Development and validation of prognostic machine learning models for short- and long-term mortality among acutely hospitalized patients.

Baker Jawad Jawad
Baker.jawad@regionh.dk
University of Copenhagen
Shakir Maytham Shaker
IT University of Copenhagen
Izzet Altintas
Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre
Jesper Eugen-Olsen
Amager and Hvidovre Hospital, The Capital Region of Denmark
Jan Nehlin
Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre
Ove Andersen
Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre
Thomas Kallemose
Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre

Article

Keywords:

Posted Date: March 30th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2722508/v1

Additional Declarations: Competing interest reported. J.E.O. is a cofounder, shareholder and Chief Scientific Officer of ViroGates A/S. J.E.O. and O.A. are named inventors on patents covering suPAR owned by Copenhagen University Hospital Amager and Hvidovre, Hvidovre, Denmark and licensed to ViroGates A/S. All remaining authors B.J, S.S, J.N, T.K declare no financial or non-financial competing interests

Version of Record: A version of this preprint was published at Scientific Reports on March 11th, 2024. See the published version at https://doi.org/10.1038/s41598-024-56638-6.

Abstract

Background: Several scores predicting mortality at the emergency department have been developed. However, all with shortcomings either simple and applicable in a clinical setting, with poor performance, or advanced, with high performance, but clinically difficult to implement. This study aimed to explore if machine learning algorithms could predict all-cause short- and long-term mortality based on the routine blood test collected at admission.

Methods: We analyzed data from a retrospective cohort study, including patients > 18 years admitted to the Emergency Department (ED) of Copenhagen University Hospital Hvidovre, Denmark between November 2013 and March 2017. The primary outcomes were 3-,10-,30-, and 365-day mortality after admission. PyCaret, an automated machine learning library, was used to evaluate the predictive performance of fifteen machine learning algorithms using the area under the receiver operating characteristic curve (AUC).

Results: Data from 48841 admissions were analyzed, of these 34190 (70%) were randomly divided into training data, and 14651 (30%) were in test data. Eight machine learning algorithms achieved very good to excellent results of AUC on test data in a of range 0.85-0.90. In prediction of short-term mortality, lactate dehydrogenase (LDH), leukocyte counts and differentials, Blood urea nitrogen (BUN) and mean corpuscular hemoglobin concentration (MCHC) were the best predictors, whereas prediction of long-term mortality was favored by age, LDH, soluble urokinase plasminogen activator receptor (suPAR), albumin, and blood urea nitrogen (BUN).

Conclusion: The findings suggest that measures of biomarkers taken from one blood sample during admission to the ED can identify patients at high risk of short-and long-term mortality following emergency admissions.

Introduction

Prognostic tools predicting all-cause mortality are crucial for decision making in Emergency Departments and Intensive Care Units (ICU). Tools predicting disease severity and mortality have been inquired for effective patient management and resource allocation to ensure appropriate treatment and evaluate medications, protocols, and interventions (1). Consequently, various scores and indices have been proposed to predict mortality, such as Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) (2), National Early Warning Score (NEWS)(3, 4), Modified Early Warning Score (MEWS) (5), Mortality Probability Models (6), Sequential Organ Failure Assessment (SOFA) (7), Emergency Severity Index (ESI) (8), and Cardiac Arrest Risk Triage score (CART) (9). Lately, Geriatric scores have also been proposed, such as the Barthel Index (10, 11), the Clinical Frailty Score (12), and FI-OutRef, a frailty index, calculated as the number of admission laboratory test results outside of the reference interval based upon blood collected at the admission time in + 65 years old acutely admitted patients (13). The majority of the existing score systems are based on a specifically defined patient cohort and target specific conditions. Furthermore, with an area under the curve (AUC) of 0.68-80 (14, 15), these scores have only moderate accuracy in predicting short-term mortality. The existing scores are typically based on physiological and laboratory parameters based on a simple linear relationship. However, considering the global aging phenomenon and the increase in the prevalence of multimorbidity and polypharmacy, the proportion of complex patients has increased. As a result, these scores are simple and cannot elucidate the complexity, and the clinical requirements for use in daily clinical practice are not met (16–18). In recent years, many studies have shown the significant potential of applying advanced machine learning (ML) algorithms in healthcare data (19–22). Several ML algorithms have been explored in healthcare to assist with diagnosis and prognosis, including the prediction of short and long-time mortality (23–30). For instance, 30-day and up to 4-year mortality risk models have been explored for medical and surgical patients discharged from the hospital with ROC-AUC of 0.95–0.96 (31, 32), and with a balanced-accuracy between 65–67% for 4 year-mortality (33). For in-hospital mortality prediction, Li et al. (2021) achieved an excellent AUC of 0.97 using ML algorithms based on 76 combined invasive and non-invasive parameters (34). For 30-day mortality risk after discharge, Blom et al. (31) achieved an excellent discrimination AUC of 0.95, using data from electronic health records, morbidity scores, information about the referred doctor, ambulance transport, previous emergency medical condition, information about radiological order, discharge time and days in the hospital, and triage priority. However, to our knowledge, most existing models use various parameters, such as demographics, patient history, morbidity, medication, and non-invasive and invasive parameters, to predict mortality, which is difficult for clinicians to interpret and implement in a flow culture setting such as the ED (35). Furthermore, very few studies have investigated ML modeling for all-cause short- and long-term mortality risk in a general population cohort at the ED. Hence, the aim of this study was to explore, develop and validate ML algorithms which can predict all-cause short – and long-term mortality based on few or easily measured routine blood samples collected at admittance at the emergency department.

Results

Table 1

Baseline characteristics of patients’ blood test results. Stratified by time of death within, 3, 10, 30, and 365 days after admission at the emergency department.

	Total	**3-day mortality**	**10-day mortality**	**30-day mortality**	**365-day mortality**
Mortality rate (%)	5632 (19.6%)	355 (1.2%)	1252 (4.4%)	2338 (8%)	4677 (16.3%)
Readmission rate, median (IQR)	0 (0:2)	0 (0:1)	1 (0:2)	1 (0:2)	1 (0:2)
Variables					
Age	65.6 (48.2: 78.5)	80.6 (71.5: 87)	80.8 (71.7: 88.2)	80.4 (71: 87.7)	79.1 (69.5: 86.5)
ALAT (U/L)	21 (15: 33)	31 (19: 82)	25 (16: 52)	22 (15: 41)	19 (13: 32)
Albumin (g/L)	34 (30: 37)	27 (22: 31)	27 (22: 31)	27 (22: 31)	29 (25: 33)
Basophils (x 10^9 /L)	0.03 (0.02: 0.05)	0.03 (0.02: 0.05)	0.03 (0.01: 0.05)	0.03 (0.01: 0.05)	0.03 (0.02: 0.05)
Alkaline Phosphatase (U/L)	76 (63: 94)	123 (90: 176)	114 (83: 152)	114 (82: 149)	98 (75: 129)
Bilirubin (µmol/L)	7 (5: 10)	9 (6: 17)	10 (6: 16)	9 (6: 14)	8 (5: 13)
BUN (mmol/L)	5.1 (3.8: 7.2)	11.5 (7.1: 19.5)	10.8 (6.7: 17.4)	9.6 (6.1: 15.8)	7.6 (5.1: 12.3)
Creatinine (µmol/L)	77 (62: 97)	123 (82: 190)	107 (72: 171)	98 (67: 153)	90 (66: 132)
CRP (mg/L)	7 (2: 39)	76 (20.5: 180)	73 (27: 160)	67 (22: 148.5)	38 (9: 95)
HB (mmol/L)	8.1 (7.2: 8.9)	7.3 (6.3: 8.5)	7.2 (6.2: 8.3)	7.1 (6.2: 8.1)	7.2 (6.3: 8.1)
INR	1 (1: 1.1)	1.2 (1: 1.4)	1.1 (1: 1.3)	1.1 (1: 1.3)	1.1 (1: 1.2)
Potassium (mmol/L)	3.9 (3.6: 4.2)	4.3 (3.8: 5.1)	4.1 (3.6: 4.6)	4 (3.6: 4.5)	4 (3.6: 4.3)
KF2710	0.91 (0.77: 1.02)	0.74 (0.55: 0.89)	0.76 (0.56: 0.91)	0.77 (0.58: 0.91)	0.83 (0.66: 0.96)
LDH (U/L)	186 (169: 214)	334 (267: 410)	289 (229: 355)	268 (219: 328)	224(188: 274)
Leukocytes (x 10^9 /L)	8.7 (6.9: 11.3)	13.7 (9.6: 19.6)	12.7 (9.1: 17.0)	11.8 (8.6: 16.0)	10.1 (7.6: 13.8)

