Original Paper

Disease-Course Adapting Machine Learning Prognostication Models in Elderly Patients Critically Ill With COVID-19: Multicenter Cohort Study With External Validation

Christian Jung[1], MD, PhD[‡]; Behrooz Mamandipoor[2], BSc; Jesper Fjølner[3], MD; Raphael Romano 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], MHA; Peter V van Heerden[9], MD, PhD; Wojciech Szczeklik[10]; Miroslaw Czuczwar[11]; Muhammed Elhadi[12]; Michael Joannidis[13]; Sandra Oeyen[14]; Tilemachos Zafeiridis[15†]; Brian Marsh[16]; Finn H Andersen[17,18]; Rui Moreno[19,20], MD, PhD; Maurizio Cecconi[21]; Susannah Leaver[22]; Dylan W De Lange[23]; Bertrand Guidet[24,25]; Hans Flaatten[26,27]; Venet Osmani[2], PhD

[1]Division of Cardiology, Pulmonology and Vascular Medicine, Medical Faculty, Heinrich-Heine-University Duesseldorf, University Hospital Duesseldorf, 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, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

[9]Department of Anesthesia, Intensive Care and Pain Medicine, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

[10]Center for Intensive Care and Perioperative Medicine, Jagiellonian University Medical College, Krakow, Poland

[11]Second Department of Anesthesiology and Intensive Care, Medical University of Lublin, Lublin, Poland

[12]Faculty of Medicine, University of Tripoli, Tripoli, Libyan Arab Jamahiriya

[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]Department of Anaesthesia and Intensive Care, Ålesund Hospital, Alesund, Norway

[18]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway

[19]Hospital de São José, Centro Hospitalar Universitário de Lisboa Central, Lisbon, Portugal

[20]Faculdade de Ciências Médicas de Lisboa, Nova Medical School - Faculdade de Ciências Médicas, Universidade da Beira Interior, Lisbon, Portugal

[21]Department of Anaesthesia, IRCCS Instituto Clínico Humanitas, Humanitas University, Milan, Italy

[22]General Intensive Care, St George's University Hospitals, NHS Foundation Trust, London, United Kingdom

[23]Department of Intensive Care Medicine, University Medical Center, Utrecht University, Utrecht, Belgium

[24]Épidémiologie Hospitalière Qualité et Organisation des Soins, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMR_S 1136, Paris, France

[25]Service de Réanimation Médicale, Assistance Publique-Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France

[26]Department of Clinical Medicine, University of Bergen, Bergen, Norway

[27]Department of Anesthesia and Intensive Care, Haukeland University Hospital, Bergen, Norway

[†]deceased[‡]COVIP Study Group

Corresponding Author:
Christian Jung, MD, PhD
Division of Cardiology, Pulmonology and Vascular Medicine
Medical Faculty, Heinrich-Heine-University Duesseldorf
University Hospital Duesseldorf
Moorenstraße 5

Duesseldorf, 40225
Germany
Phone: 49 2118118800
Fax: 49 211 81 19520
Email: christian.jung@med.uni-duesseldorf.de

Abstract

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

Objective: The aim of this study was to evaluate machine learning–based prognostication models for critically ill elderly COVID-19 patients, which dynamically incorporated multifaceted clinical information on evolution of the disease.

Methods: This multicenter cohort study (COVIP study) obtained patient data from 151 intensive care units (ICUs) from 26 countries. Different models based on the Sequential Organ Failure Assessment (SOFA) score, logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB) 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 that of the baseline group. Furthermore, we derived baseline and final models on a European patient cohort, which were externally validated on a non-European cohort that included Asian, African, and US patients.

Results: In total, 1432 elderly (≥70 years old) COVID-19–positive patients admitted to an ICU were included for analysis. Of these, 809 (56.49%) patients survived up to 30 days after admission. The average length of stay was 21.6 (SD 18.2) days. Final models that incorporated clinical events and time-to-event information provided superior performance (area under the receiver operating characteristic curve 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. This study shows that machine-learning models provide additional information and may support complex decision-making in critically ill elderly COVID-19 patients.

Trial Registration: ClinicalTrials.gov NCT04321265; https://clinicaltrials.gov/ct2/show/NCT04321265

(JMIR Med Inform 2022;10(3):e32949) doi: 10.2196/32949

KEYWORDS

machine-based learning; outcome prediction; COVID-19; pandemic; machine learning; prediction models; clinical informatics; patient data; elderly population

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 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 [9] retrospectively analyzed arterial blood gas data from septic intensive care patients from a multicenter electronic ICU database as well as from a single-center MIMIC-III (Medical Information Mart for Intensive Care) data set to predict 96-hour mortality.

Izquierdo et al [15] combined classical epidemiological methods, natural language processing, and ML 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 tachypnea predicted which patients would require ICU admission [15]. The observational study by Bolourani et al [16] had a similar aim. They used clinical and laboratory data commonly collected in the emergency department to train and validate three predictive models (two based on extreme gradient boosting [XGB] and one that used logistic regression [LR]) with cross-hospital validation. The XGB model had the highest mean accuracy to predict 48-hour respiratory failure [16]. Aktar et al [17] used ML to distinguish between healthy people and those with COVID-19 and subsequently to predict COVID-19 severity. They used decision tree, random forest (RF), variants of gradient boosting machine, support vector machine, k-nearest neighbor, and deep learning methods for blood samples. The developed analytical methods evidenced accuracy and precision scores >90% for disease severity prediction. To avoid locally aggregating raw clinical data across multiple institutions, Vaid et al [18] evaluated a federated learning ML technique using electronic health records from 5 hospitals. In brief, they used LR with L1 regularization/least absolute shrinkage and selection operator, and multilayer perceptron models that were trained using local data at each study site. The federated models outperformed the local models with regard to their accuracy in predicting the mortality in hospitalized patients with COVID-19 within 7 days. In a smaller study, Domínguez-Olmedo et al [19] selected 32 predictor laboratory features in 1823 patients with confirmed COVID-19 for an XGB algorithm. Similar to the other studies, using laboratory parameters resulted in excellent outcome prediction. Subudhi et al [20] used ensemble-based ML models to identify C-reactive protein, lactate dehydrogenase, and oxygen saturation as the most important factors for predicting ICU admission, with estimated glomerular filtration rate <60 mL/min/1.73 m^2, and neutrophil and lymphocyte percentages as the important factors for predicting mortality.

