BMJ Open

Predicting mortality of individual patients with COVID-19: a multicentre Dutch cohort

Maarten C Ottenhoff ,[1] Lucas A Ramos,[2] Wouter Potters,[3] Marcus L F Janssen,[4] Deborah Hubers,[5] Shi Hu,[6] Egill A Fridgeirsson,[7] Dan Piña-Fuentes,[3] Rajat Thomas,[7] Iwan C C van der Horst ,[5] Christian Herff,[1] Pieter Kubben,[8] Paul W G Elbers,[9] Henk A Marquering,[2] Max Welling,[6] Suat Simsek,[10,11] Martijn D de Kruif,[12] Tom Dormans,[13] Lucas M Fleuren,[14] Michiel Schinkel,[15] Peter G Noordzij ,[16] Joop P van den Bergh,[17] Caroline E Wyers ,[17] David T B Buis,[18] W Joost Wiersinga,[19,20] Ella H C van den Hout,[10] Auke C Reidinga,[21] Daisy Rusch,[22] Kim C E Sigaloff,[19] Renee A Douma,[23] Lianne de Haan,[23] Niels C Gritters van den Oever,[24] Roger J M W Rennenberg,[25] Guido A van Wingen,[26] Marcel J H Aries ,[5] Martijn Beudel,[3] on behalf of The Dutch COVID-PREDICT research group

To cite: Ottenhoff MC, Ramos LA, Potters W, *et al*. Predicting mortality of individual patients with COVID-19: a multicentre Dutch cohort. *BMJ Open* 2021;**11**:e047347. doi:10.1136/bmjopen-2020-047347

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-047347).

MJHA and MB contributed equally.

Received 27 November 2020
Accepted 16 June 2021

For numbered affiliations see end of article.

Correspondence to
Maarten C Ottenhoff;
m.ottenhoff@maastrichtuniversity.nl

ABSTRACT

Objective Develop and validate models that predict mortality of patients diagnosed with COVID-19 admitted to the hospital.

Design Retrospective cohort study.

Setting A multicentre cohort across 10 Dutch hospitals including patients from 27 February to 8 June 2020.

Participants SARS-CoV-2 positive patients (age ≥18) admitted to the hospital.

Main outcome measures 21-day all-cause mortality evaluated by the area under the receiver operator curve (AUC), sensitivity, specificity, positive predictive value and negative predictive value. The predictive value of age was explored by comparison with age-based rules used in practice and by excluding age from the analysis.

Results 2273 patients were included, of whom 516 had died or discharged to palliative care within 21 days after admission. Five feature sets, including premorbid, clinical presentation and laboratory and radiology values, were derived from 80 features. Additionally, an Analysis of Variance (ANOVA)-based data-driven feature selection selected the 10 features with the highest F values: age, number of home medications, urea nitrogen, lactate dehydrogenase, albumin, oxygen saturation (%), oxygen saturation is measured on room air, oxygen saturation is measured on oxygen therapy, blood gas pH and history of chronic cardiac disease. A linear logistic regression and non-linear tree-based gradient boosting algorithm fitted the data with an AUC of 0.81 (95% CI 0.77 to 0.85) and 0.82 (0.79 to 0.85), respectively, using the 10 selected features. Both models outperformed age-based decision rules used in practice (AUC of 0.69, 0.65 to 0.74 for age >70). Furthermore, performance remained stable when excluding age as predictor (AUC of 0.78, 0.75 to 0.81).

Conclusion Both models showed good performance and had better test characteristics than age-based decision rules, using 10 admission features readily available in Dutch hospitals. The models hold promise to aid decision-making during a hospital bed shortage.

Strengths and limitations of this study

- This study uses the largest cohort of hospital admitted patients with COVID-19 in the Netherlands.
- Proven methods, such as leave-one-hospital-out cross-validation, strengthen the reliability of the results.
- However, the models are based on only Dutch patients, so it is unknown whether it is generalisable to other countries. Nonetheless, the results are comparable with a large multicentre cohort from the UK.
- The distribution of a favourable and unfavourable outcome is skewed, given that much more patients survived than that died. This is represented in a high negative predictive value and lower positive predictive value.

INTRODUCTION

The first wave of the COVID-19 pandemic had a dramatic effect on our society and severely disrupted our daily lives, economies and healthcare systems. During the peak of the first wave, hospitals and intensive care units (ICU) throughout Europe were overwhelmed and resources were exhausted. Implementation of public health policies reduced the infection rate; however, there is a considerable risk that relaxation of these policies leads to a next pandemic wave, which is already seen throughout European countries.

Given the novelty of the virus, accurate information about the clinical course and prognosis of individual patients is still largely unknown, which led to the use of crude limits to unilaterally withhold advanced life support measures to face the large numbers of pulmonary insufficient patients during the first wave. Although criticised, several hospitals in Europe have already solely used age as a triage criterion.[1] Many publications have developed and evaluated triage selection criteria, but there remains a significant knowledge gap and the final criteria are subject to socio-ethical debate.[2–4] Preferably, triage is averted, but when necessary, the decision should be guided by evidence-based medical criteria. Since March 2020, many studies have been published regarding the clinical characteristics of patients suffering from a SARS-CoV-2 infection in both smaller (n=58,[5] n=200[6]) and larger cohorts (n>5000[7–9]). However, these studies have reported notable differences in clinical characteristics that were associated with an adverse outcome. Importantly, these studies only provide information about clinical characteristics and risk factors on the group level and therefore do not provide information about the prognosis for individual patients. Prognostics models using multivariable analysis, such as[7] and[9] could be of great value during triage, especially when tailored towards individual prediction. These models can provide information about the individual patients' chance of survival, despite largely unknown underlying risk factors. Within the ongoing socio-ethical debate in the Netherlands, whether age should be included in the triage selection criteria, a predictive model could allow to exclude age or to include it together with clinical characteristics.[10] Wynants *et al*[10] reviewed COVID-19 prediction models, identifying 145 prediction models of which 23 were tailored towards predicting mortality. The authors identified that all studies were at high risk of bias and are likely to underperform in clinical practice. However, a recent paper, not yet reviewed by Wynants *et al*, showed promising results on predicting mortality with excellent performance, using a very large cohort (n>50.000) from the UK.[11]

The uncertainty and risk of bias in almost all published COVID-19 related prognostic models, stresses the importance of thorough methodology in variable selection, internal and external model validation and performance evaluation.[10] In addition, a constant interplay between data scientists and clinicians must be in place during model development. Furthermore, studies developed and performed independently with similar methodology are more valuable than ever to reduce the uncertainty of published models and the risk of spurious publications.[12 13] Therefore, a prognostic model was developed and evaluated that predicts 21-day all-cause mortality; using data from 2273 SARS-CoV-2 infected patients from 10 hospitals across the Netherlands.

MATERIALS AND METHODS

Data collection

Data were included from 10 Dutch hospitals varying from small to large peripheral hospitals to large academic centres. For an up-to-date overview of the including centres (see wwwcovidpredictorg). Clinical data were derived from electronic health records, pseudonymised and stored in the database (Castor EDC, Amsterdam, The Netherlands) by each hospital independently. Data collection started with the first admitted patients in the included centres. This was after the first confirmed case in the Netherlands on 27 February 2020. Records were included up to an admission date up until 8 June when the Dutch admission rates sharply decreased.[14] Inclusion criteria were admission in a hospital, age ≥18 years, a positive SARS-CoV-2 PCR before or during admission, or a CO-RADS CT thorax score ≥4 at admission. All patients were included consecutively. Retrospective data collection was based on the rapid COVID-19 case report form (rCCRF) developed by the WHO.[15] After consultation with several specialist consultants and an evaluation of the COVID-19 literature (mainly from China and Italy), additional clinical and laboratory features were added to the rCCRF. All included variables can be found in online supplemental table 1.

