Disease-course adapting machine learning prognostication models in critically ill elderly COVID-19 patients: a multi-centre cohort study with external validation

Christian Jung, Behrooz Mamandipoor, Jesper Fjølner, Raphael Bruno, Bernhard Wernly, Antonio Artigas, Bernardo Bollen Pinto, Joerg C. Schefold, Georg Wolff, Malte Kelm, Michael Beil, Sigal Sviri, Peter Vernon van Heerden, Wojciech Szczeklik, Miroslaw Czuczwar, Muhammed Elhadi, Michael Joannidis, Sandra Oeyen, Tilemachos Zafeiridis, Brian Marsh, Finn H. Andersen, Rui Moreno, Maurizio Cecconi, Susannah Leaver, Dylan W. De Lange, Bertrand Guidet, Hans Flaatten, Venet Osmani

Submitted to: JMIR Medical Informatics
on: August 16, 2021

Table of Contents

Original Manuscript **5**
Supplementary Files **45**
Figures **46**
Figure 1 ..... **47**
Figure 2 ..... **48**
Figure 3 ..... **49**
Figure 4 ..... **50**
Figure 5 ..... **51**
Multimedia Appendixes **52**
Multimedia Appendix 1 ..... **53**
Multimedia Appendix 2 ..... **53**
Multimedia Appendix 3 ..... **53**
Multimedia Appendix 4 ..... **53**
Multimedia Appendix 5 ..... **53**
Multimedia Appendix 6 ..... **53**
Multimedia Appendix 7 ..... **53**
Multimedia Appendix 8 ..... **53**
Multimedia Appendix 9 ..... **53**
Multimedia Appendix 10 ..... **53**

Disease-course adapting machine learning prognostication models in critically ill elderly COVID-19 patients: a multi-centre cohort study with external validation

Christian Jung[1] MD, PhD; Behrooz Mamandipoor[2] BSc; Jesper Fjølner[3] MD; Raphael Bruno[1] MD; Bernhard Wernly[4] MD, PhD; Antonio Artigas[5] MD; Bernardo Bollen Pinto[6]; Joerg C. Schefold[7]; Georg Wolff[1] MD; Malte Kelm[1] MD; Michael Beil[8]; Sigal Sviri[8]; Peter Vernon van Heerden[9]; Wojciech Szczeklik[10]; Miroslaw Czuczwar[11]; Muhammed Elhadi[12]; Michael Joannidis[13]; Sandra Oeyen[14]; Tilemachos Zafeiridis[15]; Brian Marsh[16]; Finn H. Andersen[17]; Rui Moreno[18]; Maurizio Cecconi[19]; Susannah Leaver[20]; Dylan W. De Lange[21]; Bertrand Guidet[22]; Hans Flaatten[23]; Venet Osmani[2]

[1]University Hospital Duesseldorf Duesseldorf DE
[2]Fondazione Bruno Kessler Research Institute, Trento, Italy Trento IT
[3]Department of Intensive Care, Aarhus University Hospital, Aarhus, Denmark Aarhus DK
[4]Department of Anaesthesiology, Paracelsus Medical University, Salzburg, Austria Salzburg AT
[5]Department of Intensive Care Medicine, CIBER Enfermedades Respiratorias, Corporacion Sanitaria Universitaria Parc Tauli, Autonomous University of Barcelona, Sabadell, Spain Sabadell ES
[6]Department of Acute Medicine, Geneva University Hospitals, Geneva, Switzerland Geneva CH
[7]Department of Intensive Care Medicine, Inselspital, Universitätsspital, University of Bern, Bern, Switzerland Bern CH
[8]Department of Medical Intensive Care, Hadassah University Medical Center, Jerusalem, Israel Jerusalem IL
[9]Dept. of Anesthesia, Intensive Care and Pain Medicine Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Israel Jerusalem IL
[10]Center for Intensive Care and Perioperative Medicine, Jagiellonian University Medical College, Krakow, Poland Krakow PL
[11]2nd Department of Anesthesiology and Intensive Care, Medical University of Lublin, Staszica 16, 20-081, Lublin, Poland Lublin PL
[12]Faculty of Medicine, University of Tripoli, Tripoli, Libya Tripoli LY
[13]Division of Intensive Care and Emergency Medicine, Department of Internal Medicine, Medical University Innsbruck, Innsbruck, Austria Innsbruck AT
[14]Department of Intensive Care 1K12IC Ghent University Hospital, Ghent, Belgium Ghent BE
[15]Intensive Care Unit General Hospital of Larissa, Larissa, Greece Larissa GR
[16]Mater Misericordiae University Hospital, Dublin, Ireland; Dublin IE
[17]Dep. Of Anaesthesia and Intensive Care, Ålesund Hospital, Ålesund, Norway. Dep. of Circulation and medical imaging, Norwegian university of Science and Technology, Trondheim, Norway Alesund NO
[18]Unidade de Cuidados Intensivos Neurocríticos e Trauma. Hospital de São José, Centro Hospitalar Universitário de Lisboa Central, Faculdade de Ciências Médicas de Lisboa, Nova Médical School, Lisbon, Portugal Lisbon PT
[19]Department of Anaesthesia IRCCS Instituto Clínico Humanitas, Humanitas University, Milan, Italy Milan IT
[20]General Intensive care, St George´s University Hospitals NHS Foundation trust, London, United Kingdom London GB
[21]Department of Intensive Care Medicine, University Medical Center, University Utrecht, the Netherlands Utrecht BE
[22]Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Equipe: épidémiologie hospitalière qualité et organisation des soins, F-75012, Paris, France. Assistance Publique - Hôpitaux de Paris Paris FR
[23]Department of Clinical Medicine, University of Bergen, Department of Anaestesia and Intensive Care, Haukeland University Hospital , Bergen, Norway Bergen NO

Corresponding Author:
Christian Jung MD, PhD
University Hospital Duesseldorf
Moorenstraße 5
Duesseldorf
DE

Abstract

Background: The SARS-CoV-2 coronavirus disease (COVID-19) pandemic is challenging health care systems globally. The disease disproportionately affects the elderly population, both in terms of disease severity and mortality risk.

Objective: This study aimed to evaluate machine-learning based prognostication models for critically ill elderly COVID-19 patients, which dynamically incorporate multifaceted clinical information on the evolution of the disease.

Methods: Patient data was obtained from 151 ICUs from 26 countries (COVIP study). In total, 1,432 elderly (aged 70 years and above) COVID-19 positive patients admitted to an intensive care unit. Different models based on the Sequential Organ Failure Assessment (SOFA), Logistic Regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were derived as baseline models that included admission variables only. Then, we included clinical events and time-to-event as additional variables to derive the final models using the same algorithms and compared their performance with the baseline group. Furthermore, we derived baseline and final models on an EU patient cohort and externally validated them on a non-EU cohort that included Asian, African and Americas patients.

Results: Final models that incorporated clinical events and time-to-event provided superior performance with AUC of 0.81 (95% CI 0.804-0.811), with respect to both, the baseline models that used admission variables only, and conventional ICU prediction model (SOFA-score, p<0.001).

Conclusions: Integrating important clinical events and time-to-event information led to superior 30-day mortality prediction accuracy compared to models based on the admission information and conventional ICU prediction models. The present study shows that machine-learning models provide may support complex decision-making in critically ill elderly COVID-19 patients. Clinical Trial: NCT04321265

(JMIR Preprints 16/08/2021:32949)
DOI: https://doi.org/10.2196/preprints.32949

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Disease-course adapting machine learning prognostication models in critically ill elderly COVID-19 patients: a multi-centre cohort study with external validation

Christian Jung (1), Behrooz Mamandipoor (2), Jesper Fjølner (3), Raphael Romano Bruno (1), Bernhard Wernly (4), Antonio Artigas (5), Bernardo Bollen Pinto (6), Joerg C. Schefold (7), Georg Wolff (1), Malte Kelm (1), Michael Beil (8), Sigal Sviri (8), Peter Vernon van Heerden (9), Wojciech Szczeklik (10), Miroslaw Czuczwar (11), Muhammed Elhadi (12), Michael Joannidis (13), Sandra Oeyen (14), Tilemachos Zafeiridis (15), Brian Marsh (16), Finn H. Andersen (17), Rui Moreno (18), Maurizio Cecconi (19) Susannah Leaver (20), Dylan W. De Lange (21), Bertrand Guidet (22), Hans Flaatten (23), Venet Osmani (2)

on behalf of the COVIP study group

Affiliations

1. Heinrich-Heine-University Duesseldorf, Medical Faculty, Department of Cardiology, Pulmonology and Vascular Medicine, Duesseldorf, Germany
2. Fondazione Bruno Kessler Research Institute, Trento, Italy;
3. Department of Intensive Care, Aarhus University Hospital, Aarhus, Denmark;
4. Department of Anaesthesiology, Paracelsus Medical University, Salzburg, Austria;
5. Department of Intensive Care Medicine, CIBER Enfermedades Respiratorias, Corporacion Sanitaria Universitaria Parc Tauli, Autonomous University of Barcelona, Sabadell, Spain;
6. Department of Acute Medicine, Geneva University Hospitals, Geneva, Switzerland;
7. Department of Intensive Care Medicine, Inselspital, Universitätsspital, University of Bern, Bern, Switzerland;
8. Department of Medical Intensive Care, Hadassah University Medical Center, Jerusalem, Israel;
9. Dept. of Anesthesia, Intensive Care and Pain Medicine Hadassah Medical Center and

Faculty of Medicine, Hebrew University of Jerusalem, Israel;
10. Center for Intensive Care and Perioperative Medicine, Jagiellonian University Medical College, Krakow, Poland;
11. 2nd Department of Anesthesiology and Intensive Care, Medical University of Lublin, Staszica 16, 20-081, Lublin, Poland;
12. Faculty of Medicine, University of Tripoli, Tripoli, Libya;
13. Division of Intensive Care and Emergency Medicine, Department of Internal Medicine, Medical University Innsbruck, Innsbruck, Austria;
14. Department of Intensive Care 1K12IC Ghent University Hospital, Ghent, Belgium;
15. Intensive Care Unit General Hospital of Larissa, Larissa, Greece;
16. Mater Misericordiae University Hospital, Dublin, Ireland;
17. Dep. Of Anaesthesia and Intensive Care, Ålesund Hospital, Ålesund, Norway. Dep. of Circulation and medical imaging, Norwegian university of Science and Technology, Trondheim, Norway;
18. Unidade de Cuidados Intensivos Neurocríticos e Trauma. Hospital de São José, Centro Hospitalar Universitário de Lisboa Central, Faculdade de Ciências Médicas de Lisboa, Nova Médical School, Lisbon, Portugal;
19. Department of Anaesthesia IRCCS Instituto Clínico Humanitas, Humanitas University, Milan, Italy
20. General Intensive care, St George's University Hospitals NHS Foundation trust, London, United Kingdom;
21. Department of Intensive Care Medicine, University Medical Center, University Utrecht, the Netherlands;
22. Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Equipe: épidémiologie hospitalière qualité et organisation des soins, F-75012, Paris, France. Assistance Publique - Hôpitaux de Paris, Hôpital Saint-Antoine, service de réanimation médicale, Paris, F-75012, France;
23. Department of Clinical Medicine, University of Bergen, Department of Anaestesia and Intensive Care, Haukeland University Hospital, Bergen, Norway;

Corresponding author:

Christian Jung, M.D. PhD

Division of Cardiology, Pulmonology, and Vascular Medicine

Heinrich-Heine-University Duesseldorf

Moorenstraße 5, 40225 Duesseldorf, Germany

Phone: +49 211 81 18567

Fax: +49 211 81 015 18567

Email: christian.jung@med.uni-duesseldorf.de

Word count: 2786 (excluding methods)

References: 43

Trial registration number

NCT04321265

Abstract:

Background:

The SARS-CoV-2 coronavirus disease (COVID-19) pandemic is challenging health care systems globally. The disease disproportionately affects the elderly population, both in terms of disease severity and mortality risk.