A recent systematic review by Syeda et al [21] identified more than 400 articles that investigated the role of ML in the field of COVID-19. For example, Pan et al [22] studied 123 ICU patients and identified eight important risk factors with high recognition ability using an XGB model. A similar approach was used by Kim et al [23], who established an XGB model in 4787 patients admitted to a hospital due to COVID-19. Furthermore, Burian et al [24] estimated the need for intensive care treatment in 65 patients with confirmed COVID-19, and Shahsikumar et al [25] 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).

Patients who are very old represent 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, the aim of this study was to evaluate 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 1432 patients in a prospective multicenter 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 models. Baseline models were derived using admission variables only, whereas the final model group incorporated clinical events such as catecholamine therapy, renal replacement therapy, noninvasive 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 EU patient cohort and externally validated them on a non-EU cohort that included Asian, African, and US patients.

Clinical Data Sources and Study Population

Patient data were obtained from 151 ICUs across 26 independent countries, including European ICUs, and from ICUs in Asia, Africa, and the United States as part of the multinational COVIP trial (NCT04321265). This study was conducted in line with the European Union General Data Privacy Regulation directive. As in previous successful studies [6,26,28], national coordinators recruited the ICUs, coordinated national and local ethical permissions, and supervised patient recruitment at the national level. In the COVIP studies, ethical approval was obligatory for study participation. The electronic case report form (eCRF) and database were hosted on a secure server in Aarhus University, Denmark. Data from 1432 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.

Ethical Considerations

The primary competent ethics committee was the Ethics Committee of the University of Duesseldorf, Germany. Institutional research ethics 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 were all observational; no examinations (eg, blood sampling) or tissue sampling took place.

Study Data

Demographic data included age, gender, weight, height, and BMI. Furthermore, information on admission characteristics prior to ICU hospitalization, duration of hospital stay, day of symptom onset, and comorbidities were available. Preexisting

comorbidities were recorded in the eCRF: diabetes, ischemic heart disease, renal insufficiency, arterial hypertension, pulmonary comorbidity, and chronic heart failure.

During the ICU stay, data on bacterial coinfection were noted, in addition to Sequential Organ Failure Assessment (SOFA) subscores (respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems). Laboratory values included partial oxygen pressure and the fraction of inspired oxygen (FiO2), and their ratio. Six clinical events of interest (catecholamine therapy, renal replacement therapy, noninvasive and invasive ventilation, prone position, and tracheostomy) were recorded along with the time the event occurred.

Model Derivation and Validation

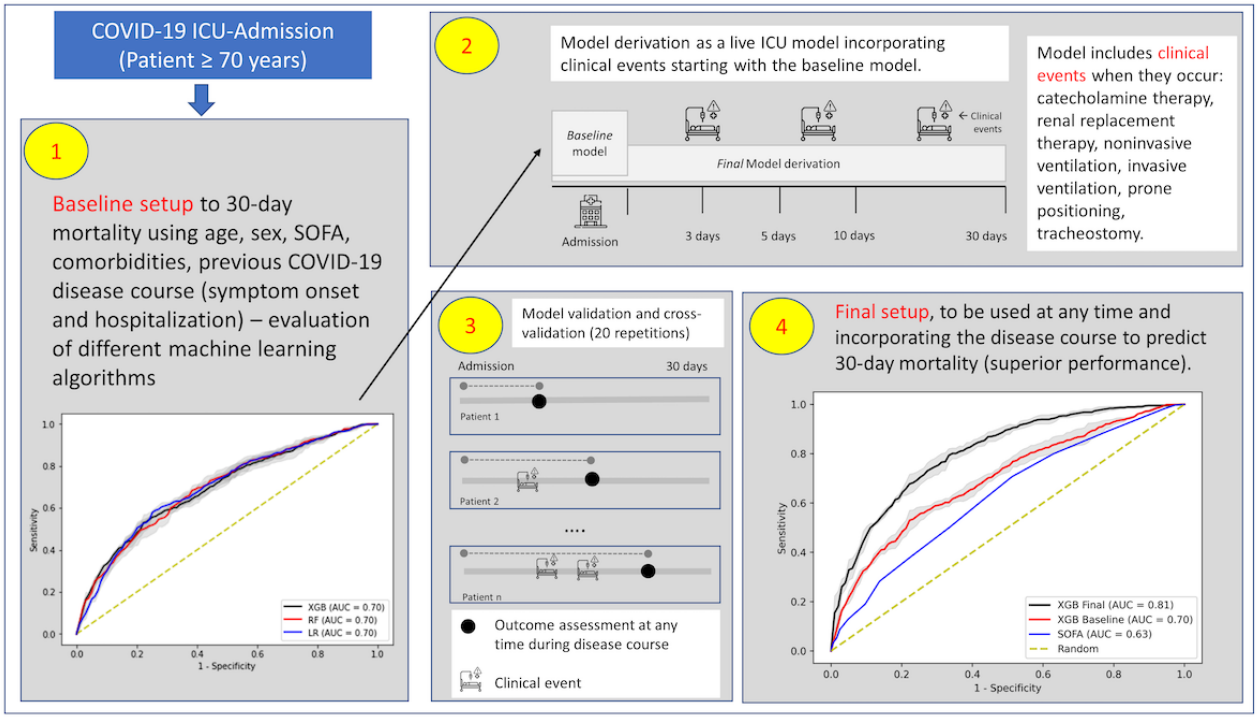
We derived models based on XGB [30], RF [31], and LR [32]. As the best-performing model, the XGB algorithm provides robust prediction results using a method where new models are added to correct the errors made by existing models. Models are added sequentially and the combination of many models in the XGB model accommodates nonlinearity between input variables [30]. Hyperparameter tuning was performed by an exhaustive grid search directed toward maximizing 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 XGB 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 a randomly generated seed. To compare the performance of the XGB model, we also derived and validated two more predictive models based on LR and RF. This decision was driven by the fact that LR is typically considered a baseline algorithm, and RF has been previously used in other research with COVID-19 data [33]. Both LR and RF were optimized by an exhaustive grid search, similar to the XGB 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 models, the issue of missing values was addressed by removing variables with >90% missing values. We then used the median and zero to impute the missing data in the remaining continuous and categorical variables, respectively. All 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 (AUC; see step 3 in Figure 1) as well as area under the precision-recall curve (PRC), also known as average precision [34].

