

Development and validation of a simplified risk score for the prediction of critical COVID-19 illness in newly diagnosed patients

Stanislas Werfel¹  | Carolin E. M. Jakob^{2,3} | Stefan Borgmann⁴ |
 Jochen Schneider^{5,6} | Christoph Spinner^{5,6}  | Maximilian Schons² |
 Martin Hower⁷ | Kai Wille⁸ | Martina Haselberger⁹ | Hanno Heuzeroth¹⁰ |
 Maria M. Rührich¹¹ | Sebastian Dolff¹²  | Johanna Kessel¹³ | Uwe Heemann¹ |
 Jörg J. Vehreschild^{2,3,13} | Siegbert Rieg¹⁴ | Christoph Schmaderer¹ | on behalf of the
 LEOSS study group

¹Department of Nephrology, School of Medicine, Technical University of Munich, Klinikum rechts der Isar, Munich, Germany

²Department I for Internal Medicine, University Hospital of Cologne, University of Cologne, Cologne, Germany

³German Centre for Infection Research (DZIF), Partner Site Bonn-Cologne, Cologne, Germany

⁴Department of Infectious Diseases and Infection Control, Ingolstadt Hospital, Ingolstadt, Germany

⁵Department of Internal Medicine II, School of Medicine, Technical University of Munich, University Hospital rechts der Isar, Munich, Germany

⁶German Centre for Infection Research (DZIF), Partner Site Munich, Munich, Germany

⁷Department of Pneumology, Infectious Diseases and Internal Medicine, Klinikum Dortmund gGmbH, Dortmund, Germany

⁸University Clinic for Haematology, Oncology, Haemostaseology and Palliative Care, Johannes Wesling Medical Centre Minden UKRUB, University of Bochum, Minden, Germany

⁹Department of Internal Medicine I, Klinikum Passau, Passau, Germany

¹⁰Department of Emergency and Intensive Care Medicine, Klinikum Ernst von Bergmann, Potsdam, Germany

¹¹Department of Internal Medicine II, Hematology and Medical Oncology, University Hospital Jena, Jena, Germany

¹²Department of Infectious Diseases, University Hospital Essen, University Duisburg-Essen, Essen, Germany

¹³Department of Internal Medicine, Hematology and Oncology, Goethe University Frankfurt, Frankfurt, Germany

¹⁴Department of Medicine II, University of Freiburg, Freiburg, Germany

Correspondence

Stanislas Werfel and Christoph Schmaderer,
 Department of Nephrology, School of
 Medicine, Technical University of Munich,
 Klinikum rechts der Isar, Ismaninger Straße 22,
 81675 Munich, Germany.

Email: stanislas.werfel@tum.de and
christoph.schmaderer@mri.tum.de

Funding information

German Centre for Infection Research; Willy
 Robert Pitzer Foundation

Abstract

Scores to identify patients at high risk of progression of coronavirus disease (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), may become instrumental for clinical decision-making and patient management. We used patient data from the multicentre Lean European Open Survey on SARS-CoV-2-Infected Patients (LEOSS) and applied variable selection to develop a simplified scoring system to identify patients at increased risk of critical illness or death. A total of 1946 patients who tested positive for SARS-CoV-2 were included

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Journal of Medical Virology* Published by Wiley Periodicals LLC

in the initial analysis and assigned to derivation and validation cohorts ($n = 1297$ and $n = 649$, respectively). Stability selection from over 100 baseline predictors for the combined endpoint of progression to the critical phase or COVID-19-related death enabled the development of a simplified score consisting of five predictors: C-reactive protein (CRP), age, clinical disease phase (uncomplicated vs. complicated), serum urea, and D-dimer (abbreviated as CAPS-D score). This score yielded an area under the curve (AUC) of 0.81 (95% confidence interval [CI]: 0.77–0.85) in the validation cohort for predicting the combined endpoint within 7 days of diagnosis and 0.81 (95% CI: 0.77–0.85) during full follow-up. We used an additional prospective cohort of 682 patients, diagnosed largely after the “first wave” of the pandemic to validate the predictive accuracy of the score and observed similar results (AUC for the event within 7 days: 0.83 [95% CI: 0.78–0.87]; for full follow-up: 0.82 [95% CI: 0.78–0.86]). An easily applicable score to calculate the risk of COVID-19 progression to critical illness or death was thus established and validated.