Given the exceptional circumstances related to the COVID-19 crisis and in accordance with national guidelines and European privacy law, the need for informed consent was waived and an opt-out procedure was communicated by press release. Despite this, individual centres used local guidelines to obtain consent retrospectively from patients or representatives. In all centres, measures were taken to ensure adequate and safe data pseudonymisation and storage.

Outcome definition

To support the decision of (ICU) treatment during scarcity at hospital admission, we aim to predict the unfavourable outcome of patients with COVID-19 at hospital admission. Given the amount of data, predicting each possible outcome, such as mortality, palliative care, discharge and hospitalisation, could increase the risk of biased models and overfitting. Therefore, the prediction goal was modelled as a binary classification problem, where an unfavourable outcome corresponds to patients that either died or were discharged for palliative care within 21 days after hospital admission. Palliative discharge is end-of-life care that focuses on patient comfort rather than treatments with curative intentions. A favourable outcome corresponds to patients that are discharged to home, nursing homes or rehabilitation units within 21 days and patients that are alive and still hospitalised at 21 days after hospital admission. Patients that were still hospitalised but shorter than 21 days, transferred to other hospitals (including transfers to participating hospitals), readmitted or have an unknown outcome were excluded from further analysis.

Data processing and quality

The rCCRF was filled in manually by a large team of researchers and doctors because the electronic patient dossiers in the different hospitals could not be coupled to the Castor database. The rCCRF and additional features resulted in a large number of features (>400). A consensus meeting with clinicians was held (18 April 2020) to remove features that were not available at hospital admission, not within the standard admission laboratory values or at risk of bias. This resulted in a feature set of 80 features. These 80 features were then divided into six sets: (1) premorbid characteristics (age, gender, occupation and medical history, n=24), (2) clinical characteristics at admission (n=14), (3) laboratory and radiology findings at admission (n=42), (4) the combination of set 1 and 2 (n=38), (5) all features (n=80) and (6) a data-driven selection from all features (n=10). The process of data-driven selection is described further in the modelling process section. The decision to use 10 variables was a practical one, in an attempt to balance fewer variables for easier application in practice and more variables to inform about important features. A complete overview of all features per set is shown in online supplemental table 1 and numerical characteristics per set are shown in online supplemental table 2. The resulting features were checked for physiologically implausible outliers by two authors (MJHA/DH). Some features contained high but plausible values and were therefore not removed (eg, creatine kinase). Furthermore, collinearity was assessed by a Pearson correlation matrix (online supplemental figure 1). No variables were removed due to high collinearity.

Predictive modelling

Ultimately, the obtained models could change the clinicians' decision and thus could directly influences the life of a patient. It is therefore of utmost importance that the obtained models are both robust and interpretable.[16] To comply with these requirements, two models with a fundamentally different modelling approach were selected: a logistic regression (LR) that fits the data linearly, and a tree-based gradient boosting algorithm that fits the data non-linearly. The models were implemented using the Python 3 libraries Scikit-learn[17] and XGBoost (extreme gradient boosting (XGB)),[18] respectively. Both models can be interpreted relatively easy and XGB often shows state-of-the-art results in multiple tasks. The models were trained and validated using leave-one-hospital-out cross-validation (LOHO-cv). By iteratively training the models on all but one hospital and performance testing on the left-out hospital, the performance of the model represents the ability to predict the outcome on independent data and thereby incorporate possible data heterogeneity between hospitals. To prevent skewed performance on individual folds due to a small number of samples, we combined the data from the two hospitals with the smallest number of samples and considered them as a single hospital in LOHO-cv for further analysis. Additional to LOHO-cv, internal 10-fold random subsampling cross-validation using data from all hospitals was performed to facilitate a comparison of the results to other studies that typically only perform internal cross-validation.

Modelling process

Features that had more than 50% missing values and subsequently patient records that had more than 80% missing values were removed. The remaining missing values were imputed using Bayesian ridge regression, which is inspired by the Multivariate Imputation by Chained Equations (MICE) method,[19] and implemented using the IterativeImputer from the Sci-Kit Learn library. Only one dataset per imputation was used since the disadvantages of single imputation are most apparent in small datasets with less than 100 events.[20] This imputation method models the missing values in each feature as a function of all other features and therefore provides a more sophisticated approach than the traditional imputation methods, such as using mean, median or mode imputation.[19] After imputation, each feature was scaled to its IQR. IQR scaling is known to be robust to outliers and often gives better results than *z*-score or minmax scaling.[21] Non-linear interactions between continuous variables can be taken into account by a non-linear model like XGB, thus splines were not included to prevent an unnecessary increase of the feature space.[22]

The data were then split into folds using LOHO-cv, where each iteration consists of a training fold with eight hospitals and a test fold with one hospital. The data-driven feature selection of set (6) was performed on the training fold by selecting the 10 features showing the highest ANOVA F value. Because for each iteration, the training fold consists of eight different hospitals, the selected features with the highest F values can differ due to heterogeneity between hospitals. To be able to describe the 10 most predictive features in further analysis, the features selected most often overall iterations are presented. If two feature sets are selected equally often, the set with the highest summed F values was chosen. Both missing value imputation and feature selection were performed independently on the training and test set. After feature selection, both models were fitted and parameters optimised by a 50-iteration randomised grid search using a stratified shuffle split cross-validation. A schematic overview of all the processing steps is shown in figure 1 and the grid search parameters are shown in online supplemental table 3. All code in the pipeline was implemented using the Scikit-learn python package.[17] To adhere to guidelines on transparent reporting of multivariable prediction models, the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis checklist is included in online supplemental table 4.[23] All code used in this paper, the final model and a calculator is available in online (DOI:10.5281/zenodo.4077342). A screenshot of the calculator is shown in online supplemental figure 2.

Performance analysis

Model discrimination was assessed by area under the curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Except for AUC, the metrics require a binary classification instead of

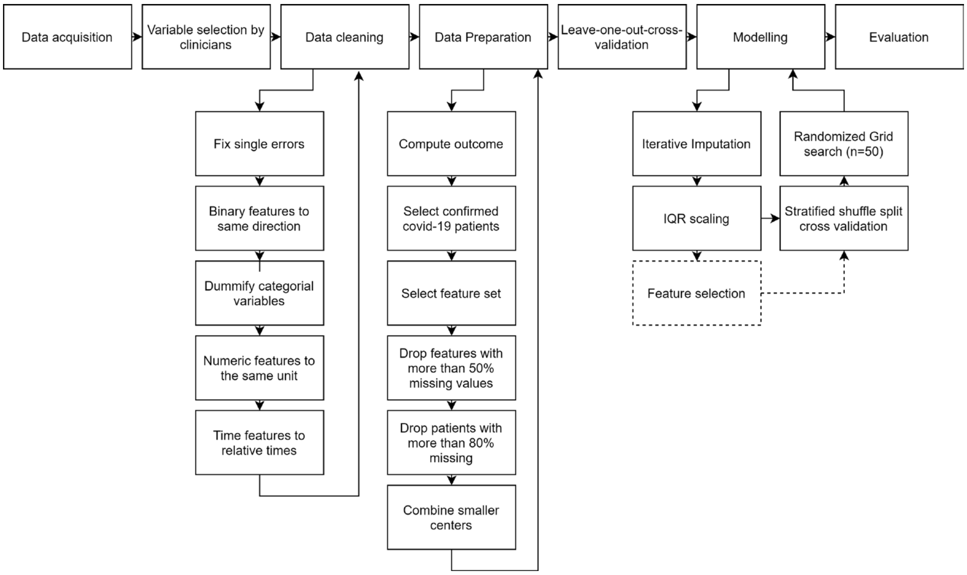
Data acquisition
Variable selection by clinicians
Data cleaning
Data Preparation
Leave-one-out-cross-validation
Modelling
Evaluation
Fix single errors
Binary features to same direction
Dummify categorial variables
Numeric features to the same unit
Time features to relative times
Compute outcome
Select confirmed covid-19 patients
Select feature set
Drop features with more than 50% missing values
Drop patients with more than 80% missing
Combine smaller centers
Iterative Imputation
IQR scaling
Feature selection
Randomized Grid search (n=50)
Stratified shuffle split cross validation

Figure 1 A schematic overview of all steps involved data acquisition to model evaluation. The dotted line depicts the step only used during feature selection of the 10 best features.

likelihood and therefore the cut-off threshold was tuned to the shortest distance to the upper-left corner in the receiver operating curve (ROC) plot, which was named as the 'optimal' threshold in further analysis. In addition, a confusion matrix was derived over the complete dataset and for each centre, also tuned to the optimal threshold. Furthermore, model calibration is shown in online supplemental figure 3.