Figure 1. Graphical methods. (1) Study design, from admission to derivation and validation of baseline setup. (2) Derivation and validation of six models incorporating clinical events individually.Performance of individual models is shown in Multimedia Appendix 2-5. (3) Derivation of the final model, including baseline variables as well as clinical events. (4) Evaluation of the final model in predicting 30-day outcomes. SOFA: Sequential Organ Failure Assessment; ICU: intensive care unit.



The PRC shows the relationship between the 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. The area under the PRC is typically more informative than the AUC in the presence of imbalanced outcomes [34]. Additional performance metrics are detailed in Multimedia Appendix 2-5, including the positive predictive value (PPV), negative predictive value, F1 score (the balance between PPV and sensitivity), Matthews correlation coefficient (used to measure the quality of classification between

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 shapely additive explanation (SHAP) scores. SHAP scores are a game-theoretic approach to model interpretability; they provide explanations of global model structures based on 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 SHA*P* values were calculated for each variable across all observations in both the baseline model and the final model based on XGB. We also plotted SHAP interaction values that capture the 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, 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" indicates that a tracheostomy was performed within the 10-30–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.

Results

Study Population

Out of the total 1432 patients in the COVIP cohort, 809 (56.49%) patients survived up to 30 days after admission, with an average length of stay of 21.6 (SD 18.2) days. Patient baseline characteristics are given in Table 1, with distribution of mortality and length of stay detailed in Multimedia Appendix 6.

Table 1. Demographic characteristics, vital signs, and clinical events of patient cohorts (N=1432).

Variables	Alive at 30 days (n=809)	Dead at 30 days (n=623)	*P* value
Sex (male), n (%)	587 (72.6%)	463 (74.6%)	.18
Age (years), mean (SD)	75.0 (4.2)	76.5 (4.8)	<.001
Weight (kg), mean (SD)	81.3 (14.7)	81.0 (14.8)	.42
Height (cm), mean (SD)	169.7 (10.7)	169.8 (10.5)	.06
BMI (kg/m²), mean (SD)	28.5 (6.5)	28.4 (5.7)	.02
Hospital stay prior to ICU[a] admission (days), mean (SD)	3.8 (5.7)	3.5 (6.3)	.002
Symptoms prior to hospital admission (days), mean (SD)	7.2 (5.2)	6.6 (4.5)	.10
PaO2[b] (mmHg), mean (SD)	87.3 (44.2)	84.3 (57.5)	.003
FiO2[c] (%), mean (SD)	62.3 (31.0)	73.0 (24.0)	<.001
SOFA[d] score (points), mean (SD)	5.2 (3.0)	6.7 (3.4)	<.001
ICU treatment and outcome			
Mechanical ventilation, n (%)	561 (69.3)	510 (81.9)	<.001
Vasopressors, n (%)	525 (64.9)	515 (82.7)	<.001
Prone positioning, n (%)	309 (38.2)	279 (44.8)	.10
Tracheostomy, n (%)	227 (28.1)	64 (10.3)	<.001
Noninvasive ventilation, n (%)	169 (20.9)	119 (19.1)	.32
Renal replacement therapy, n (%)	121 (15.0)	119 (19.1)	.01
Length of ICU stay (days), mean (SD)	21.6 (18.2)	10.6 (7.6)	<.001
Preexisting comorbidities, n (%)			
Diabetes mellitus	268 (33.1)	240 (38.5)	.01
Ischemic heart disease	151 (18.7)	152 (24.4)	.007
Chronic renal insufficiency	91 (11.2)	130 (20.9)	<.001
Arterial hypertension	527 (65.1)	431 (69.2)	.03
Pulmonary disease	175 (21.6)	145 (23.3)	.07
Chronic heart failure	98 (12.1)	103 (16.5)	.01

[a]ICU: intensive care unit.

[b]PaO2: partial oxygen pressure.

[c]FiO2: fraction of inspired oxygen.

[d]SOFA: Sequential Organ Failure Assessment.

