on admission to ICU include the Acute Physiology and Chronic Health Evaluation II and its variations (APACHE II) [8] as well as the Simplified Acute Physiology Score (SAPS II) [9]. In addition to describing the severity of a patient's disease, Multiple Organ Dysfunction Score (MODS) [10] and Sequential Organ Failure Assessment Score (SOFA) [11] were also developed to specifically evaluate patients' organ function or determine the rate of organ failure. The MODS score (see the full list of its components in Appendix Table A3) is scaled from 0 to 24, and it is constructed from six organ systems and demonstrates a strong correlation with the risk of ICU mortality [10]. The Nine Equivalents Nursing Manpower Use Score (NEMS) [12] (see the full list of its components in Appendix Table A4) was developed from the Therapeutic Intervention Scoring System (TISS) [13] to measure the nursing workload in ICU. NEMS is based on nine life support interventions ranging from 0 to 56, and has been validated in an adult 30-bed medical-surgical ICU in a tertiary care university hospital. Its good agreement is further confirmed with TISS-28 [14].

With advances in information technology and data science, statistical models and machine learning methods have been applied to ICU data for length-of-stay (LOS) and mortality predictions [1, 15–18]. In LOS prediction, Zimmerman [7] developed a multivariate linear regression procedure using

APACHE IV to estimate ICU stay using data across ICUs in the United States. Results in [7] show that the accuracy and utility of the predictions based on the APACHE IV model were unsatisfactory. Zimmerman and Andrew [19] improved patient LOS prediction by firstly optimizing a threshold for a prolonged stay and building a multivariate linear regression with the severity score information on day five, achieving a better prediction than that based on ICU day one information alone. A linear regression model based on the APACHE III score was applied in [20] for ICU LOS prediction of patients in Australia and New Zealand. They discovered that the ICU level (level 1, level 2, etc.) and locality significantly affected predicting length of stay and obtained a high R^2 of 0.88 in the validation set. Most often, scholars and practitioners are interested in binary classification for LOS prediction (long-stay or short-stay). Therefore classification methods, including logistic regression, support vector machine (SVM), random forests (RF), and neural networks (NN), were also implemented in predicting prolonged LOS or short LOS [21–23]. Neural networks were developed as predictive instruments for ICU LOS for the first time in 1992 in Tu and Guerriere's study [22]. They defined the prolonged LOS as a stay longer than two days and found that the NN model performed well with an area under the receiving operating characteristic curve (AUC) of 0.696 in the validation set. Rein et al. [21] and Rodrigues [23] performed similar work on LOS prediction by applying machine learning methods but based on different severity scores (SOFA score in [21] while MODS in [23]). In [21], prolonged LOS (stay longer than ten days) prediction was modelled via classification techniques, including neural networks, k-nearest neighbors (KNN), SVM, classification trees, and RF, while the continuous patient LOS was predicted by regression. The result was that SVM outperformed other machine learning methods in short or long-LOS prediction with an AUC value of 0.82 in the validation set. In Rodrigues [23], machine learning methods were applied, and a parametric Accelerated Failure Time model was developed for clinical LOS prediction, which resulted in an efficient estimation of the expected time to LOS discharge for each patient upon arrival in the ICU.

In ICU mortality prediction (discharged or deceased), a binary classification problem, logistic regression was implemented in [20, 23–27]. The authors of [25] conducted the first study using ICU data from Ontario, Canada, where MODS and NEMS were measured during the first 24 h of ICU admission. They developed a logistic regression model and found that age, source of admission, ICU admitting diagnosis, MODS and NEMS were significant covariates in predicting ICU mortality and obtained an AUC value of 0.760 in the validation set. Machine learning methods are also widely used in mortality prediction [23, 28, 29]. Following [25], Rodrigues [23] also analyzed ICU data from Ontario and proposed different kinds of supervised learning models, including SVM, NN, RF, and super learner (SL). The best performers were RF and SL, with AUC values ranging from 0.75 to 0.76 in the validation set. Recently, a new study using data from Medical Information Mart for Intensive Care III database was conducted by Ahmed et al. [30] to predict mortality in trauma

patients admitted to ICU in Israel. They first selected the risk factors by univariate Cox regression and multivariate analyses. They then applied deep neural network (DNN) and other traditional machine learning models, including Linear Discriminant Analysis, Naive Bayes, tree-based model, and KNN for prediction using the selected predictors. Their findings are that the DNN model outperformed all the other methods with a high accuracy value of 0.923 and an AUC value of 0.91.

In this work, we expand the work of [23, 25] by adding the components of MODS and NEMS as predictors in a logistic regression model and applying machine learning methods (random forests and neural networks) for mortality and LOS predictions. Model optimization in neural networks and model stability investigation via five-fold cross-validation (CV) have been proposed to validate the work of [23]. We also further analyze the effects of MODS and NEMS on mortality and ICU LOS by logistic regression along with its corresponding 95% confidence interval. To the best of our knowledge, we are the first ones to combine mortality outcome with LOS to create a new categorical health outcome called LMClass (meaning LOS-Mortality Class) and develop multinomial regression and random forests to predict LMClass using MODS and NEMS scales.

The rest of the paper is structured as follows. Section 2 presents the material and the statistical methodology. Then, Section 3 describes our analysis results. Finally, the conclusion and discussion are presented in Section 4.

2 Material and methods

2.1 Data source and data management

Our research is a retrospective study conducted at two teaching hospital ICUs in southwestern Ontario, which specialize in the care of various patient populations, including neurosurgical, cardiovascular surgery, and transplantation patients. Data were collected from Jan. 1, 2015, to May. 31, 2021 and stored in four separate data sets called MODS, NEMS, Source and Awaiting Transfer. MODS is the data set containing the MODS score along with its components measured upon patient admission to ICU. NEMS is another important set containing patients' NEMS scores and their components measured daily in ICU. ICU discharge time and destination are also provided in MODS and NEMS data sets. The Source set includes de-identified patient general characteristics (e.g., age, sex) and admission characteristics (e.g., admission source, admission diagnosis, patient category, referring service). The last data set, Awaiting Transfer, provides the admission time and the awaiting transfer discharge start date time, both of which were used to calculate the clinical LOS.