KEYWORDS

COVID-19, logistic models, machine learning, risk factors

1 | INTRODUCTION

The first human cases of coronavirus disease (COVID-19) were described in December 2019 in Wuhan.¹ COVID-19 subsequently developed into one of the most disastrous pandemics experienced in our civilization since the Spanish flu at the beginning of the 20th century.^{2,3} The exponential spread of the disease-causing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), as happened throughout Europe during the first wave of the pandemic, can result in excessive hospital overload and a shortage of healthcare resources, which may negatively impact patient outcomes.⁴ This experience underpinned the importance of an effective process to allocate limited healthcare resources to the COVID-19 patients most likely to benefit. To guarantee functional patient care, disease severity assessment for patients presenting at the emergency department (ED) may prove useful and guide frontline physicians in the decision-making process. A considerable number of patients deteriorate rapidly following hospital admission and require transfer to the intensive care unit (ICU). Conversely, the clinical conditions of other COVID-19 patients improve rapidly. Therefore, a prediction model can guide physicians in determining whether patients require hospital admission or can be followed up in outpatient care.

A risk assessment score may additionally be a useful tool to estimate the individual risk-benefit trade-off for therapeutic interventions.

This study aimed to develop a simplified risk prediction model based on clinical and demographic characteristics and laboratory findings at the time of COVID-19 diagnosis to estimate the risk of clinical deterioration to critical illness. We used data from the Lean European Open Survey on SARS-CoV-2 (LEOSS) project, a prospective European multicenter cohort study.⁵

2 | METHODS

2.1 | Study design and patient cohort

This analysis included patients who received care at a LEOSS partner site (inpatient or outpatient) beginning March 16, 2020. Cases documented in the LEOSS registry up till August 6, 2020 comprised the initial cohort, which was split into derivation and validation sets. Cases entered from August 7, 2020 to November 18, 2020 comprised the additional test sets (Figure 2A). The design of the LEOSS study and data acquisition was previously described.⁵

Data were recorded anonymously, and no patient-identifying data were stored. The requirement for written informed consent was therefore waived. Continuous parameters were categorized. To ensure anonymity at all stages of the analysis process, an individual LEOSS Scientific Use File (SUF) was created, which is based on the LEOSS public use file (PUF) principles, as described previously.⁶ Following these principles, a minor portion of patients and variable values were removed from the data set and set to missing values to ensure anonymization. Approval for LEOSS was obtained by the applicable local ethics committees of participating centers, and the study was registered at the publicly accessible German Clinical Trials Register (DRKS, No. DRKS00021145).

All predictors included in the stability selection are listed in Tables 1 and S1. We predefined a combined endpoint of progression to critical disease or COVID-19-related death. The definitions of the disease phases are summarized in Figure 1. The baseline (Day 0) was defined as the day of the first positive SARS-CoV-2 test. Only baseline predictors were included in the analysis (laboratory values collected within 48 h of diagnosis). If no CT was conducted within 48 h of positive testing, an exception was made and those CT-scan