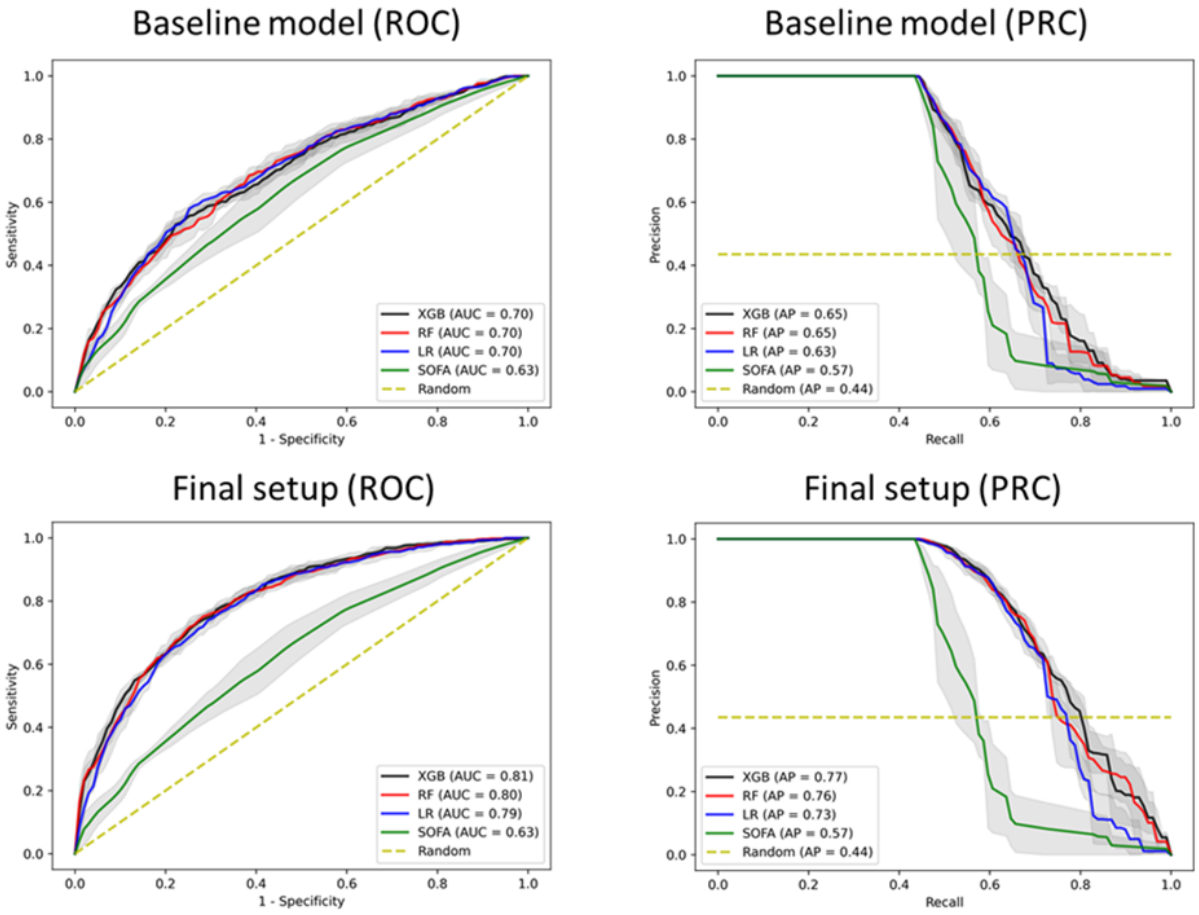
Model Derivation and Validation

We evaluated the performance of *baseline setup* risk prognostication that included baseline variables only (see step 1 in Figure 1) and 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, noninvasive ventilation, invasive ventilation, prone positioning, and tracheostomy (step 2 in Figure 1). 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 EU cohort of patients that were then externally evaluated using a non-EU cohort composed of Asian, African, and US patients.

Three risk prognostication models were derived from ML-based algorithms: LR and, for comparison, RF and XGB algorithms, as outlined in the Methods section [30,31].

The XGB 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 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 with respect to the area under the receiver operating characteristic (ROC) curve (AUC) and area under the precision-recall curve (PRC). The PRC shows the relationship between the positive predictive value (precision) and sensitivity (recall) at all thresholds. XGB: extreme gradient boosting; RF: random forest; LR: logistic regression; SOFA: Sequential Organ Failure Assessment.