Since we have four separate data sets, several new variables were created in each set before merging them into a single data set. In the MODS set, we created *Mortality* as a binary response which can be constructed from discharge destination: 1 if the patient is deceased at the end of the ICU stay, otherwise 0 for being discharged alive. Besides, we calculated the *total LOS*,

defined as the period between patients' admission to and exit from ICU, which is used to detect extreme values of stay in data cleaning procedure as done in [23]. In the Source set, we edited the admission source and admission diagnosis by combining some of their categories in the same way proposed in [23]. For admission sources, we kept the Emergency Department, Operating Room, and Unit/Ward/Stepdown while combining the other levels to Outside Hospital/Other. For admission diagnosis, we kept the Cardiovascular/Cardiac/Vascular, Gastrointestinal, Neurological, Respiratory, and Trauma while combining other levels to Other. In the Awaiting Transfer set, we calculated the *clinical LOS*, defined as the period between patient admission to ICU and the physician's disposition decision (i.e., transfer or discharge). Then the prolonged LOS called *IsLong* is defined as a stay longer than five days, based on the empirical distribution of LOS as discussed in Section 3.1. In other words, *IsLong* takes the value of 1 if clinical LOS > 5 days and 0 otherwise. Besides, *LMClass*, a categorical response with three levels, was also created by combining *Mortality* and *IsLong*: short stay & discharged, short stay & deceased, or long stay without specifying mortality outcomes.

To build predictive models for each health outcome, we need to combine these four separate data sets into a single one. Patients' ID and admission time can link these four data sets. We first extracted the MODS score with its components from the MODS set and used patients' IDs and admission time to merge the NEMS score with its components on admission day in the NEMS set. To obtain the clinical LOS, patient characteristics, and admission characteristics, we merged the latest combined data set with the Awaiting Transfer set and the Source set, resulting in a single data set with 15,474 cases. Similar to [23], some cases with large total LOS (≥ 60 *days*, 90 cases), unknown or missing sex (15 cases), and unusual age (≥ 110 *years*, 19 cases) were removed from our merged data set for further analysis, resulting in a finalized data set with 15,350 cases. A flow chart of this process is provided in Figure 1.

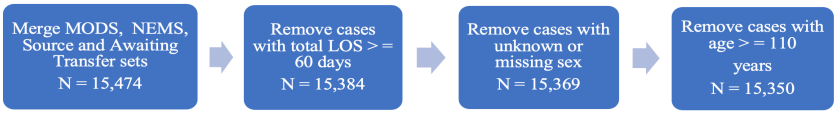

Fig. 1: Flow chart of data cleaning

2.2 Statistical analysis

The finalized data set ($N = 15,350$) was split into a training set ($N = 10,745$) and a validation set ($N = 4,605$) with a ratio of 7:3. Each proposed model was built on the training set and validated on the validation set.

Logistic and multinomial regression are two widely used generalized linear models for modelling binary and multi-class responses, respectively [31]. We built logistic regression models for mortality and LOS predictions while multinomial regression for LMClass prediction. These models were fitted on the training set using all the predictors available, including the components of MODS and NEMS followed by the best model selection based on the Akaike information criterion (AIC) in both directions [31].

Random forests (RF), proposed in [32], is another popular model for classification by constructing a multitude of decision trees. The two most critical parameters in the random forests model are the number of trees to be built and the number of variables randomly sampled as candidates at each split. In the R package *RandomForest* [33], these parameters are represented by function arguments *ntree* and *mtry*, respectively. In our data set, there are 23 predictors, so *mtry* can be $1, 2, ...,$ or 23. For the number of trees, we chose from 100, 500 and 1000. Combining both *ntree* and *mtry*, we have 69 (i.e., 23×3) alternative models built on the training set using all predictors available. To analyze the black-box mechanisms of random forests, one of the most efficient variable importance measures, mean decrease accuracy (MDA) introduced in [34], can be applied, which is a method of computing the predictor importance on permuted out-of-bag samples based on the mean decrease in the accuracy. In other words, if MDA is high for a predictor, this predictor is important. Visualization of MDA for all predictors is provided after fitting an RF model in the same R package, *RandomForest*.

Neural networks (NN) were built based on the resilient back-propagation with weight backtracking algorithm proposed by Riedmiller M. in 1994 [35]. Before modelling, we conducted data preprocessing for numeric predictors (e.g., age) and ordinal categorical predictors (e.g., components of MODS) by min-max normalization and for nominal categorical predictors (e.g., admission diagnosis) by a one-hot encoding scheme [36]. The two most important parameters of the NN are the number of hidden layers and the number of neurons on each layer. We consider one or two hidden layers with one to five neurons, and as a result, we need to find the optimal NN model from 30 (i.e., $5 + (5 \times 5)$) alternative combinations of different numbers of hidden layers and neurons.

For optimization of RF and NN models, we use average AUC as a criterion from five replicates in mortality and LOS predictions, while accuracy as a criterion in LMClass prediction. In mortality and LOS binary predictions, model discrimination performance was assessed by AUC from the receiver operating characteristic (ROC) curve, sensitivity (Sen), specificity (Spe), accuracy (Acc), Matthews correlation coefficient (MCC), positive predictive value (PPV), negative predictive value (NPV) and F1 score. For LMClass prediction, accuracy and balanced accuracy were used to evaluate the performance of each proposed model. All statistical analyses were performed using software R version 4.2.1.

3 Results

In this section, we first present the descriptive analysis results of our data set, and then the results regarding mortality, LOS, and LMClass prediction, respectively. We also quantitatively elaborate on how MODS and NEMS affect the prediction of health outcomes via the odds ratio (i.e., the relative risk ratio) from regression-based models.

3.1 Descriptive analysis

In our dataset, 11,963 admitted patients were discharged alive when exiting the ICU while 3,387 patients died. Most patients stayed less than 5 days in ICU, accounting for 71.26% of the study population. Table 1 shows the general characteristics of patients (excluding the components of MODS and NEMS) in the training and validation sets. The median with interquartile interval (IQI) was presented for numeric variables, and for categorical variables, raw counts and percentages were presented. We can observe that the median values of MODS and NEMS scores are the same in the training and validation sets (5 points with IQI 3-7 and 32 points with IQI 27-39, respectively). The mortality rate in training is 21.83%, 0.78% lower than that in the validation set. Median clinical LOS in training and validation is 2.545 days (IQI 1.082-5.776) and 2.537 days (IQI 1.115-5.878), respectively.

Figure 2 shows the histogram with the estimated density function (the red curve) of clinical LOS. There are only a few cases with clinical LOS longer than 20 days and most of cases have clinical LOS between 0 and 4 days, resulting in a right-skewed distribution. Therefore, to avoid issues with imbalanced data, we defined our prolonged LOS as a stay longer than 5 days instead of 7 or 21 days in [23].

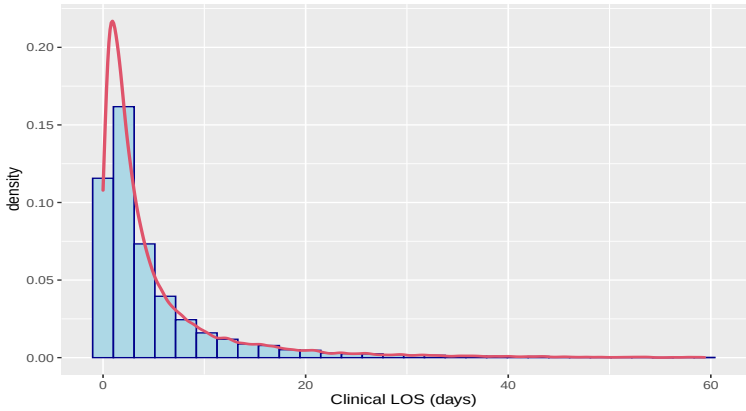

Fig. 2: Histogram plot with estimated density (red curve) of clinical LOS.