	Total	3-day mortality	10-day mortality	30-day mortality	365-day mortality
Lymphocytes (x 10^9 /L)	1.7 (1.1: 2.3)	1.2 (0.7: 1.9)	1 (0.6: 1.6)	1.1 (0.7: 1.6)	1.2 (0.8: 1.8)
MCHC (mmol/L)	20.7 (20.1: 21.2)	19.8 (19.1: 20.5)	20 (19.3: 20.6)	20 (19.4: 20.7)	20.2 (19.5: 20.8)
MCV (fL)	89 (86: 93)	93 (88: 98)	92 (87: 97)	91 (87: 96)	91 (87: 95)
Monocytes (x 10^9 /L)	0.7 (0.5: 0.9)	0.8 (0.5: 1.3)	0.8 (0.5: 1.16)	0.8 (0.51: 1.1)	0.8 (0.53: 1.02)
Neutrocytes (x 10^9 /L)	5.8 (4.1: 8.3)	10.9 (7.4: 15.4)	10.3 (6.9: 14.2)	9.5 (6.4: 13.4)	7.6 (5.3: 11.2)
Promm (x 10^9 /L)	0.03 (0.02: 0.06)	0.11 (0.05: 0.29)	0.09 (0.04: 0.19)	0.08 (0.04: 0.16)	0.05 (0.03: 0.11)
suPAR (ng/ml)	3.3 (2.3: 5.0)	7.3 (5.2: 10.8)	7.0 (4.9: 10.2)	6.7 (4.7: 9.7)	5.7 (4.0: 8.2)
Thrombocytes (x 10^9 /L)	247 (201: 302)	266 (196: 350)	259 (193: 350)	266 (200: 354)	260 (199: 338)
Eosinophils (x 10^9 /L)	0.11 (0.04: 0.19)	0.01 (0: 0.05)	0.01 (0: 0.07)	0.028 (0: 0.10)	0.07 (0.01: 0.17)
eGFR (mL/min)	80 (60: 90)	42 (25: 69)	51 (29: 77)	56 (34: 83)	62 (40: 86)
Sodium (mmol/L)	139 (136: 141)	138 (134: 142)	138 (134: 142)	138 (134: 142)	138 (135: 141)
Sex (female)	25537 (52.3%)	184 (51.8%)	668 (52%)	1376 (51.5%)	4224 (49%)

Results are expressed as median (IQR interquartile range) for continuous variables. For categorical variables, results are expressed

as number of participants (percentage). ALAT: Alanine-aminotrasferase; BUN: Blood urea nitrogen; CRP: C-reactive protein; HB: Hemoglobin; INR: Prothrombin Time and International Normalized Ratio; KF2710: coagulation factors 2,7,10; LDH: Lactate dehydrogenase; MCV: mean corpuscular volume; MCHC: mean corpuscular hemoglobin concentration; Promm: Metamyelo-, Myelo. – Promyelocytes; suPAR: soluble urokinase plasminogen activator receptor; eGFR: estimated glomerular filtration rate.

Figure 1 shows the flow of data. Between 18 November 2013 and 17 March 2017, a total of 51007 ED admissions occurred during this period. Of these 2166 patient records were excluded due to missing data on more than 50% of variables, resulting in a study cohort of 48841 admissions obtained from 28671 unique patients. Randomly, 34193 (70%) patient records were allocated to training data and 14651 (30%) patient records were allocated to test data. Table 1 shows the baseline characteristics of patients, at admission, median age was 65.6 (IQR: 48.2–78.5) years and 52.3% were female. The median readmission rate was 1 (IQR:0–2) after 30-days and 365-days, and 0 (IQR: 0–2) during the entire follow-up. A total of 5632 (19.6%) patients did not survive during the follow-up (see Methods). The differences between not-survived patient admissions at different times follow-up, are shown in Table 1. The mortality rates were 1.2%, 4.4%, 8% and 16.3% at 3-day, 10-day, 30-day and 365-day follow-up, respectively.

Model performance

In Figs. 2a-d and 3a-d, the performance, as denoted by AUC and sensitivity of all fifteen ML models are shown (models described in methods section), respectively. The datasets used in the models included all 26 biomarkers from the routine blood tests and sex as an additional variable. Feature selection (see method section) ranked the most important biomarkers, removing seven to one variable in every iteration, resulting in models ranging from 27 to 1 variable. Based on training data, the AUC of all models ranged between 0.5–0.93 and the sensitivity ranged between 0.00-0.91 (Fig. 2). Eight of the fifteen models achieved very good to excellent results on training data with an AUC of 0.85–0.93 with a sensitivity > 0.80 using more ten variables (Fig. 2, and supplementary Table 2). Six of the ML algorithms, the Gradient Boosting Classifier (GBC), Light Gradient Boosting Machine (LightGBM), Linear Discriminant Analysis (LDA), Logistic regression (LR), Naïve Bayes (NB) and Quadratic Discriminant Analysis (QDA) had particularly high AUCs > 0.85 and sensitivity > 0.80, even when using only ten variables (Fig. 2). After reducing the number of variables to five, the performance in AUC showed very good performance and high sensitivity > 0.80 in two specific ML models the Gradient Boosting Classifier and the Quadratic Discriminant Analysis (Fig. 2a-c). The ML algorithm Gradient Boosting Classifier achieved an AUC of 0.89 for prediction of 3-day, 10-day, and 30-day mortality, with a sensitivity of 0.85, 0.83, and 0.83, respectively. For prediction of 365-day mortality, the ML algorithm Quadratic Discriminant Analysis had the highest AUC of 0.86, with a sensitivity of 0.80 (Fig. 2d). Using fewer than five variables resulted in a significant decrease in all models to below 0.85 in AUC and below 0,80 in sensitivity (Fig. 2a-d). Further performance metrics for all models can be found in supplementary (Table 2).

