Prospective and External Validation of Prognostic Machine Learning Models for Short- and Long-Term Mortality Among Acutely Admitted Patients Based on Blood Tests.

Baker Nawfal Jawad

baker_jawad@hotmail.com

Amager and Hvidovre Hospital, The Capital Region of Denmark https://orcid.org/0000-0002-6915-744X

Izzet Altintas

Amager and Hvidovre Hospital, The Capital Region of Denmark

Jesper Eugen-Olsen

Amager and Hvidovre Hospital, The Capital Region of Denmark

Siar Niazi

North Zealand Hospital, Hillerød

Abdullah Mansouri

Emergency Medical Services, Capital Region

Line Jee Hartmann Rasmussen

Amager and Hvidovre Hospital, The Capital Region of Denmark

Martin Schultz

Amager and Hvidovre Hospital, The Capital Region of Denmark

Kasper Iversen

Department of Cardiology, Herlev University Hospital, Herlev

Nikolaj Normann Holm

Amager and Hvidovre Hospital, The Capital Region of Denmark

Thomas Kallemose

Amager and Hvidovre Hospital, The Capital Region of Denmark

Ove Andersen

Amager and Hvidovre Hospital, The Capital Region of Denmark

Jan Nehlin

Amager and Hvidovre Hospital, The Capital Region of Denmark

Article

Keywords:

Posted Date: April 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4277483/v1

Additional Declarations: Yes there is potential Competing Interest. 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 declare no financial or non-financial competing interests.

Abstract

The application of machine learning (ML) models in emergency departments (EDs) to predict short- and long-term mortality encounters challenges, particularly in balancing simplicity with performance. This study addresses this gap by developing models that uses a minimal set of biomarkers, derived from a single blood sample at admission, to predict both short-term and long-term mortality. Our approach utilizes biomarkers representing vital organs and the immune system, offering a comprehensive view of both acute and chronic disease states. Moreover, by integrating explainable machine learning methods, we ensured that clinicians can easily interpret the model's outputs. Our Analysis included 65,484 admissions from three cohorts at two large Danish university hospitals, demonstrating the models' efficacy with high accuracy, with AUC values between 0·87 and 0·93. These results underscore that a single assessment of routine clinical biochemistry upon admission can serve as a powerful tool for both short-term and long-term mortality prediction in ED admissions.

Introduction

Effectively identifying patients at low and high risk of adverse outcomes is crucial for optimal resource allocation and treatment prioritization in healthcare systems. With the global population aging[1], the demand on emergency departments (EDs), is expected to rise significantly[2,3]. Consequently, this underlies the urgent need for innovative, personalized care strategies to ensure efficient resource use and patient care. However, in advancing these novel strategies, finding a balance between simplicity and performance becomes critical. High-performance solutions often come with complexity that may limit their practicality in dynamic settings like EDs. Addressing this challenge is critical to developing effective yet manageable tools that can adapt to the fast-paced nature of emergency care.

In clinical practice, predicting all-cause mortality and assessing risk have consistently been crucial outcomes for clinicians[4–9]. Various scores and indices based on simple linear relationship have been proposed and used to predict mortality; however, their accuracy is only moderate[4,8,10]. In recent years, applying ML to healthcare data has surpassed these traditional methods, offering enhanced accuracy in predicting outcomes for various patient groups, including those with conditions like cardiac disease[11], COVID-19[12], trauma[13], sepsis[14], and those in intensive care units (ICU)[15]. While ML algorithms offer significant advantages, their integration into clinical practice presents challenges, primarily due to the incorporation of numerous clinical and non-clinical variables. This complexity poses a challenge for clinicians in terms of comprehension and practical application within the fast-paced environment of EDs[16]. Despite advancements in ML for mortality prediction, significant research gaps persist: most models are either designed as triage tools for short-term outcomes or as risk assessment tools for long-term mortality, typically focusing on specific patient cohorts. Additionally, there is a lack of models that address both short- and long-term outcomes simultaneously across a diverse patient population, and few of these models are interpretable. The effectiveness of ML models in clinical settings is closely

linked to their transparency and interpretability, highlighting the need for predictive and comprehensible models for clinicians.

In this study, we aim to develop and externally validate an easily adaptable prognostic machine learning tool, the Short and Long-Term Mortality Models (SLTM), which are designed to predict both short-term and long-term mortality among patients acutely admitted to the ED. These models utilize a single blood sample for routine clinical biochemistry analysis, including biomarkers for vital organs and the immune system. We also aim to incorporate explainable ML techniques to more clearly explain how the models use input variables to make predictions, thereby assisting clinicians in understanding the ML model's predicted outcomes.

Results

Description of the cohorts used in the study.

In this study, we included a total of 65,484 admissions from the EDs of two different Danish hospitals, the Copenhagen University Hospital Amager and Hvidovre (AHH), and North Zealand University Hospital (NZH). From AHH we included ED data from both retro and prospective cohorts: 29K (2013–2017) and RESPOND-COVID (2020–2022), respectively, in our analysis. From NZH, we included ED data from a prospective cohort, TRIAGE (2013). The hospital and patient characteristics are summarized in Table 1.

For the 29K dataset, there were 51,007 admissions at the ED of AHH involving 28,683 unique patients during the study period. After excluding 2,166 patient records for missing data, the study cohort consisted of 48,841 admission from 28,671 unique patients (Fig. 1). The cohort was herof named 29K. Out of these, 34,187 (70%) were allocated as training data, 7,327 (15%) as validation data, and 7,327 (15%) as internal test data. The median age of 29K patients at admission was 65·6 years (IQR: 48·2–78·5), with 52·3% being female.

The TRIAGE cohort from NZH included 6,383 admissions in the ED, involving 6,356 unique patients. After excluding 233 patient records with missing data, the study cohort comprised 6,150 admissions involving 6,124 unique patients (Fig. 1). All TRIAGE data were used for external data validation. The median age of the TRIAGE patients at admission was 63·0 years (IQR: 46·0–76·0), with 50·6% being female.

The RESPOND-COVID cohort from AHH consisted of 28,210 patient records from 8853 unique patients; however, only 10,493 admissions from 8,451 unique patients with a suPAR measurement were included. The median age of RESPOND-COVID patients at admission was 66·0 years (IQR 49·1–78·2), with 51·2% being female.

Patients from the two AHH cohorts (29K and RESPOND-COVID) were slightly older ($p < 0·0001$), had a higher proportion of females ($p < 0·01$), and exhibited higher mortality rates ($p < 0·0001$), 4·0% and 4·4%, respectively, compared to patients from NZH, who had a mortality rate of 2·9% (Table 1). In general, distributions of all variables were significantly different in the three datasets. Patients excluded based on

missing data from the 29K and TRIAGE datasets showed no significant differences compared to those included. However, in the RESPOND-COVID cohort, the excluded patients were predominantly older, mostly female, and showed a lower mortality rate at 10 to 365 days.