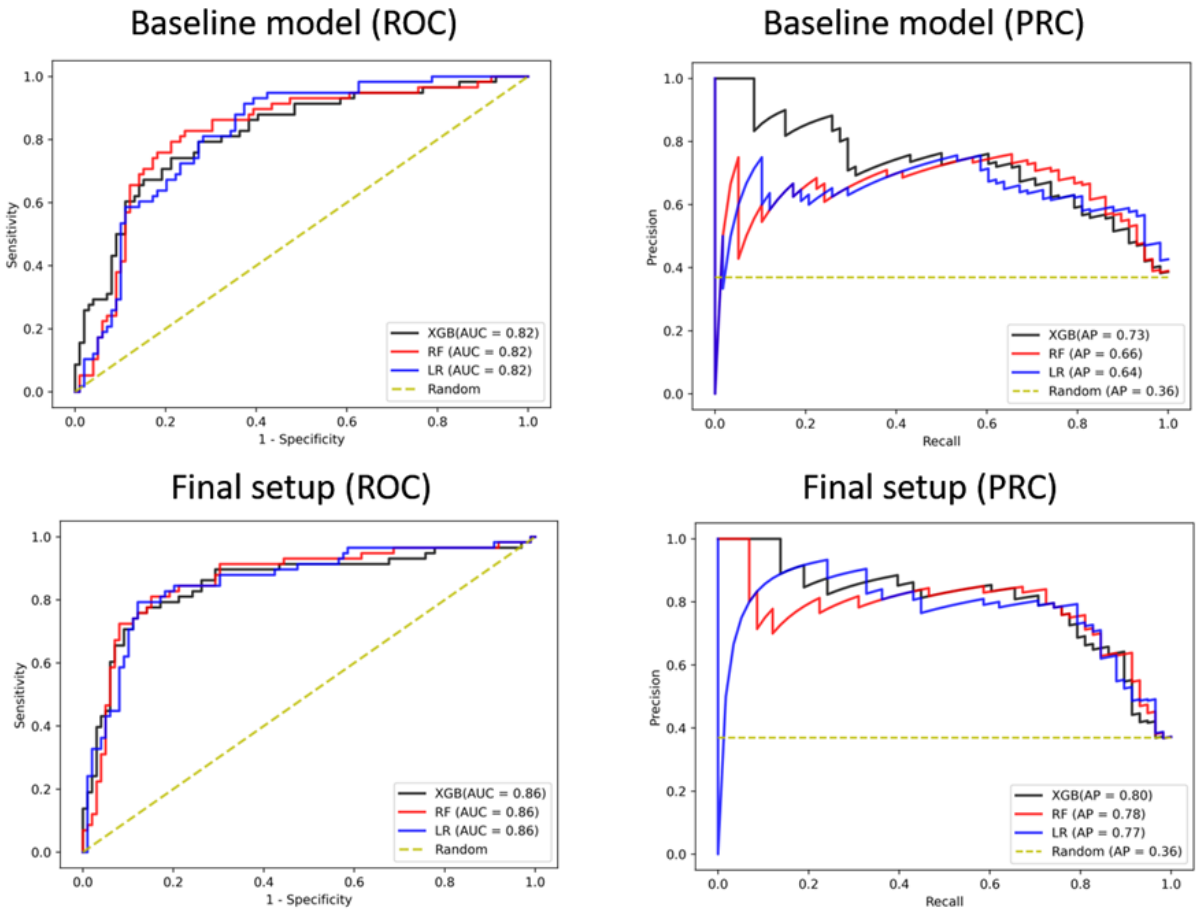


The LR (AUC 0.79, 95% CI 0.788-0.796) and RF (AUC 0.80, 95% CI 0.798-0.805) algorithms showed similar performance in the *baseline model* and improvement in the *final model*, comparable to XGB performance (see step 4 in Figure 1). The final XGB model provided superior performance compared to both the baseline model and SOFA score (both $P<.001$).

Experimental Evaluation

In the external validation of the EU patient cohort, all three models achieved similar performance in the baseline and the final setup with an AUC of 0.82 and 0.86, respectively, when evaluated on predicting the 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 toward two opposing health states of either being 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 4). As shown in Multimedia Appendix 7, the distribution for both outcomes was significantly skewed toward 21% for survivors and toward 100% for nonsurvivors.

Figure 3. Performance of the final model derived using the EU patient cohort and externally validated on a non-EU patient cohort, comprising Asian, African, and US patients. Model performance is measured using area under the receiver operating characteristic (ROC) curve (AUC) and area under the precision-recall curve (PRC). XGB: extreme gradient boosting; RF: random forest; LR: logistic regression.



We also assessed the calibration of each model to ensure that the distribution of predicted outcomes matches the distribution of observed outcomes in our patient cohort. Baseline and final models were, in general, well calibrated (Figure 5), matching the estimated risk of outcome with observed risk. The final setup for each algorithm was better calibrated (Brier score of 0.17) with respect to the baseline setup (Brier score 0.22). Full details of Brier scores for each algorithm are detailed in Multimedia Appendix 1.

Figure 4. Ranking of input variables of the final setup derived from the extreme gradient boost algorithm, using the shapely additive explanation (SHAP) method.

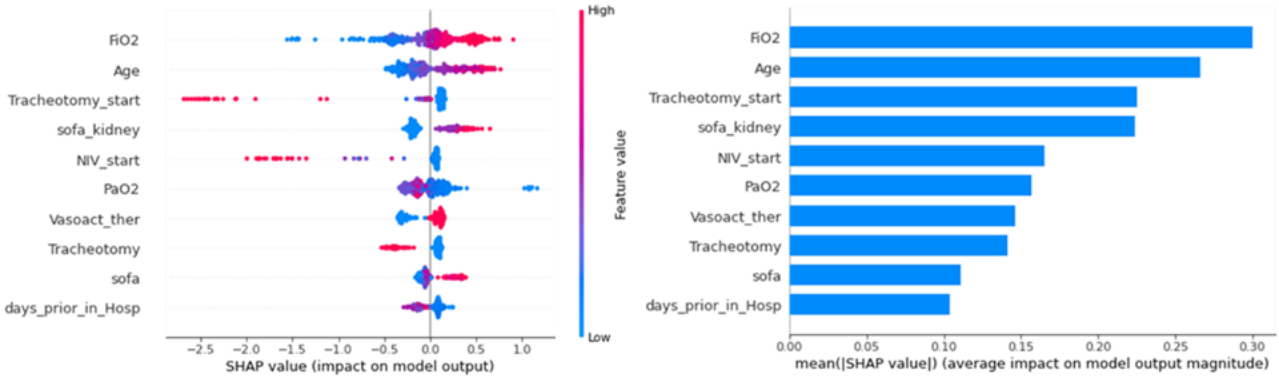
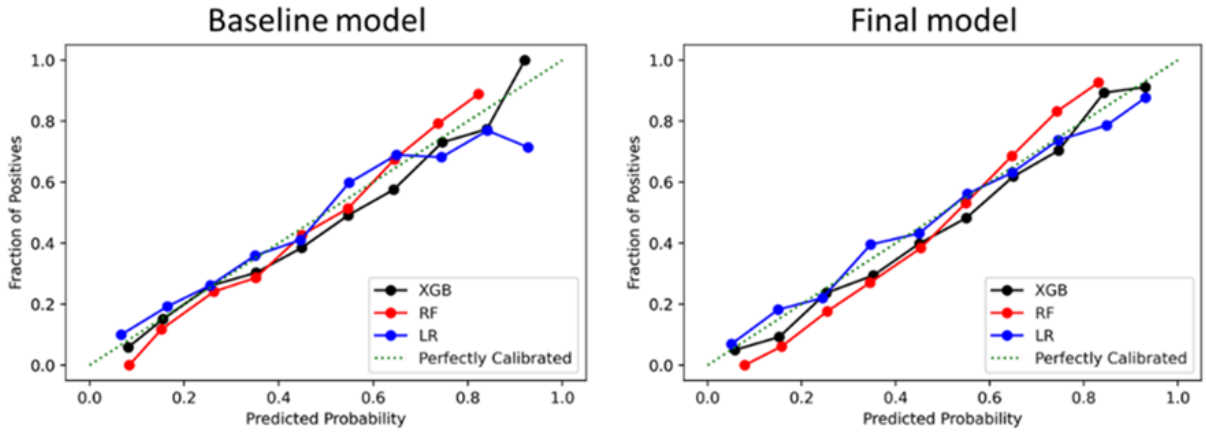