Table 2
Results from test data for top 3 models predicting short- and long- term mortality.

	AUC	Sensitivity	Specificity	PPV	NPV	Number of variables
3-day Mortality						
Naive Bayes	0.91 [0.91–0.91]	0.92 [0.91–0.92]	0.78 [0.78–0.79]	0.03 [0.3–0.3]	0.99 [0.99–0.99]	15
Linear Discriminant Analysis	0.93 [0.93–0.93]	0.89 [0.86–0.89]	0.83 [0.82–0.83]	0.04 [0.4–0.5]	0.99 [0.99–0.99]	15
Logistic Regression	0.93 [0.93–0.93]	0.85 [0.83–0.86]	0.85 [0.84–0.89]	0.04 [0.4–0.4]	0.99 [0.99–0.99]	15
10-day mortality						
Linear Discriminant Analysis	0.91 [0.90–0.91]	0.90 [0.87–0.93]	0.78 [0.78–0.79]	0.1 [0.09–0.11]	0.99 [0.99–0.99]	10
Logistic Regression	0.91 [0.89–0.93]	0.90 [0.87–0.93]	0.79 [0.79–0.79]	0.1 [0.09–0.11]	0.99 [0.99–0.99]	10
Quadratic Discriminant Analysis	0.90 [0.90 – 0.90]	0.91 [0.87–0.93]	0.77 [0.76–0.77]	0.1 [0.08–0.10]	0.99 [0.99–0.99]	10
30-day Mortality						
Linear Discriminant Analysis	0.90 [090 – 0.90]	0.90 [0.87–0.92]	0.78 [0.77–0.79]	0.19 [0.18–0.21]	0.99 [0.99–0.99]	10
Quadratic Discriminant Analysis	0.91 [0.89–0.91]	0.89 [0.86–0.91]	0.76 [0.75–077]	0.18 [0.17–0.19]	0.99 [0.99–0.99]	10
Gradient Boosting Classifier	0.92 [0.92–0.92]	0.86 [0.84–0.89]	0.82 [0.82–0.83]	0.22 [0.21–0.24]	0.99 [0.99–0.99]	10
365-day mortality						
Gradient Boosting Classifier	0.88 [0.88–0.89]	0.82 [0.81–0.83]	0.77 [0.76–0.77]	0.44 [0.43–0.46]	0.96 [0.95–0.99]	10
Light Gradient Boosting Machine	0.89 [0.89–0.89]	0.80 [0.80–0.81]	0.81 [0.80–0.82]	0.46 [0.44–0.49]	0.95 [0.95–0.98]	15

	AUC	Sensitivity	Specificity	PPV	NPV	Number of variables
3-day Mortality						
Quadratic Discriminant Analysis	0.87 [0.87–0.89]	0.85 [0.84–0.89]	0.74 [0.73–0.75]	0.40 [0.40–0.41]	0.96 [0.95–0.99]	15

AUC: mean area under receiver operating curve based on 10-fold cross-validation. The numbers are presented as mean with 95%-confidence. PPV: Positive predictive value, NPV: Negative predictive value.

Table 2 shows the performance metrics for the top three ML models for prediction of 3-, 10-, 30- and 365-day mortality on test data based on the highest AUC and sensitivity performance. The best models were models with ten to fifteen variables. Performance metrics between training and test data were similar. The ML algorithms Naive Bayes, Linear Discriminant Analysis, and Logistic Regression had the highest mean AUC of 0.91–0.93 and sensitivity of 0.85–0.92, for 3-day mortality using 15 variables in the models (Table 2). For 10-day mortality, the ML algorithms Linear Discriminant Analysis and Quadratic Discriminant Analysis had the highest mean AUC of 0.90–0.91 and sensitivity of 0.90–91 using 10 variables. For 30-day mortality, the Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Gradient Boosting Classifier had the highest mean AUC of 0.90–0.92 and sensitivity of 0.86–0.90 using 10 variables. Lastly, for 365-day mortality, the ML algorithms Gradient Boosting Classifier, the Light Gradient Boosting Machine, and Quadratic Discriminant Analysis had the highest mean AUC of 0.87–0.89 and a sensitivity of 0.80–0.85 using 10 to 15 variables (Table 2).

Biomarker importance

Based on feature selection technique used on the IDA, LR, GBC, ADA and LightGBM models (ML algorithms in Methods), the biomarkers with the most importance for prediction of mortality were identified. Figure 4. shows the top-ranked biomarkers for 3-,10-,30-, and 365-day mortality. Biomarkers like age, LDH, albumin, BUN, MCHC, are repeatedly ranked among the top variables in all models. Even when excluding age as a biomarker, the remaining variables where still top predictors and the predicted mortality for 3-,10-,30-, and 365-day remained showing very good performance AUC of > 0.80. Biomarkers like basophiles, INR, bilirubin, and monocytes are ranked in repeatedly among the lowest five in all models. Eosinophils, leukocytes, and neutrophils are among the biomarkers that move from top to bottom of the rank as follow-up time increases. In contrast, suPAR initially was ranked low at 3-day mortality outcome but rises to the top, at 365-day mortality, with an increase in follow-up time (Fig. 4).

Discussion

The aim of this study was to develop and validate machine learning algorithms for finding high-mortality patients admitted to Emergency Departments using the results from routine blood testing and age. With as few as five biomarkers, machine learning-based algorithms provided very good performance predicting mortality in acutely admitted patients with AUC of 0.89 and 0.86, sensitivity of 0.83 and 0.80 for short and long-term mortality, respectively. Top three models, used between ten and fifteen biomarkers achieved an AUC of 90–93 and 87–89, sensitivity of 0.86-92 and 0.80–85 for short and long-term mortality, respectively. However, most models did not see an improvement from adding additional biomarkers. The models in this study were trained on original data that required minimal modification. In this regard, data for these algorithms are easy to obtain in clinical practice as only a blood sample and age is needed, with no need for multiple measurements of vital signs, medication and disease history. A similar study by Xie et al. (14) developed and validated scores to predict the risk of death for ED patients using five to six biomarkers. These biomarkers included age, heart rate, respiration rate, diastolic blood pressure, systolic blood pressure, and cancer history. The 30-day score by Xie et al. achieved the best performance for mortality prediction, with an AUC of 0.82 (95%CI, 0.81–0.83). However, similar to the very good discriminative performance of the scores, we further have demonstrated an excellent performance of > 0.90 by only using one routine blood sample. In this study, we argue that clinically, it is easy to interpret and understand algorithms that can predict mortality based on the use of biomarkers, such as LDH, albumin, BUN, leukocyte and differential counts, and suPAR. An increase or decrease indicates underlying clinically pathological conditions which clinicians can comprehend, such as sever tissue damage, kidney disease and infection, and the levels of such biomarkers are stable overtime with only minor fluctuations. In contrast, abnormal values of vital signs as heart rate, respiration rate and blood pressure are either indicators of acute failure of the body's most essential physiological functions, or an indication of compensatory physiological mechanisms in the heart or lungs and can fluctuate suddenly and significantly over minutes. Furthermore, clinically abnormal vital sign values need multiple recordings and re-evaluations ranging from four times per hour to two times per day to determine patients at risk of any deterioration.