Table 1
Patient baseline characteristics and mortality rates

	Retrospective	Prospective	Prospective	P-value
	AHH (29K) 2013–2017	NZH (Triage) 2013	AHH (RESPOND-COVID) 2020–2022	
Number of unique patients	28671	6124	8451	
Number of admissions	48841	6150	10493	
Variables				
Age	65·6 (48·2–78·5)	63·0 (46·0–76·0)	66·0 (49·1–78·2)	P < 0·0001*
Sex (female), n (%)	(52·3%)	(50·5%)	(51·2%)	P < 0·0001
ALAT (U/L)	21·0 (15·0–33·0)	20·8 (14·8–31·5)	23·0 (16·0–35·0)	P > 0·05
Albumin (g/L)	34·0 (30·0–37·0)	37·2 (33·5–39·8)	34·0 (30·0–37·0)	P < 0·0001[a,c]
Alkaline Phosphatase (U/L)	75·6 (63·0–94·0)	84·2 (69·3–105·9)	79·0 (63·0–103·0)	P < 0·0001*
Bilirubin (µmol/L)	7·0 (5·0–10·1)	7·9 (5·7–11·2)	8·0 (5·0–11·0)	P < 0·0001*
BUN (mmol/L)	5·1 (3·8–7·2)	5·2 (4·0–7·1)	5·3 (3·9–7·7)	P < 0·0001[b,c]
Creatinine (µmol/L)	77·0 (62·0–97·0)	71·0 (59·0–88·0)	77·0 (62·0–98·0)	P < 0·0001*
CRP (mg/L)	7·0 (2·0–39·0)	5·2 (2·9–23·2)	12·0 (2·6–54·0)	P < 0·0001*
HB (mmol/L)	8·1 (7·2–8·9)	8·4 (7·6–9·0)	8·2 (7·3–9·0)	P < 0·0001*
INR	1·0 (1·0–1·1)	1·0 (0·9–1·1)	1·0 (1·0–1·1)	P < 0·0001*
Potassium (mmol/L)	3·9 (3·6–4·2)	4·0 (3·8–4·3)	3·9 (3·6–4·2)	P < 0·0001*
KF2710	0·9 (0·8–1·0)	0·9 (0·8–1·1)	0·8 (0·7–0·9)	P < 0·0001*

	Retrospective	Prospective	Prospective	P-value
	AHH (29K) 2013–2017	NZH (Triage) 2013	AHH (RESPOND-COVID) 2020–2022	
LDH (U/L)	186·0 (169·0–214·0)	182·6 (157·6–217·3)	214·0 (184·0–260·0)	P < 0·0001[b,c]
Leukocytes (x 10^9 /L)	8·7 (6·9–11·3)	8·2 (6·5–10·6)	8·7 (6·6–11·8)	P < 0·0001*
Lymphocytes (x 10^9 /L)	1·7 (1·1–2·3)	1·6 (1·2–2·0)	1·4 (0·9–2·1)	P < 0·0001[a,b]
Monocytes (x 10^9 /L)	0·7 (0·5–0·9)	0·6 (0·5–0·8)	0·7 (0·5–0·9)	P < 0·0001
Neutrophils (x 10^9 /L)	5·8 (4·1–8·3)	5·7 (4·0–7·9)	6·0 (4·1–8·9)	P < 0·0001
suPAR (ng/mL)	3·3 (2·3–5·0)	4·5 (3·5–6·4)	4·1 (2·9–5·9)	P < 0·0001
Thrombocytes (x 10^9 /L)	247·0 (201·0–302·0)	238·0 (196·0–288·0)	248·0 (196·0–310·0)	P < 0·0001
Eosinophils (x 10^9 /L)	0·1 (0·0–0·2)	0·2 (0·1–0·4)	0·1 (0·0–0·2)	P < 0·0001
eGFR (mL/min)	80·0 (60·0–90·0)	86·3 (67·6–90·0)	77·0 (58·0–84·5)	P < 0·0001
Sodium (mmol/L)	139·0 (136·0–141·0)	139·0 (136·9–140·6)	138·0 (135·0–140·0)	P < 0·0001[b,c]
Mortality rate 10 days, n (%)	1252 (4·4%)	177 (2·9%)	341 (4%)	P < 0·0001*
Mortality rate 30 days, n (%)	2338 (8·2%)	284 (4·6%)	712 (8·4%)	P < 0·0001[a,c]
Mortality rate 90 days, n (%)	3394(11·8%)	475 (7·8%)	1052 (12·4%)	P < 0·0001*
Mortality rate 365 days, n (%)	4677 (16·3%)	729 (11·9%)	1560 (18·5%)	P < 0·0001*

Table 1. *Results are expressed as median (IQR, interquartile range) for continuous variables. For categorical variables, results are expressed as number of participants (percentage). ALAT: Alanine-aminotransferase; BUN: Blood urea nitrogen; CRP: C-reactive protein; eGFR: estimated glomerular filtration rate; HB: Hemoglobin; INR: Prothrombin Time and International Normalized Ratio; KF2710: coagulation factors 2, 7, 10; LDH: Lactate dehydrogenase; suPAR: soluble urokinase plasminogen activator receptor. 29K: Emergency Department at the Copenhagen University Hospital Amager and Hvidovre (AHH), 2013–2017. TRIAGE: Emergency Department at North Zealand University Hospital*

*(NZH), 2013. RESPOND-COVID: Emergency Department at AHH, 2020–2022. * P-value: significant difference between all datasets; a between 29K dataset and TRIAGE dataset; b between 29K dataset and RESPOND-COVID dataset; c between RESPOND-COVID dataset and TRIAGE dataset.*

Model performance

Figure 2 illustrates the predictive performance of the LightGBM models, assessed by the AUC, for mortality predictions at 10, 30, 90, and 365 days. Additional performance metrics are detailed in Table 2. For the 29K test dataset, the LightGBM model demonstrated high predictive accuracy, exhibiting an AUC of 0·93 (95% CI: 0·92–0·94) for 10-day mortality predictions and an MCC of 0·30 (95% CI: 0·28–0·32). The model maintained high performance for 30-day mortality predictions with an AUC of 0·92 (95% CI: 0·90–0·92) and an MCC of 0·40 (95% CI: 0·38–0·42). For 90-day mortality, the AUC was 0·91 (95% CI: 0·90–0·92) alongside an MCC of 0·51 (95% CI: 0·49–0·53), and for 365-day mortality, the AUC was 0·91 (95% CI: 0·91–0·91) with an MCC of 0·53 (95% CI: 0·51–0·55).

In the RESPOND-COVID dataset, the AUCs were 0·88 (95% CI: 0·86–0·89) for 10-day, 0·88 (95% CI: 0·87–0·89) for 30-day, 0·87 (95% CI: 0·86–0·88) for 90-day, and 0·88 (95% CI: 0·86–0·90) for 365-day mortality predictions. The MCC values corresponded to 0·22 (95% CI: 0·20–0·24), 0·32 (95% CI: 0·30–0·33), 0·38 (95% CI: 0·36–0·40), and 0·43 (95% CI: 0·41–0·45), respectively. Lastly, for the TRIAGE dataset, the AUCs were 0·87 (95% CI: 0·85–0·89) for 10-day mortality, 0·88 (95% CI: 0·86–0·90) for 30-day, 0·88 (95% CI: 0·86–0·90) for 90-day, and 0·90 (95% CI: 0·89–0·91) for 365-day mortality. The MCCs were 0·25 (95% CI: 0·23–0·27), 0·34 (95% CI: 0·32–0·36), 0·40 (95% CI: 0·38–0·42), and 0·43 (95% CI: 0·41–0·45), respectively.