Table 1: Characteristics of ICU admitted patients in the training and validation sets.

Variables	Training	Validation
MODS Score (median [interquartile interval] [1])	5 [3, 7]	5 [3, 7]
NEMS Score (median [interquartile interval])	32 [27, 39]	32 [27, 39]
Mortality, yes, (N, [%] [2])	2346 [21.83%]	1041 [22.61%]
Total LOS, days, (median, [interquartile interval])	3.361 [1.671, 6.913]	3.372 [1.645, 6.885]
Clinical LOS, days, (median, [interquartile interval])	2.545 [1.082, 5.064]	2.537 [1.115, 5.878]
Age, years, (median, [interquartile interval])	63.39 [51.14, 73.46]	63.55 [50.94, 73.78]
Sex, Male, (N, [%])	6269 [58.34%]	2690 [58.41%]
Patient Category		
Medical, (N, [%])	6769 [63.00%]	2855 [62.00%]
Surgical, (N, [%])	3976 [37.00%]	1750 [38.00%]
Admission Source		
Emergency Department, (N, [%])	3740 [34.81%]	1551 [33.68%]
Operating Room, (N, [%])	2045 [19.03%]	922 [20.02%]
Unit/Ward/Stepdown, (N, [%])	2214 [20.60%]	960 [20.85%]
Outside Hospital/Other, (N, [%])	2746 [25.56%]	1172 [25.45%]
Admission Diagnosis		
Cardiovascular/Cardiac/Vascular, (N, [%])	1754 [16.32%]	750 [16.29%]
Gastrointestinal, (N, [%])	988 [9.19%]	417 [9.06%]
Neurological, (N, [%])	1473 [13.71%]	561 [12.18%]
Respiratory, (N, [%])	3111 [28.95%]	1336 [29.01%]
Trauma, (N, [%])	832 [7.74%]	370 [8.03%]
Other, (N, [%])	2587 [24.08%]	1171 [25.43%]

[1]for numeric variables, the median with interquartile interval was presented;

[2]for categorical variables, raw counts and percentages were presented.

3.2 Mortality prediction

The best RF model in mortality prediction is with $mtry = 3$ and $ntree = 1000$. In other words, the number of variables randomly sampled as candidates at each split on the tree is three, and 1000 trees were built to construct the forests. The best NN model has one hidden layer on which there are two neurons.

Performance of logistic regression, RF and NN on the validation set for mortality prediction is shown in Table 2. ROC curves for all models in the validation set are also presented in Figure 3 which shows no big difference among different models. Logistic regression outperforms RF and NN with the highest scores in AUC (0.795), accuracy (0.705), F1 score (0.532) and PPV (0.415). RF performs the best in achieving the highest sensitivity of 0.748 and NPV of 0.904. Logistic regression has the second highest sensitivity (0.743), while NN provides a relatively lower sensitivity (0.732). The MCC values are low, ranging from 0.364 to 0.373 and the F1 scores are better, ranging from 0.527 to 0.532. All models are stable since the standard deviations of AUC values from five-fold CV are small, ranging from 0.005 to 0.011.

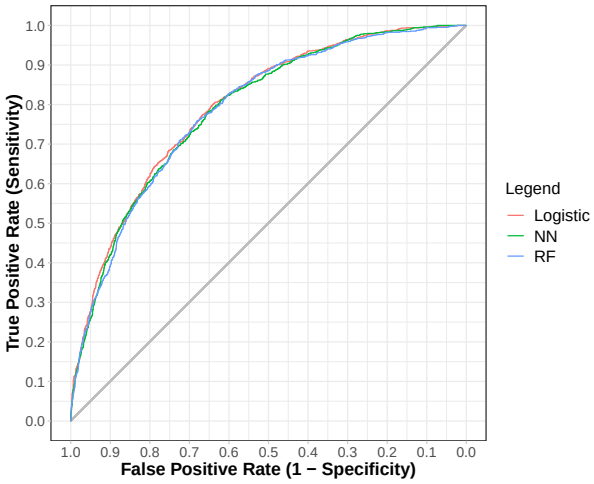

Fig. 3: ROC curves for mortality prediction in the validation set.

Red: logistic regression; Green: random forests; Blue: neural networks

Table 3 shows the results of selected predictors for logistic regression based on AIC. Those selected predictors include six components of MODS (Haematologic, Hepatic, Renal, Cardiovascular, Neurologic and Respiratory), NEMS score, six components of NEMS (Basic Monitoring, Intracranial Pressure Monitor, Dialysis, Intra-Aortic Balloon Pump, Other Interventions Within this Unit and Interventions Outside this Unit), Age, Sex, Patient Category, Admission Source and Admission Diagnosis. The likelihood ratio test (LRT) was applied to all the selected predictors to assess their significance, and the corresponding

Table 2: Performance of logistic regression, RF and NN on the validation set for mortality prediction.

Model	AUC	95% CI	Cutoff	Sen	Spe	Acc	MCC	F1	PPV	NPV
Logistic	0.795	(0.780, 0.810)	0.206	0.743	0.694	0.705	0.372	0.532	0.415	0.902
RF	0.788	(0.773, 0.803)	0.186	0.748	0.690	0.703	0.373	0.532	0.413	0.904
NN	0.789	(0.774, 0.804)	0.221	0.732	0.695	0.703	0.364	0.527	0.412	0.899

AUC: area under the curve; 95% CI: 95% confidence interval for AUC; Cutoff: optimal threshold for determining mortality obtained via Youden's J statistic [37]; Sen: sensitivity; Spe: specificity; Acc: accuracy; MCC: Matthews correlation coefficient; F1: F1 score; PPV: positive predictive value; NPV: negative predictive value.

p-values were reported in the table. Except for one of the NEMS components, Intra-Aortic Balloon Pump, we find that all other selected predictors are statistically significant and useful for mortality prediction. A visualization of predictor importance using MDA from the RF model is presented in the left of Figure 4, which shows that NEMS, MODS and ICU admission source are the three most important predictors in mortality prediction with an MDA higher than 50%.

The odds ratio and its 95% confidence interval (CI) were also calculated for each selected predictor. We find that MODS score is not selected in the optimal model but its six components, which is reasonable when correlation exists among predictors (MODS and its components) [38]. As a result, the significance of MODS score is represented by its components. However, NEMS is selected in the model and has an odd ratio of 1.07 (95% CI = [1.06, 1.08]), which indicates that one point increase in NEMS score would increase the relative risk of mortality (i.e., death) by 7% (95% CI = [6%, 8%]), holding the other covariates fixed.