Biomarkers

Biomarker selection was essential since the practical use of algorithms with many clinical biomarkers are not feasible. All models ranked age, albumin, LDH, and BUN as key predictive factors. However, the ranks were different for short- and long-term mortality. In our study, the best predictors of short-term mortality are LDH, leukocyte counts and differential, BUN and MCHC while the best predictors of long-term mortality are age, LDH, suPAR, albumin and BUN. The biomarkers identified have previously been shown and used as prognostic and monitoring tools for diseases such as anemia, heart attack, bone fractures, muscle trauma, cancers, infections, inflammatory disorders, and hepatic-, renal-, and congestive heart failure (36–42). These diseases are often found among frailty patients admitted to ED. Our results show that combining these biomarkers in one algorithm makes them valuable predictors for mortality.

ML algorithms: new resources to find high-risk patients.

The findings of this study have provided a basis for developing ML models based on a few biomarkers. These models can be used in the future to identify patients at risk following emergency admissions. Considering an aging population and crowded emergency departments worldwide, we see a broad opportunity to use such tools to determine patients' health status more accurately and allocate appropriate resources to high-risk patients. In several clinical settings, such mortality algorithms can be used to increase patient safety and reduce preventable mistakes and hospital mortality, for instance, when triaging patients. Similarly, we suggest that these algorithms may also be helpful as a decision-making tool in challenging decisions in order to prevent overtreatment, and provision of care that does not correspond to the patient's wishes and recovery capacity.

Limitations and future research

To our knowledge, this is the first published study that has applied machine learning methods to predict acutely admitted emergency patients based on a few routine blood samples with excellent performance. There are, however, some limitations to this study. First, this was a retrospective study conducted at a single clinical center, introducing issues of generalizability. Second, 4.3% of the total amount of patients with more than 50% missing data were excluded from the study, which could result in selection bias for the performance estimates. thirdly, in this study we have used a probability threshold of 0.5, a more comprehensive analysis of the consequences of different thresholds is required to determine the right threshold. Last, but not least, machine learning techniques have also been criticized as black boxes by critics, so clinicians are skeptical of their use. This issue may be reduced by using interpretable biomarkers and using explaining ML tools or educating clinicians in ML concepts. Future work would need to focus on determining which algorithm should in the end be used, additional external validation would be needed to verify the robustness of this algorithm. Implementation and prospective randomized trials would also be necessary to ensure the use and effectiveness of the algorithm.

Conclusion

This study has demonstrated that high-risk of death in patients following admission can be identified by a routine blood sample, using a combination of five to fifteen biomarker measures. Eight of the fifteen evaluated ML algorithms achieved very good to excellent results of AUC (0.85–0.93). The ML algorithms Gradient Boosting Classifier, Light Gradient Boosting Machine, Linear Discriminant Analysis, Logistic regression, Naïve Bayes and Quadratic Discriminant Analysis showed the best performance on AUCs and sensitivity, even using only five biomarkers.

Methods

Study Design and Settings

In this study, we analyzed data from a retrospective cohort study from the Emergency Department at the Copenhagen University Hospital, Amager and Hvidovre. The cohort included all patients admitted to the Acute Medical Unit of the Emergency Department with at least one available blood sample and suPAR measurement during the follow-up between 18 November 2013 and 17 March 2017, whose follow-up data are available in the Danish National Patient Registry (DNPR). The Acute Medical Unit receives patients within all specialties, except children, gastroenterological patients, and obstetric patients. The follow-up period began from admission and extending to 90 days after discharge for the last patient was included, corresponding to a median follow-up time of 2 years: a range of 90-1.301 days. During the study period, patients who left the country for an extended length of time were censored at the time they were last admitted.

This study was reported in accordance with the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement(43).

Biomarkers

On admission, blood samples were taken, and a standard panel of markers was measured at the Department of Clinical Biochemistry, including C-reactive protein (CRP), Soluble urokinase plasminogen activator receptor (suPAR), Alanine Aminotransferase (ALAT), Albumin (ALB), International Normalized Ratio (INR), coagulation factors 2,7,10 (KF2710), total Bilirubin (BILI), Alkaline Phosphatase, Creatinine, Lactate dehydrogenase (LDH), Blood urea nitrogen (BUN), Potassium (K), Sodium (NA), Estimated Glomerular Filtration Rate (eGFR), Hemoglobin (HB), mean corpuscular volume (MCV) and mean corpuscular hemoglobin concentration (MCHC), number of leukocytes, lymphocytes, neutrocytes, monocytes, thrombocytes, eosinophils, basophils, and Metamyelo-, Myelo. - Promyelocytes (PROMM) (44). Age and sex were also included as variables in the algorithms (Table 1).

From The Danish Civil Registration System demographic information, including age, sex readmissions, and death time was collected. All methods were carried out in accordance with relevant guidelines and regulations. The study was approved by the Danish Data Protection Agency (ref. HVH-2014-018, 02767), the Danish Health and Medicines Authority (ref. 3-3013-1061/1) and The Capital Region of Denmark, Team for Journaldata (ref. R-22041261).

Outcomes

In this study, the primary outcomes were 3-,10-,30-, and 365-day mortality, defined as deaths within 3, 10, 30, and 365 days after admission at the emergency department, resulting in binary outcomes (0 = survive, 1 = dead).

Statistical analysis:

R version (4.1.0) and Python (version 3.8.0) was used for statistical analysis in the demographic statistics part of this study. Categorical variables were described as numbers and percentages (%) and continuous variables were described as medians with interquartile range (IQR) for the groups.

Data preparation:

First the data format was unified. Secondly, admissions with more than 50% missing data were dropped. For missing values, iterative imputations were used from scikit-learn package (45). For the unequal distribution of our target outcome (imbalance data), several resampling methods were explored, including the random undersampling, the random oversampling, and SMOTE (46, 47).

In this study, we used the random oversampling from imbalanced-learn package (48) to handle the imbalanced classification distribution best. Outliers were identified and removed through principal component analysis linear dimensionality reduction using the Singular Value Decomposition technique. The default setting is 0.05, resulting as 0.025 of the values on each side of the distribution's tail were dropped from the training set. To reduce the impact of magnitude in the variance, we normalized the values of all variables in the data by z-score. To make all variables more normal-distributed like, we power transformed the data by the Yeo-Johnson method (49).

Model Construction.

In this study we used the PyCaret's classification module to train fifteen different algorithms, resulting in a total of 480 models for the four outcomes with a set of 27, 20, 15, 10, 5, 3, 2, 1 biomarker(s). PyCaret (version 2.2.6) (50), is an automated machine learning low-code library in Python that automates the ML workflow. For all models Python (version 3.8.0) were used. By default, the random selection method was used to split the data into training and test sets of 70% and 30%, respectively. For hyperparameter tuning, a random grid search was used in PyCaret. There was no significant difference between training and test sets after split considering variable values.

Algorithm selection and performance measures

The fifteen machine learning algorithms (Random Forest (RF), SVM-Radial Kernel (RBFSVM), Extra Trees Classifier (ET), Extreme Gradient Boosting (XGBOOST), Decision Tree Classifier (DT), neural network (MLP), Light Gradient Boosting Machine(LIGHTBM), K Neighbors Classifier (KNN), Gradient Boosting Classifier (GBC), CatBoost Classifier (CATBOOST), Ada Boost Classifier (ADA), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Naive Bayes(NB)), were trained and evaluated first on 10-fold cross-validation, then on test data. Model selection was based on the Area under the receiver operating characteristic curve (AUC) measure. Additionally, sensitivity, specificity, positive predictive value, and negative predictive value for the complete data, based on probability threshold of 0.5, were estimated for the training and test data and evaluated between them.