Feature importance

Feature importance of models is described using SHAP (SHapley Additive exPlanations),[24] a game-theoretic approach to explain the output of any machine learning model. SHAP computes the average contribution of all features by permuting all of them and subsequently evaluating the error in the prediction for when a given feature is either included or not in the model. With SHAP, the impact of low and high values of a given feature on the models' predictions can be evaluated, as well as how impactful the feature is in predicting the correct class.[24]

Subgroup analysis for ICU admitted patients

During a large influx of patients suffering from life threatening lung infections, it is most likely that the ICU is exhausted first due to the low bed count and invasive ventilation capacity. It is therefore important to analyse whether the model also performs well on ICU admitted patients, as triage might be dependent on ICU capacity. In the Netherlands, triage was prevented by distributing patients to districts with fewer admissions or German hospitals. However, possible bias may already be present in the selection of patients, because, for example, certain patients might not be admitted to the ICU because of old age, premorbid characteristics, presentation with multi-organ failure and patients' own treatment restraints wishes. For these reasons, both LR and XGB performances were assessed by training on the complete dataset and the ICU patient subgroup.

Age as feature

To compare the models to clinical practice, the performance was compared with two age-based decision rules that have been applied in practice during the crisis.[1] The rules were translated as follows: (1) if age is above 70 then the outcome is considered unfavourable and (2) if age is above 80 then the outcome is considered unfavourable.

Furthermore, it was assessed whether age is important for the final prediction to be able to contribute to the ongoing socio-ethical debate in the Netherlands. In July 2020, a discussion between ethicists, medical professionals and policy-makers was started about criteria for triage to decide which patients receive ICU care during acute hospital care shortage. The main point of discussion was that the Dutch government was firmly opposed to using an age-based decision rule because it violates the Dutch constitution, which states that everyone should be treated equal and discrimination on any ground is illegal. To contribute to this discussion, the effect of age on the best performing model was assessed, by retraining the model on the same feature set, while excluding age as a feature.

Patient and public involvement

This study was a rapid response to an international public health emergency. Patients were not involved in any stage of this study.

RESULTS

Patient population

The database included 2527 patients from 10 different hospitals on 8 June 2020. Two hundred and twenty-three

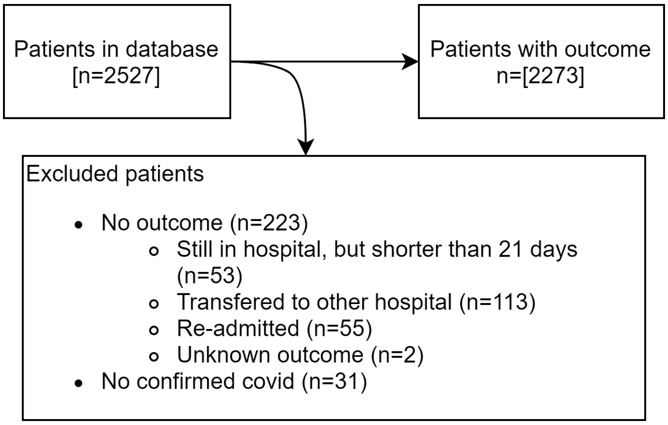

Figure 2 Flow diagram of patients excluded for further analysis.

patients were excluded because it was not possible to retrieve an outcome for these patients: patients that were still in the hospital, but less than 21 days (n=53), patients transferred to another hospital (n=113), patients that were discharged and re-admitted (n=55) and patients where the outcome was listed as unknown (n=2). In addition to these 223 excluded patients, 31 patients were excluded because they did not have a confirmed COVID-19 infection. After exclusion, 2273 patient remained to be included in further modelling and analysis.

Of these 2273 included patients, 1757 had a favourable outcome and 516 had an unfavourable outcome. Of the 1757 patients with a favourable outcome, 1195 were discharged home and not re-admitted patients, 76 were discharged to a nursing home and 232 were discharged to a rehabilitation unit. In addition, 254 were still in the hospital at 21 days after admission (112 at the ward or medium care and 142 in the ICU). Of the 516 patients with an unfavourable outcome, 509 patients died and 7 patients were discharged to palliative care. See figure 2 for an overview.

To better balance the samples per hospital, the two smallest hospitals (n=59 and n=70) were combined. The resulting ratio of unfavourable outcome/total patients per hospital is 19% (n=261), 14% (n=169), 10% (n=118), 31% (n=317), 14% (n=113), 21% (n=401), 27% (n=325), 27% (n=440) and 19% (n=129).

Feature description

Two features, history of smoking and alcohol abuse, were removed because of multi-interpretable questions in the rCCRF. One feature was removed from the Clinical presentation feature set and eleven features were removed from laboratory and radiology feature set for missing more than 50% values. No patient records were excluded for missing more than 80% values. After preprocessing, Premorbid and Clinical presentation features had 2.8% and 4.0% missing values, respectively. The admission laboratory and radiology features showed 21.6% missing values. See online supplemental tables 1 and 2 for a complete overview of features and missing values. Descriptive statistics of a selection of features are shown in table 1.

Overall model performance

XGB and LR performed equally on the premorbid set with an AUC of 0.77 (95% CI 0.73 to 0.81) and 0.77 (95% CI 0.72 to 0.81), respectively. On all other feature sets, XGB performed better than LR, although most 95% CIs overlapped. Both XGB and LR achieved the highest AUC on the 10 best features (0.82, 95% CI 0.79 to 0.85 and 0.81, 95% CI 0.77 to 0.85, respectively). Figure 3A shows a comparison of the AUCs per feature set, and figure 3B the confusion matrix of XGB trained on the 10 best features. Sensitivity and specificity were comparable between the algorithms. Overall, the NPV was high and the PPV was low, as the number of patients with a favourable outcome was considerably higher than the number of patients with an unfavourable outcome. This implies that the model can make accurate predictions of favourable outcomes, but less accurate predictions of unfavourable outcomes. All results are shown in table 2. For an in-depth overview of the results per fold, see online supplemental table 5. The results from internal cross-validation were comparable and shown in online supplemental table 6.

The between-hospital performance variation was small for both algorithms, shown by the small 95% CIs in AUC of 0.02 to 0.06 and a low SD (0.01). LR showed larger CIs (0.04 to 0.07) with equal SD (0.01).

The overall SD for all folds is small, and the comparison between internal cross-validation and LOHO-cv shows only minor differences in results between these approaches, reducing the risk of over-optimistic results. Between models, XGB fitted the data more robustly than LR, supported by the relatively equal ratios between correct and incorrect predictions, as shown in figure 4, which shows the confusion matrix per hospital for XGB-10 best predicting features using the optimal threshold derived from the complete dataset.