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

	N	AUC	Sensitivity	Specificity	PPV	NPV	MCC
10-day Mortality							
29K	7·327 (272)	0·93 (0·92–0·94)	0·90 (0·86–0·93)	0·82 (0·81–0·83)	0·12 (0·11–0·14)	1·0 (1·0–1·0)	0·30 (0·28–0·32)
RESPOND-COVID	10·493 (341)	0·88 (0·86–0·89)	0·88 (0·84–0·91)	0·70 (0·69–0·71)	0·09 (0·08–0·10)	0·99 (0·99–1·0)	0·22 (0·22–0·24)
TRIAGE	6·150 (177)	0·87 (0·85–0·89)	0·72 (0·65–0·79)	0·84 (0·84–0·85)	0·12 (0·10–0·14)	0·99 (0·99–0·99)	0·25 (0·22–0·29)
30-day mortality							
29K	7·327 (537)	0·92 (0·90–0·92)	0·89 (0·86–0·91)	0·83 (0·82–0·83)	0·23 (0·21–0·24)	0·99 (0·99–0·99)	0·40 (0·38–0·42)
RESPOND-COVID	10·493 (712)	0·88 (0·87–0·89)	0·89 (0·86–0·91)	0·68 (0·68–0·69)	0·18 (0·17–0·19)	0·99 (0·98–0·98)	0·32 (0·30–0·33)
TRIAGE	6·150 (284)	0·88 (0·86–0·90)	0·76 (0·71–0·81)	0·84 (0·83–0·85)	0·18 (0·16–0·21)	0·99 (0·98–0·99)	0·34 (0·30–0·37)
90-day Mortality							
29K	7·327 (982)	0·91 (0·90–0·92)	0·84 (0·82–0·86)	0·85 (0·84–0·86)	0·38 (0·36–0·40)	0·98 (0·98–0·98)	0·51 (0·49–0·53)
RESPOND-COVID	10·493 (1052)	0·87 (0·86–0·88)	0·84 (0·82–0·86)	0·73 (0·72–0·74)	0·28 (0·26–0·29)	0·97 (0·97–0·97)	0·38 (0·36 – 0·40)
TRIAGE	6·150 (475)	0·88 (0·86–0·90)	0·77 (0·73–0·81)	0·84 (0·83–0·85)	0·30 (0·27–0·32)	0·98 (0·97–0·98)	0·40 (0·37 – 0·43)
365-day mortality							
29K	7·327 (1812)	0·91 (0·91–0·91)	0·87 (0·86–0·88)	0·79 (0·79–0·79)	0·47 (0·46–0·48)	0·97 (0·96–0·97)	0·53 (0·51–0·55)
RESPOND-COVID	10·493 (1569)	0·88 (0·86–	0·85 (0·83–	0·78 (0·77–	0·45 (0·44–	0·96 (0·96–	0·43 (0·42–

		0·90)	0·87)	0·79)	0·47)	0·97)	0·45)
TRIAGE	6·150 (729)	0·90 (0·89–0·91)	0·87 (0·84–0·89)	0·75 (0·74–0·76)	0·32 (0·30–0·34)	0·98 (0·97–0·98)	0·43 (0·41–0·45)

Table 2. N denotes the total number of admissions in the RESPOND-COVID and TRIAGE cohorts, and the number of test admissions in the 29K cohort, with the number in parentheses indicating the number of deaths. Results are based on the calibrated models, presented as the mean with 95% confidence intervals. AUC: area under receiver operating curve based on test data. PPV: Positive predictive value, NPV: Negative predictive value, MCC: Matthews Correlation Coefficient.

The calibrated LightGBM models, showed varying levels of sensitivity and specificity across the datasets for mortality prediction intervals (Table 2). In the 29K dataset, sensitivity for predicting mortality ranged from 84–90%, while specificity was between 79–83%. Within the RESPOND-COVID dataset, model sensitivity was between 84–88%, with specificity ranging from 70–78%. For the TRIAGE dataset, sensitivity varied from 72–87%, and specificity showed a narrow range of 75–84%.

In this analysis, we utilized Explainable Artificial Intelligence (XAI) techniques, particularly using estimated SHAP values, to analyze the LGBM model's predictions of 10-day mortality risk for two specific patient cases within the TRIAGE cohort. These plots were based on the calibrated model **3a**. Case 1: a notable case of an elderly male patient who was admitted to the Emergency Department at North Zealand Hospital. X-axis: Displays the percentage contribution of biomarkers taken at the ED to the prediction of mortality risk. Below, these biomarkers are categorized by function: inflammatory, infection-related, liver, and kidney markers. **3b**. Case 2: This case involves an elderly female patient who presented at an Emergency Department in TRIAGE with a suspected infection. The model estimated a relatively low 10-day mortality risk of 4.2% for her. In case 1 and 2 variable sex is not shown (value 0.002) **3c.** Calibration plots on TRIAGE cohort. X-axis: Mean Predicted Probability; Y-axis: Observed Frequency. The diagonal dash-line stretching from the bottom left to the top right represents perfect agreement between predicted probabilities and actual outcomes. The blue line is the model´s predicted probabilities and actual outcomes.

Machine Learning Predictions Application

From our prospective cohort dataset TRIAGE at NZH, upon which we tested the prediction model, we analyzed a case of an elderly male patient, with a history of cardiovascular and neurological conditions (Fig. 3a). This patient was brought to the ED with symptoms suggesting an infection in the respiratory system. Upon arrival, the patient was categorized as with moderate urgency based on the triage (Early Warning Score). Laboratory tests revealed elevated levels of inflammatory markers, indicating a bacterial infection. The patient was treated with standard outpatient antibiotics and discharged the same day. Unfortunately, the patient’s condition deteriorated, leading to the patient's passing a few days later. The short-term mortality model, predicting 10-day mortality, had estimated a high mortality risk of 97·5% for this patient. The contribution plot created using Explainable Artificial Intelligence (XAI), depicted in Fig.

3a, provides insights into how the patients impaired organ-specific biomarkers contributed to the model's overall output of mortality risk. In this case, the most important contributors to mortality risk prediction were identified as markers of inflammatory response to the infection (27%), markers of immune system (14%), markers of liver function (21%), and kidney function (17%). These insights shed light on the specific factors driving the elevated risk for this patient. With this knowledge, this patient should have been hospitalized. Subsequent independent reviews by two specialists suggested that a hospital admission might have been necessary for this patient.

Moving to a different scenario, in another case, we examined an elderly female patient, with a history of cardiovascular conditions and metabolic disease, presenting also with symptoms indicative of an infection in the respiratory system (Fig. 3b). The patient was categorized as urgent upon triage. Laboratory tests indicated signs of infection, with elevated inflammatory markers and abnormal blood cell counts. Other notable lab results included imbalances in electrolytes and liver enzymes. Despite these findings, the LGBM model, analyzed through XAI, predicted a lower 10-day mortality risk of just 4·3%. The patient was hospitalized and treated with IV antibiotics for three days. However, subsequent reviews by specialists suggested that hospital admission might have been unnecessary.