Biomarker selection

In this study we aimed to use few biomarkers for predicting mortality. This can reduce the risk of over-fitting, improve accuracy, and reduce the training time (51). Biomarker selection (Feature Selection) was achieved in PyCaret using various permutation importance techniques depending on the type of model being evaluated. These included Random Forest, Adaboost, and linear correlation with the mortality outcome to select the subset of the most relevant biomarkers for modeling. By default, the threshold used for feature selection was 0.8 (52). During iteration, all biomarkers were fed into each of the models, the best biomarkers were kept, and seven to one biomarker were removed, resulting in models starting with 27 variables and decreasing to 1.

Declarations

AUTHORS’ CONTRIBUTIONS

B.J, and O.A were responsible for the research idea. T.K, O.A and J.E.O were responsible for the study design. B.J, S.S and T.K. were responsible for the statistical analysis and algorithm training and evaluation. B.K, and T.K were responsible for the interpretation of results. All authors contributed important intellectual content during manuscript drafting or revision, accept personal accountability for their own contributions and agree to ensure that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved.

Acknowledgements

The authors thank the Department of Clinical Biochemistry at Amager and Hvidovre Hospital for all analyses used in this study. This study received no external funding.

Competing interests

J.E.O. is a cofounder, shareholder and Chief Scientific Officer of ViroGates A/S. J.E.O. and O.A. are named inventors on patents covering suPAR owned by Copenhagen University Hospital Amager and Hvidovre, Hvidovre, Denmark and licensed to ViroGates A/S. All remaining authors B.J, S.S, J.N, T.K declare no financial or non-financial competing interests.

Data Availability

The datasets analyzed during the current study are not publicly available due to privacy (data use agreements) or ethical restrictions but are available from the corresponding author on reasonable request.

Code availability

The underlying code for this study [and training/validation datasets] is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

References

1. Silva I, Moody G, Scott DJ, Celi LA, Mark RG. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in cardiology challenge 2012. In: Computing in Cardiology. 2012.
2. Knaus WA. APACHE 1978–2001: The development of a quality assurance system based on prognosis: Milestones and personal reflections. Vol. 137, Archives of Surgery. 2002.
3. Silcock DJ, Corfield AR, Gowens PA, Rooney KD. Validation of the National Early Warning Score in the prehospital setting. Resuscitation. 2015;89(C).

4. Mahmoodpoor A, Sanaie S, Saghaleini S, Ostadi Z, Hosseini M-S, Sheshgelani N, et al. Prognostic value of National Early Warning Score and Modified Early Warning Score on intensive care unit readmission and mortality: A prospective observational study. Front Med. 2022 Aug 4;9.
5. Burch VC, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and inhospital mortality. Emerg Med J. 2008;25(10).
6. Lemeshow S, Gehlbach SH, Klar J, Avrunin JS, Teres D, Rapoport J. Mortality Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients. JAMA J Am Med Assoc. 1993;270(20).
7. Toma T, Abu-Hanna A, Bosman RJ. Discovery and inclusion of SOFA score episodes in mortality prediction. J Biomed Inform. 2007;40(6).
8. Phungoen P, Khemtong S, Apiratwarakul K, Ienghong K, Kotruchin P. Emergency Severity Index as a predictor of in-hospital mortality in suspected sepsis patients in the emergency department. Am J Emerg Med. 2020;38(9).
9. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs. Crit Care Med. 2012;40(7).
10. Walsh M, O'Flynn B, O'Mathuna C, Hickey A, Kellett J. Correlating Average Cumulative Movement and Barthel Index in Acute Elderly Care. In: Communications in Computer and Information Science. 2013.
11. Higuchi S, Kabeya Y, Matsushita K, Taguchi H, Ishiguro H, Kohshoh H, et al. Barthel Index as a Predictor of 1-Year Mortality in Very Elderly Patients Who Underwent Percutaneous Coronary Intervention for Acute Coronary Syndrome: Better Activities of Daily Living, Longer Life. Clin Cardiol. 2016;39(2).
12. Torsney KM, Romero-Ortuno R. The clinical frailty scale predicts inpatient mortality in older hospitalised patients with idiopathic parkinson's disease. J R Coll Physicians Edinb. 2018;48(2).
13. Klausen HH, Petersen J, Bandholm T, Juul-Larsen HG, Tavenier J, Eugen-Olsen J, et al. Association between routine laboratory tests and long-term mortality among acutely admitted older medical patients: a cohort study. BMC Geriatr [Internet]. 2017;17(1):62. Available from: https://doi.org/10.1186/s12877-017-0434-3
14. Xie F, Ong MEH, Liew JNMH, Tan KBK, Ho AFW, Nadarajan GD, et al. Development and Assessment of an Interpretable Machine Learning Triage Tool for Estimating Mortality after Emergency Admissions. JAMA Netw Open. 2021;4(8).
15. Suwanpasu S, Sattayasomboon Y. Accuracy of Modified Early Warning Scores for Predicting Mortality in Hospital: A Systematic Review and Meta-analysis. J Intensive Crit Care. 2016;02(02).
16. Strand K, Flaatten H. Severity scoring in the ICU: A review. Vol. 52, Acta Anaesthesiologica Scandinavica. 2008.
17. Moreno R, Matos R. New issues in severity scoring: Interfacing the ICU and evaluating it. Vol. 7, Current Opinion in Critical Care. 2001.
18. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. Crit Care Med.

2013;41(4).

19. Nguyen NH, Picetti D, Dulai PS, Jairath V, Sandborn WJ, Ohno-Machado L, et al. Machine Learning-based Prediction Models for Diagnosis and Prognosis in Inflammatory Bowel Diseases: A Systematic Review. J Crohn's Colitis. 2022;16(3).
20. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson J V., Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Vol. 13, Genome Medicine. 2021.
21. Nordin N, Zainol Z, Mohd Noor MH, Chan LF. Suicidal behaviour prediction models using machine learning techniques: A systematic review. Artif Intell Med. 2022 Oct 1;132:102395.
22. Singh DP, Kaushik B. A systematic literature review for the prediction of anticancer drug response using various machine-learning and deep-learning techniques. Chem Biol Drug Des [Internet]. 2023 Jan 1;101(1):175–94. Available from: https://doi.org/10.1111/cbdd.14164
23. Wang G, Liu X, Shen J, Wang C, Li Z, Ye L, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. Nat Biomed Eng [Internet]. 2021;5(6):509–21. Available from: https://doi.org/10.1038/s41551-021-00704-1
24. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA - J Am Med Assoc. 2016;316(22).
25. Shouval R, Labopin M, Bondi O, Mishan-Shamay H, Shimoni A, Ciceri F, et al. Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: A European group for blood and marrow transplantation acute leukemia working party retrospective data mining study. J Clin Oncol. 2015;33(28).
26. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. JAMA Netw Open [Internet]. 2019 Oct 25;2(10):e1915997–e1915997. Available from: https://doi.org/10.1001/jamanetworkopen.2019.15997
27. Naemi A, Schmidt T, Mansourvar M, Naghavi-Behzad M, Ebrahimi A, Wiil UK. Machine learning techniques for mortality prediction in emergency departments: a systematic review. BMJ Open [Internet]. 2021 Nov 1;11(11):e052663. Available from: http://bmjopen.bmj.com/content/11/11/e052663.abstract
28. Caires Silveira E, Mattos Pretti S, Santos BA, Santos Corrêa CF, Madureira Silva L, Freire de Melo F. Prediction of hospital mortality in intensive care unit patients from clinical and laboratory data: A machine learning approach. World J Crit Care Med. 2022;11(5):317–29.
29. Iwase S, Nakada T, Shimada T, Oami T, Shimazui T, Takahashi N, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. Sci Rep [Internet]. 2022;12(1):12912. Available from: https://doi.org/10.1038/s41598-022-17091-5
30. Ning Y, Li S, Ong MEH, Xie F, Chakraborty B, Ting DSW, et al. A novel interpretable machine learning system to generate clinical risk scores: An application for predicting early mortality or unplanned