Objectives:

This study aimed to evaluate machine-learning based prognostication models for critically ill elderly COVID-19 patients, which dynamically incorporated multifaceted clinical information on the evolution of the disease.

Methods:

This multi-centre cohort study obtained patient data from 151 ICUs from 26 countries (COVIP study). Different models based on the Sequential Organ Failure Assessment (SOFA), Logistic Regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were derived as baseline models that included admission variables only. We subsequently included clinical events and time-to-event as additional variables to derive the final models using the same algorithms and compared their performance with the baseline group. Furthermore, we derived baseline and final models on a European patient cohort and externally validated them on a non-European cohort that included Asian, African and American patients.

Results:

In total, 1,432 elderly (≥70 years) COVID-19 positive patients were admitted to an intensive care unit. Of these 809 (56.5%) patients survived up to 30 days after admission. The average length of stay was 21.6 (±18.2) days. Final models that incorporated clinical events and time-to-event provided superior performance with AUC of 0.81 (95% CI 0.804-0.811), with respect to both, the baseline models that used admission variables only and conventional ICU prediction models (SOFA-score, p<.001). The average precision increased from 0.65 (95% CI

0.650-0.655) to 0.77 (95% CI 0.759-0.770).

Conclusions:

Integrating important clinical events and time-to-event information led to a superior accuracy of 30-day mortality prediction compared with models based on the admission information and conventional ICU prediction models. The present study shows that machine-learning models provide additional information and may support complex decision-making in critically ill elderly COVID-19 patients.

Trial registration:

NCT04321265

Key words:
Machine-based learning; outcome prediction; Covid-19

MANUSCRIPT

Introduction

The SARS-CoV-2 coronavirus (COVID-19) pandemic is continuing to challenge health care systems globally [1]. The disease disproportionately affects the elderly population, both in terms of disease severity and mortality risk [2]. In many countries, intensive care unit (ICU) capacity was increased during the pandemic to meet demand. In addition, novel treatment modalities were introduced [3]. A key challenge in clinical outcome prediction in a dynamic disease is that the response to a given treatment varies considerably from patient to patient – especially in the elderly population [4]. Baseline data alone are inadequate to predict prognosis with sufficient accuracy for an individual patient, as they cannot capture the dynamic nature of the underlying critical illness [5]. It is well established that various factors provide prognostic information that should be taken into consideration [6]. More elaborate methods are thus urgently needed for both sophisticated and concise risk stratification of severely affected individual ICU patients [7]. Biomarkers, frailty, and severity scores are validated in elderly critically ill patients [8-11]. However, all of these have important limitations as they do not reflect the dynamics of the underlying disease pathophysiology and as a result have limited prognostic power. Ultimately it remains up to the physician to integrate all baseline data, the changing course of the disease and subjective experience into a clinical decision [12]. However, physicians do not assess dynamically evolving processes perfectly, as they are influenced by numerous factors, including fatigue and other human factors, resulting in less objective and reproducible decision making[13]. This aspect is especially relevant for new diseases, such as COVID-19, where physician experience is lacking.

Therefore, a supportive prognostication model that can integrate baseline data with complex, dynamic processes in an objective manner is necessary. Machine learning (ML) algorithms could be used to address this need as some have successfully been evaluated in

clinical settings such as in cardiovascular, intensive care [14]: Wernly *et al*. retrospectively analysed the arterial blood gas analyses from septic intensive care patients from a multi-centre eICU database as well as from a single centre MIMIC-III dataset to predict 96 hours mortality [9].

Izquierdo *et al*. combined classical epidemiological methods, natural language processing, and machine learning to examine the electronic health records of 10,504 patients with COVID-19. According to their analysis, the combination of easily obtainable clinical variables such as age, fever, and tachypnoea predicted which patients would require ICU admission [15]. The observational study by Bolourani *et al*. had a similar aim. They used clinical and laboratory data commonly collected in the emergency department in order to train and validate three predictive models (two based on XGBoost and one that used logistic regression) using cross-hospital validation. The XGBoost model had the highest mean accuracy to predict 48-hour respiratory failure[16]. Another study by Aktar *et al*. used machine learning to distinguish between healthy people and those with COVID-19 and subsequently to predict COVID-19 severity. They used decision tree, random forest, variants of gradient boosting machine, support vector machine, k-nearest neighbour, and deep learning methods for blood samples. The developed analytical methods evidenced accurate and precise scores >90% for disease severity prediction [17]. To avoid locally aggregating raw clinical data across multiple institutions, Vaid *et al*. evaluated a federated learning machine learning technique using electronic health records from 5 hospitals. In brief, they used a logistic regression with L1 regularisation/least absolute shrinkage and selection operator and multilayer perceptron models that were trained by using local data at each study site. The federated models outperformed the local models with regards to their accuracy in predicting the mortality in hospitalised patients with COVID-19 within 7 days [18]. In a smaller study, Domínguez-Olmedo et al. selected 32 predictor laboratory features in 1823 confirmed patients with COVID-19 for

an extreme gradient boosting algorithm. Similar to the other studies, using laboratory parameters resulted in an excellent outcome prediction [19]. Subudhi *et al.* used ensemble-based machine learning models to identify CRP, LDH, and O2 saturation as the most important factors for ICU-admission, and eGFR <60 ml/min/1.73 m2, and neutrophil and lymphocyte percentages for mortality [20].

A recent systematic review by Syeda *et al*. identified more than 400 articles that investigated the role of machine learning in the field of Covid-19 [21]. For example, Pan *et al*. studied 123 ICU patients and identified eight important risk factors with high recognition ability using an eXtreme Gradient Boosting (XGBoost) model [22]. A similar approach was used by Kim *et al*., who established an XGBoost model in 4787 patients admitted to a hospital due to COVID-19 [23]. Furthermore, Burian *et al*. estimated the need for intensive care treatment in 65 patients with confirmed COVID-19 infection[24], while Shahsikumar *et al*. investigated the performance of an algorithm to predict the need for mechanical ventilation on 402 patients with Covid-19, using cohorts with a wide age range (48 to 74 years)[25].

Very old intensive care patients are the most vulnerable intensive care subgroup [26]. However, to date, there are no studies investigating the role of ML models in this specific subgroup exclusively. To address this lack of evidence, this study evaluates whether ML models can reliably improve mortality prognostication in critically ill elderly patients with COVID-19 – based on clinical baseline information, biomarkers, accumulating events, and time-to-event information during the disease course.

Methods

Study design

This was a retrospective analysis that included data from 1,432 patients in a prospective multi-centre study. The primary outcome was 30-day mortality. We also used the 3-month outcome to ensure consistency of the primary outcome and allay concerns of censoring bias [27]. We derived two groups of models: baseline and final. Baseline models are derived using admission variables only, while the final model group incorporates clinical events such as catecholamine therapy, renal replacement therapy, non-invasive ventilation, invasive ventilation, prone position, and tracheostomy, in addition to the baseline variables. We evaluated both model groups using stratified 3-fold cross validation, to mitigate the variability of a single derivation-validation random split. Furthermore, we derived baseline and final models on an E.U. patient cohort and externally validated them on a non-EU cohort that included Asian, African and American patients.

Clinical data sources and study population

Patient data were obtained from 151 ICUs from 26 independent countries, including European ICUs and from ICUs in Asia, Africa, and America as part of the multinational COVIP trial (NCT04321265). This study was in line with the European Union General Data Privacy Regulation (GDPR) directive. As in previous successful studies [6, 26, 28], national coordinators recruited the intensive care units (ICUs), coordinated national and local ethical permissions, and supervised patient recruitment at the national level. In the VIP studies ethical approval was obligatory for study participation. The eCRF and database were hosted on a secure server in Aarhus University, Denmark. Data from 1,432 elderly (aged 70 years and above) COVID-19 positive patients admitted to a participating ICU between February 4 and May 26, 2020 were recorded. The study protocol is available from the COVIP study website [29] . Patients were followed up until hospital discharge and survival at 3 months using telephone interviews.

Study data

Demographic data included age, gender, weight, and height, and BMI. Furthermore, information on admission characteristics prior to ICU hospitalisation, duration of hospital stay, day of symptom onset and co-morbidities were available. Pre-existing co-morbidities were recorded in the eCRF: Diabetes, Ischemic heart disease, renal insufficiency, arterial hypertension, pulmonary co-morbidity, and chronic heart failure.

During the ICU stay, data on bacterial co-infection was noted, in addition to SOFA subscores (respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems). Laboratory values included partial pressures of oxygen (PaO2, P) and the fraction of inspired oxygen (FiO2, F), including the P/F ratio. Six clinical events of interest (catecholamine therapy, renal replacement therapy, non-invasive and invasive ventilation, prone position, and tracheostomy) were recorded with the time when the event occurred.

Model derivation and validation

We derived models based on Extreme Gradient Boosting (XGBoost) [30], Random Forest (RF) [31], and Logistic Regression (LR)[32]. As the best performing model, XGBoost algorithm provides robust prediction results using a method where new models are added to correct the errors made by existing models. Models are added sequentially and the combination of many models in the XGBoost model accommodates nonlinearity between input variables [30]. Hyperparameter tuning was performed by an exhaustive grid search directed toward maximising the F1 score metric. Three-fold cross validation was performed inside each grid option, and the optimal hyperparameter set was chosen based on the model in the grid search with the highest F1 score. Hyperparameters of the final model of the XGBoost are listed in Multimedia Appendix 1. To generate confidence intervals for the baseline and the final models, 3-fold cross validation was performed with 20 times repetition with randomly generated seed. To compare the performance of the XGBoost model, we also derived and validated two more

predictive models based on Logistic Regression and Random Forest. The decision was driven by the fact that LR is typically considered a baseline algorithm, while RF has been previously used in other research work with COVID-19 data [33]. Both LR and RF were optimised by an exhaustive grid search, similarly to the XGBoost method.