Discussion

In this study, we utilized routine clinical biochemistry data from a single time point upon admission, representing vital organ and immune system function, to predict mortality risk in acutely admitted patients. By incorporating explainable ML methods, we ensured that the model's outputs could be interpreted, thereby aiding clinicians in understanding the predicted ML outcomes. Our results, for both Short and Long-Term Mortality Models demonstrated very good to excellent performance metrics, achieving high AUC values ranging from 0·87 to 0·93. Although a small decline in AUC values in the TRIAGE and RESPOND-COVID datasets was observed compared to the 29K dataset, this was anticipated due to significant differences in patient characteristics and mortality rates across cohorts. Performance metrics, especially AUC and MCC, showed overlapping confidence intervals for the RESPOND-COVID and TRIAGE datasets. This overlap indicates that the models performed similarly across these datasets. Nonetheless, we observed variability in the models' sensitivity and specificity across the different cohorts.

Overall, the models demonstrated low PPVs ranging from 9–47%, indicating a large proportion of false positives, while showing very high NPVs ranging from 96–100%. A trend of increase in PPV and MCC values was observed from short-term to long-term mortality prediction, indicating a higher probability to predict the outcome over the length of time. The low PPV, in short-term mortality prediction, could be attributed to the low mortality prevalence in the studied patient populations. Additionally, it is possible that the model identifies patients (false positive) as being at high risk of mortality, but upon readmission and/or subsequent treatment after their initial discharge, these patients survive. As regards the high NPV, the results should be interpreted considering the dataset's overall low mortality rate, or conversely, its high survival rate.

In clinical practice, screening tools that offer high sensitivity and high NPV are preferred and well-justified[17,18], as these tools align with clinicians' needs for safely excluding individuals at low risk of adverse outcomes in the future. This approach is preferred due to the low pretest probability, and the goal of the diagnostic test will be “ruling out” the condition, emphasizing high sensitivity where a negative result effectively excludes the condition. This is in contrast with diagnostic tools, where a high pretest probability of a condition leads to the goal of “ruling-in” the condition, emphasizing high Specificity and PPV, as a positive result effectively confirms the condition.

Our ML models embody this clinical principle, providing reliable decision support that matches the preferences of healthcare practitioners. This alignment with clinical practices not only supports the models' utility but also sets a foundation for their potential development and application in healthcare settings.

Comparing our models with existing ML models, in terms of short-term mortality prediction, our models achieved an AUC of 0·87 to 0·93 for 10-day mortality predictions across the studied cohorts. This performance, when compared to other promising models, seems to be either on par or clearly superior, as explained below. Nevertheless, it's crucial to acknowledge that comparing results from data across diverse populations can be complex, given the multifaceted nature of socioeconomic, health factors, and other variables. Furthermore, mortality rates can be different in each population. Despite these complexities, when reviewing the literature, we find notable results. For instance, Trentino et al. conducted a study analyzing data from three adult tertiary care hospitals in Australia[19]. This study achieved a remarkable AUC of 0.93 for predicting in-hospital mortality among all admitted patients, regardless of whether their cases were medical or surgical. The predictive model used in this study incorporated various variable, including demographic, diagnosis code, administrative information and Charlson comorbidity Index. Similarly, an ED triage tool, the Score for Emergency Risk Prediction (SERP), to predict mortality within 2 to 30 days for ED patients was initially applied in a cohort from a Singaporean ED and subsequently underwent external validation in a South Korean ED[20,21]. These studies demonstrated AUCs of 0·81 – 0·82 for in-hospital mortality and 0·80 – 0·82 for 30-day mortality prediction. The SERP scores incorporate variables, including age, vital signs, and comorbidities. Additionally, in a study conducted on hospitalized patients in the U.S. by Brajer et al., reported an AUCs between 0·86 and 0·89 based on 57 electronic health record data variables[22]. In contrast, our models, performed competitively, achieving comparable or superior results for short-term mortality prediction using just 15 biomarkers measured from a single routine blood sample collected upon arrival in the ED. For 30-day mortality, our models consistently maintained high AUCs (0·88–0·92) in both internal and external evaluations. Likewise, the Long-Term Mortality models showed near-excellent performance, with AUCs ranging from 0·87 to 0·91 for 90-day mortality prediction and 0·88 to 0·91 for 365-day mortality prediction. The performance of this model is either superior or comparable to similar studies in the field.

The Random Forest models developed by Sahni et al. achieved an AUC of 0·86 for 1-year mortality predictions[23]. Their model incorporated various variable, including demographic, physiological,

biochemical factors, and comorbidities. Similarly, Woodman et al. developed a ML model trained on a patient cohort aged > 65 years. Their model achieved an AUC of 0·76, incorporated variables including demographic, BMI, anticholinergic risk score, biochemical markers, and a comprehensive geriatric assessment.

In this study, we have adopted a streamlined biomarker approach that aligns with the latest recommendations for AI deployment in healthcare settings, prioritizing consistency and reduced error susceptibility[24]. This approach, which is centered on a single blood sample routinely analyzed for a select set of standard vital organ and immune system biomarkers, presents significant advantages. Unlike existing tools that primarily focus on triage, our models extend its utility to encompass resource allocation, treatment planning, discharge, and potentially preventing overtreatment and ensuring that care aligns with the patient's preferences and recovery potential. Specifically, it provides a stable and chronic disease-oriented perspective, which is crucial for uncovering underlying pathologies that might not be apparent with other data types.

In stark contrast to other models that depend on various inputs—such as continuous vital sign monitoring, administrative variables, medical history, comorbidities, and medication profiles—our model's simplicity integrates more fluidly into clinical workflows and mitigates the 'black box' nature that often accompanies complex AI systems, where the intricacy of ML models and the use of non-clinical features can make it challenging to understand the rationale behind the model output. Our methodology, with its deliberately limited parameters, enhances the models' output transparency and interpretability, thereby building confidence and trust among clinicians in AI-assisted decision-making.

Limitations

The exclusion of specific patient groups from the cohorts, including children, and obstetric patients, limits the trained model applicability of our models to these populations. Furthermore, the retrospective design of our cohort introduces inherent limitations, such as the potential for selection and information biases. These biases can impact the validity of our findings and their applicability to broader, more diverse populations. There are also several limitations to SHAP values. SHAP values are used for interpreting predictions of ML models, specifically by quantifying the contribution of each feature to a particular prediction. However, they do not provide causal insights. This means that while SHAP values can tell us which variables were important in the model's decision-making process, they do not imply a cause-and-effect relationship between these variables and the prediction. Lastly, the models were primarily validated within the same geographical region and governing clinical jurisdiction. While they were evaluated across different cohorts, this regional focus might constrain the generalizability of our findings.

Future research

The present models are only meant as a proof-of-concept study. Refining and validating these models with diverse datasets remain a priority. Future research should focus on enhancing the PPV and incorporating more comprehensive patient data before implementation in clinical practice.