readmission in a retrospective cohort study. PLOS Digit Heal [Internet]. 2022;1(6):e0000062. Available from: http://dx.doi.org/10.1371/journal.pdig.0000062

31. Blom MC, Ashfaq A, Sant'Anna A, Anderson PD, Lingman M. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: A retrospective, population-based registry study. BMJ Open. 2019;9(8).
32. Gao J, Merchant AM. A Machine Learning Approach in Predicting Mortality Following Emergency General Surgery. Am Surg. 2021;87(9).
33. Krasowski A, Krois J, Kuhlmey A, Meyer-Lueckel H, Schwendicke F. Predicting mortality in the very old: a machine learning analysis on claims data. Sci Rep [Internet]. 2022;12(1):1–9. Available from: https://doi.org/10.1038/s41598-022-21373-3
34. Li C, Zhang Z, Ren Y, Nie H, Lei Y, Qiu H, et al. Machine learning based early mortality prediction in the emergency department. Int J Med Inform [Internet]. 2021;155(June):104570. Available from: https://doi.org/10.1016/j.ijmedinf.2021.104570
35. Kirk JW, Nilsen P. Implementing evidence-based practices in an emergency department: Contradictions exposed when prioritising a flow culture. J Clin Nurs. 2016;25(3–4):555–65.
36. Amulic B, Cazalet C, Hayes GL, Metzler KD, Zychlinsky A. Neutrophil Function: From Mechanisms to Disease. Annu Rev Immunol [Internet]. 2012 Mar 26;30(1):459–89. Available from: https://doi.org/10.1146/annurev-immunol-020711-074942
37. Meier S, Henkens M, Heymans S, Robinson EL. Unlocking the Value of White Blood Cells for Heart Failure Diagnosis. J Cardiovasc Transl Res [Internet]. 2021;14(1):53–62. Available from: https://doi.org/10.1007/s12265-020-10007-6
38. Swirski FK, Nahrendorf M. Leukocyte Behavior in Atherosclerosis, Myocardial Infarction, and Heart Failure. Science (80-) [Internet]. 2013 Jan 11;339(6116):161–6. Available from: https://doi.org/10.1126/science.1230719
39. Rasmussen LJH, Petersen JEV, Eugen-Olsen J. Soluble Urokinase Plasminogen Activator Receptor (suPAR) as a Biomarker of Systemic Chronic Inflammation. Front Immunol. 2021;12(December):1–22.
40. Huang YL, Hu Z De. Lower mean corpuscular hemoglobin concentration is associated with poorer outcomes in intensive care unit admitted patients with acute myocardial infarction. Ann Transl Med. 2016;4(10):1–8.
41. LaRosa DF, Orange JS. 1. Lymphocytes. J Allergy Clin Immunol. 2008;121(2 SUPPL. 2).
42. Eugen-Olsen J, Giamarellos-Bourboulis EJ. SuPAR: The unspecific marker for disease presence, severity and prognosis. Vol. 46, International Journal of Antimicrobial Agents. 2015.
43. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. Eur Urol. 2015;67(6).
44. Nehlin JO, Andersen O. Molecular Biomarkers of Health BT - Explaining Health Across the Sciences. In: Sholl J, Rattan SIS, editors. Cham: Springer International Publishing; 2020. p. 243–70. Available from: https://doi.org/10.1007/978-3-030-52663-4_15

45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12.

46. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16.

47. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020;7(1).

48. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18.

49. Yeo I, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika [Internet]. 2000 Dec 1;87(4):954–9. Available from: https://doi.org/10.1093/biomet/87.4.954

50. Moez A. PyCaret: An open source, low-code machine learning library in Python [Internet]. 2020 [cited 2023 Mar 8]. Available from: https://www.pycaret.org

51. Afshar M, Usefi H. Optimizing feature selection methods by removing irrelevant features using sparse least squares. Expert Syst Appl. 2022 Aug 1;200:116928.

52. Moez A. Feature Selection - PyCaret Official [Internet]. 2020 [cited 2023 Mar 8]. Available from: https://pycaret.gitbook.io/docs/get-started/preprocessing/feature-selection