Performance stability over time

With increased duration of stay within the hospital, the uncertainty of the patients' outcome may also increase. The patient's chance of survival might change because patients that have a longer hospital stay are likely to have a more complicated clinical course and/or get different types of treatments. Additionally, prolonged hospital stay simply allows more events to happen. To assess whether the models' performance changes based on the duration of hospital stay, the patients were split per duration of stay and subsequently, the performance per group was assessed. The result, presented in figure 5, shows that model performance does not deteriorate as the hospital duration increases, as the relative correct predictions remain between 0.6 and 0.9 and no trend is shown.

Table 1 Patients characteristics per outcome group and a selection of features. P values were calculated using a t-test and corrected for multiple comparisons by Bonferroni correction

Variables	Missing	Overall	Favourable outcome	Unfavourable outcome	Adjusted p value
Total patients		2273	1758	516	
Age, median (Q1, Q3)	19	69.0 (58.0,78.0)	65.0 (55.0,75.1)	77.1 (71.0,83.1)	p<0.001***
Gender, n (%)	0				
Female		858 (37.7)	690 (39.3)	168 (32.6)	
Male		1415 (62.3)	1067 (60.7)	348 (67.4)	
History of hypertension, n (%)	30				p<0.001***
No		1207 (53.8)	998 (57.7)	209 (40.8)	
Yes		1036 (46.2)	733 (42.3)	303 (59.2)	
History of diabetes with complications, n (%)	64				p<0.001***
No		2044 (92.5)	1608 (94.4)	436 (86.3)	
Yes		165 (7.5)	96 (5.6)	69 (13.7)	
History of diabetes without complications, n (%)	69				p<0.001***
No		1789 (81.2)	1412 (83.0)	377 (75.1)	
Yes		415 (18.8)	290 (17.0)	125 (24.9)	
History of asthma, n (%)	55				p>0.05
No		1988 (89.6)	1524 (89.0)	464 (91.7)	
Yes		230 (10.4)	188 (11.0)	42 (8.3)	
History of liver disease, n (%)	57				p>0.05
No		2194 (99.0)	1693 (99.0)	501 (99.0)	
Yes		22 (1.0)	17 (1.0)	5 (1.0)	
History of rheumatological disorder, n (%)	43				p<0.05*
No		1981 (88.8)	1549 (89.9)	432 (85.2)	
Yes		249 (11.2)	174 (10.1)	75 (14.8)	
History of autoimmune and/or inflammatory diseases, n (%)	62				p<0.05
No		2027 (91.7)	1559 (91.5)	468 (92.3)	
Yes		184 (8.3)	145 (8.5)	39 (7.7)	
History of chronic cardiac disease, n (%)	36				p<0.001***
No		1539 (68.8)	1271 (73.6)	268 (52.4)	
Yes		698 (31.2)	455 (26.4)	243 (47.6)	
History of chronic haematological disease, n (%)	50				p<0.05
No		2133 (96.0)	1648 (96.0)	485 (95.7)	
Yes		90 (4.0)	68 (4.0)	22 (4.3)	
History of chronic kidney disease, n (%)	45				p<0.001***
No		1987 (89.2)	1566 (91.3)	421 (82.2)	
Yes		241 (10.8)	150 (8.7)	91 (17.8)	
History of chronic neurological disorder, n (%)	45				p<0.001***
No		1921 (86.2)	1519 (88.4)	402 (79.0)	

Continued

Table 1 Continued

Variables	Missing	Overall	Favourable outcome	Unfavourable outcome	Adjusted p value
Yes		307 (13.8)	200 (11.6)	107 (21.0)	
History of chronic pulmonary disease (not asthma), n (%)	47				p<0.001***
No		1790 (80.4)	1419 (82.5)	371 (73.2)	
Yes		436 (19.6)	300 (17.5)	136 (26.8)	

***p<0.001, **p<0.01, *p<0.05.

Feature importance

The 10 features selected most often were, in order of highest F value to lowest F value: age, urea nitrogen, number of home medications, oxygen saturation (%), history of chronic cardiac disease, oxygen saturation is measured on room air, oxygen saturation is measured on oxygen therapy, blood lactate dehydrogenase (LDH), blood albumin and blood gas pH value. Blood gas pH is measured from arterial, venous and capillary samples, of which 90.7% of the pH values are arterial measurements. The two 'oxygen is measured on' features are binary features that determine whether the oxygen saturation (%) is measured on room air or during oxygen therapy. The features were chosen independently of the choice of the model; therefore, the selected features were the same for both LR and XGB. Figure 6A,B shows the SHAP values per feature based on XGB trained on all features. For readability, only the top 20 features are shown. The features selected by the ANOVA in pretraining are also present in the top features computed by the SHAP values in post-training, which strengthens the likelihood of these features being the most important features within this dataset. This is also shown by the fact that LR scored notably higher by using the 10-best features than using all features and XGB showing equal performance using 10-best or all features. Analysis of SHAP values for LR on all features (online supplemental figure 4) showed that the linear LR model was not able to capture the non-linear predictive value of the age feature, as it was ranked as fourth. Nonetheless, the highest-ranked features for LR show importance and a direction of association consistent with the literature.[25 26]

Subgroup analysis for ICU patients

Of the 2273 included patients, 384 (17%) were admitted to the ICU at any time during the hospitalisation. LR showed the highest overall performance on ICU patients with an AUC of 0.71 (0.66 to 0.76). XGB showed the

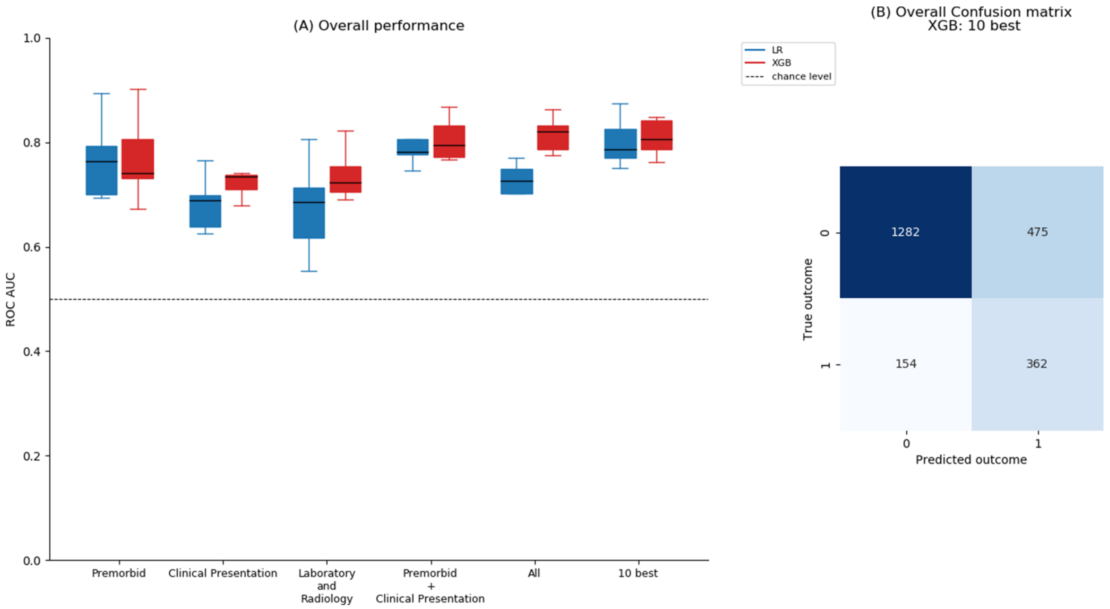

Figure 3 (A) Overall performance of both models per feature set. All models perform well above chance level. XGB generally performs better than LR, except on the premorbid feature set, where both models performed equally. The highest performance was achieved by XGB on both all features and the 10 selected features. (B) The confusion matrix of the best performing models, XGB trained on the 10 selected features. The prediction threshold was tuned to the shortest distance to the upper left corner of the AUC plot to create the 'optimal' binary prediction. AUC, area under the curve; LR, logistic regression; ROC, receiver operating curve; XGB, extreme gradient boosting.