In this study, sensitivity might be the main criterion used to evaluate the best model, because the ICU practitioners usually care more about identifying the mortality group for clinical reasons. Specifically, they are concerned about the proportion of correct prediction for those patients who deceased at the end of ICU stay. In mortality prediction, the sensitivity coming from RF is 0.748, meaning that it works relatively well in predicting the mortality outcome in those patients who deceased in the end and 74.8% could be correctly predicted. NPV is also an important index in mortality prediction. RF has the highest NPV of 0.904, meaning that if we predict that someone will be discharged alive at the end of the stay, he would probably be with a probability of 90.4%.

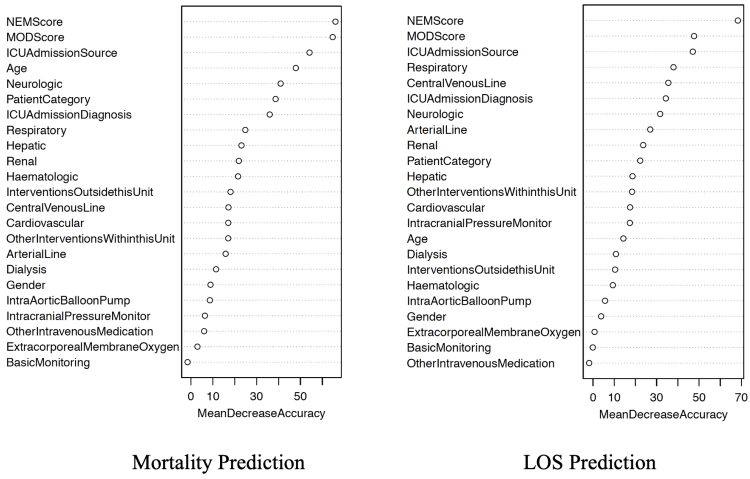

Fig. 4: Importance of predictors based on RF model in mortality and LOS predictions.

Table 3: Selected predictors in logistic regression for mortality prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
Haematologic			< 0.0001
None (reference)	-	-	
Minimal	0.38	1.46 [1.25, 1.71]	
Mild	0.45	1.57 [1.25, 1.96]	
Moderate	0.85	2.35 [1.82, 3.02]	
Severe	1.22	3.40 [2.22, 5.20]	
Hepatic			< 0.0001
None (reference)	-	-	
Minimal	0.13	1.14 [0.94, 1.39]	
Mild	0.27	1.30 [0.94, 1.80]	
Moderate	0.87	2.39 [1.57, 3.61]	
Severe	1.20	3.32 [2.14, 5.15]	
Renal			< 0.0001
None (reference)	-	-	
Minimal	0.30	1.36 [1.20, 1.53]	
Mild	0.49	1.64 [1.38, 1.95]	
Moderate	0.57	1.77 [1.37, 2.27]	
Severe	0.13	1.14 [0.86, 1.51]	
Cardiovascular			< 0.0001
None (reference)	-	-	
Minimal	0.27	1.31 [1.16, 1.47]	
Mild	0.57	1.77 [1.44, 2.17]	
Moderate	0.53	1.70 [1.24, 2.32]	
Severe	0.71	2.03 [1.22, 3.42]	
Neurologic			< 0.0001
None (reference)	-	-	
Minimal	0.28	1.32 [1.09, 1.61]	
Mild	0.13	1.14 [0.93, 1.40]	
Moderate	0.06	1.06 [0.84, 1.34]	
Severe	0.89	2.45 [2.13, 2.81]	
Respiratory			< 0.0001
None (reference)	-	-	
Minimal	-0.07	0.94 [0.80, 1.09]	
Mild	0.13	1.14 [0.98, 1.33]	
Moderate	0.33	1.39 [1.20, 1.60]	
Severe	0.52	1.68 [1.37, 2.05]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table 3 continued: Selected predictors in logistic regression for mortality prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
NEMS Score	0.07	1.07 [1.06, 1.08]	< 0.0001
Basic Monitoring			0.005
No (reference)	-		
Yes	-2.23	0.11 [0.03, 0.48]	
Intracranial Pressure Monitor			< 0.0001
No (reference)	-	-	
Yes	0.74	2.09 [1.46, 2.95]	
Dialysis			0.002
No (reference)	-	-	
Yes	-0.44	0.65 [0.49, 0.85]	
Intra-Aortic Balloon Pump			0.119
No (reference)	-	-	
Yes	-0.40	0.67 [0.40, 1.11]	
Other Interventions			0.001
No (reference)	-	-	
Yes	-0.21	0.81 [0.72, 0.92]	
Interventions Outside this Unit			< 0.0001
No (reference)	-	-	
Yes	-0.43	0.65 [0.57, 0.75]	
Age	0.03	1.03 [1.03, 1.04]	< 0.0001
Sex			0.001
Female (reference)	-	-	
Male	-0.17	0.84 [0.76, 0.94]	
Patient Category			0.001
Medical (reference)	-	-	
Surgical	-0.25	0.78 [0.67, 0.91]	
Admission Source			< 0.0001
Emergency (reference)	-	-	
Operating Room	-0.94	0.39 [0.31, 0.48]	
Outside Hospital/Other	-0.01	0.99 [0.86, 1.13]	
Unit/Ward/Stepdown	0.36	1.43 [1.23, 1.66]	
Admission Diagnosis			< 0.0001
Cardiovascular (reference)	-	-	
Gastrointestinal	-0.45	0.64 [0.51, 0.80]	
Neurological	-0.06	0.94 [0.79, 1.14]	
Other	-0.58	0.56 [0.47, 0.66]	
Respiratory	-0.48	0.62 [0.53, 0.72]	
Trauma	-0.11	0.89 [0.68, 1.16]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

3.3 LOS prediction

In LOS prediction, the optimal RF model has the same parameter structure as that in mortality prediction where three predictors were randomly sampled at each split on the tree and 1000 trees were built. The best NN model has one hidden layer with one neuron. This simplest NN model was also presented and investigated as the best model in [39] for survival prediction in the ICUs, which is consistent with the empirical analysis result that reducing the complexity of a neural network structure may provide a better performance of prediction for health outcomes in ICUs [40, 41].

Table 4 presents the evaluation measures of all models for LOS prediction in the validation set. The AUC values of all models are lower than those in mortality prediction by around 10%. The MCC values ranging from 0.247 to 0.251 are very low while the F1 scores ranging from 0.503 to 0.508 are still acceptable. On the whole, logistic regression and RF outperform NN. Specifically, logistic regression has the highest score in sensitivity (0.673), MCC (0.251) and NPV (0.820), while RF has the highest score in AUC (0.689), specificity (0.622), accuracy (0.630) and PPV (0.411). The corresponding ROC curves were shown in Figure B1 in Appendix B and once again, no big difference among models can be seen from the plot. All models are stable since the standard deviations of AUC values from five-fold CV are small, ranging from 0.016 to 0.020.