Figures

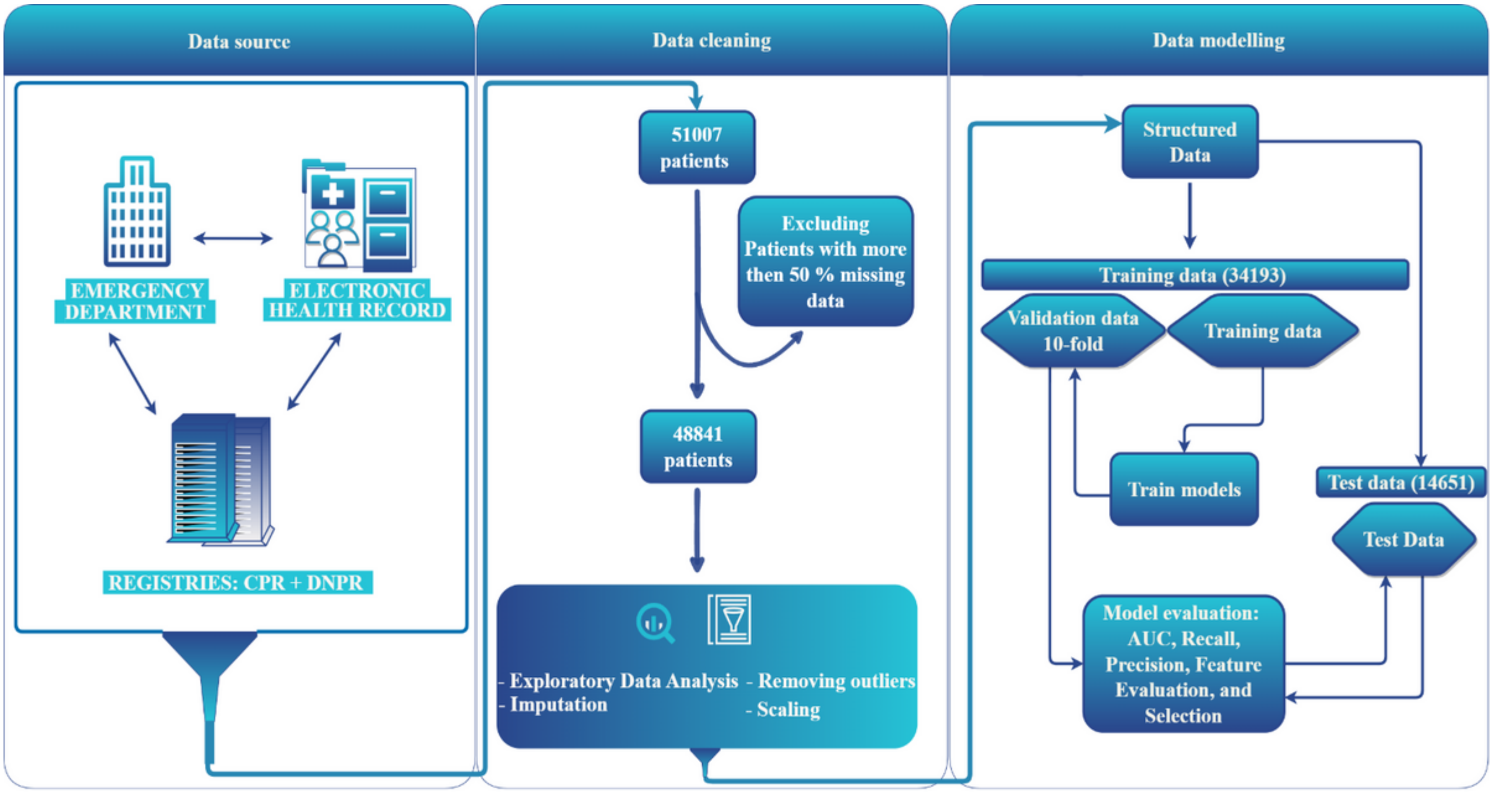

Figure 1

Flowchart of data. We used a cohort study of 51007 acute patient admissions at the emergency department with laboratory and demographical data. At data pre-processing we excluded 2166 records with more the 50% missing in data (4.3% of the total), removed outliers, imputed and scaled the data. In total 48841 records were structured data, where 34190 (70%) patient records were allocated to training data and 14651 (30%) patient records were allocated to validation and test data.

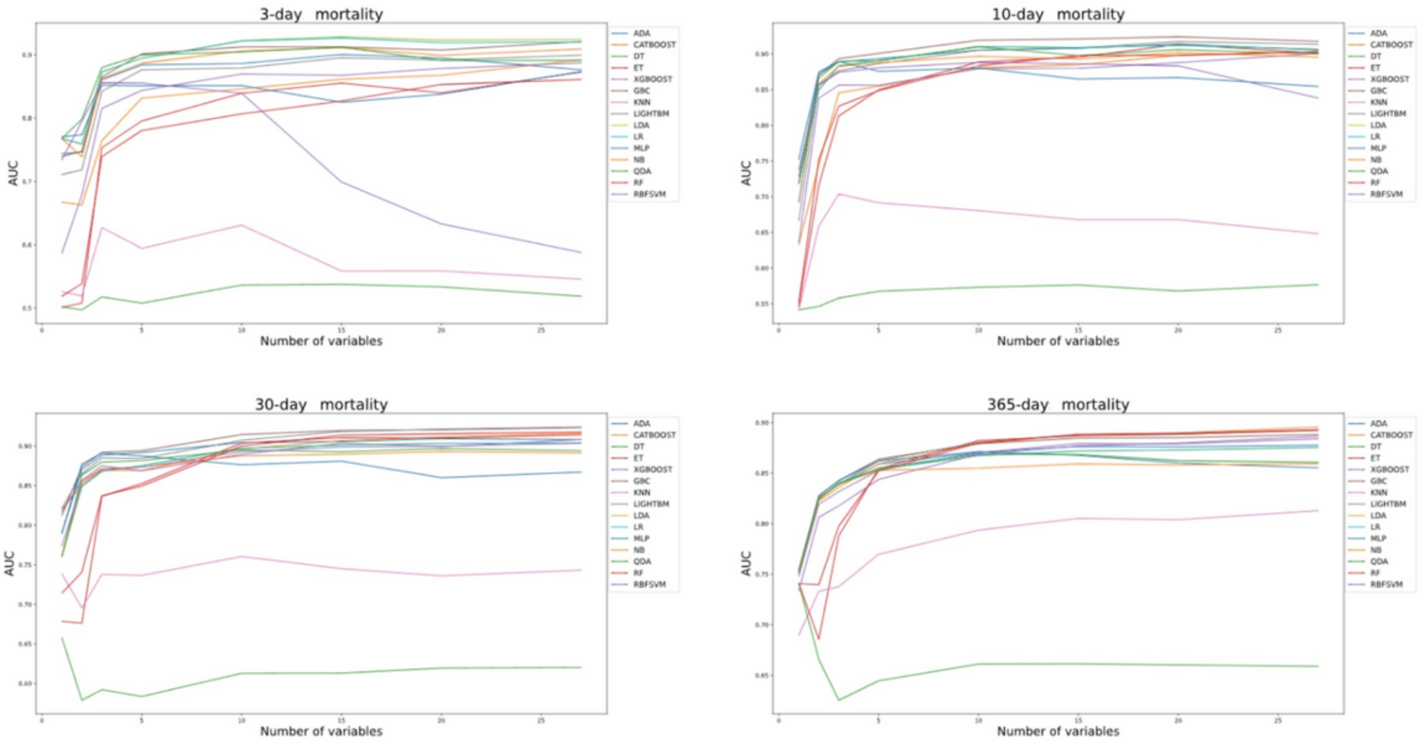

Figure 2

Predictive performance of fifteen ML algorithms on training data, as measured by the Area Under the Receiver Operating Characteristic Curve (AUC) when using 1 to 27 variables as predictors in the machine learning algorithms. Figures 2a-2d demonstrate the predictive performance for 3-day, 10-day, 30-day and 365-days mortality, respectively. Among all models, the highest predictive performances in AUC are shown between 0.90-0.93 in figures 2a-2d. When using five variables, the top 3 models achieved an AUC of 0.89 in figures 2a-2c, and an AUC of 0.86 or above when using five variables in figure 2d. The AUC falls below 0.85 when using fewer than three variables for all models in figures 2a-2d.