Table 2 Evaluation metrics for both classifiers for each feature set

Classifiers	Feature set	AUC	Sensitivity	Specificity	PPV	NPV
LR	Premorbid	0.77 (0.72 to 0.81)	0.73 (0.61 to 0.84)	0.71 (0.64 to 0.78)	0.39 (0.35 to 0.44)	0.91 (0.88 to 0.95)
	Clinical presentation	0.67 (0.62 to 0.71)	0.60 (0.51 to 0.68)	0.63 (0.57 to 0.69)	0.30 (0.22 to 0.38)	0.86 (0.83 to 0.90)
	Laboratory and radiology	0.66 (0.59 to 0.73)	0.65 (0.47 to 0.83)	0.54 (0.34 to 0.73)	0.25 (0.16 to 0.34)	0.83 (0.74 to 0.91)
	Premorbid+clinical presentation	0.79 (0.75 to 0.83)	0.71 (0.62 to 0.80)	**0.71 (0.66 to 0.75**)	0.38 (0.32 to 0.43)	0.91 (0.89 to 0.93)
	All	0.71 (0.67 to 0.76)	0.62 (0.52 to 0.73)	0.70 (0.62 to 0.78)	0.36 (0.28 to 0.44)	0.88 (0.85 to 0.92)
	Ten best	**0.81 (0.77 to 0.85**)	**0.77 (0.68 to 0.85**)	0.71 (0.65 to 0.77)	**0.41 (0.36 to 0.45**)	**0.93 (0.90 to 0.95**)
XGB	Premorbid	0.77 (0.73 to 0.81)	0.68 (0.54 to 0.81)	0.60 (0.39 to 0.82)	0.36 (0.29 to 0.43)	0.68 (0.44 to 0.92)
	Clinical presentation	0.73 (0.71 to 0.74)	0.69 (0.61 to 0.77)	0.64 (0.59 to 0.69)	0.33 (0.26 to 0.40)	0.89 (0.87 to 0.92)
	Laboratory and radiology	0.72 (0.66 to 0.77)	0.68 (0.60 to 0.75)	0.63 (0.57 to 0.68)	0.31 (0.27 to 0.35)	0.88 (0.84 to 0.92)
	Premorbid+clinical presentation	0.81 (0.78 to 0.83)	**0.76 (0.67 to 0.85**)	0.62 (0.47 to 0.78)	0.36 (0.29 to 0.44)	0.81 (0.62 to 1.00)
	All	**0.82 (0.79 to 0.85**)	0.66 (0.54 to 0.78)	0.77 (0.65 to 0.89)	**0.47 (0.42 to 0.52**)	0.91 (0.88 to 0.95)
	Ten best	**0.82 (0.79 to 0.85**)	0.67 (0.57 to 0.77)	**0.75 (0.63 to 0.86**)	0.44 (0.40 to 0.48)	**0.91 (0.88 to 0.94**)

The average and 95% CIs over all leave-onehospital-out cross-validation iterations are presented. Values in bold represent the best performance for each metric per classifier. The premorbid feature set includes age, gender, occupation and medical history.
AUC, area under the curve; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; XGB, extreme gradient boosting.

highest performance on both premorbid and premorbid+clinical presentation features (0.69, 0.59 to 0.79). See table 3 for all results. For non-ICU patients, LR showed highest performance on the 10-best features (AUC 0.85, 0.81 to 0.88) and XGB on all features (AUC 0.86, 0.82 to 0.89). Compared with the results on the complete dataset, the performance dropped notably on ICU patients, decreasing in AUC by 0.04 to 0.20. The CIs also increased, overall ranging from 0.03 to 0.18. The decreased discriminative power of the models is considered acceptable, as the initially best-performing feature sets decreased only slightly and retained small CIs. The decrease was expected, given that performance on a smaller subgroup is inevitably lower. In addition, the prognosis of the outcome of ICU admitted patients might change, for example, due to receiving distinct interventions only available at the ICU. Inspection of sensitivity and specificity indicates that the lower performance was due to a decrease in sensitivity rather than specificity (see online supplemental table 7). The main objective in times of ICU admission at the time of ICU bed shortage is to correctly identify those patients that would benefit from intensive care. Therefore, the models may still be considered for application in practice, despite lower overall performance. Nonetheless, a more tailored approach might capture the unique characteristics of ICU patients better.

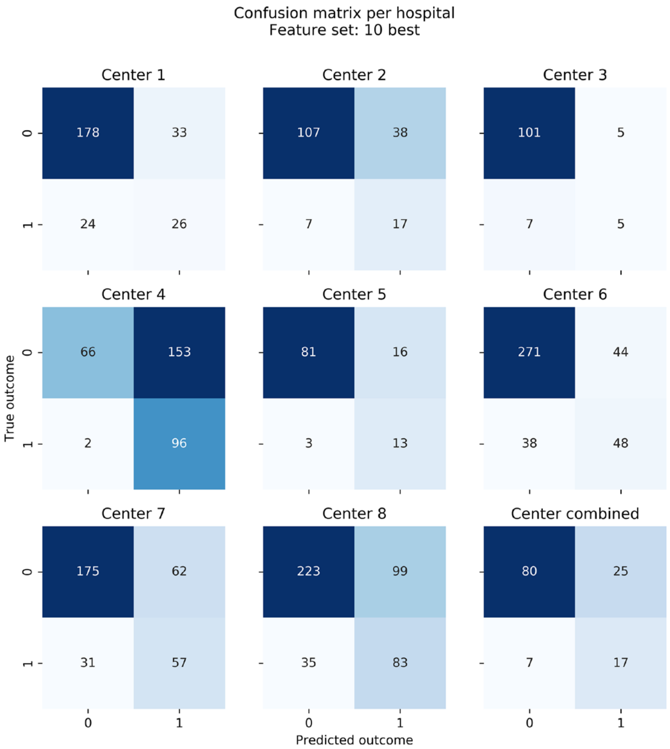

Figure 4 Confusion matrix per centre as predicted by extreme gradient boosting trained on the 10 selected features. The prediction threshold is optimised by the shortest distance to the upper-left corner in the receiver operating curve plot of the complete dataset. All matrices show comparable distributions, though centre 4 shows relatively many false positives.

Comparison with age-based rules for whole cohort

Of the 2273 patients, the age of 19 patients was missing and these were thus excluded from this analysis. Of the remaining 2254 patients, 1061 were older than 70 and 415 were older than 80. The age-based decision criteria therefore 'predicted' that of age >70, 1193 will survive and 1061 will die. For age >80 the prediction was 1839 and 415, respectively. Age >70 showed an AUC of 0.69 (0.65 to 0.74) whereas age >80 showed a lower AUC (0.61, 0.57 to 0.65). Figure 7 shows the confusion matrices of LR and XGB trained on the 10-best features and both age-based decision criteria. To compare both models with