Figure 5. Calibration curves for each model and individual algorithms used to derive the model. XGB, extreme gradient boosting; RF: random forest; LR: logistic regression.



Model Interpretation

The SHAP 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 (XGB), where the FiO2, age, and tracheostomy had the highest impact on outcome prediction (Figure 4 and Multimedia Appendix 7).

We also report the model interpretability analysis for the RF- and LR-based models in Multimedia Appendix 8 and 9, respectively. The top three variables remained common between XGB and RF, whereas for LR, only tracheostomy appeared in the top three, with the other two high-ranking variables being weight and BMI.

Discussion

Principal Findings and Comparison With Related Studies

This 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 ML 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 [20] compared the ability of 18 different ML algorithms to predict the rate of admission and mortality of patients suffering from COVID-19. In their analysis, ensemble-based models were superior to other algorithms (including LR and XGB). Specific laboratory values and oxygen saturation were the most important factors for ICU admission, whereas impaired kidney function and differential blood count best predicted mortality [20]. However, this previous study primarily used data from patients, of all ages, presenting to the emergency room.

Domínguez-Olmedo et al [19] used data from 1823 patients with confirmed COVID-19 and established an XGB model. Their model found lactate dehydrogenase activity, C-reactive protein level, neutrophil count, and urea level to be the most important variables, 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.

Pan et al [22] used data from 123 patients with COVID-19 admitted to an ICU to construct an XGB model, and identified eight factors (albumin level, creatinine, eosinophil percentage, lactate dehydrogenase, lymphocyte percentage, neutrophil percentage, prothrombin time, and total bilirubin) that were predictive for ICU mortality.

Vaid et al [18] utilized a different approach based on federated learning of electronic health records from five different hospitals, providing robust predictive models without compromising patient privacy.

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

Applications

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 offer objective support for complex allocation decisions. However, ML algorithms would mainly access data at presentation and few dynamic parameters, limiting the predictive power. ML algorithms could also be used in the context of time-limited trials (TLTs), which are common clinical practice in ICUs in some countries. This may be particularly helpful in patients for 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 patients with sepsis [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 emphasized 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, and generalize 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 form part of a holistic patient assessment.

Limitations

This 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

This 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.

Acknowledgments

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 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 Heinrich-Heine-University Düsseldorf (grant 2018-32 to GW and grant 2020-21 to RB for a Clinician Scientist Track). The complete list of COVIP collaborators is provided in Multimedia Appendix 10.

Authors' Contributions

BW, BM, JF, RB, VO, and CJ analyzed the data and wrote the first draft of the manuscript. AA, BBP, JCS, and GW contributed to the statistical analysis and improved the paper. MK, MB, SS, PVH, WS, MC, ME, MJ, SO, TZ, BM, FA, RM, MC, SL, DWDL, BG, and HF gave guidance and improved the paper. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Hyperparameters for each algorithm found through an exhaustive grid search.
[DOCX File , 14 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Performance of the baseline model in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), extreme gradient boosting (XGB).
[DOCX File , 14 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Performance of the final model in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), extreme gradient boosting (XGB).
[DOCX File , 14 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Performance of the baseline model derived using the EU patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB).
[DOCX File , 14 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Performance of the final model derived using the EU patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB).
[DOCX File , 14 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Distribution of deaths over time and length of intensive care unit stay.
[DOCX File , 56 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Distribution of fraction of inspired oxygen (FiO2) for outcomes of survivors (left) and nonsurvivors (right). FiO2 was chosen as it was the variable that had the highest impact on the performance prediction, based on SHAP analysis.
[DOCX File , 66 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Ranking of input variables of the final setup derived using the random forest–based model.
[DOCX File , 161 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Ranking of input variables of the final setup derived using the logistic regression–based model.
[DOCX File , 126 KB-Multimedia Appendix 9]

Multimedia Appendix 10

List of COVIP-collaborators.
[DOCX File , 56 KB-Multimedia Appendix 10]