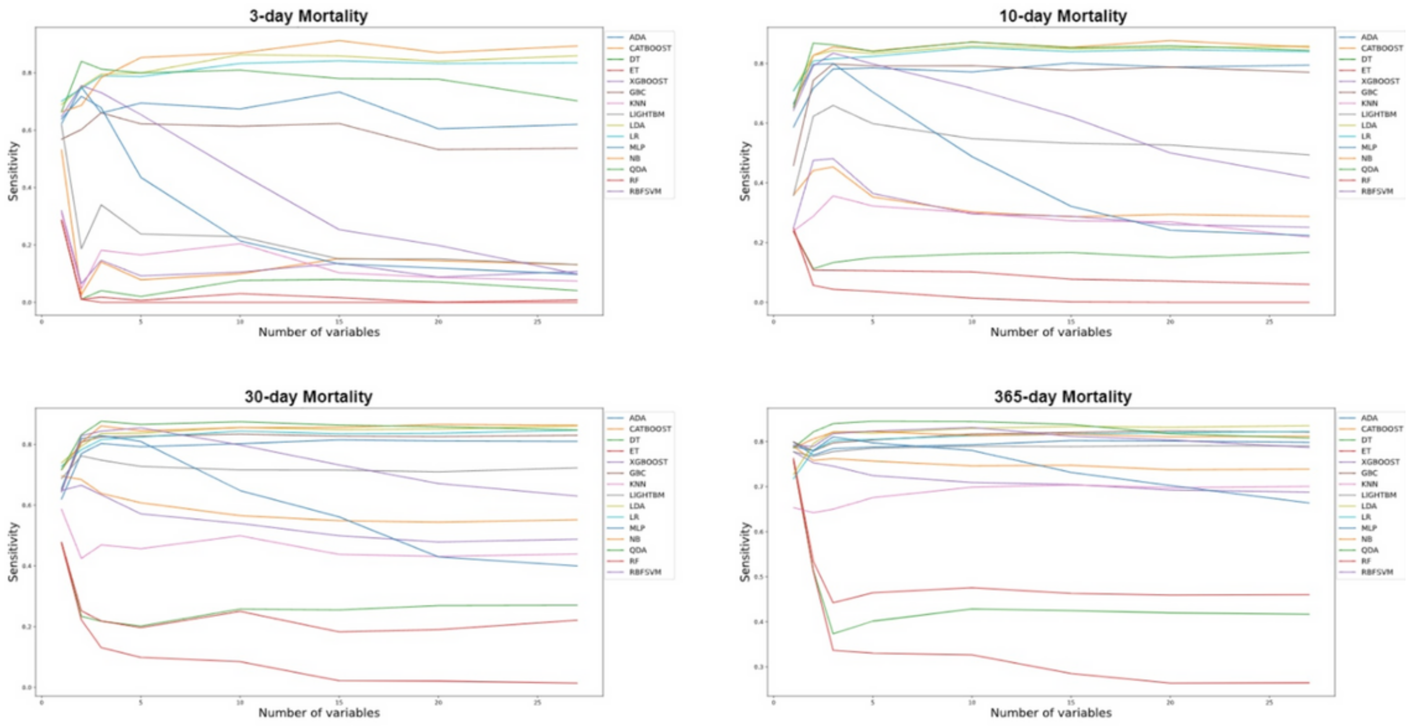

Figure 3

Sensitivity of fifteen ML algorithms on training data, as measured by the Area Under the Receiver Operating Characteristic Curve (AUC) when using 1 to 27 variables as predictors in the machine learning algorithms. Figures 3a-3d demonstrate the sensitivity for 3-day, 10-day,30-day and 365-day mortality on training data, respectively. Among all models, the highest sensitivity is shown between 0.88-0.91 in figures 3a-3c. In Figure 3d, the highest sensitivity reached is 0.85. When using ten variables, the top 3 models achieved a sensitivity above 0.85-91 in figures 3a-3d, The Sensitivity falls below 0.8 when using fewer than three variables for all models in figures 3a-3d.

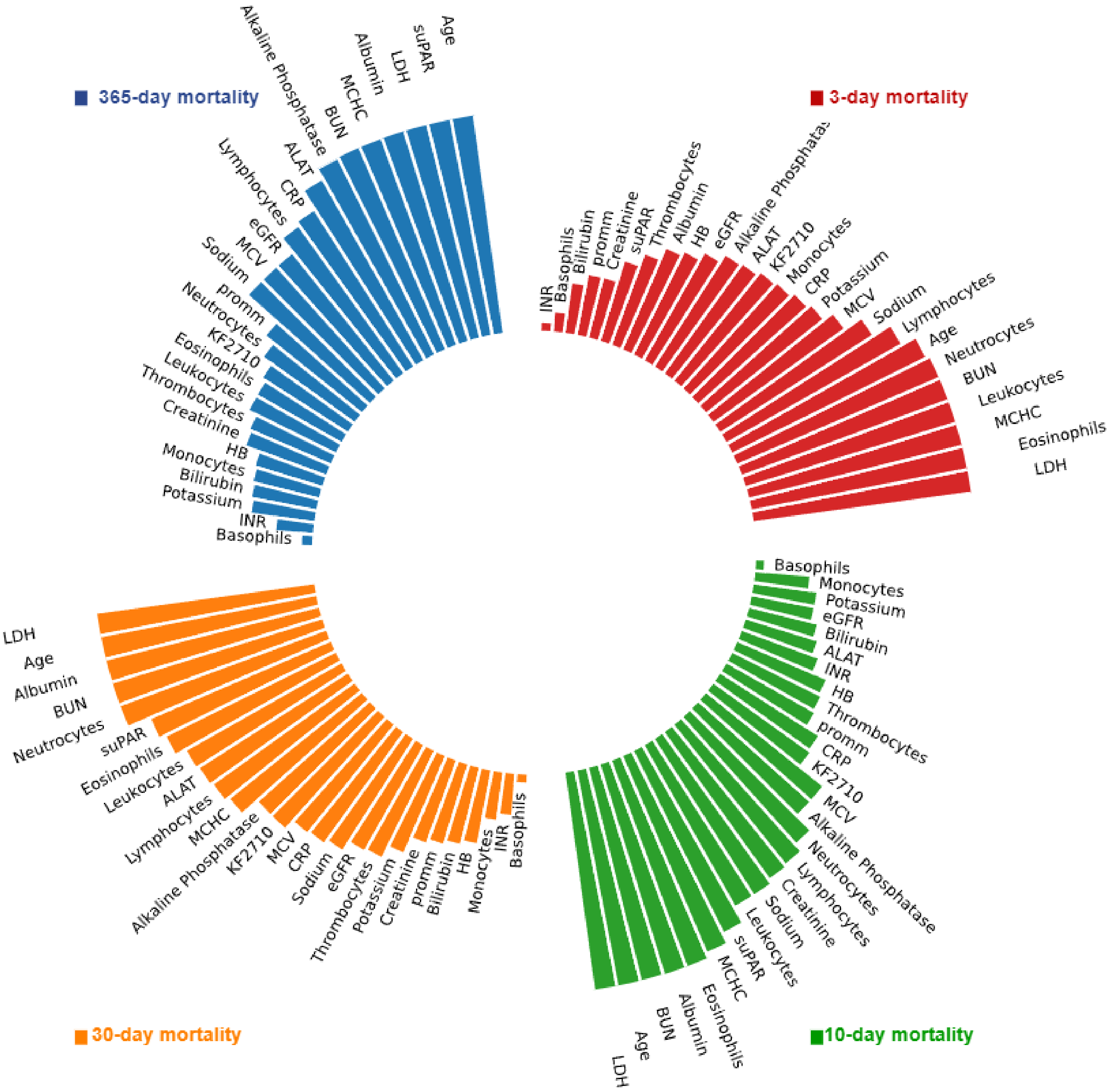

Figure 4

Ranking of importance of biomarkers in the IDA, LR, GBC, ADA and LightGBM models for prediction of 3-,10, 30, and 365-day mortality.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarytables12.pdf