In LOS prediction, sensitivity is one of the most important indices for the choice of models. Logistic regression has the highest sensitivity of 0.673, indicating that 67.3% of patients who stayed more than 5 days could be correctly predicted. The negative predictive values (NPV) from logistic regression is also the highest (0.82), meaning that if we predict one stays less than 5 days, he would probably stay less than 5 days with a probability of 0.82.

Table A1 in Appendix A contains information of the selected predictors for logistic regression in LOS prediction. MODS score with its two components (Cardiovascular and Respiratory), NEMS score with its five components (Central Venous Line, Arterial Line, Intracranial Pressure Monitor, Dialysis and Interventions Outside this Unit), Sex, Admission Source and Admission Diagnosis are selected in the best model. The odds ratios for MODS and NEMS are 1.04 (95% CI = [1.02, 1.06]) and 1.04 (95% CI = [1.04, 1.05]), respectively. This indicates that one point increase in MODS score or in NEMS score would increase the relative risk of staying more than 5 days by 4%, holding the other predictors fixed. Furthermore, from the right plot of Figure 4, we can see NEMS and MODS are both important predictors in prolonged LOS prediction, which is consistent with results of LRT for testing the significance of MODS and NEMS based on logistic regression.

3.4 LMClass prediction

In LMClass prediction, as we did in mortality and LOS predictions, we tried to apply the regression based, RF and NN models. However, we found that due to the data structure complexity, the NN model failed to learn the pattern

Table 4: Performance of logistic regression, RF and NN on the validation set for LOS Prediction.

Model	AUC	95% CI	Cutoff	Sen	Spe	Acc	MCC	F1	PPV	NPV
Logistic	0.681	(0.665, 0.698)	0.275	0.673	0.604	0.624	0.251	0.508	0.408	0.820
RF	0.689	(0.673, 0.705)	0.272	0.650	0.622	0.630	0.247	0.503	0.411	0.814
NN	0.682	(0.665, 0.698)	0.299	0.672	0.604	0.624	0.250	0.508	0.408	0.819

AUC: area under the curve; 95% CI: 95% confidence interval for AUC; Cutoff: optimal threshold for determining mortality obtained via Youden's J statistic [37]; Sen: sensitivity; Spe: specificity; Acc: accuracy; MCC: Matthews correlation coefficient; F1: F1 score; PPV: positive predictive value; NPV: negative predictive value.

using the training data and could not converge, even when we considered a network of two layers with five neurons on each layer. Therefore, we present below the results from multinomial regression and RF for LMClass prediction.

The number of random predictors considered at each split and the number of trees built of the optimal RF model for LMClass prediction are three and 500, respectively. Table 5 presents each class's accuracy and balanced accuracy from the fitted multinomial regression and RF on the validation set. Multinomial regression, with a higher accuracy of 0.630, outperforms slightly than RF which has an accuracy of 0.624. Except for a lower balanced accuracy in the class of long stay without specifying mortality outcomes, multinomial regression also returns a higher balanced accuracy in the other two classes, short stay & deceased and short stay & discharged. Five-fold CV results show that both the fitted multinomial regression and RF model are stable with standard deviations of accuracy ranging from 0.010 to 0.014.

The information of selected predictors in the multinomial regression model for LMClass prediction is provided in Table A2 in Appendix A . In multinomial regression, we set the baseline class to be short stay & discharged, and the odds ratio with 95% CI for each selected predictor was collected for another two classes (e.g. short stay & deceased and long stay without specifying mortality outcomes) with respect to the baseline. The odds ratios of MODS for the class short stay & deceased and the class long stay without specifying mortality outcomes are 1.40 (95% CI = [1.34, 1.46]) and 1.15 (95% CI = [1.11, 1.19]), respectively. This means if one point increase in MODS score would increase the relative risk of short stay & discharged over short stay & deceased and long stay without specifying mortality outcomes by 40% (95% CI = [34%, 46%]) and 15% (95% CI = [11%, 19%]), respectively. Similarly, the odds ratios of NEMS for short stay & deceased and long stay without specifying mortality outcomes are 1.11 (95% CI = [1.10, 1.13]) and 1.07 (95% CI = [1.06, 1.08]), respectively. This means that one point increase in NEMS score would increase the relative risk of short stay & discharged over short stay & deceased and long stay without specifying mortality outcomes by 11% (95% CI = [10%, 13%]) and 7% (95% CI = [6%, 8%]), respectively, holding the other predictors fixed. As a reference, importance of predictors based on MDA is visualized in Figure B2 in the Appendix B which shows that both NEMS and MODS are important in LMClass prediction.

4 Conclusions and Discussion

In this work, we developed several models for health outcomes prediction in intensive care units. Compared with [25], adding the components of MODS and NEMS in the logistic regression for mortality prediction has an improvement of 3.5% in AUC values in the validation set (see Table 6). This study also demonstrates that MODS and NEMS with their components measured upon patient arrival significantly contribute to health outcome prediction in ICUs. In mortality prediction, achieving the highest sensitivity and NPV, RF

Table 5: Performance of multinomial regression, RF and NN on the validation set for LMClass Prediction.

Model	Accuracy	*L.*[1]	*S. and Dec.*[2]	*S. and Dis.*[3]
Multinomial	0.630	0.594	0.621	0.679
RF	0.624	0.597	0.575	0.660

[1]Balanced accuracy for long stay without specifying mortality outcomes

[2]Balanced accuracy for short stay & deceased

[3]Balanced accuracy for short stay & discharged

outperforms logistic regression and NN, but logistic regression achieves the highest AUC. In LOS prediction, no big difference in the performance appears in the logistic regression and NN model. In practice, we need to evaluate the pros and cons of each model, and choose the best according to the type and goals of the analysis. For example, if we are concerned about the accuracy of mortality prediction, logistic regression is our first choice, while we choose RF if we emphasize on the sensitivity. Explanation power may also play a role, especially with respect to the implications of such predictions. As an example, the predictors of long stays may help inform capacity planning and resource scheduling.

Furthermore, in terms of the definition of prolonged stay in the ICU, random forests and neural networks have greatly improved LOS prediction when we cut the short and long LOS at 5 days instead of 7 or 21 days as in [23]. A comparison on AUC values in LOS prediction between previous works and our study is provided in Table 7. However, as in [23] for LOS prediction, we find it is harder to classify a short or long stay than to detect mortality status. The underlying reason could be the definition of prolonged LOS as a binary health outcome. To improve the prediction accuracy, survival models can be developed for LOS prediction, and in this scenario LOS can be considered as a continuous time-to-event response.