References

1. European Society of Intensive Care Medicine (ESICM), Global Sepsis Alliance (GSA), Society of Critical Care Medicine (SCCM). Reducing the global burden of sepsis: a positive legacy for the COVID-19 pandemic? Intensive Care Med 2021 Jul 16;47(7):733-736. [doi: 10.1007/s00134-021-06409-y] [Medline: 34132841]
2. Maltese G, Corsonello A, Di Rosa M, Soraci L, Vitale C, Corica F, et al. Frailty and COVID-19: a systematic scoping review. J Clin Med 2020 Jul 04;9(7):2106 [FREE Full text] [doi: 10.3390/jcm9072106] [Medline: 32635468]
3. Alkuzweny M, Raj A, Mehta S. Preparing for a COVID-19 surge: ICUs. EClinicalMedicine 2020 Aug;25:100502 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100502] [Medline: 32835188]
4. Chopra V, Flanders SA, Vaughn V, Petty L, Gandhi T, McSparron JI, et al. Variation in COVID-19 characteristics, treatment and outcomes in Michigan: an observational study in 32 hospitals. BMJ Open 2021 Jul 23;11(7):e044921 [FREE Full text] [doi: 10.1136/bmjopen-2020-044921] [Medline: 34301650]
5. Mudatsir M, Fajar JK, Wulandari L, Soegiarto G, Ilmawan M, Purnamasari Y, et al. Predictors of COVID-19 severity: a systematic review and meta-analysis. F1000Res 2020 Sep 9;9:1107. [doi: 10.12688/f1000research.26186.1]
6. Flaatten H, De Lange DW, Morandi A, Andersen FH, Artigas A, Bertolini G, VIP1 study group. 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 Dec 21;43(12):1820-1828. [doi: 10.1007/s00134-017-4940-8] [Medline: 28936626]
7. Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PLoS One 2020 Jul 30;15(7):e0236618 [FREE Full text] [doi: 10.1371/journal.pone.0236618] [Medline: 32730358]
8. Jung C, Bruno RR, Wernly B, Wolff G, Beil M, Kelm M. Frailty as a prognostic indicator in intensive care. Dtsch Arztebl Int 2020 Oct 02;117(40):668-673. [doi: 10.3238/arztebl.2020.0668] [Medline: 33357351]
9. Wernly B, Mamandipoor B, Baldia P, Jung C, Osmani V. Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation. Int J Med Inform 2021 Jan;145:104312. [doi: 10.1016/j.ijmedinf.2020.104312] [Medline: 33126059]