TABLE 1 Characteristics of patients in the derivation and validation data sets

Predictor	Deriv.	Valid.	Test, f.	Test, l.	Predictor	Deriv.	Valid.	Test, f.	Test, l.
Total patients					CRP (mg/L)				
	1297	649	682	219	<3	181 (14%)	101 (16%)	97 (14%)	37 (17%)
Event during follow-up (7d/all)					3-29				
No	1095/1036 (84%/80%)	555/522 (86%/80%)	613/597 (90%/88%)	198/190 (90%/87%)	30-69	266 (21%)	132 (20%)	140 (21%)	47 (21%)
Yes	202/261 (16%/20%)	94/127 (14%/20%)	69/85 (10%/12%)	21/29 (10%/13%)	70-119	166 (13%)	85 (13%)	92 (13%)	28 (13%)
Type of patient care (not used for analyses)					120-179				
Outpatient	16 (1%)	11 (2%)	9 (1%)	1 (0%)	180-249	52 (4%)	26 (4%)	18 (3%)	6 (3%)
Inpatient	1255 (97%)	627 (97%)	648 (95%)	207 (95%)	>249	32 (2%)	17 (3%)	6 (1%)	0 (0%)
Missing	26 (2%)	11 (2%)	25 (4%)	11 (5%)	Missing	22 (2%)	11 (2%)	12 (2%)	3 (1%)
Age (year)					PCT (ng/ml)				
≤25	22 (2%)	17 (3%)	36 (5%)	9 (4%)	<0.005	78 (6%)	28 (4%)	27 (4%)	12 (5%)
26-35	78 (6%)	42 (6%)	64 (9%)	29 (13%)	0.005-0.5	562 (43%)	282 (43%)	367 (54%)	161 (74%)
36-45	105 (8%)	50 (8%)	86 (13%)	29 (13%)	0.51-2	58 (4%)	35 (5%)	28 (4%)	10 (5%)
46-55	189 (15%)	98 (15%)	104 (15%)	38 (17%)	2.1-10	0 (0%)	0 (0%)	13 (2%)	5 (2%)
56-65	244 (19%)	117 (18%)	120 (18%)	45 (21%)	>10	10 (1%)	6 (1%)	4 (1%)	1 (0%)
66-75	214 (16%)	118 (18%)	89 (13%)	25 (11%)	Missing	589 (45%)	298 (46%)	243 (36%)	30 (14%)
76-85	317 (24%)	140 (22%)	133 (20%)	30 (14%)	D-dimer (LN)				
>85	110 (8%)	59 (9%)	47 (7%)	13 (6%)	Normal	232 (18%)	123 (19%)	158 (23%)	83 (38%)
Missing	18 (1%)	8 (1%)	3 (0%)	1 (0%)	>1x, ≤2x	211 (16%)	109 (17%)	126 (18%)	72 (33%)
Sex					>2x, ≤5x	159 (12%)	69 (11%)	72 (11%)	34 (16%)
Male	768 (59%)	360 (55%)	390 (57%)	133 (61%)	>5x, ≤10x	39 (3%)	27 (4%)	24 (4%)	9 (4%)
Female	529 (41%)	289 (45%)	292 (43%)	86 (39%)	>10x, ≤20x	20 (2%)	11 (2%)	8 (1%)	2 (1%)
Disease phase					>20x	21 (2%)	12 (2%)	6 (1%)	4 (2%)
Uncompl.	876 (68%)	430 (66%)	488 (72%)	162 (74%)	Missing	615 (47%)	298 (46%)	288 (42%)	15 (7%)
Compl.	421 (32%)	219 (34%)	194 (28%)	57 (26%)	Neutrophils (×1000/μl)				
Any cardiovascular comorbidity					<0.1	11 (1%)	3 (0%)	4 (1%)	1 (0%)
Yes	727 (56%)	370 (57%)	346 (51%)	104 (47%)	0.1 to <0.3	14 (1%)	3 (0%)	2 (0%)	0 (0%)
No	545 (42%)	262 (40%)	326 (48%)	113 (52%)	0.3 to <0.5	22 (2%)	10 (2%)	2 (0%)	0 (0%)
Missing	25 (2%)	17 (3%)	10 (1%)	2 (1%)	0.5 to <2	118 (9%)	62 (10%)	47 (7%)	15 (7%)
Malignant neoplasia					2 to <5	524 (40%)	262 (40%)	275 (40%)	105 (48%)
No	1263 (97%)	635 (98%)	678 (99%)	218 (100%)	5 to <9	262 (20%)	139 (21%)	144 (21%)	54 (25%)
Yes	34 (3%)	14 (2%)	4 (1%)	1 (0%)	≥9	71 (5%)	40 (6%)	39 (6%)	6 (3%)
LDH (LN)					Missing	275 (21%)	130 (20%)	169 (25%)	38 (17%)
<Normal	0 (0%)	0 (0%)	8 (1%)	2 (1%)	Lymphocytes (×1000/μl)				
Normal	439 (34%)	218 (34%)	249 (37%)	98 (45%)	<0.1	16 (1%)	8 (1%)	7 (1%)	1 (0%)
>1x, ≤2x	596 (46%)	312 (48%)	305 (45%)	95 (43%)	0.1 to <0.3	56 (4%)	30 (5%)	18 (3%)	1 (0%)
>2x, ≤5x	87 (7%)	51 (8%)	38 (6%)	11 (5%)	0.3 to <0.5	95 (7%)	43 (7%)	33 (5%)	9 (4%)
					0.5 to <0.8	230 (18%)	124 (19%)	118 (17%)	39 (18%)
					0.8 to <1.5	421 (32%)	212 (33%)	231 (34%)	94 (43%)