To address noise and outliers in the data, we defined a clinically valid interval for each variable and the values out of the valid scope were considered as missing values. For all the models, the issue of missing values was addressed by removing variables with >90% of missing values. We then used median and zero to impute the missing data in the remaining continuous and categorical variables. All the analyses were carried out using open-source software based on Python 3.6.8 with scikit-learn version 0.23.2.

Experimental evaluation

Performance evaluation of the models was based on 3-fold, stratified cross-validation with 20 repetitions using the area under the receiver operating characteristic curve (area-under-the-curve (AUC), GA Step 3) as well as area under the precision-recall curve (PRC), also known as average precision (AP) [34].

The precision-recall curve shows the relationship between positive predictive value (precision) and sensitivity (recall), measuring the performance of the model in correctly predicting mortality in patients with a high probability of dying (Figure 2). It is typically more informative than the AUC in presence of imbalanced outcomes [34]. Additional performance metrics are detailed in the appendix (Multimedia Appendix 2-5), including Positive Predictive Value (PPV), Negative Predictive Value (NPV), F-1 score (the balance between PPV and sensitivity), Matthews correlation coefficient MCC (used to measure the quality of classification between our algorithms) and Brier score. Calibration quality was evaluated using Brier scores, where a lower score indicates a higher calibration quality, and we also present calibration plots (also known as reliability curves). The models were compared based on their AUC and PRC

performance metrics for both the baseline data as well as the final models incorporating clinical events.

Model Interpretation

We evaluated the ranking of variables that contributed toward the model description using SHAP scores. SHAP scores are a game-theoretic approach to model interpretability; they provide explanations of global model structures based upon combinations of several local explanations for each prediction [35]. To interpret and rank the significance of input variables toward the final prediction of the model, mean absolute SHAP values were calculated for each variable across all observations in both, the baseline model, and the final model based on XGBoost. We also plotted SHAP interaction values that capture contribution of pairwise interactions between unique features to model prediction. To improve interpretability, especially in terms of the impact of clinical events, we defined a clinically meaningful day interval (0-3 days, 3-5, 5-10 and 10-30 days) and added a variable for each clinical event based on when the clinical event occurred, for example 'Tracheostomy-10-30' indicating that a tracheostomy was performed within the 10-to-30-day period. This allowed us to evaluate not only the importance of clinical events, but also the time-to-event information. Naturally, these variables were only available in the final model.

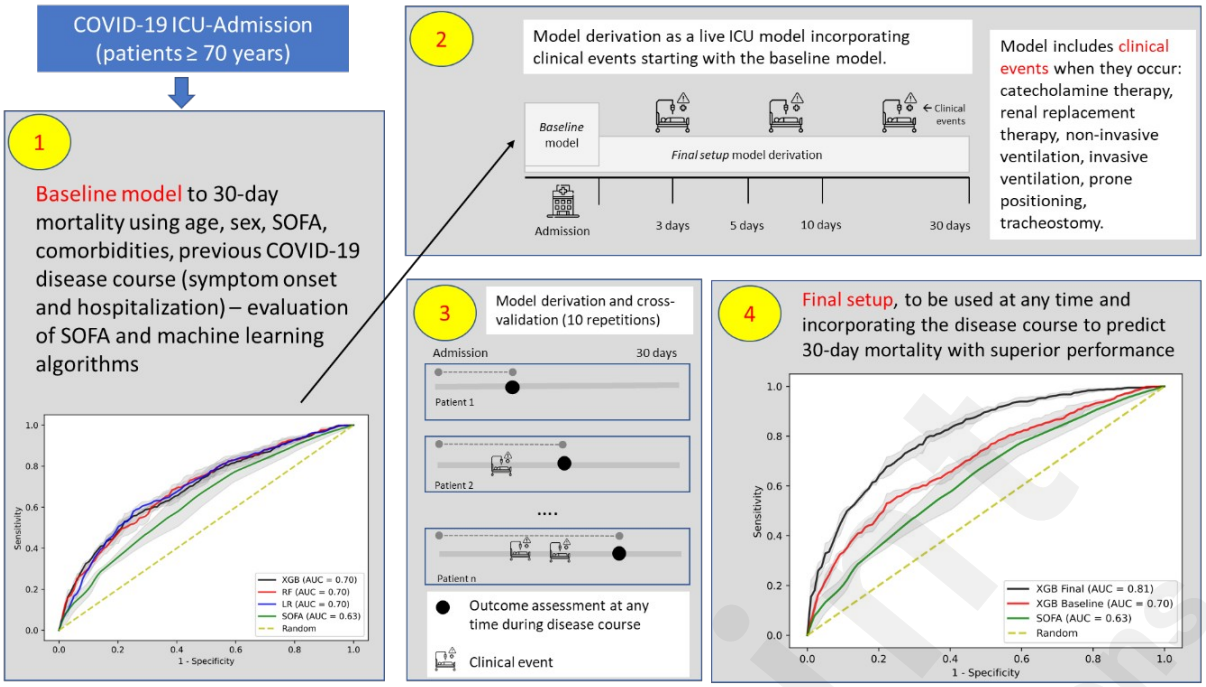


Figure 1: Graphical methods: Study design, from admission to derivation and validation of baseline setup (1); Derivation and validation of six models incorporating clinical events individually (2) (note: performance of individual models is shown in the Multimedia Appendix 2). Derivation of the final model including baseline variables as well as clinical events (3) and its evaluation in predicting 30-day outcomes as final setup (4).

Results

Study population

Out of 1,432 patients in the COVIP cohort, 809 (56.5%) patients survived up to 30 days after admission, with an average length of stay of 21.6 (±18.2) days. Patient baseline characteristics are given (Table 1) with distribution of mortality and length of stay detailed (Multimedia Appendix 6).

Table 1 Demographic characteristics, vital signs, and clinical events of patient cohorts

Variables	Alive	Dead (30 day)	*p*-Value
Sex (male %)	809 (72.5%)	623 (74.6%)	.18
Age (years)	75.0 ± 4.2	76.5 ± 4.8	<.01
Weight (kg)	81.3 ± 14.7	81.0 ± 14.8	.42
Height (cm)	169.7 ± 10.7	169.8 ± 10.5	.06

Body-mass index (kg/m^2)	28.5 ± 6.5	28.4 ± 5.7	.02
Hospital stay prior to ICU admission (days)	3.8 ± 5.7	3.5 ± 6.3	<.01
Symptoms prior to hospital admission (days)	7.2 ± 5.2	6.6 ± 4.5	.10
PaO_2 (mmHg)	87.3 ± 44.2	84.3 ± 57.5	<.01
FiO_2 (%)	62.3 ± 31.0	73.0 ± 24.0	<.01
SOFA score (points)	5.2 ± 3.0	6.7 ± 3.4	<.01
ICU-Treatment and Outcome			
Mechanical ventilation (number, %)	561 (69%)	510 (82%)	<.01
Vasopressors (number, %)	525 (65%)	515 (83%)	<.01
Prone positioning (number, %)	309 (38%)	279 (45%)	<.01
Tracheostomy (number, %)	227 (28%)	64 (10%)	<.01
Non-invasive ventilation (number, %)	169 (21%)	119 (19%)	.32
Renal replacement therapy (number, %)	121 (15%)	119 (19%)	<.01
ICU length of stay (days)	21.6 ± 18.2	10.6 ± 7.6	<.01
Pre-existing co-morbidities			
Diabetes mellitus (number, %)	268 (33%)	240 (38%)	<.01
Ischemic heart disease (number, %)	151 (19%)	152 (24%)	<.01
Chronic renal insufficiency (number, %)	91 (11%)	130 (21%)	<.01
Arterial Hypertension (number, %)	527 (65%)	431 (69%)	.03
Pulmonary disease (number, %)	175 (22%)	145 (23%)	.07
Chronic heart failure (number, %)	98 (12%)	103 (17%)	<.01

Model derivation and validation

We evaluated the performance of *baseline setup* risk prognostication that included baseline variables only (see graphical abstract, GA step 1); the *final* setup, which – in addition to baseline variables – included six key clinical events that occurred during the disease course and their time-to-event information: catecholamine therapy, renal replacement therapy, non-

invasive ventilation, invasive ventilation, prone positioning, and tracheostomy (GA step 2). The final set of selected variables is shown in Table 1. Furthermore, the baseline and the final setup were used to derive models on the E.U. cohort of patients that were then externally evaluated using a non-EU cohort composed of Asian, African and American patients.

Three risk prognostication models were derived from machine-learning based algorithms: Logistic Regression (LR) and – for comparison – Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) algorithms, as outlined in the Methods section. [30, 31].

The XGBoost algorithm achieved the numerically highest increase in discrimination performance from the *baseline setup* (AUC 0.70; 95% CI 0.692-0.701) to the *final setup* (AUC 0.81; 95% CI 0.804-0.811); average precision (AP) increased from 0.65 (95% CI 0.650-0.655) to 0.77 (95% CI 0.759-0.770, Figure 2).