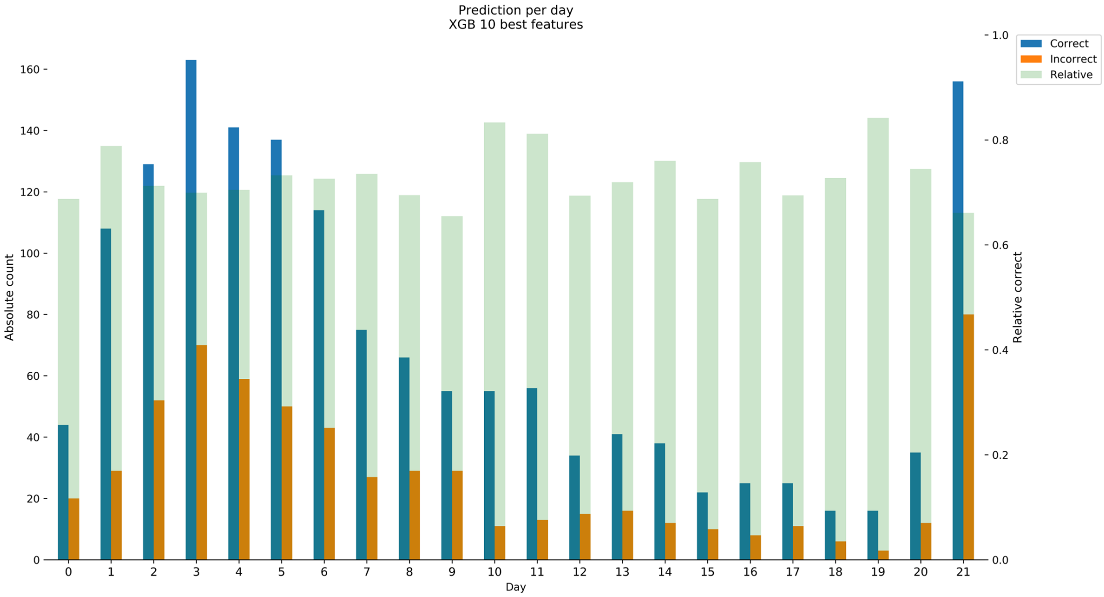

Figure 5 Performance per day for the extreme gradient boosting (XGB) trained on the 10 selected features. The left y-axis shows the absolute number of correct predictions and the right y-axis the relative number of correct predictions. Relative performance was calculated by *correct/(correct+incorrect*) and was well above chance level (0.5) for all days. The results indicate robust performance as the relative performance showed no decrease over time while varying between 0.6 and 0.9. The absolute performance shows that most patients have an outcome (both favourable and unfavourable within 1 week after admission. A high number of patients is seen on day 21, which is caused by the aggregation of all patients that are in the hospital 21 days or longer. Logistic regression on the 10 best features shows similar performance (figure not shown).

the age-based rules, the results were tuned to the shortest distance to the upper left corner in the ROC plot. Both LR and XGB show a higher AUC than either age-based decision criteria. The results show that the presented models can outperform earlier applied triage rules during crises and can thus provide better information based on individual medical data.

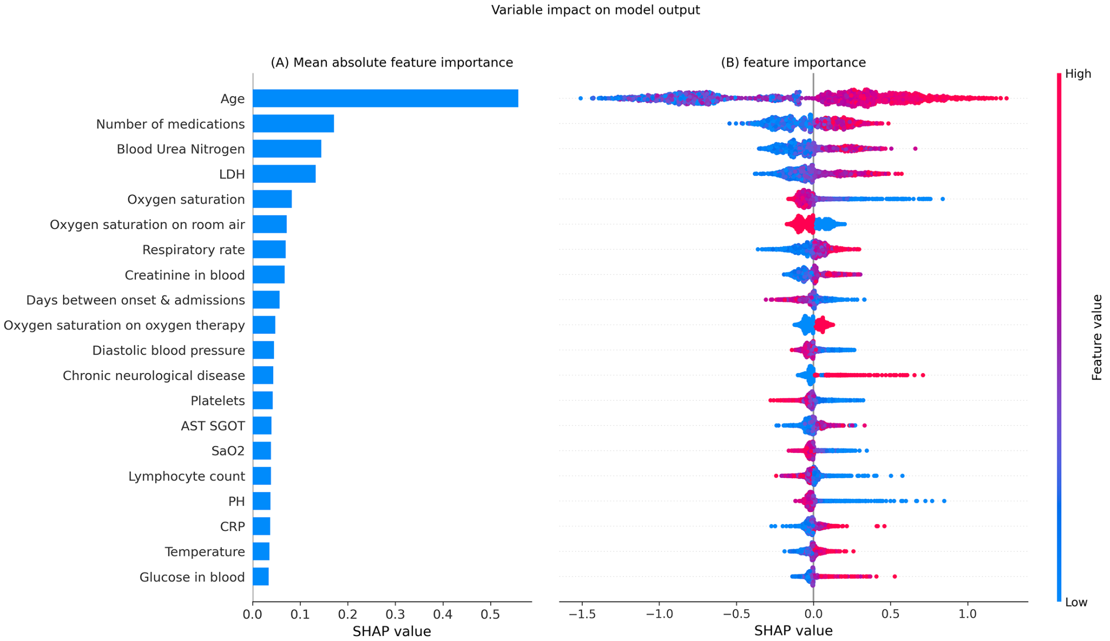

Figure 6 SHAP values of XGB trained on all features. To prevent readability issues, only the top 20 features are shown and the SHAP value range is set from −1.5 to 1.5, visually cutting of a few outliers. The colour of each data points depicts the height of the value, where red corresponds to high values and blue to low values. SHAP values above 0 suggest a positive association with the outcome. Given the outcome is defined as mortality within 21 days, the positive SHAP values translate to association with higher mortality. AST SGOT, aspartate aminotransferase / serum glutamic-oxaloacetic transaminase; LDH, lactate dehydrogenase; SHAP, SHapley Additive exPlanations; XGB, extreme gradient boosting.

Table 3 Model performance on (non-)ICU subgroup

Classifiers	Feature set	AUC—ICU patients	AUC—non-ICU patients
LR	Premorbid	**0.71 (0.66 to 0.76**)	0.81 (0.77 to 0.84)
	Clinical presentation	0.51 (0.37 to 0.66)	0.68 (0.64 to 0.72)
	Laboratory and radiology	0.54 (0.45 to 0.63)	0.69 (0.61 to 0.76)
	Premorbid+clinical presentation	0.60 (0.42 to 0.78)	0.83 (0.80 to 0.86)
	All	0.63 (0.50 to 0.76)	0.75 (0.72 to 0.79)
	10 best	0.62 (0.44 to 0.80)	**0.85 (0.81 to 0.88**)
XGB	Premorbid	**0.69 (0.59 to 0.79**)	0.80 (0.76 to 0.83)
	Clinical presentation	0.57 (0.41 to 0.72)	0.75 (0.72 to 0.77)
	Laboratory and radiology	0.59 (0.52 to 0.66)	0.76 (0.69 to 0.83)
	Premorbid+clinical presentation	**0.69 (0.59 to 0.79**)	0.84 (0.81 to 0.87)
	All	0.68 (0.58 to 0.78)	**0.86 (0.82 to 0.89**)
	10 best	0.68 (0.57 to 0.79)	0.85 (0.82 to 0.88)

Values in bold represent the best performance per classifier per subgroup. The premorbid feature set includes age, gender, occupation and medical history.
AUC, area under the curve; ICU, intensive care unit; LR, logistic regression; XGB, extreme gradient boosting.

Sensitivity analysis of age as feature

The best performing model, XGB-10, was retrained and evaluated without age as a feature. While expecting the performance to drop significantly, given that age was the most predictive feature by both the feature selection and SHAP analysis (figure 6), the performance decreased only slightly from an AUC of 0.82 (0.79 to 0.85) to 0.78 (0.75 to 0.81). Even though there were no signs of troublesome collinearity (online supplemental figure 1), age did show high multicollinearity (variance inflation factor; VIF >20). However, during model development, it was decided not to exclude features beforehand. Nonetheless, the high VIF indicates that the information present in age is latently present in two or more other features, which could explain the retained performance.