A trade-off between interpretation power and accuracy of prediction usually exists in predictive modelling. Logistic and multinomial regression models provide an interpretation for quantitative relationships between predictors and health outcomes using odds ratio (i.e., relative risk). Compared with regression-based models, RF provides qualitative relationships using MDA, while NN is a black box whose statistical theoretical justifications are still under investigation in different frameworks [42, 43].

To our best of knowledge, we are the first to combine mortality with prolonged LOS to construct a new categorical health outcome and develop MODS and NEMS based predictive models for its prediction. In our expectation, more complexity occurs in this three-level outcome, making it more challenging to achieve high prediction accuracy. More complex deep learning models such as convolutional neural networks [44] and recurrent neural networks [45] can be applied but with higher computational costs.

It is important to note that, our data, with two main intensive care scoring systems, MODS and NEMS, were collected from two ICUs in London, Ontario, Canada, and the results may not be consistent with those in other ICUs outside of London, Ontario. For future work, a larger data set including the cases in several different ICUs from the Critical Care Information System in Ontario will be obtained for further analysis and validation based on [46]. Covid-19 patients will also be included in the new data set for predictive modelling.

Table 6: Comparison on AUC values in Mortality prediction between previous works and our study.

Model	AUC value
Logistic regression in [25]	0.760
Logistic regression in [23]	0.767
Logistic regression in our study	0.795
Random forest (RF) in [23]	0.751
Random forest in our study	0.788
Neural network (NN) in [23]	0.638
Neural network (NN) in our study	0.789

Table 7: Comparison on AUC values in LOS prediction between work in [23] and our study.

	Logistic regression	RF	NN
LOS cutting at 7 days in [23]	0.701	0.677	0.606
LOS cutting at 21 days in [23]	0.635	0.622	0.526
LOS cutting at 5 days in our study	0.681	0.689	0.682

CRediT authorship contribution statement

Chengqian Xian: Conception and design of the study, Implementation of statistical analyses, Writing – original draft, Writing – review & editing. **Camila P.E. de Souza:** Conception and design of the study, Writing – review & editing. **Felipe F. Rodrigues** Conception and design of the study, Writing – review & editing.

Acknowledgments

This research work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). All authors approved the version of the manuscript to be published.

Appendix A Tables

Table A1: Selected predictors in logistic regression for LOS prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
MODS Score	0.04	1.04 [1.02, 1.06]	0.012
Cardiovascular			0.052
None (reference)	-	-	
Minimal	0.09	1.09 [0.98, 1.21]	
Mild	-0.19	0.83 [0.68, 1.00]	
Moderate	-0.22	0.81 [0.59, 1.09]	
Severe	-0.29	0.75 [0.45, 1.22]	
Respiratory			< 0.0001
None (reference)	-	-	
Minimal	0.12	1.12 [0.98, 1.28]	
Mild	0.29	1.34 [1.16, 1.53]	
Moderate	0.40	1.50 [1.30, 1.73]	
Severe	0.17	1.19 [0.97, 1.46]	
NEMS Score	0.04	1.04 [1.04, 1.05]	< 0.0001
Central Venous Line			< 0.0001
No (reference)	-	-	
Yes	0.46	1.59 [1.41, 1.79]	
Arterial Line			0.041
No (reference)	-	-	
Yes	0.23	1.25 [1.12, 1.41]	
Intracranial Pressure Monitor			0.001
No (reference)	-	-	
Yes	0.67	1.96 [1.43, 2.66]	
Dialysis			0.001
No (reference)	-	-	
Yes	-0.22	0.80 [0.64, 0.99]	
Interventions Outside this Unit			0.007
No (reference)	-	-	
Yes	-0.13	0.88 [0.78, 0.98]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table A1 continued: Selected predictors in logistic regression for LOS prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
Sex			0.026
Female (reference)	-	-	
Male	0.11	1.12 [1.02, 1.23]	
Admission Source			< 0.0001
Emergency (reference)	-	-	
Operating Room	-0.33	0.72 [0.62, 0.83]	
Outside Hospital/Other	0.47	1.61 [1.43, 1.81]	
Unit/Ward/Stepdown	0.34	1.40 [1.23, 1.60]	
Admission Diagnosis			< 0.0001
Cardiovascular (reference)	-	-	
Gastrointestinal	0.21	1.23 [1.01, 1.49]	
Neurological	0.23	1.25 [1.06, 1.49]	
Other	0.18	1.20 [1.03, 1.39]	
Respiratory	0.55	1.73 [1.51, 1.99]	
Trauma	0.86	2.37 [1.94, 2.90]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table A2: Selected predictors in multinomial regression for LMClass prediction.

Predictor	short & deceased		long stay	
	Odds Ratio	95% CI	Odds Ratio	95% CI
MODS Score	1.40	[1.34, 1.46]	1.15	[1.11, 1.19]
Renal				
None (reference)	-	-	-	-
Minimal	1.03	[0.87, 1.21]	0.99	[0.87, 1.13]
Mild	0.99	[0.78, 1.26]	1.02	[0.84, 1.25]
Moderate	0.83	[0.59, 1.18]	1.21	[0.92, 1.60]
Severe	0.30	[0.21, 0.45]	0.69	[0.52, 0.92]
Neurologic				
None (reference)	-	-	-	-
Minimal	1.06	[0.83, 1.36]	0.97	[0.81, 1.17]
Mild	0.54	[0.41, 0.72]	0.98	[0.81, 1.18]
Moderate	0.40	[0.29, 0.55]	0.79	[0.64, 0.99]
Severe	0.85	[0.68, 1.07]	0.95	[0.79, 1.14]
Respiratory				
None (reference)	-	-	-	-
Minimal	0.67	[0.56, 0.81]	0.96	[0.84, 1.11]
Mild	0.61	[0.49, 0.75]	1.11	[0.95, 1.30]
Moderate	0.60	[0.48, 0.75]	1.28	[1.08, 1.52]
Severe	0.60	[0.45, 0.82]	1.10	[0.85, 1.42]
NEMS Score	1.11	[1.10, 1.13]	1.07	[1.06, 1.08]
Basic Monitoring				
No (reference)	-	-	-	-
Yes	0.07	[0.01, 0.35]	0.70	[0.08, 5.94]
Central Venous Line				
No (reference)	-	-	-	-
Yes	1.04	[0.87, 1.23]	1.54	[1.36, 1.74]
Arterial Line				
No (reference)	-	-	-	-
Yes	0.94	[0.80, 1.11]	1.22	[1.08, 1.38]
Intracranial Pressure Monitor				
No (reference)	-	-	-	-
Yes	2.05	[1.29, 3.25]	2.34	[1.65, 3.33]
Dialysis				
No (reference)	-	-	-	-
Yes	0.58	[0.41, 0.82]	0.70	[0.54, 092]
Other Interventions Within Unit				
No (reference)	-	-	-	-
Yes	0.74	[0.64, 0.86]	0.94	[0.84, 1.05]
Interventions Outside Unit				
No (reference)	-	-	-	-
Yes	0.56	[0.47, 0.66]	0.76	[0.67, 0.86]

Table A2 continued: Selected predictors in multinomial regression for LMClass prediction.