Conclusion

In this study, we have successfully developed and externally validated machine learning models that predict both short-term and long-term mortality in acutely admitted patients based on single set of routine blood tests. With AUC scores ranging from 0·87 to 0·93, we have demonstrated that a simplified approach can achieve sufficient high predictive accuracy, with the potential to warrant investigation into its applicability as an additional tool in clinical decision-making.

Methods

Study Design and Settings

In this study, we evaluated data from three study cohorts. First, the retrospective 29K cohort study from the ED at the Copenhagen University Hospital Amager and Hvidovre (AHH), Denmark was included. The 29K cohort included all patients admitted to the Acute Medical Unit with an available blood sample. The 29K cohort consisted of 51,007 patient records from ED admissions between 18 November 2013 and 17 March 2017. The Acute Medical Unit at the ED receives patients across all specialties, except children (patients under 18 years), gastroenterological patients, and obstetric patients. Second, a prospective observational cohort, the TRIAGE Study, from the ED at North Zealand University Hospital (NZH), Hilleroed, Denmark was included[25]. The TRIAGE cohort included all patients admitted to the ED with an available blood sample and consisted of 6,383 patient records from ED admissions between 5 September 2013 and 6 December 2013[25]. Children and obstetric patients were excluded. Third, a prospective observational cohort from the specialized COVID-19 Unit of the ED at AHH was included. This RESPOND-COVID cohort (Respiratory Emergency Surveillance and suPAR Outcome in COVID and Non-COVID Disease) included patients admitted with respiratory symptoms, suspected of having COVID-19. The RESPOND-COVID cohort consisted of 28,210 patient records from admissions between 10 March 2020 and 31 March 2022. The follow-up data were retrieved from the Central public server at Statistics Denmark and The Danish Civil Registration System. During the study period, patients who left the country had their last admission data censored.

This study was reported in accordance with the *transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD) statement[26]. The research conducted in this study was in strict compliance with both regional and national guidelines and received the necessary approvals from the appropriate Danish authorities. The Danish Data Protection Agency (29K: HVH-2014-018, 02767; TRIAGE: J. 2007-58-0015) and the Danish Health and Medicines Authority (29K: 3-3013-1061/1; TRIAGE: 31-1522-102) approved the studies and analyses. Furthermore, the Legal Office at the Capital Region of Denmark, specifically the Team for patient medical records, issued a permit with reference numbers R-

22041261 and R-20064514 for the usage and management of healthcare data within the region. The study adhered to all relevant ethical and legal standards required for research within Denmark.

Medical Records

Information regarding hospital admissions and discharges for the 29K cohort was extracted from the Danish National Patient Registry (NPR). In the RESPOND-COVID and TRIAGE cohorts, information related to patients' health conditions, vital signs, triage, events in-hospital requiring continued care at the ED, and duration of hospital stay was retrieved from the patient´s health records in OPUS Arbejdsplads (version 2.5.0.0 Computer Sciences Corporation [CSC]) and “Sundhedsplatformen” (Epic). The TRIAGE study also involved the review of each patient record by two experts in internal or emergency medicine, who evaluated whether each admission was necessary or could have been avoided. An admission was deemed unnecessary only if both specialists agreed on this assessment. These reviewers were blinded to their own roles as either primary or secondary evaluator of each case and to the other reviewer's decisions. An unnecessary admission was defined as one where the patient's condition could have been adequately managed by a general practitioner or in an outpatient setting within 1–2 weeks. For all three cohorts, blood test results were obtained from the LABKA II national database[27]. Using each person’s unique personal identification number from the Danish Civil Registration System, we linked the data, encompassing biochemistry, diagnoses, records of hospital admissions, readmissions, and mortality.

Biomarkers

For all three cohorts, blood samples were collected upon admission to the ED, and routine whole blood counts along with clinical biochemistry were analyzed by AHH’s and NZH’s respective Departments of Clinical Biochemistry[28]. The results were extracted from the LABKA database. The admission blood tests included 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 (ALP), creatinine, lactate dehydrogenase (LDH), blood urea nitrogen (BUN), potassium (K), sodium (NA), estimated glomerular filtration rate (eGFR), hemoglobin (HB), and counts of leukocytes, lymphocytes, neutrophils, monocytes, thrombocytes, eosinophils, and basophils. All procedures were executed following the appropriate guidelines and standards.

Outcomes

In this study, the outcomes were 10-, 30-, 90-, and 365-day mortality after admission at the ED. We define short-term mortality as death within 30 days, intermediate-term mortality as death within 90 days, and long-term mortality as death within 365 days.

Data preparation

A standard format was applied to all data. Patient admissions with over 50% data (routine clinical biochemistry results) missing were dropped. The median percentage missing blood biochemistry results for each study was: 29K: 2.6%; TRIAGE: 3.0%, and RESPOND-COVID: 2.4%, with an overall interquartile

range (IQR) ranging from 1.6–7.7%. The 29K cohort was split into training and test sets. To avoid information leakage, as some patients were readmitted multiple times, we ensured that the same patient ID did not appear in both training and test sets. The TRIAGE and RESPOND-COVID datasets were exclusively used as test data. To handle missing values, we employed iterative imputations from the Python scikit-learn package[29]. Training sets and test sets were fitted and handled separately. To handle class-imbalance in our target outcome, we used the random oversampling technique from the Imbalanced-Learn Package (Python)[30] during training.

To reduce the impact of magnitude on the variance, we normalized the values of all variables in the data by z-score. This process was applied to the training data. For the test data, normalization was based on the mean and standard deviation estimated from the training data. To achieve a normal distribution approximation for our variables, we employed the Yeo-Johnson transformation[31] to the training data. For the test data, the transformation was based on the estimates derived from the training set. These preprocessing steps significantly enhanced the model's performance on the validation set.

Model Construction

All models were crafted using Python (version 3.8.0). We employed the classification module from PyCaret (version 2.2.6)[32] to develop models using four distinct algorithms predicting 10-,30-, 90-. And 365-day mortality. PyCaret is a low-code machine learning library that streamlines the entire machine learning process. To optimize hyperparameters during cross-validation, we utilized a random grid search with 100 iterations within PyCaret and AUC as metric to evaluate the hyperparameter tuning.

Algorithm selection and performance measures

Fifteen machine learning algorithms were trained and evaluated using PyCaret through 10-fold cross-validation ((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))[33]. These models were developed using fifteen biomarkers (ALAT, ALB, ALP, BUN, creatinine, CRP, eosinophils, KF2710, leukocytes, LDH, lymphocytes, neutrophils, platelets, sodium, suPAR) from a single routine blood test along with age and sex as additional variables. The best machine learning algorithm (Light Gradient Boosting Machine [LIGHTGBM]) was fine-tuned and evaluated through a 10-fold cross-validation on the 29K cohort, then tested on an unseen test sample of the same cohort, and on TRIAGE and RESPOND-COVID data. Model selection was based on the area under the receiver operating characteristic curve (AUC). Additionally, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Matthew's correlation coefficient (MCC) for the complete data, were estimated for the validation and test data and evaluated between them. To evaluate the reliability of our predictive model's performance, we used a bootstrap resampling technique to calculate confidence intervals (CIs) for key

metrics including accuracy, AUC, sensitivity, specificity, PPV, NPV and MCC. We performed 1000 bootstrap iterations.