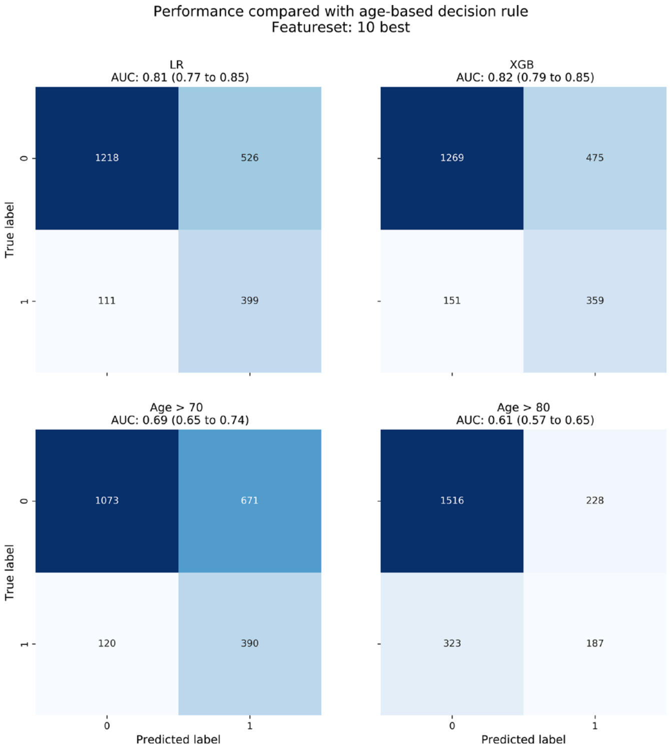

Figure 7 LR and XGB trained on the 10 selected features compared with two age-based decision rules. Both LR and XGB showed a higher AUC than both age-based rules. Nineteen patients did not have a value for age and were excluded for this analysis. AUC, area under the curve; LR, logistic regression; XGB, extreme gradient boosting.

DISCUSSION

We have shown that the mortality of individual patients with COVID-19 can be predicted at hospital admission with good discrimination using both linear (LR; AUC 0.81, 0.77 to 0.85) and non-linear (XGB; 0.82, 0.79 to 0.85) models with 10 features that are readily available in most hospitals. Both models showed improved discrimination over age-based decision rules, used in practice during acute hospital bed shortage.[2] XGB trained on all 80 features and the 10 best features performed comparable, but the latter model may be preferred for easier translation to clinical practice.

The presented models were trained on a large cohort, representing approximately 16% of the total COVID-19 related hospital admissions in the Netherlands during the first wave (Nationale Intensive Care Evaluatie (NICE), consulted 7 October).[24] Wynants *et al* reported that most models were at severe risk of bias due to poor patient selection, predictor description and methodology.[10] The present study has addressed these issues by aiming to clearly describe the patient inclusion process of a large cohort, clearly defining an outcome measure and by using a standardised predictor format (rCCRF) from the WHO, expanded with potentially predictive variables curated by clinicians working in the COVID-19 field. The models were calibrated by using nested cross-validation to prevent data leakage and validated by LOHO-cv. This location-based external cross-validation shows better results than classic cross-validation,[27] although validation on an independent dataset remains preferred. Additional to LOHO-cv, the risk of overfitting was further reduced by regularising both models, where regularisation parameters were optimised using nested cross-validation on the training set. Furthermore, the internal cross-validation results shown in online supplemental table 6 are similar to the results of LOHO-cv validation, indicating that the risk of overfitting on specific centres is low. An important note is that a good model fit was not shown on all feature sets, for example, laboratory and radiology features or LR on all features. Additionally, analysis of the SHAP values of LR (online supplemental figure 4), showed that the predictive value of age was not well captured by LR, as the feature was only ranked as the fourth most predictive feature. Combined, XGB-10 is the recommended model, as it showed a good fit and capture the non-linear predictive value of age well.

The results shown in this study are similar to a large-scale study by Knight *et al* that used a UK cohort 10-fold larger than this cohort and external validation.[11] The authors presented similar methods with similar results as this study, which strengthens the reliability of both models and reducing the risk of reporting over-optimistic results.

However, before application in practice the models need to be validated by an independent research group and for data in other countries. We have identified several uncertainties that may limit the current reliability of the models. First, the skewed outcome distribution (516 events over 2273 records) in our cohort limits the calibration of our models (online supplemental figure 3). This becomes apparent in the decreased calibration for higher-risk patients, though it must be noted that the recommended model (XGB-10) retains good calibration. Second, the cohort represents Dutch hospitalised patients with COVID-19 during the first wave of infections and might differ from current patients with COVID-19 due to the availability of therapies like steroids or vaccination. The included features could be improved by adding some features that are known to be highly predictive such as d-dimer, presence of infiltrates on the chest X-ray and duration of symptoms before hospital admission. These variables were initially included in our data but had to be removed due to too many missing values. Additionally, the duration of symptoms before admission was anamnestic, decreasing its reliability due to the retrospective data collection.

Finally, some uncertainties arise from the outcome definition, defined as the chance of death or discharge to palliative care within 21 days after hospital admission. The outcome was defined as all-cause mortality instead of COVID-19 related mortality, and this might result in an overestimation of the predictive power of specific comorbidities. Furthermore, the cut-off point of 21 days was considered as a balanced choice between early outcome and eventual outcome. A shorter timeframe might result in inaccurate outcomes and extending it would not have resulted in many more cases. However, some patients that were still at the hospital on day 21 might have an unfavourable outcome shortly after, resulting in a mislabelling of the patient, which overall might lead to an underestimation of mortality. Moreover, no follow-up of patients discharged to palliative care was implemented, possibly labelling patients with an erroneous unfavourable outcome. However, given that only seven patients (0.3% of all patients) were discharged to palliative care, we consider the risk negligible. Finally, no bed shortage was experienced in the Netherlands and it is therefore unlikely that prioritisation towards specific patients biased the cohort.

Implications for clinicians and policymakers

The presented models show that a reliable prediction can be made based on 10 features readily available in all Dutch and most worldwide hospitals: age, number of home medications, admission blood values urea nitrogen/LDH/albumin, oxygen saturation (%), blood gas pH and history of chronic cardiac disease. The models are thus easily applicable in practice and can improve the triage decision by providing a more objective medical foundation. We also showed that age as a feature is contributing towards a better prediction, but is not crucial. This implicates that policymakers can decide to exclude age when using these models.

Unanswered questions and future research

This work shows a promising step towards a triage tool during a hospital bed shortage. However, given the rapidly improving medical care for patients with COVID-19 and the lack of external validation, the data used during development are likely less representational of the current hospitalised patients with COVID-19. Additionally, the models are trained on a Dutch cohort and cannot be generalised to other countries. Finally, it should be evaluated how the prediction of the models compare with the clinician expertise. Altogether, a validation study evaluating these unanswered questions would be the next step towards clinical implication.

CONCLUSION AND RECOMMENDATION

Both LR and XGB showed good performance using the 10 best features, and outperformed age-based rules, with or without age included in the features. The results suggest that XGB using the 10 best features can improve decision making during an acute hospital bed shortage during a COVID-19 crisis and this model holds promise to be developed into a clinical tool.