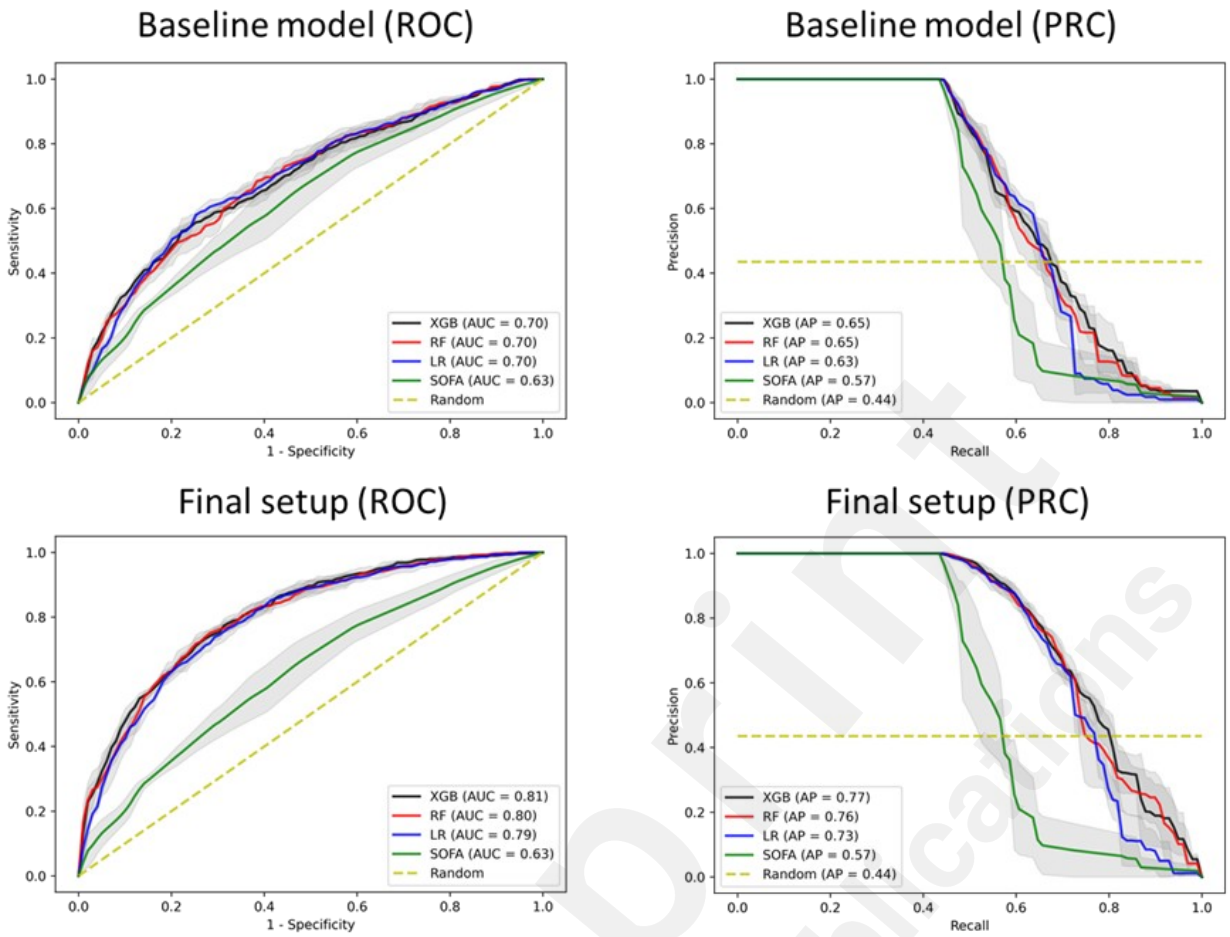


Figure 2: Performance of the baseline model (top) and improved performance in the final model (bottom) in response to clinical events detailing area under the curve of Receiver Operating Characteristics (ROC) and area under the Precision Recall Curve (PRC). PRC shows the relationship between positive predictive value (precision) and sensitivity (recall) at all thresholds.

LR (AUC 0.79; 95% CI 0.788-0.796) and RF (AUC 0.80; 95% CI 0.798-0.805) algorithms showed a similar performance in the *baseline model* and comparable improvement in the *final model*, comparable to XGBoost performance (GA step 4). The final XGBoost model provided superior performance compared to both the baseline model and SOFA score (both $p<.001$).

Experimental evaluation

In the external validation of the E.U. patient cohort, all three models achieved a similar performance in the baseline and the final setup with AUC of 0.82 and 0.86 respectively, when evaluated on predicting mortality of non-EU patients (Figure 3). One explanation for this

performance on the external validation cohort might be that the patients in the non-EU cohort tended to gravitate towards two opposing health states: either they were quite stable or very sick, making it easier for the model to discriminate between the two outcomes. To investigate this further, we plotted the distribution of the variable that had the highest impact on outcome prediction (FiO2), based on SHAP analysis (see Figure 5). As shown in Multimedia Appendix 7, the distribution for both outcomes are significantly skewed, towards the 21% for survivors and towards 100% for non-survivors.

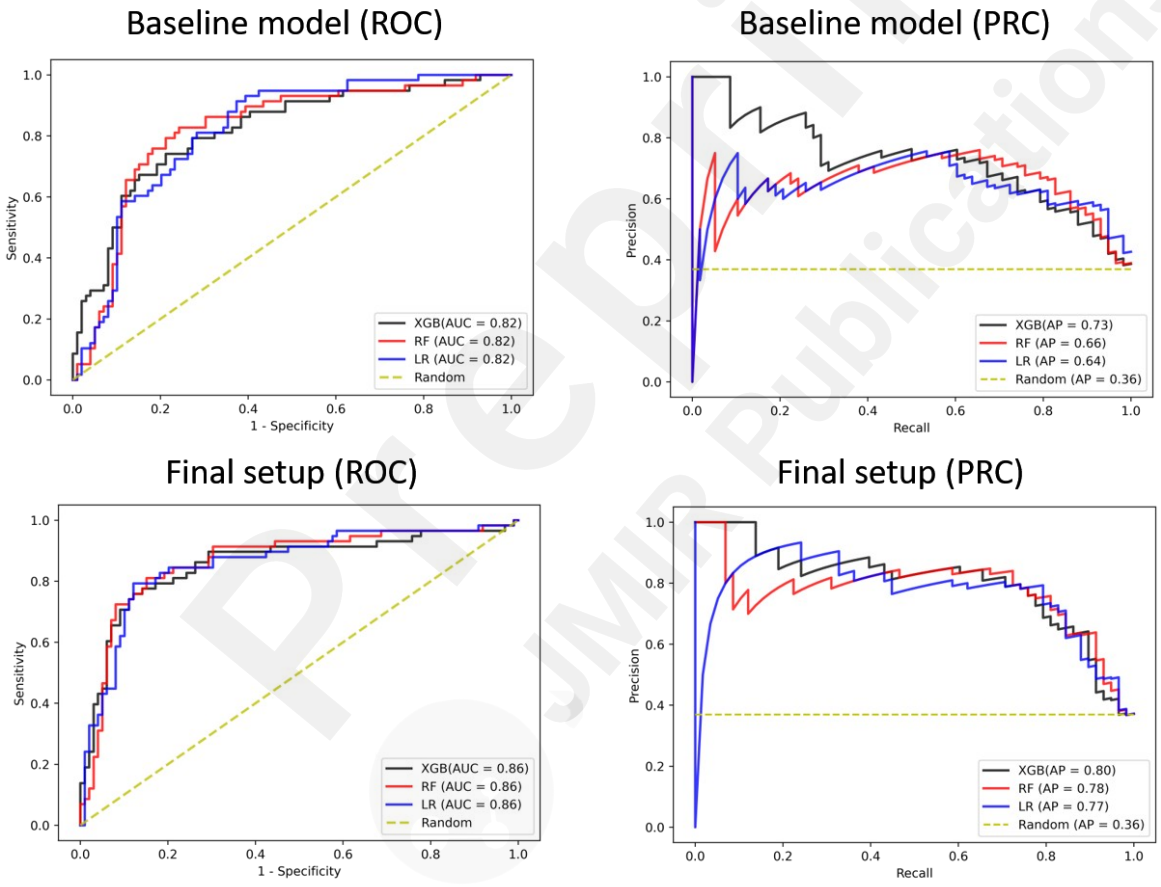


Figure 3: Performance of the final model derived using the E.U. patient cohort and externally validated on a non-EU patient cohort, comprising Asian, African and American patients. Model performance is measured using area under the curve of Receiver Operating Characteristics (ROC) and area under the Precision Recall Curve (PRC).

We also assessed calibration of each model to ensure that distribution of predicted outcomes matches distribution of observed outcomes in our patient cohort. Baseline and final

models were, in general, well calibrated (Figure 4), matching estimated risk of outcome with observed risk. The final setup for each algorithm was better calibrated (Brier score of 0.17) with respect to baseline setup (Brier score 0.22). Full details of Brier scores for each algorithm are detailed (Multimedia Appendix 1).

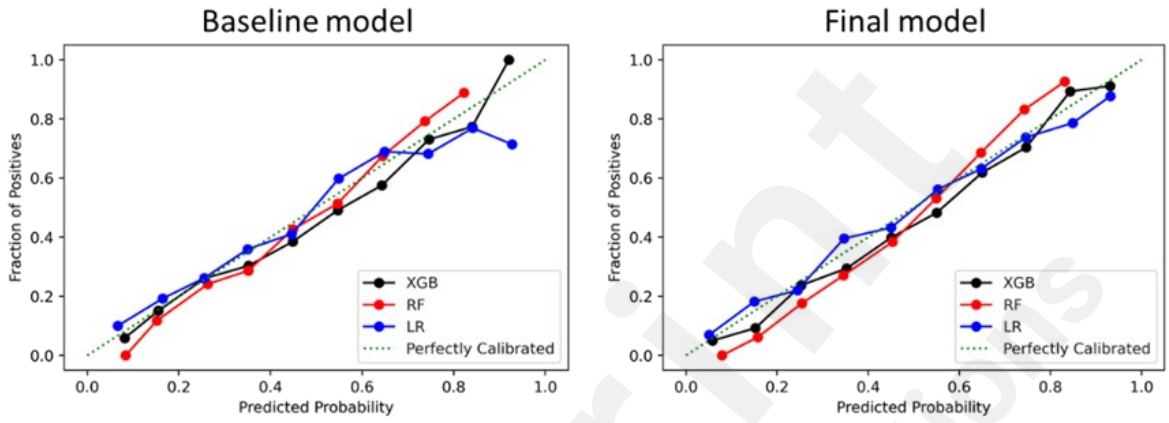


Figure 4: Calibration curves for each model and individual algorithms used to derive the model, XGBoost (XGB), Random Forest (RF), and Logistic Regression (LR).

Model Interpretation

The SHAP (Shapley Additive exPlanations) method was used to perform interpretability analysis, which explains model output by computing the contribution of each variable to the prediction. Among others, the SHAP method was applied on the best performing model (XGBoost), where the fraction of inspired oxygen (FiO2), age, and tracheostomy had the highest impact on outcome prediction (Figure 5 and Multimedia Appendix 7).

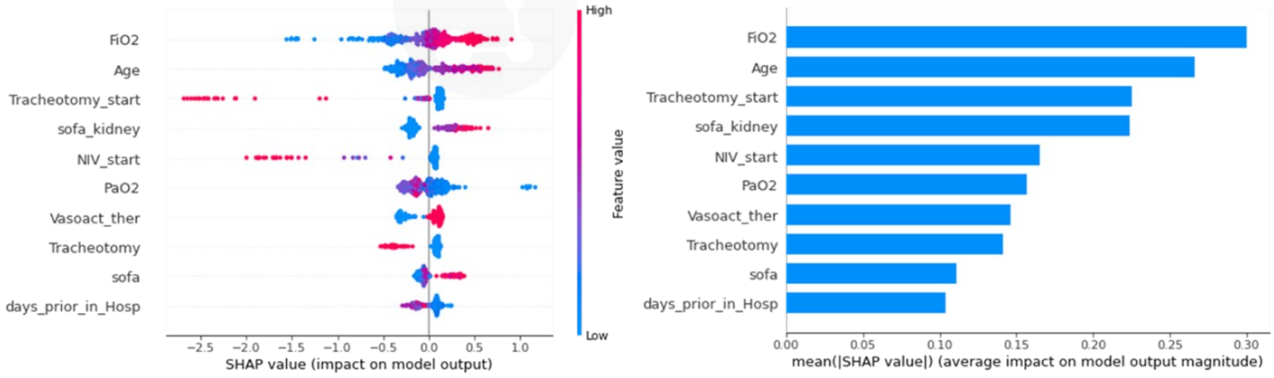


Figure 4: Ranking of input variables of the final setup derived from XGBoost algorithm, using the

SHAP method.

We also report model interpretability analysis for the RF and LR-based models in the appendix (Multimedia Appendix 8 and 9, respectively). The top three variables remained common between XGBoost and RF, while for LR, only tracheostomy appeared in the top three, while the two others were weight and BMI.