Calibration

To increase the accuracy of our predictive model in estimating probabilities that accurately reflect actual outcomes, we implemented probability calibration using isotonic regression from the scikit-learn package (version 1.4) in Python[29]. The data used for the calibration process was unseen test data. To evaluate the model's precision, we employed the Brier score as an accuracy metric, comparing its values both before and after the calibration process. After Calibration, the Youden Index for 29K dataset was employed for all outcomes as a method to empirically determine the optimal dichotomous cutoff post-calibration, allowing us to assess the model's sensitivity, specificity, positive predictive value, and negative predictive value.

Explaining Model Predictions

We used TreeExplainer[34], a method based on the approximation of the set of trees to calculate the SHapley Additive exPlanations (SHAP)[35] values, to interpret our machine learning models' predictions, quantifying the impact of each variable on individual predictions. SHAP values were calculated using the SHAP library package in Python, which helped assess variable contributions in terms of magnitude and direction. To visualize these contributions, we employed waterfall plots. These plots illustrated the influence of individual variables on individual model prediction.

Statistical analysis

Statistical analyses were performed using R (version 4.1.0) and Python (version 3.8.0). Categorical variables are presented as frequencies (N) and percentages (%), whereas continuous variables are represented by their median values and IQR. To evaluate the differences between datasets, we employed the Student's t-test, setting our significance level at 5% for continues variables, and the chi-squared test for categorical variables.

Declarations

Data and Code availability

The datasets analyzed during the current study are not publicly available due to privacy considerations (data use agreements) or ethical restrictions. However, they can be made available from the corresponding author upon reasonable request and after obtaining the necessary approvals from the relevant authorities.

The underlying code for this study, as well as the training/validation datasets, is not publicly available, for proprietary reasons. However, qualified researchers may request access from the corresponding author, and they can be provided upon reasonable request.

Authors´ Contributions

B.J., J.N. and O.A. were responsible for the research idea. T.K, O.A, L.J.H.R, J.E.O, K.I, and M.S were responsible for the study design. B.J, S.Z, and T.K. were responsible for the statistical analysis and algorithm training and evaluation. B.J, A.M, T.K, A.M, S.Z, N.N.H and J.N 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 and at North Zealand University Hospital for all analyses used in this study. We thank the contribution from Associate Professor Anders Stockmarr for valuable input. This study received no external funding.

Declaration of 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 declare no financial or non-financial competing interests.

References

1. ONU World population, ageing. *Suggest Cit United Nations, Dep Econ Soc Aff Popul Div (2015) World Popul Ageing*; United Nat
2. Veser A, Sieber F, Groß S et al The demographic impact on the demand for emergency medical services in the urban and rural regions of bavaria, 2012–2032. *J Public Heal*; 23. Epub ahead of print 2015. 10.1007/s10389-015-0675-6
3. Businger AP, Kaderli R, Burghardt LR et al (2012) Demographic Changes and Their Implications in a Nonacademic Emergency Department in Switzerland: An 11-Year Trend Analysis (2000–2010) of 104,510 Patients. *ISRN Emerg Med*; Epub ahead of print 2012. 10.5402/2012/865861
4. Lemeshow S, Gehlbach SH, Klar J et al Mortality Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients. JAMA J Am Med Assoc ; 270. Epub ahead of print 1993. 10.1001/jama.1993.03510200084037
5. Toma T, Abu-Hanna A, Bosman RJ Discovery and inclusion of SOFA score episodes in mortality prediction. *J Biomed Inform*; 40. Epub ahead of print 2007. 10.1016/j.jbi.2007.03.007
6. Burch VC, Tarr G, Morroni C Modified early warning score predicts the need for hospital admission and inhospital mortality. *Emerg Med J*; 25. Epub ahead of print 2008. 10.1136/emj.2007.057661

7. Klausen HH, Petersen J, Bandholm T et al (2017) Association between routine laboratory tests and long-term mortality among acutely admitted older medical patients: a cohort study. BMC Geriatr 17:62
8. Phungoen P, Khemtong S, Apiratwarakul K et al Emergency Severity Index as a predictor of in-hospital mortality in suspected sepsis patients in the emergency department. *Am J Emerg Med*; 38. Epub ahead of print 2020. 10.1016/j.ajem.2020.06.005
9. Mahmoodpoor A, Sanaie S, Saghaleini S et al (2022) 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 9 Epub ahead of print. 10.3389/fmed.2022.938005
10. Knaus WA APACHE 1978–2001: The development of a quality assurance system based on prognosis: Milestones and personal reflections. *Archives of Surgery*; 137
11. Sherazi SWA, Jeong YJ, Jae MH et al (2020) A machine learning-based 1-year mortality prediction model after hospital discharge for clinical patients with acute coronary syndrome. Health Inf J 26:1289–1304
12. Yadaw AS, Li Y-C, Bose S et al (2020) Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Heal 2:e516–e525
13. Bonde A, Bonde M, Troelsen A et al (2023) Assessing the utility of a sliding-windows deep neural network approach for risk prediction of trauma patients. Sci Rep 13:5176
14. Kijpaisalratana N, Sanglertsinlapachai D, Techaratsami S et al (2022) Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study. Int J Med Inf 160:104689
15. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP et al (2020) Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Heal 2:e179–e191
16. Kirk JW, Nilsen P (2016) Implementing evidence-based practices in an emergency department: Contradictions exposed when prioritising a flow culture. J Clin Nurs 25:555–565
17. Bartol T (2015) Thoughtful use of diagnostic testing: Making practical sense of sensitivity, specificity, and predictive value. Nurse Pract 40:10–12
18. Galvin R, Gilleit Y, Wallace E et al (2017) Adverse outcomes in older adults attending emergency departments: a systematic review and meta-analysis of the Identification of Seniors At Risk (ISAR) screening tool. Age Ageing 46:179–186
19. Trentino KM, Schwarzbauer K, Mitterecker A et al (2022) Machine Learning-Based Mortality Prediction of Patients at Risk During Hospital Admission. J Patient Saf 18:494–498
20. Xie F, Ong MEH, Liew JNMH et al Development and Assessment of an Interpretable Machine Learning Triage Tool for Estimating Mortality after Emergency Admissions. JAMA Netw Open ; 4. Epub ahead of print 2021. 10.1001/jamanetworkopen.2021.18467
21. Yu JY, Xie F, Nan L et al (2022) An external validation study of the Score for Emergency Risk Prediction (SERP), an interpretable machine learning-based triage score for the emergency