Author affiliations
[1]Department of Neurosurgery, Maastricht University, Maastricht, The Netherlands
[2]Department of Biomedical Engineering and Physics/Department of Epidemiology & Data Science, Amsterdam University Medical Centres, Duivendrecht, The Netherlands

[3]Department of Neurology, Amsterdam University Medical Centres, Duivendrecht, The Netherlands
[4]Department of Clinical Neurophysiology, Maastricht University Medical Centre+, Maastricht, The Netherlands
[5]Department of Intensive Care, Maastricht Universitair Medisch Centrum+, Maastricht, The Netherlands
[6]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[7]Department of Psychiatry, Amsterdam University Medical Centres, Duivendrecht, The Netherlands
[8]Department of Neurosurgery, Maastricht Universitair Medisch Centrum+, Maastricht, The Netherlands
[9]Department of Intensive Care, Amsterdam UMC - Locatie VUMC, Amsterdam, The Netherlands
[10]Department of Internal Medicine, Noordwest Ziekenhuisgroep, Alkmaar, The Netherlands
[11]Department of Internal Medicine, Section of Endocrinology, Amsterdam UMC - Locatie VUMC, Amsterdam, The Netherlands
[12]Department of Pulmonary Medicine, Zuyderland Medical Centre Heerlen, Heerlen, The Netherlands
[13]Vascular Medicine, Amsterdam University Medical Centres, Duivendrecht, The Netherlands
[14]Department of Intensive Care, Amsterdam University Medical Centres, Duivendrecht, Noord-Holland, The Netherlands
[15]Center for Experimental and Molecular Medicine (C.E.M.M.), Amsterdam University Medical Centres, Duivendrecht, The Netherlands
[16]Department of Anesthesiology and Intensive Care, Sint Antonius Hospital, Nieuwegein, The Netherlands
[17]Department of Internal Medicine, VieCuri Medical Centre, Venlo, The Netherlands
[18]Department of Internal Medicine, Amsterdam UMC Locatie VUmc, Amsterdam, The Netherlands
[19]Department of Internal Medicine, Amsterdam University Medical Centres, Duivendrecht, The Netherlands
[20]Center for Experimental and Molecular Medicine (C.E.M.M.), Amsterdam UMC Locatie AMC, Amsterdam, The Netherlands
[21]Department of Intensive Care, Martini Ziekenhuis, Groningen, The Netherlands
[22]Research, Martini Ziekenhuis, Groningen, The Netherlands
[23]Department of Internal Medicine, Flevoziekenhuis, Almere, Flevoland, The Netherlands
[24]Department of Intensive Care, Treant Zorggroep, Hoogeveen, The Netherlands
[25]Department of Internal Medicine, Maastricht Universitair Medisch Centrum+, Maastricht, The Netherlands
[26]Department of Psychiatry, University of Amsterdam, Amsterdam, The Netherlands

Twitter Maarten C Ottenhoff @mcottenhoff and Rajat Thomas @rajatthomas

Contributors Conceptualisation: MCO, LAR, WP, MJ, RT, DP, CH, PK, PWGE, MS, JW, GAvW, MJHA, MB. Data curation: MCO, LAR, WP, MJ, DH, DP, SS, MDdK, TD, MS, PN, JPvdB, CEW, DTPB, EHCvdH, ACR, DR, KCES, RD, LdH, NCGvdO, GAvW, MJHA, MB. Formal Analysis: MCO, LAR, WP, SH, EAF, RT, CH, HAM, MW, GW, MJHA, MB. Investigation: MCO, LAR, SH, EAF, WP, GAvW, MJHA, MB. Methodology: MCO, LAR, SH, EG, RT, WP, GAvW, MJHA, MB. Project administration: WP, MJ, PK, PWGE, LMF, MS, GAvW, MJHA, MB. Software: MCO, LAR, WP. Supervision: GAvW, MJHA, MB. Validation: MCO, LAR, GW, MJHA, MB. Visualisation: MCO, LAR. Writing-original draft: MCO, LAR, WP, MJ, GAvW, MA, MB. Writing-review and editing: MCO, LAR, IH, DP, CH, PK, PWGE, HAM, SS, MDdK, TD, PN, JPvdB, CEW, JW, EHCvdH, ACR, DR, KCES, RD, LdH, NCGvdO, RJMWR, GAvW, MA, MB.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests The COVID-predict consortium declare to have received non-financial support from Castor, who provided access and use of their database free of charge. Pacmed occasionally provided scientific support for methodology and analysis.

Patient consent for publication Not required.

Ethics approval The ethical boards of the Amsterdam University Medical Centres (20.131) and Maastricht University Medical Centre approved the study protocol (MUMC: 2020-1323).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Not all patients provided active informed consent, and therefore data cannot be shared. The code used in this study is made publicly available and can be found at DOI:10.5281/zenodo.4077342

ORCID iDs
Maarten C Ottenhoff http://orcid.org/0000-0001-7676-1920
Iwan C C van der Horst http://orcid.org/0000-0003-3891-8522
Peter G Noordzij http://orcid.org/0000-0002-7115-8249
Caroline E Wyers http://orcid.org/0000-0001-7662-3990
Marcel J H Aries http://orcid.org/0000-0002-2155-688X

REFERENCES

1 Herreros B, Gella P, Real de Asua D. Triage during the COVID-19 epidemic in Spain: better and worse ethical arguments. *J Med Ethics* 2020;46:455–8.
2 Robert R, Kentish-Barnes N, Boyer A, *et al*. Ethical dilemmas due to the Covid-19 pandemic. *Ann Intensive Care* 2020;10:1–9.
3 Dahine J, Hébert PC, Ziegler D, *et al*. Practices in triage and transfer of critically ill patients: a qualitative systematic review of selection criteria. *Crit Care Med* 2020;48:e1147–57.
4 Sprung CL, Joynt GM, Christian MD, *et al*. Adult ICU triage during the coronavirus disease 2019 pandemic: who will live and who will die? recommendations to improve Survival*. *Crit Care Med* 2020;48:1196–202.
5 Yang X, Yu Y, Xu J, *et al*. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* 2020;8:475–81.
6 Zhou F, Yu T, Du R, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;395:1054–62.
7 Petrilli CM, Jones SA, Yang J, *et al*. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. *BMJ* 2020;369:m1966.
8 Richardson S, Hirsch JS, Narasimhan M, *et al*. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the new York City area. *JAMA* 2020;323:2052–9.
9 Docherty AB, Harrison HM, Green CA. Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC who clinical characterisation protocol. *medRxiv* 2020.
10 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
11 Knight SR, Ho A, Pius R, *et al*. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC who clinical characterisation protocol: development and validation of the 4C mortality score. *BMJ* 2020;370:m3339.
12 Mehra MR, Desai SS, Kuy S, *et al*. Cardiovascular disease, drug therapy, and mortality in Covid-19. *N Engl J Med* 2020;382:e102.
13 Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020;395:1820.
14 Fang Y, Zhang H, Xie J, *et al*. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* 2020;296:E115–7.
15 World Health Organization,. Novel coronavirus (covid-19) - Rapid version,, 2020. Available: https://apps.who.int/iris/rest/bitstreams/1274888/retrieve

16 Biobank UK, Hospital ME. Towards trustable machine learning. *Nat Biomed Eng* 2018;2:709–10.
17 Pedregosa F, Weiss R, Brucher M. Scikit-learn : Machine Learning in Python. *Journal of machine learning research* 2011;12:2825–30.
18 Chen T, Guestrin C. XGBoost : A Scalable Tree Boosting System," KDD '16. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
19 van Buuren S, Groothuis-oudshoorn K. mice : Multivariate Imputation by Chained. *J Stat Softw* 2011;45.
20 Neeman T. Clinical prediction models: a practical approach to development, validation, and updating by Ewout W. Steyerberg. *Int Stat Rev* 2009;77:320–1.
21 Iglewicz JW, Hoaglin B, Mosteller DC. Robust scale estimators and confidence intervals for location. In: *Understanding robust and exploratory data analysis*, 1983.
22 Friedman JH. Multivariate adaptive regression splines. *Hands-On Machine Learning with R* 2020;19:141–56.
23 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation* 2015;131:211–9.
24 Lundberg SM, Erion G, Chen H, *et al*. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;2:56–67.
25 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245–7.
26 Wool GD, Miller JL. The impact of COVID-19 disease on platelets and coagulation. *Pathobiology* 2021;88:15–27.
27 König IR, Malley JD, Weimar C, *et al*. Practical experiences on the necessity of external validation. *Stat Med* 2007;26:5499–511.