Discussion

The present study demonstrates that individual prognostication accuracy based on patient baseline characteristics can be considerably improved with ML algorithms that incorporate occurrence and time-to-event information of clinical events along the course of a disease, such as COVID-19, in elderly critically ill patients. These results align with many previous studies that investigated machine learning approaches in patients suffering from COVID-19. The major difference between this COVIP study and others published previously lies in its focus on the especially vulnerable subgroup of very old intensive care patients [21]. The second important difference is that the current approach includes the risk for clinical events such as tracheostomy.

Subudhi *et al*. compared 18 different machine learning algorithm's ability to predict the rate of admission and mortality of patients suffering from COVID-19 [20]. In their analysis, ensemble-based models were superior to other algorithms (including LR and XGB). Specific laboratory values and O2-saturation were the most important factors for ICU admission, while impaired kidney function and differential blood count best predicted mortality [20]. This study primarily used data from patients, of all ages, presenting to the emergency room.

Domínguez-Olmedo *et al*. used data from 1823 patients with confirmed COVID-19 and established the XGBoost model. Their model found lactate dehydrogenase activity, C-reactive protein level, neutrophil count, and urea level to be the most important values reaching an AUC of 0.93 (95% CI 0.89-0.98) for sensitivity, and 0.91 (95% CI 0.86-0.96) for specificity [19].

Pan *et al.* used the data from 123 patients with COVID-19 admitted to an ICU [22] and after constructing a XGBoost model, they identified eight factors (albumin level; creatinine;

eosinophil percentage; lactate dehydrogenase; lymphocyte percentage; neutrophil percentage; prothrombin time; total bilirubin) that were predictive for ICU mortality.

Vaid *et al*. utilised a different approach by using federated learning of electronic health records from five different hospitals [18]. In brief, their models provided robust predictive models without compromising patient privacy.

Other studies focused primarily on peripheral blood samples. Aktar *et al.* developed machine learning and deep learning algorithms to predict the disease severity [17]. Similarly, Kim *et al*. established an XGBoost model in 4787 hospital-admitted patients to predict their intensive care treatment requirements [23]. Their model was significantly superior to the established CURB-65(confusion, urea, respiratory rate, blood pressure) score.

Immediate clinical applications are conceivable, especially given the limited number of ICU beds available. Our models may be used in several ways: ML could be used before ICU admission to give objective support to complex allocation decisions. However, ML algorithms would mainly access data at presentation and few dynamic parameters, limiting predictive power. ML algorithms could also be used in the context of time-limited trials (TLT), which are common clinical practice in ICUs in some countries. This may be particularly helpful in patients in whom realistic therapeutic goals/outcomes are unclear at presentation. These patients could be admitted to the ICU under the premise of gaining more information about the patient and the initial response to treatment. This additional information could then be evaluated using ML algorithms [36] as already shown in septic patients [9]. The ideal temporal combination of a TLT and ML should be the subject of future, prospective studies [36, 37].

In terms of practical applications, ML algorithms provide a potential strategy to improve decision confidence and predictive power over time. They are applicable at various time points during the disease course, predicting outcomes in a continuous manner. This approach is especially applicable when considering that the model was well calibrated in estimating

outcomes. However, evaluation of the model with a diverse patient population would provide further evidence of its clinical applicability.

Clinical evaluations such as assessment of wakefulness, mobility, responsiveness, and independence are subjective and subject to interrater variability. Therefore, advances in digital technologies may support but not replace physicians' skills. ML can support physicians, especially in estimations on prognosis and achievement of therapy goals. Importantly, ethical problems become evident when ML is involved in matters of life and death [38], and it must be emphasised that ML should only support and aid medical decision-making. Our data show that dedicated modern algorithms can incrementally improve certainty during TLTs in elderly patients with COVID-19 disease and generalise well in an external patient cohort. These tools can enhance our ability to improve guidance of treatment and optimally allocate ICU resources. However, such a strategy can only be viewed as complementary to clinical judgment and individual treatment goals and part of a holistic patient assessment.

The present study has some methodological limitations in common with the other COVIP-studies [11, 26, 39-42]. COVIP did not contain a control group of younger COVID-19 patients for comparison or a comparable age cohort of patients who were not or could not be admitted to the ICU. In addition, the COVIP database does not include information on pre-ICU care and triage decisions. These treatment limitations might also affect the care of older ICU patients [43]. Furthermore, COVIP recruited patients in 26 countries, and thus the participating countries varied widely in their care structure, resulting in considerable heterogeneity in treatments given.

Conclusion

The present study demonstrates that, in the particularly vulnerable subgroup of very old intensive care patients suffering from COVID-19, individual prognostication accuracy based on patient baseline characteristics can be improved with ML algorithms. These algorithms capture

the dynamic course of the disease by including the occurrence and time-to-event information of clinical events and thus reflect both disease severity and the need for intensive care treatment.

List of abbreviations

AP	average precision
AUC	Area under the curve
COVID-19	SARS-CoV-2 coronavirus disease
ICU	Intensive care unit
LR	Logistic Regression
ML	Machine learning
RF	Random Forest
SOFA	Sequential Organ Failure Assessment
TLT	time-limited trials
XGBoost	Extreme Gradient Boosting

DECLARATIONS

Ethics approval and consent to participate

The primary competent ethics committee was the Ethics Committee of the University of Duesseldorf, Germany. Institutional research ethic board approval was obtained from each study site. This was a prerequisite for participation in the study. All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the local institutional and/or licensing committees. Informed consent was obtained from all subjects if not omitted by the ethics vote. The studies conducted were observational studies. No examinations (e.g. blood sampling) or tissue sampling took place.

Consent for publication

The manuscript does not contain any individual person's data in any form.

Availability of data and material

The anonymised data can be requested from the authors if required. The datasets analysed during the current study are not publicly available due to the different local institutional and/or licensing committees but are available from the corresponding author on reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Funding

The support of the study in France by a grant from Fondation Assistance Publique-Hôpitaux de Paris pour la recherche is greatly appreciated. In Norway, the study was supported by a grant from the Health Region West. In addition, EOSCsecretariat.eu provided support and has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-05-

2018-2019, grant Agreement number 831644. This work was supported by the Forschungskommission of the Medical Faculty of the Heinrich-Heine-University Düsseldorf, No. 2018-32 to G.W. and No. 2020-21 to RRB for a Clinician Scientist Track.

Author Contributions

B.W., B.M., J.F., RRB, V.O. and C.J. analysed the data and wrote the first draft of the manuscript. A.A. and BBP and JCS and G.W. contributed to statistical analysis and improved the paper. M.K. and M.B. and S.S. and PVH and W.S. and MC and M.E. and M.J. and SO and T.Z. and B.M. and FHA and R.M. and MC and S.L. and DL and B.G. and H.F. gave guidance and improved the paper. All authors read and approved the final manuscript.

Authors' information

N/A

Acknowledgements

N/A

Financial Disclosure statement

No (industry) sponsorship has been received for this investigator-initiated study.

Figures

Figure 1

Graphical methods: Study design, from admission to derivation and validation of baseline setup (1); Derivation and validation of six models incorporating clinical events individually (2) (note: performance of individual models is shown in the Multimedia Appendix 2). Derivation of the final model including baseline variables as well as clinical events (3) and its evaluation in predicting 30-day outcomes as final setup (4).

Figure 2

Performance of the baseline model (top) and improved performance in the final model (bottom) in response to clinical events detailing area under the curve of Receiver Operating Characteristics (ROC) and area under the Precision Recall Curve (PRC). PRC shows the relationship between positive predictive value (precision) and sensitivity (recall) at all thresholds.

Figure 3

Performance of the final model derived using the E.U. patient cohort and externally validated on a non-EU patient cohort, comprising Asian, African and Americas patients. Model performance is measured using area under the curve of Receiver Operating Characteristics (ROC) and area under the Precision Recall Curve (PRC).

Figure 4

Calibration curves for each model and individual algorithms used to derive the model, XGBoost (XGB), Random Forest (RF) and Logistic Regression (LR).

Figure 5

Ranking of input variables of the final setup derived from XGBoost algorithm, using the SHAP method.

Multimedia Appendix

Multimedia Appendix 1
Table showing hyperparameters for each algorithm found through exhaustive grid search.

Multimedia Appendix 2
Table showing the performance of the baseline model in terms of various performance metrics and 95% CI. (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration)

Multimedia Appendix 3
Table showing the performance of the final model in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration with lower values indicating better calibration)

Multimedia Appendix 4
Table showing performance of the baseline model derived using the E.U. patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration)

Multimedia Appendix 5
Table showing the performance of the baseline model derived using the E.U. patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration)

Multimedia Appendix 6
*Figure about the d*istribution of deaths over time and length of ICU stay

Multimedia Appendix 7
Figure about the distribution of FiO2 for both outcomes of survivors (left) and non-survivors (right) patients. FiO2 was chosen as it was the variable that had the highest impact on the performance prediction, based on SHAP analysis.

Multimedia Appendix 8
Figure showing the ranking of input variables of the final setup derived using RF-based model.

Multimedia Appendix 9
Figure showing the ranking of input variables of the final setup derived using LR-based model

References:

1. European Society of Intensive Care, M., A. Global Sepsis, and M. Society of Critical Care, *Reducing the global burden of sepsis: a positive legacy for the COVID-19 pandemic?* Intensive Care Med, 2021. **47**(7): p. 733-736.
2. Maltese, G., et al., *Frailty and COVID-19: A Systematic Scoping Review.* J Clin Med, 2020. **9**(7).
3. Alkuzweny, M., A. Raj, and S. Mehta, *Preparing for a COVID-19 surge: ICUs.* EClinicalMedicine, 2020. **25**: p. 100502.
4. Chopra, V., et al., *Variation in COVID-19 characteristics, treatment and outcomes in Michigan: an observational study in 32 hospitals.* BMJ Open, 2021. **11**(7): p. e044921.
5. Mudatsir, M., et al., *Predictors of COVID-19 severity: a systematic review and meta-analysis.* F1000Research, 2020. **9**: p. 1107-1107.
6. Flaatten, H., et al., *The impact of frailty on ICU and 30-day mortality and the level of care in very elderly patients (>/= 80 years).* Intensive Care Med, 2017. **43**(12): p. 1820-1828.
7. Zhao, Z., et al., *Prediction model and risk scores of ICU admission and mortality in COVID-19.* PLoS One, 2020. **15**(7): p. e0236618.
8. Jung, C., et al., *Frailty as a Prognostic Indicator in Intensive Care.* Dtsch Arztebl Int, 2020. **117**(40): p. 668-673.
9. Wernly, B., et al., *Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation.* Int J Med Inform, 2021. **145**: p. 104312.
10. Masyuk, M., et al., *Prognostic relevance of serum lactate kinetics in critically ill patients.* Intensive Care Med, 2019. **45**(1): p. 55-61.
11. Bruno, R.R., et al., *Lactate is associated with mortality in very old intensive care patients suffering from COVID-19: results from an international observational study of 2860 patients.* Ann Intensive Care, 2021. **11**(1): p. 128.
12. Leeuwenberg, A.M. and E. Schuit, *Prediction models for COVID-19 clinical decision making.* Lancet Digit Health, 2020. **2**(10): p. e496-e497.
13. Perrotta, F., et al., *COVID-19 and the elderly: insights into pathogenesis and clinical decision-making.* Aging Clinical and Experimental Research, 2020. **32**(8): p. 1599-1608.
14. Quer, G., et al., *Machine Learning and the Future of Cardiovascular Care.* Journal of the American College of Cardiology, 2021. **77**(3): p. 300-313.
15. Izquierdo, J.L., et al., *Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing.* J Med Internet Res, 2020. **22**(10): p. e21801.
16. Bolourani, S., et al., *A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation.* J Med Internet Res, 2021. **23**(2): p. e24246.
17. Aktar, S., et al., *Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development.* JMIR Med Inform, 2021. **9**(4): p. e25884.
18. Vaid, A., et al., *Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach.* JMIR Med Inform, 2021. **9**(1): p. e24207.
19. Dominguez-Olmedo, J.L., et al., *Machine Learning Applied to Clinical Laboratory Data in Spain for COVID-19 Outcome Prediction: Model Development and Validation.* J Med Internet Res, 2021. **23**(4): p. e26211.
20. Subudhi, S., et al., *Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19.* NPJ Digit Med, 2021. **4**(1): p. 87.