Predictor	short & deceased		long stay	
	Odds Ratio	95% CI	Odds Ratio	95% CI
Age	1.03	[1.03, 1.04]	1.01	[1.00, 1.01]
Sex				
Female (reference)	-	-	-	-
Male	0.76	[0.67, 0.86]	1.04	[0.94, 1.14]
Admission Source				
Emergency (reference)	-	-	-	-
Operating Room	0.27	[0.21, 0.33]	0.53	[0.46, 0.62]
Outside Hospital/Other	1.03	[0.87, 1.22]	1.63	[1.43, 1.85]
Unit/Ward/Stepdown	1.50	[1.26, 1.80]	1.58	[1.37, 1.83]
Admission Diagnosis				
Cardiovascular (reference)	-	-	-	-
Gastrointestinal	0.58	[0.44, 0.77]	1.00	[0.81, 1.23]
Neurological	1.19	[0.95, 1.48]	1.21	[1.00, 1.46]
Other	0.63	[0.52, 0.77]	1.00	[0.85, 1.18]
Respiratory	0.66	[0.54, 0.79]	1.45	[1.24, 1.69]
Trauma	1.24	[0.91, 1.67]	2.28	[1.83, 2.85]

Table A3: MODS Components (Adapted From Marshal et al 1995[10]).

Organ System	Indicator of Dysfunction	Degree of Dysfunction				
		None (0)	Minimal (1)	Mild (2)	Moderate (3)	Severe (4)
Respiratory	PaO_2/F_IO_2	> 300	$226 - 300$	$151 - 225$	$76 - 150$	≤ 75
Renal	Creatinine ($mmol/L$)	≤ 100	$101 - 200$	$201 - 350$	$351 - 500$	> 500
Cardiovascular	Pressure-adjusted rate	≤ 10.0	$10.1 - 15.0$	$15.1 - 20.0$	$20.1 - 30.0$	> 30.0
Hematological	Platelets ($\times 10^3/mm^3$)	> 120	$80 - 120$	$50 - 80$	$20 - 50$	< 20
Hepatic	Bilirubin ($\mu mol/L$)	≤ 20	$21 - 60$	$61 - 120$	$121 - 240$	> 240
Neurological	Glasgow Coma Score	15	$13 - 14$	$10 - 12$	$7 - 9$	≤ 6

Preprint

Table A4: NEMS Components (Adapted From Miranda et al 1997 [12]).

Item	Points
1. *Basic monitoring*: hourly vital signs, regular record and calculation of fluid balance	9
2. *Intravenous medication*: bolus or continuously, not including vasoactive drugs	6
3. *Mechanical ventilatory support*: any form of mechanical/assisted ventilation, with or without PEEP (e. g., continuous positive airway pressure), with or without muscle relaxants	12
4. *Supplementary ventilatory care*: breathing spontaneously through endotracheal tube; supplementary oxygen any method, except if (3) applies	3
5. *Single vasoactive medication*: any vasoactive drug	7
6. *Multiple vasoactive medication*: more than one vasoactive drug, regardless of type and dose	12
7. *Dialysis techniques*: all	6
8. *Specific interventions in the ICU*: such as endotracheal intubation, introduction of pacemaker, cardioversion, endoscopy, emergency operation in the past 24 h, gastric lavage; routine interventions such as X-rays, echocardiography, electrocardiography, dressings, introduction of venous or arterial lines, are not included	5
9. *Specific interventions outside the ICU*: such as surgical intervention or diagnostic procedure; the intervention/procedure is related to the severity of illness of the patient and makes an extra demand upon manpower efforts in the ICU	6
Total	56

Appendix B Figures

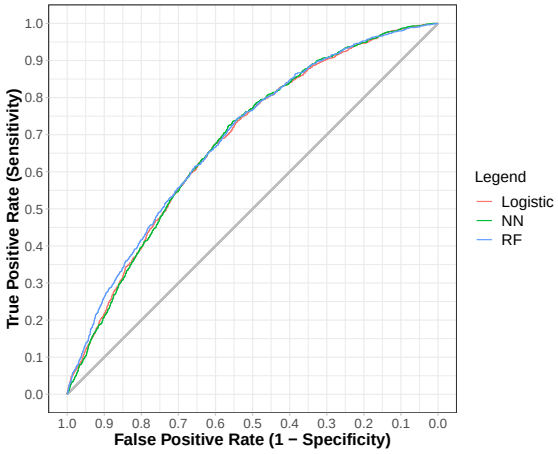

Fig. B1: ROC curves for LOS prediction in validation set.

Red: logistic regression; Green: random forests; Blue: neural networks

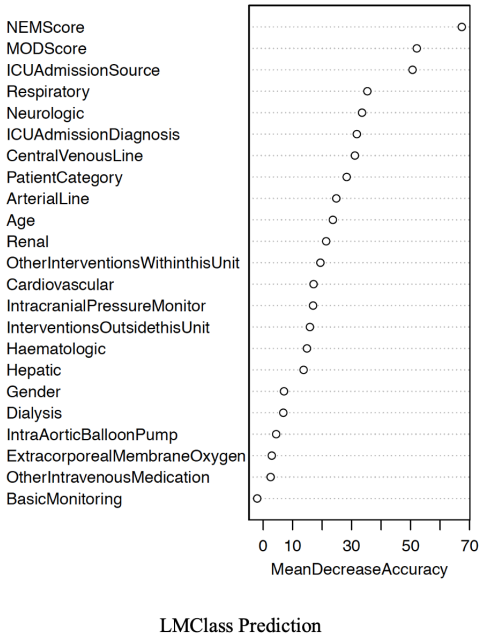

Fig. B2: Importance of predictors based on RF model in LMClass prediction.

References

[1] Salluh J, Chiche JD, Reis CE. New perspectives to improve critical care benchmarking. Ann Intensive Care. 2018;8(1):17.

[2] Silva Ramos F, Figueira Salluh JI. Data-driven management for intensive care units. ICU Management & Practice. 2019;19.

[3] Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: A compendium. Indian journal of critical care medicine. 2014;18(4):220–228.

[4] Salluh J, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. Current Opinion in Critical Care. 2014;20(5):557–565.

[5] Le Gall JR, A N, F H, et al. Mortality prediction using SAPS II: an update for French intensive care units. Crit Care. 2005;9:R645.