department. Sci Rep 12:1–8

22. Brajer N, Cozzi B, Gao M et al (2020) Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. JAMA Netw Open 3:1–14
23. Sahni N, Simon G, Arora R (2018) Development and Validation of Machine Learning Models for Prediction of 1-Year Mortality Utilizing Electronic Medical Record Data Available at the End of Hospitalization in Multicondition Patients: a Proof-of-Concept Study. J Gen Intern Med 33:921–928
24. Kristiansen TB, Kristensen K, Uffelmann J et al Erroneous data: The Achilles' heel of AI and personalized medicine. Front Digit Heal ; 4. Epub ahead of print 2022. 10.3389/fdgth.2022.862095
25. Plesner LL, Iversen AKS, Langkjaer S et al The formation and design of the TRIAGE study - Baseline data on 6005 consecutive patients admitted to hospital from the emergency department. Scand J Trauma Resusc Emerg Med ; 23. Epub ahead of print 2015. 10.1186/s13049-015-0184-1
26. Collins GS, Reitsma JB, Altman DG et al Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. Eur Urol ; 67. Epub ahead of print 2015. 10.1016/j.eururo.2014.11.025
27. Arendt JFH, Hansen AT, Ladefoged SA et al (2020) Existing data sources in clinical epidemiology: Laboratory information system databases in Denmark. Clin Epidemiol 12:469–475
28. Nehlin JO, Andersen O Molecular Biomarkers of Health BT - Explaining Health Across the Sciences. In: Sholl J, Rattan SIS (eds). Cham: Springer International Publishing, pp. 243–270
29. Pedregosa F, Varoquaux G, Gramfort A et al Scikit-learn: Machine learning in Python. *J Mach Learn Res*; 12
30. 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*; 18
31. Yeo I, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87:954–959
32. Moez A, PyCaret (2020) An open source, low-code machine learning library in Python, https://www.pycaret.org accessed March 8, 2023)
33. Jawad BN, Shaker SM, Altintas I et al (2024) Development and validation of prognostic machine learning models for short- and long-term mortality among acutely admitted patients based on blood tests. Sci Rep 14:5942
34. Lundberg SM, Erion G, Chen H et al (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2:56–67
35. Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., pp. 4768–4777

Figures

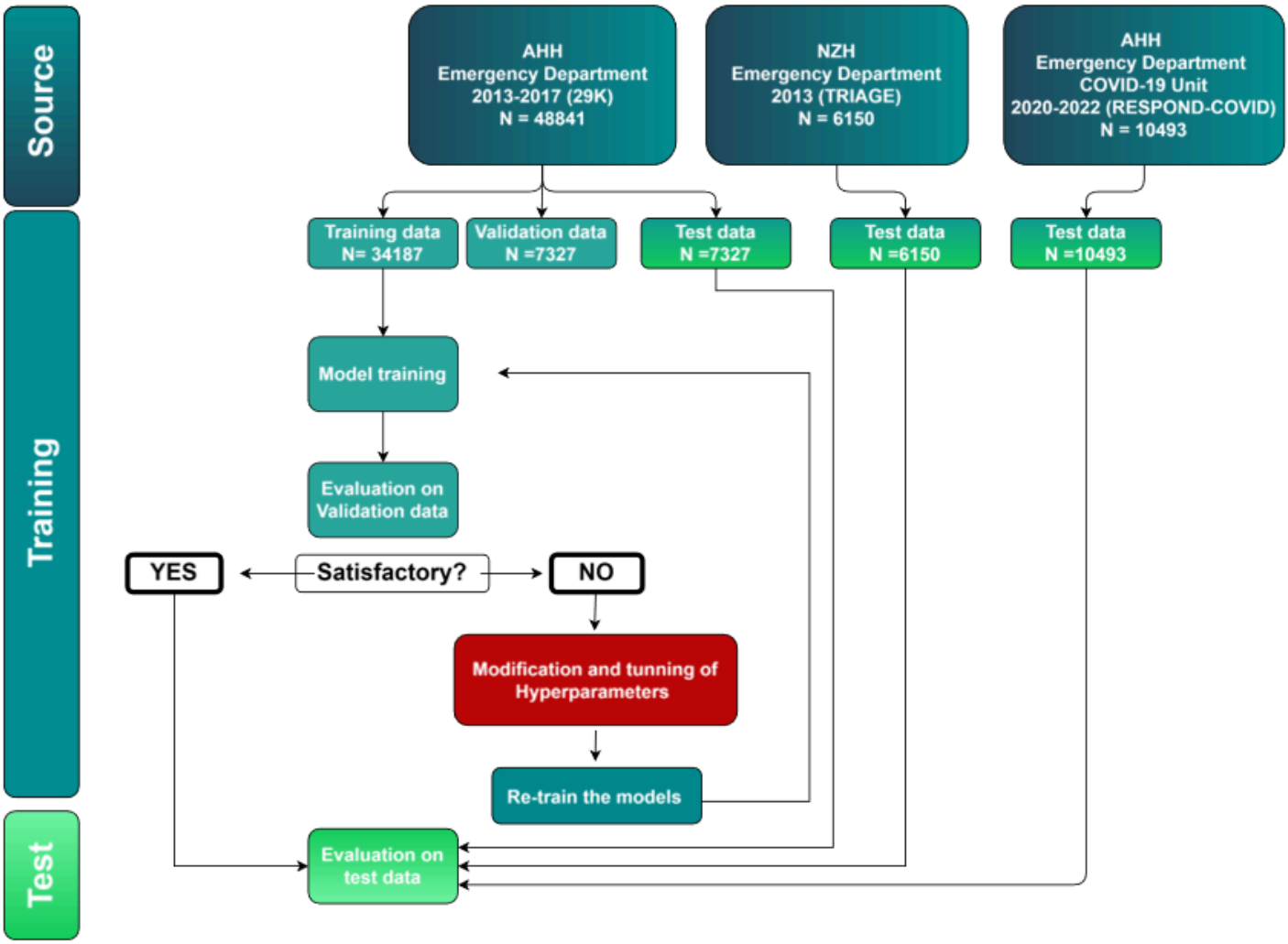

Figure 1

Flowchart of data training, validation, and testing using two machine learning algorithms.

Patient medical records originated from three cohorts at two Danish hospitals: 1) 29K: Emergency Department at the Copenhagen University Hospital Amager and Hvidovre (AHH), 2013-2017, 2) TRIAGE: Emergency Department at North Zealand University Hospital (NZH), 2013, and 3) RESPOND-COVID: Emergency Department at AHH, 2020-2022. N= Number of admissions.