10. Masyuk M, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig JM, et al. Prognostic relevance of serum lactate kinetics in critically ill patients. Intensive Care Med 2019 Jan 26;45(1):55-61. [doi: 10.1007/s00134-018-5475-3] [Medline: 30478622]
11. Bruno RR, Wernly B, Flaatten H, Fjølner J, Artigas A, Bollen Pinto B, COVIP Study Group. 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 Aug 21;11(1):128 [FREE Full text] [doi: 10.1186/s13613-021-00911-8] [Medline: 34417919]
12. Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. Lancet Digit Health 2020 Oct;2(10):e496-e497 [FREE Full text] [doi: 10.1016/S2589-7500(20)30226-0] [Medline: 32984794]
13. Perrotta F, Corbi G, Mazzeo G, Boccia M, Aronne L, D'Agnano V, et al. COVID-19 and the elderly: insights into pathogenesis and clinical decision-making. Aging Clin Exp Res 2020 Aug 16;32(8):1599-1608 [FREE Full text] [doi: 10.1007/s40520-020-01631-y] [Medline: 32557332]
14. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. J Am Coll Cardiol 2021 Jan 26;77(3):300-313 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.030] [Medline: 33478654]
15. Izquierdo JL, Ancochea J, Savana COVID-19 Research Group, Soriano JB. 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 Oct 28;22(10):e21801 [FREE Full text] [doi: 10.2196/21801] [Medline: 33090964]
16. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, Northwell COVID-19 Research Consortium. 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 Feb 10;23(2):e24246 [FREE Full text] [doi: 10.2196/24246] [Medline: 33476281]
17. Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad A, Uddin S, Kamal A, 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 Apr 13;9(4):e25884 [FREE Full text] [doi: 10.2196/25884] [Medline: 33779565]
18. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, 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 Jan 27;9(1):e24207 [FREE Full text] [doi: 10.2196/24207] [Medline: 33400679]
19. Domínguez-Olmedo JL, Gragera-Martínez Á, Mata J, Pachón Álvarez V. Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: model development and validation. J Med Internet Res 2021 Apr 14;23(4):e26211 [FREE Full text] [doi: 10.2196/26211] [Medline: 33793407]
20. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med 2021 May 21;4(1):87. [doi: 10.1038/s41746-021-00456-x] [Medline: 34021235]
21. Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, et al. Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR Med Inform 2021 Jan 11;9(1):e23811 [FREE Full text] [doi: 10.2196/23811] [Medline: 33326405]
22. Pan P, Li Y, Xiao Y, Han B, Su L, Su M, 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 Nov 11;22(11):e23128 [FREE Full text] [doi: 10.2196/23128] [Medline: 33035175]
23. Kim H, Han D, Kim J, Kim D, Ha B, Seog W, 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 Nov 09;22(11):e24225 [FREE Full text] [doi: 10.2196/24225] [Medline: 33108316]
24. Burian E, Jungmann F, Kaissis G, Lohöfer FK, Spinner CD, Lahmer T, 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 May 18;9(5):1514 [FREE Full text] [doi: 10.3390/jcm9051514] [Medline: 32443442]
25. Shashikumar SP, Wardi G, Paul P, Carlile M, Brenner LN, Hibbert KA, et al. Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation. Chest 2021 Jun;159(6):2264-2273 [FREE Full text] [doi: 10.1016/j.chest.2020.12.009] [Medline: 33345948]
26. Jung C, Flaatten H, Fjølner J, Bruno RR, Wernly B, Artigas A, COVIP study group. The impact of frailty on survival in elderly intensive care patients with COVID-19: the COVIP study. Crit Care 2021 Apr 19;25(1):149 [FREE Full text] [doi: 10.1186/s13054-021-03551-3] [Medline: 33874987]
27. Li Y, Sperrin M, Ashcroft DM, van Staa TP. 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 Nov 04;371:m3919 [FREE Full text] [doi: 10.1136/bmj.m3919] [Medline: 33148619]
28. Guidet B, de Lange DW, Boumendil A, Leaver S, Watson X, Boulanger C, VIP2 study group. 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 Jan 29;46(1):57-69 [FREE Full text] [doi: 10.1007/s00134-019-05853-1] [Medline: 31784798]
29. COVIP Study. VIPSTUDY. URL: https://vipstudy.org/covip-study/ [accessed 2021-10-11]
30. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2016; Machineryan Francisco, CA. [doi: 10.1145/2939672.2939785]
31. Ho TK. Random decision forests. 1995 Presented at: Third International Conference on Document Analysis and Recognition; August 14-16, 1995; Montreal p. 278-282. [doi: 10.1109/icdar.1995.598994]
32. McCullagh P, Nelder JA. Generalized Linear Models. 2nd edition. Milton Park, England: Routledge; 1989.
33. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
34. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
35. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. 2017. URL: https://arxiv.org/abs/1705.07874 [accessed 2022-02-22]
36. Vink EE, Azoulay E, Caplan A, Kompanje EJO, Bakker J. Time-limited trial of intensive care treatment: an overview of current literature. Intensive Care Med 2018 Sep 22;44(9):1369-1377. [doi: 10.1007/s00134-018-5339-x] [Medline: 30136140]
37. Shrime MG, Ferket BS, Scott DJ, Lee J, Barragan-Bradford D, Pollard T, et al. Time-limited trials of intensive care for critically ill patients with cancer: how long is long enough? JAMA Oncol 2016 Jan 01;2(1):76-83 [FREE Full text] [doi: 10.1001/jamaoncol.2015.3336] [Medline: 26469222]
38. Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med Exp 2019 Dec 10;7(1):70 [FREE Full text] [doi: 10.1186/s40635-019-0286-6] [Medline: 31823128]
39. Jung C, Fjølner J, Bruno RR, Wernly B, Artigas A, Bollen Pinto B, COVIP Study Group. Differences in mortality in critically ill elderly patients during the second COVID-19 surge in Europe. Crit Care 2021 Sep 23;25(1):344 [FREE Full text] [doi: 10.1186/s13054-021-03739-7] [Medline: 34556171]
40. Bruno RR, Wernly B, Hornemann J, Flaatten H, FjØlner J, Artigas A, COVIP study group. Early evaluation of organ failure using MELD-XI in critically ill elderly COVID-19 patients. Clin Hemorheol Microcirc 2021;79(1):109-120. [doi: 10.3233/CH-219202] [Medline: 34487039]
41. Jung C, Bruno RR, Wernly B, Joannidis M, Oeyen S, Zafeiridis T, COVIP study group. Inhibitors of the renin-angiotensin-aldosterone system and COVID-19 in critically ill elderly patients. Eur Heart J Cardiovasc Pharmacother 2021 Jan 16;7(1):76-77 [FREE Full text] [doi: 10.1093/ehjcvp/pvaa083] [Medline: 32645153]
42. Jung C, Wernly B, Fjølner J, Bruno RR, Dudzinski D, Artigas A, the COVIP study group. Steroid use in elderly critically ill COVID-19 patients. Eur Respir J 2021 Oct 25;58(4):2100979 [FREE Full text] [doi: 10.1183/13993003.00979-2021] [Medline: 34172464]
43. Flaatten H, deLange D, Jung C, Beil M, Guidet B. The impact of end-of-life care on ICU outcome. Intensive Care Med 2021 May 19;47(5):624-625. [doi: 10.1007/s00134-021-06365-7] [Medline: 33604761]

Abbreviations

AUC: area under the receiver operating characteristic curve
CURB-65: confusion, urea, respiratory rate, blood pressure
eCRF: electronic case report form
FiO2: fraction of inspired oxygen
ICU: intensive care unit
LR: logistic regression
MIMIC-III: Medical Information Mart for Intensive Care
ML: machine learning
PPV: positive predictive value
PRC: precision-recall curve
RF: random forest
SHAP: shapely additive explanation
SOFA: Sequential Organ Failure Assessment
TLT: time-limited trials
XGB: extreme gradient boosting

Edited by C Lovis; submitted 16.08.21; peer-reviewed by F Velayati, H Ayatollahi; comments to author 10.10.21; revised version received 22.10.21; accepted 04.12.21; published 31.03.22

Please cite as:
Jung C, Mamandipoor B, Fjølner J, Bruno RR, Wernly B, Artigas A, Bollen Pinto B, Schefold JC, Wolff G, Kelm M, Beil M, Sviri S, van Heerden PV, Szczeklik W, Czuczwar M, Elhadi M, Joannidis M, Oeyen S, Zafeiridis T, Marsh B, Andersen FH, Moreno R, Cecconi M, Leaver S, De Lange DW, Guidet B, Flaatten H, Osmani V
Disease-Course Adapting Machine Learning Prognostication Models in Elderly Patients Critically Ill With COVID-19: Multicenter Cohort Study With External Validation
JMIR Med Inform 2022;10(3):e32949
URL: https://medinform.jmir.org/2022/3/e32949
doi: 10.2196/32949
PMID: 35099394