[6] Rubenfeld G, Angus D, Pinsky M, Curtis J, Connors AJ, GR B. Outcomes research in critical care: Results of the American Thoracic Society Critical Care Assembly Workshop on Outcomes Research. Am J Respir Crit Care Med. 1999;160:358–367.

[7] Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) $IV^{\star}$. Critical Care Medicine. 2006;34(10):2517–2529.

[8] Le GJ, Lemeshow S, Saulnier F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA. 1993;270(24):2957–2963.

[9] Lemeshow S, Teres D, Klar J, Avrunin J, Gehlbach S, J R. Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients. JAMA. 1993;270(20):2478–2486.

[10] Marshall JC, Cook DJ, Christou NV, R BG, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Critical Care Medicine. 1995;23(10):1638–1652.

[11] Vincent J, de Mendonça A, Cantraine F, Moreno R, Takala J, Suter P, et al. Use of the SOFA score to assess the incidence of organ dysfunction/-failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problem" of the European Society of Intensive Care Medicine. Critical care medicine. 1998;26(11):1793–800.

[12] Reis MD, Moreno R, Iapichino G. Nine equivalents of nursing manpower use score (NEMS). Intensive Care Medicine. 1997;23(7):760–765.

[13] Cullen D, Civetta J, Briggs B, Ferrara L. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. Crit Care Med. 1974;2(2):57–60.

[14] Rothen HU, Küng V, Ryser DH, Zürcher R, Regli B. Validation of "nine equivalents of nursing manpower use score" on an independent data sample. Intensive Care Medicine. 1999;25(6):606–611.

[15] Rush B, Celi LA, Stone DJ. Applying machine learning to continuously monitored physiological data. Journal of Clinical Monitoring and Computing. 2019;33(5):887–893. https://doi.org/10.1007/s10877-018-0219-z.

[16] Vellido A, Ribas V, Morales C, et al. Machine learning in critical care: state-of-the-art and a sepsis case study. Biomed Eng Online. 2018;17(1):135.

[17] Xia H, Daley BJ, Petrie A, Zhao X. A neural network model for mortality prediction in ICU. In: 2012 Computing in Cardiology; 2012. p. 261–264.

[18] Asteris PG, Gavriilaki E, Touloumenidou T, Koravou EE, Koutra M, Papayanni P, et al. Genetic prediction of ICU hospitalization and mortality in COVID-19 patients using artificial neural networks. Journal of Cellular and Molecular Medicine. 2022;26(5):1445–1455. https://doi.org/https://doi.org/10.1111/jcmm.17098.

[19] Kramer AA, Zimmerman JE. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. BMC Med Inform Decis Mak. 2010;10:27.

[20] John LM, Peter B, Patricia JS, Carol G, Graeme KH. Mortality and length-of-stay outcomes, 1993-2003, in the binational Australian and New Zealand intensive care adult patient database. Critical Care Medicine. 2008;p. 46—61.

[21] Houthooft R, Ruyssinck J, van der Herten J, Stijven S, Couckuyt I, Gadeyne B, et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artificial Intelligence in Medicine. 2015;p. 191–207.

[22] Tu JV G. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comput Biomed Res. 1992;26:666–672.

[23] Rodrigues FF. Three Essays on Intensive Care Unit Capacity Planning. Electronic Thesis and Dissertation Repository. 2018;5984.

[24] Zampieri FG, et al. Customization and external validation of the Simplified Mortality Score for the Intensive Care Unit (SMS-ICU) in Brazilian critically ill patients. Journal of Critical Care. 2020;59:94–100.

[25] Kao R, Priestap F, Donner A. To develop a regional ICU mortality prediction model during the first 24—h of ICU admission utilizing MODS and NEMS with six other independent variables from the Critical Care Information System (CCLS) Ontario, Canada. Journal of Intensive Care. 2016;4(16).

[26] Lemeshow S, Teres D, Avrunin JS, Gage RW. Refining intensive care unit outcome prediction by using changing probabilities of mortality. Critical Care Medicine. 1988;16(5):470–477.

[27] Niskanen M, Reinikainen M, Pettilä V. Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. Intensive Care Med. 2009;35:1060–1067.

[28] Norrie J. Mortality prediction in ICU: a methodological advance. Lancet Respir Med. 2015;3(1):5–6.

[29] Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42–52.

[30] Ahmed FS, Ali L, Joseph BA, FACS; Ikram A, Ul Mustafa R, Bukhari SAC. A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit. Journal of Trauma and Acute Care Surgery. 2020;89(4):736–742.

[31] Faraway J. Extending the Linear Model with R. New York: Chapman and Hall/CRC; 1981. Available from: https://doi.org/10.1201/9781315382722.

[32] Ho TK. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;20(8):832–844.

[33] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

[34] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[35] Riedmiller MA. Rprop - Description and Implementation Details; 1994. .

[36] Harris D, Harris S. Digital Design and Computer Architecture 2nd Edition. San Francisco, Calif.: Morgan Kaufmann; 2012.

[37] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.

[38] Abraham B, Ledolter J. Introduction to Regression Modeling. Belmont, CA: Thomson Brooks/Cole; 2006.

[39] Goss EP, Ramchandani H. Survival prediction in the intensive care unit: a comparison of neural networks and binary logit regression11Study supported by a grant from Bishop-Clarkson Hospital and Applied Information Management Institute. Socio-Economic Planning Sciences. 1998;32(3):189–198. https://doi.org/https://doi.org/10.1016/S0038-0121(97)00039-6.

[40] Trigg H. An investigation of methods to enhance the performance of artificial neural networks used to estimate medical outcomes. University of New Brunswick, Electrical Engineering Department; 1997.

[41] Frize M, Ennett CM, Stevenson M, Trigg HCE. Clinical decision support systems for intensive care units: using artificial neural networks. Medical Engineering Physics. 2001;23(3):217–225. https://doi.org/https://doi.org/10.1016/S1350-4533(01)00041-8.

[42] Schmidt-Hieber J. Nonparametric regression using deep neural networks with ReLU activation function. The Annals of Statistics. 2020;48(4):1875 – 1897. https://doi.org/10.1214/19-AOS1875.

[43] Wu H, Fan Y, Lv J. Statistical insights into deep neural network learning in subspace classification. Stat. 2020;9(1):e273. https://doi.org/https://doi.org/10.1002/sta4.273.

[44] Fukushima K, Miyake S. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In: Amari Si, Arbib MA, editors. Competition and Cooperation in Neural Nets. Berlin, Heidelberg: Springer Berlin Heidelberg; 1982. p. 267–285.

[45] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences. 1982;79(8):2554–2558. https://doi.org/10.1073/pnas.79.8.2554.

[46] Priestap F, Kao R, Martin CM. External validation of a prognostic model for intensive care unit mortality: a retrospective study using the Ontario Critical Care Information System. Can J Anesth/J Can Anesth. 2020;67:981–991.