21. Syeda, H.B., et al., *Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review.* JMIR Med Inform, 2021. **9**(1): p. e23811.
22. Pan, P., et al., *Prognostic Assessment of COVID-19 in the Intensive Care Unit by Machine Learning Methods: Model Development and Validation.* J Med Internet Res, 2020. **22**(11): p. e23128.
23. Kim, H.J., et al., *An Easy-to-Use Machine Learning Model to Predict the Prognosis of Patients With COVID-19: Retrospective Cohort Study.* J Med Internet Res, 2020. **22**(11): p. e24225.
24. Burian, E., et al., *Intensive Care Risk Estimation in COVID-19 Pneumonia Based on Clinical and Imaging Parameters: Experiences from the Munich Cohort.* J Clin Med, 2020. **9**(5).
25. Shashikumar, S.P., et al., *Development and Prospective Validation of a Deep Learning Algorithm for Predicting Need for Mechanical Ventilation.* Chest, 2021. **159**(6): p. 2264-2273.
26. Jung, C., et al., *The impact of frailty on survival in elderly intensive care patients with COVID-19: the COVIP study.* Crit Care, 2021. **25**(1): p. 149.
27. Li, Y., et al., *Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar.* BMJ, 2020. **371**: p. m3919.
28. Guidet, B., et al., *The contribution of frailty, cognition, activity of daily life and comorbidities on outcome in acutely admitted patients over 80 years in European ICUs: the VIP2 study.* Intensive Care Med, 2020. **46**(1): p. 57-69.
29. COVIP. https://vipstudy.org/covip-study/. [cited 2021 10/11/2021].
30. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.
31. Ho, T.K., *Random decision forests*, in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*. 1995, IEEE Computer Society. p. 278.
32. McCullagh, P., & Nelder, J.A., *Generalized Linear Models (2nd ed.)*. 1983: Routledge.
33. Wynants, L., et al., *Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal.* BMJ, 2020. **369**: p. m1328.
34. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PLoS One, 2015. **10**(3): p. e0118432.
35. Lundberg, S. and S.-I. Lee, *A unified approach to interpreting model predictions.* arXiv preprint arXiv:1705.07874, 2017.
36. Vink, E.E., et al., *Time-limited trial of intensive care treatment: an overview of current literature.* Intensive Care Med, 2018. **44**(9): p. 1369-1377.
37. Shrime, M.G., et al., *Time-Limited Trials of Intensive Care for Critically Ill Patients With Cancer: How Long Is Long Enough?* JAMA Oncol, 2016. **2**(1): p. 76-83.
38. Beil, M., et al., *Ethical considerations about artificial intelligence for prognostication in intensive care.* Intensive Care Med Exp, 2019. **7**(1): p. 70.
39. Jung, C., et al., *Differences in mortality in critically ill elderly patients during the second COVID-19 surge in Europe.* Crit Care, 2021. **25**(1): p. 344.
40. Bruno, R.R., et al., *Early evaluation of organ failure using MELD-XI in critically ill elderly COVID-19 patients.* Clin Hemorheol Microcirc, 2021.
41. Jung, C., et al., *Inhibitors of the renin-angiotensin-aldosterone system and COVID-19 in critically ill elderly patients.* Eur Heart J Cardiovasc Pharmacother, 2021. **7**(1): p. 76-77.
42. Jung, C., et al., *Steroid use in elderly critically ill COVID-19 patients.* Eur Respir J, 2021.

43. Flaatten, H., et al., *The impact of end-of-life care on ICU outcome.* Intensive Care Med, 2021.

Multimedia Appendix

Multimedia Appendix 1:
Hyperparameters for each algorithm found through exhaustive grid search

XGBoost			**Random Forest**			**Logistic Regression**	
Parameter	Baseline Model	Final model	**Parameter**	Baseline Model	Final model	**Parameter**	Both models
eta	0.2	0.05	estimators	400	400	penalty	l2
max_depth	2	3	max_depth	9	8	solver	iblinear
min_child_weigh t	1	1	min_samples_split	9	8	class_weight	0:1,1:1.2
gamma	0.2	0.3	class_weight	0:1,1:1.25	0:1,1:1.3	C	1.0
colsample_bytree	0.7	0.4	criterion	gini	gini	-	-
scale_pos_weight	1.25	1.25	-	-	-	-	-

Multimedia Appendix 2:
Table showing the performance of the baseline model in terms of various performance metrics and 95% CI.

	AUC	AP	PPV	NPV	MCC	F1	Brier
LR	0.70 [0.696-0.701]	0.63 [0.627-0.633]	**0.65** [0.651-0.659]	0.67 [0.671-0.675]	**0.31** [0.305-0.318]	0.56 [0.560-0.568]	0.22 [0.216-0.218]
RF	0.70 [0.692-0.701]	0.65 [0.638-0.650]	0.61 [0.601-0.616]	0.67 [0.668-0.676]	0.28 [0.265-0.285]	0.57 [0.563-0.575]	0.22 [0.216-0.218]
XGB	0.70 [0.692-0.701]	**0.65** [0.650-0.655]	0.61 [0.600-0.606]	**0.69** [0.688-0.692]	0.29 [0.287-0.297]	**0.60** [0.594-0.601]	0.22 [0.217-0.219]

(AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration).

Multimedia Appendix 3:
Table showing the performance of the final model in terms of various performance metrics and 95% CI

	AUC	AP	PPV	NPV	MCC	F1	Brier
LR	0.79 [0.788-0.796]	0.73 [0.721-0.731]	**0.69** [0.685-0.694]	0.75 [0.748-0.755]	0.44 [0.433-0.444]	0.68 [0.675-0.682]	0.18 [0.182-0.186]
RF	0.80 [0.798-0.805]	0.76 [0.748-0.762]	0.68 [0.676-0.681]	0.77 [0.766-0.778]	0.44 [0.446-0.457]	0.69 [0.690-0.699]	0.18 [0.182-0.185]
XGB	**0.81** [0.804-0.811]	**0.77** [0.759-0.770]	0.67 [0.668-0.671]	**0.78** [0.771-0.783]	**0.45** [0.443-0455]	**0.70** [0693-0.703]	**0.17** [0.176-0.179]

(AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration with lower values indicating better calibration).

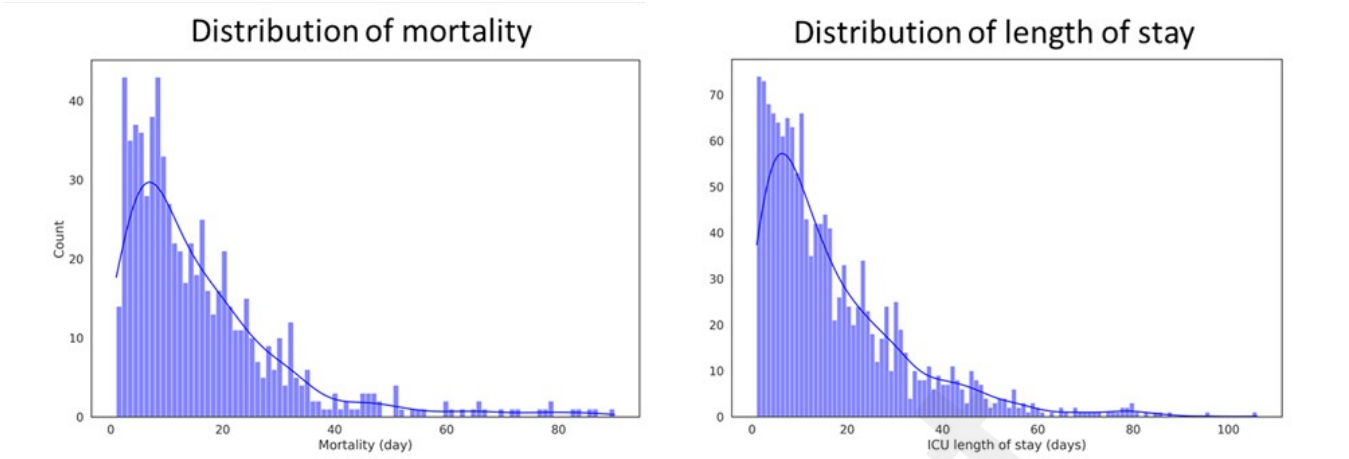
Multimedia Appendix 4:
Table showing performance of the baseline model derived using the E.U. patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CI

	AUC	AP	PPV	NPV	MCC	F1	Brier
LR	82	64	64	80	44	65	**18**
RF	82	66	**71**	**83**	**54**	**71**	19
XGB	82	**73**	69	**83**	51	70	**18**

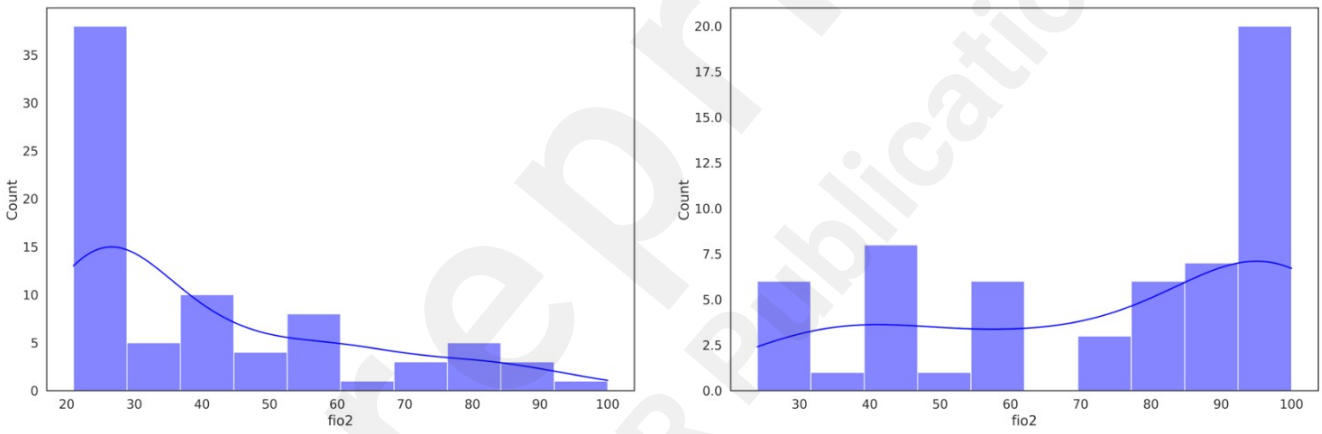
(AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration).