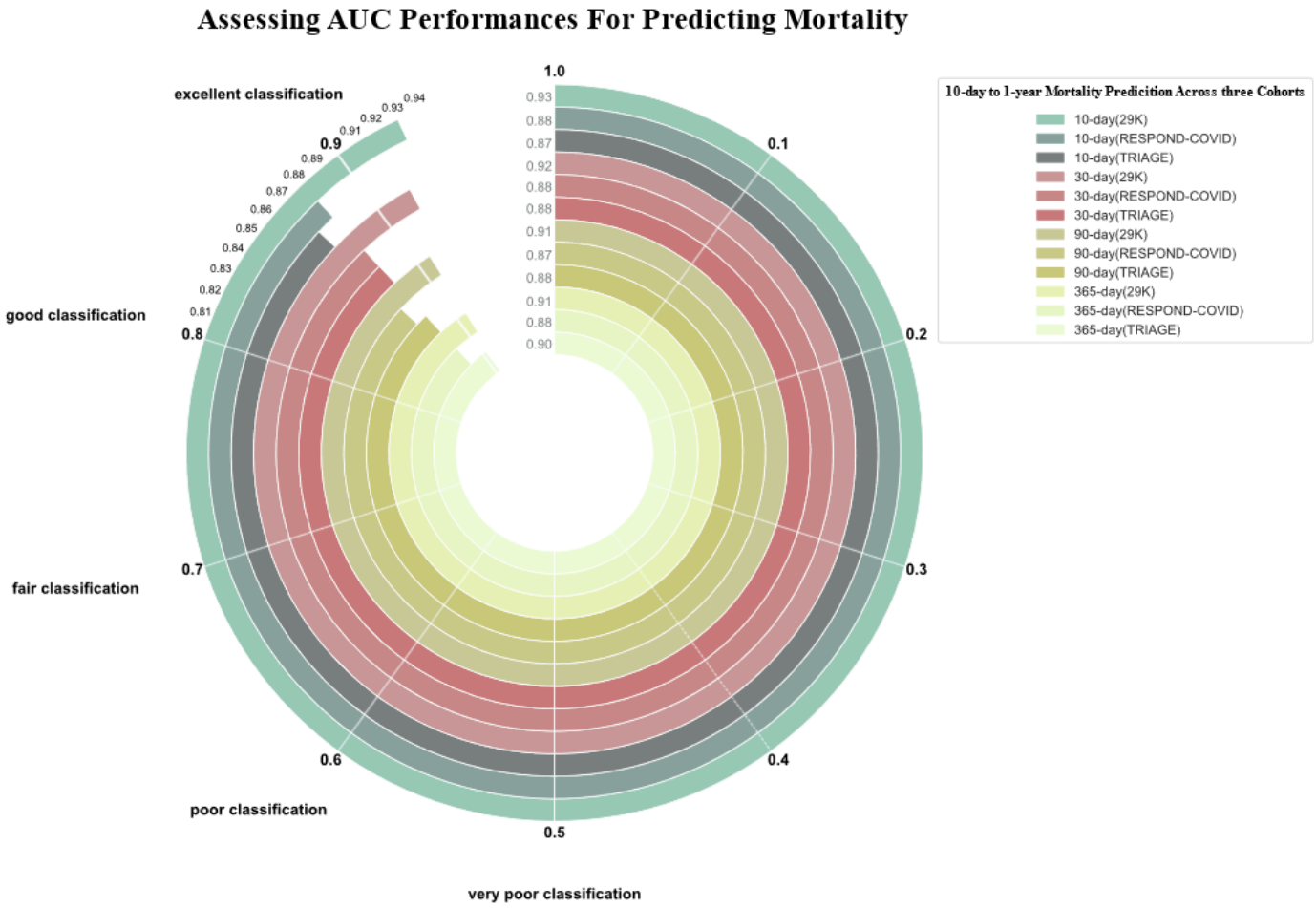

Figure 2

Area under the receiver operating characteristic curve (AUC) performance for predicting 10-day (green shaded circles), 30-day (peach-shaded circles), 90-day (dark yellow shaded circles), and 365-day (light yellow shaded circles) mortality in each of the three cohorts, 29K, TRIAGE, and RESPOND-COVID, examined.

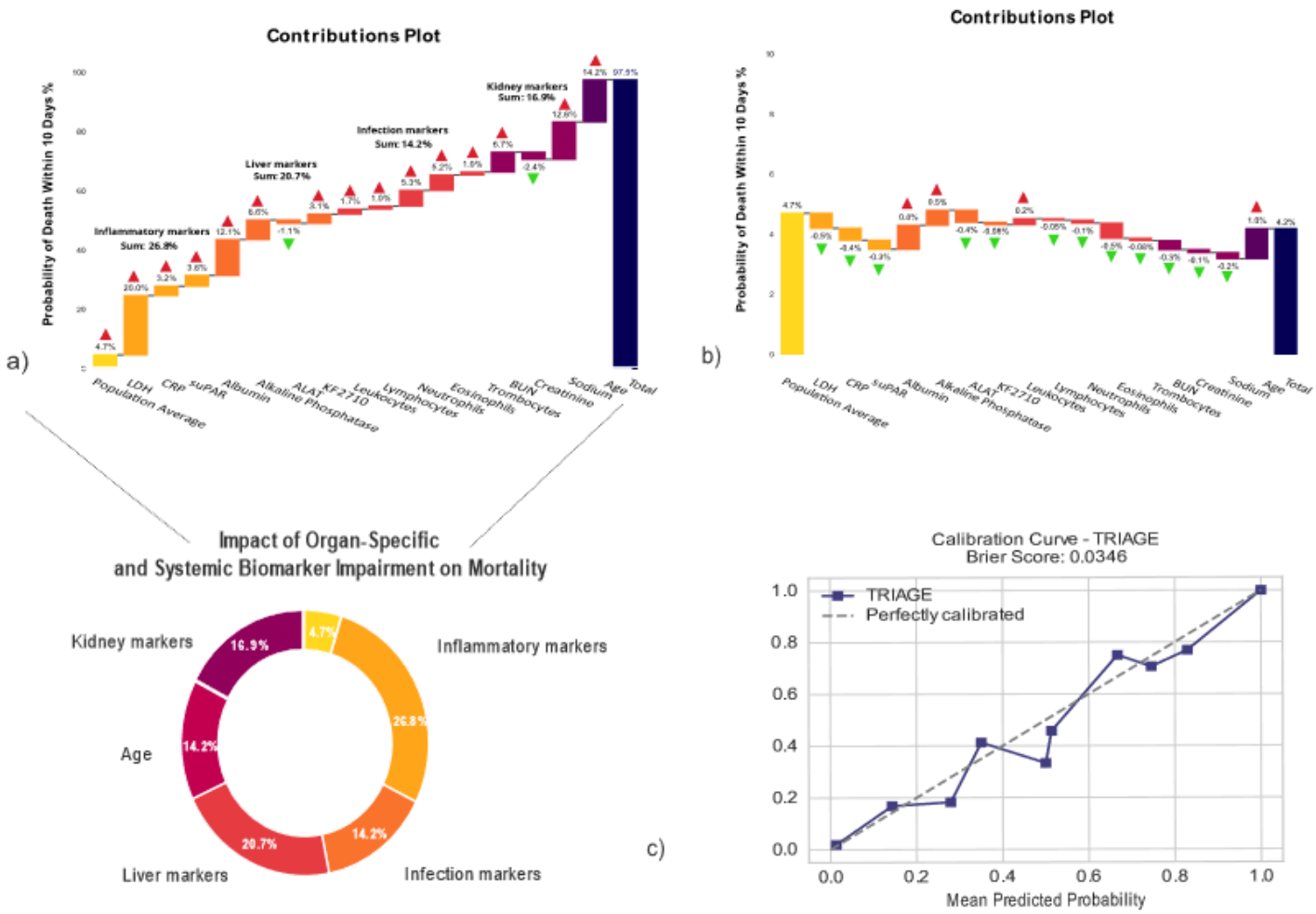

Figure 3

The Light Gradient Boosting Machine (LIGHTBM) model's individualized predictions of 10-day mortality risk for two specific patient cases in the prospective TRIAGE cohort.

Supplementary Files

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

- TripodChecklistPredictionModelDevelopmentandValidationWordnatComm.pdf
- Supplementaryfigures.docx