Multimedia Appendix 5:
Table showing the performance of the baseline model derived using the E.U. patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration).

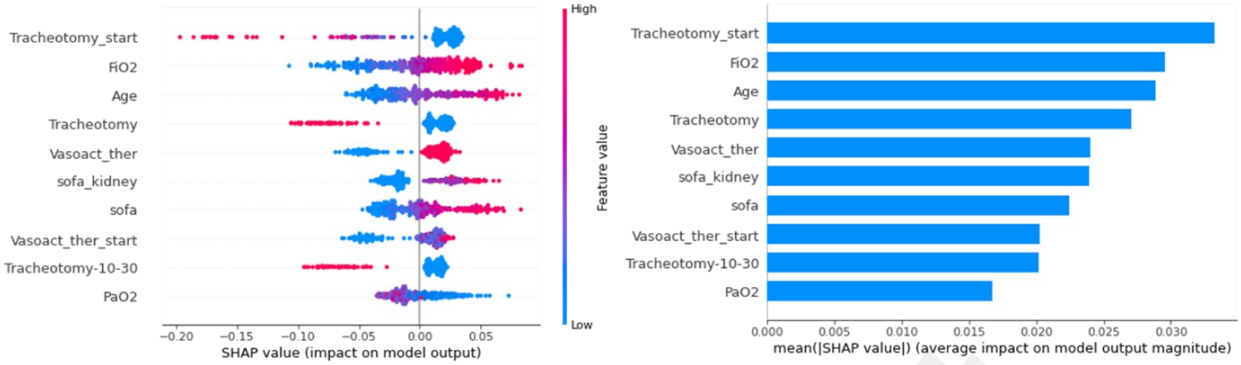
	AUC	AP	PPV	NPV	MCC	F1	Brier
LR	86	77	**79**	**87**	**66**	**78**	**15**
RF	86	78	76	86	62	76	16
XGB	86	**80**	76	**87**	62	76	**15**



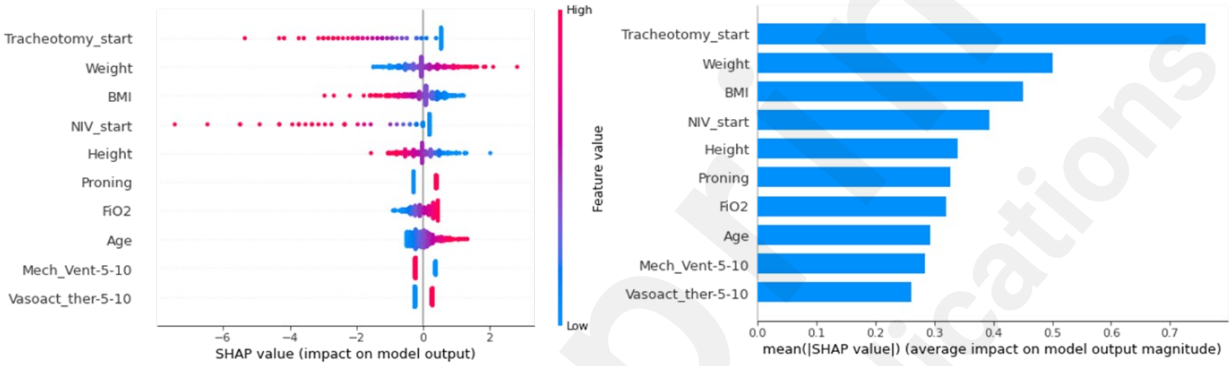
Multimedia Appendix 6: Figure about the distribution of deaths over time and length of ICU stay



Multimedia Appendix 7: Figure about the distribution of FiO2 for both outcomes of survivors (left) and non-survivors (right) patients. FiO2 was chosen as it was the variable that had the highest impact on the performance prediction, based on SHAP analysis.



Multimedia Appendix 8: Ranking of input variables of the final setup derived using RF-based model

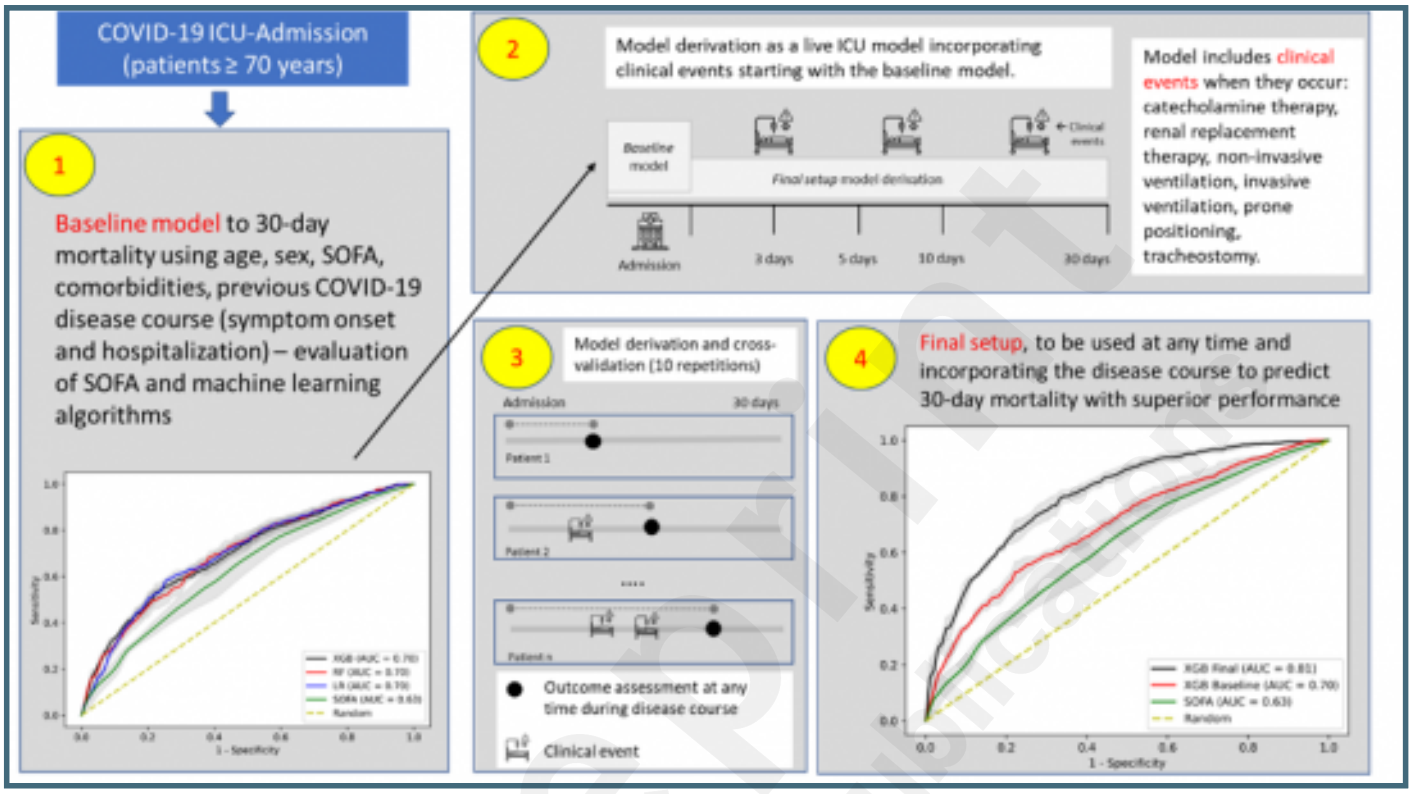


Multimedia Appendix 9: Ranking of input variables of the final setup derived using LR-based model

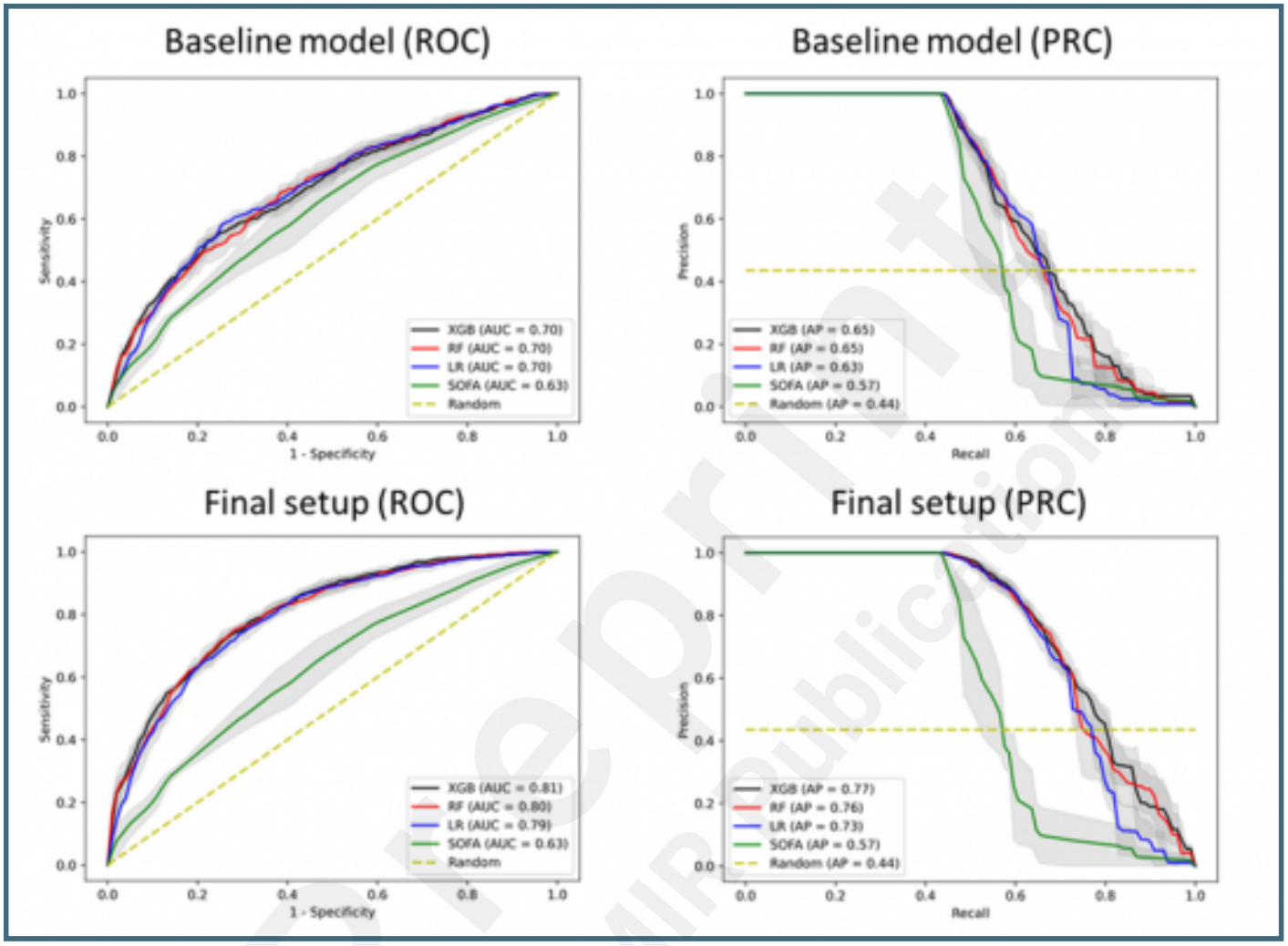
Supplementary Files

Figures

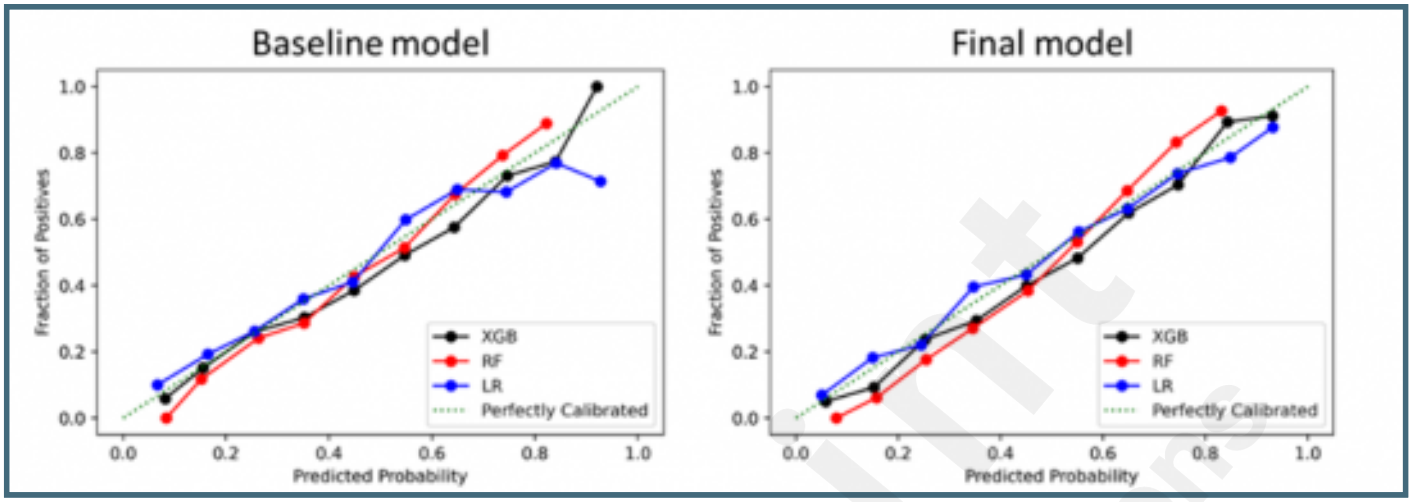
Graphical methods: Study design, from admission to derivation and validation of baseline setup (1); Derivation and validation of six models incorporating clinical events individually (2) (note: performance of individual models is shown in the Multimedia Appendix 2). Derivation of the final model including baseline variables as well as clinical events (3) and its evaluation in predicting 30-day outcomes as final setup (4).



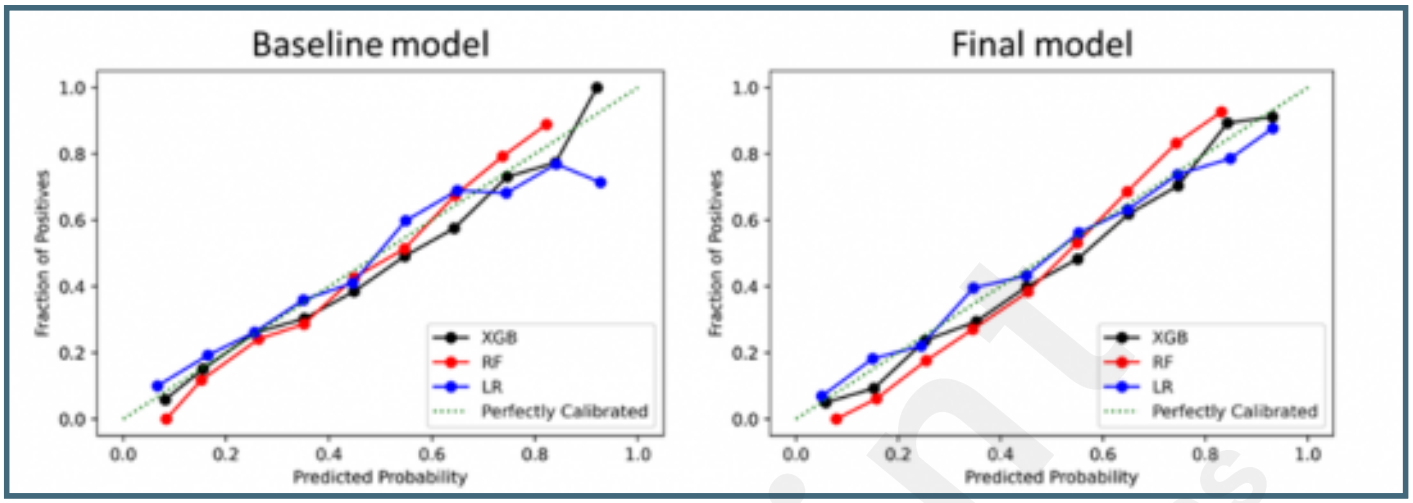
Performance of the baseline model (top) and improved performance in the final model (bottom) in response to clinical events detailing area under the curve of Receiver Operating Characteristics (ROC) and area under the Precision Recall Curve (PRC). PRC shows the relationship between positive predictive value (precision) and sensitivity (recall) at all thresholds.



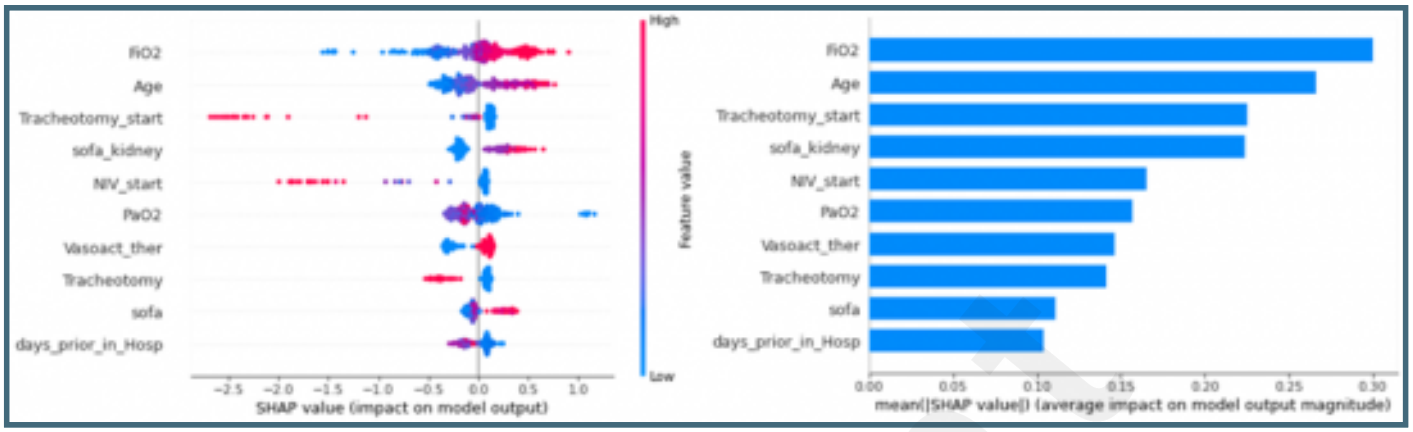
Performance of the final model derived using the E.U. patient cohort and externally validated on a non-EU patient cohort, comprising Asian, African and Americas patients. Model performance is measured using area under the curve of Receiver Operating Characteristics (ROC) and area under the Precision Recall Curve (PRC).



Calibration curves for each model and individual algorithms used to derive the model, XGBoost (XGB), Random Forest (RF), and Logistic Regression (LR).



Ranking of input variables of the final setup derived from XGBoost algorithm, using the SHAP method.



Multimedia Appendixes

Table showing hyperparameters for each algorithm found through exhaustive grid search.
URL: http://asset.jmir.pub/assets/8af4d2c3368322136c4fa00d91c2b546.docx

Table showing the performance of the baseline model in terms of various performance metrics and 95% CI. (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration).
URL: http://asset.jmir.pub/assets/f5be32f5858184d1ecfcc6d854329414.docx

Table showing the performance of the final model in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration with lower values indicating better calibration).
URL: http://asset.jmir.pub/assets/2edf38f75d19f202b675cda04499d814.docx

Table showing performance of the baseline model derived using the E.U. patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration).
URL: http://asset.jmir.pub/assets/b19418f78fd1f80bea9bb24960be4c7b.docx

Table showing the performance of the baseline model derived using the E.U. patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CI (AUC - area under the ROC curve; AP - average precision; PPV – positive predictive value; NPV – negative predictive value; MCC – Matthews correlation coefficient; F1 - harmonic mean of precision and recall and Brier score measuring quality of calibration, with lower values indicating better calibration).
URL: http://asset.jmir.pub/assets/919ac6c4045fa19d7097fa0be5c2dcb3.docx

Distribution of deaths over time and length of ICU stay.
URL: http://asset.jmir.pub/assets/43f0361644bb7d54e3fb236f8b1d7af9.docx

Distribution of FiO2 for both outcomes of survivors (left) and non-survivors (right) patients. FiO2 was chosen as it was the variable that had the highest impact on the performance prediction, based on SHAP analysis.
URL: http://asset.jmir.pub/assets/af158d2c710c56cc7c9c89273fc123e0.docx

Ranking of input variables of the final setup derived using RF-based model.
URL: http://asset.jmir.pub/assets/70e473472e6746c2ac294be7a616e012.docx

Ranking of input variables of the final setup derived using LR-based model.
URL: http://asset.jmir.pub/assets/b8c50762e6df38f2cd5fcd13b89a2d0b.docx

List of COVIP-collaborators.
URL: http://asset.jmir.pub/assets/5f5e8070b057fd0dc68b4d67ce602fb9.docx