(Continues)

TABLE 1 (Continued)

Predictor	Deriv.	Valid.	Test, f.	Test, l.	Predictor	Deriv.	Valid.	Test, f.	Test, l.
>5x	4 (0%)	1 (0%)	3 (0%)	2 (1%)	1.5 to <3	198 (15%)	104 (16%)	100 (15%)	34 (16%)
Missing	171 (13%)	67 (10%)	79 (12%)	11 (5%)	≥3	15 (1%)	13 (2%)	17 (2%)	4 (2%)
Urea (LN)					Missing	266 (21%)	115 (18%)	158 (23%)	37 (17%)
<Normal	8 (1%)	9 (1%)	33 (5%)	8 (4%)					
Normal	846 (65%)	408 (63%)	445 (65%)	173 (79%)					
>1x, ≤2x	195 (15%)	106 (16%)	89 (13%)	26 (12%)					
>2x	63 (5%)	32 (5%)	30 (4%)	8 (4%)					
Missing	185 (14%)	94 (14%)	85 (12%)	4 (2%)					

Abbreviations: 7d, event (critical phase or COVID-19-related death) within 7 days of diagnosis; CRP, C-reactive protein; LDH, lactate dehydrogenase; LN, laboratory normal range, "x" indicates multiples of the upper limit of the normal range; PCT, procalcitonin; Test, f., full test set (as shown in Figure 2); Test, l., limited test set (as shown in Figure 2).

Uncomplicated phase	Complicated phase	Critical phase
<ul style="list-style-type: none"> - asymptomatic - symptoms of upper respiratory tract infection - nausea, emesis, diarrhea - fever 	<ul style="list-style-type: none"> - need for new oxygen supplementation - clinically meaningful increase of prior oxygen home therapy - PaO₂ at room air <70 mmHg - sO₂ at room air <90 % - AST or ALT > 5x ULN - new cardiac arrhythmia - new pericardial effusion > 1 cm - new heart failure with pulmonary edema - congestive hepatopathy or peripheral edema 	<ul style="list-style-type: none"> - need for catecholamines - life-threatening cardiac arrhythmia - need for unplanned mechanical ventilation (invasive or non-invasive) - prolongation (>24h) of planned mechanical ventilation - liver failure with Quick < 50 % or INR > 3.5 - qSOFA ≥ 2 - acute renal failure in need of dialysis

FIGURE 1 Definition of COVID-19 disease phases in the LEOSS registry. Patients were assigned to the highest phase for which at least one characteristic was fulfilled. ALT, alanine transaminase; AST, aspartate transaminase; INR, international normalized ratio of prothrombin time; PaO₂, partial pressure of oxygen in arterial blood; qSOFA, quick sequential organ failure assessment score; sO₂, blood oxygen saturation; ULN, upper limit of normal

variables collected after this time but during the same clinical phase that was present at baseline were included. We additionally calculated a separate predictor describing whether the patient had cardiovascular (CV) comorbidities, defined as any of the following: history of (H/O) myocardial infarction, aortic stenosis, atrioventricular block, carotid artery disease, chronic heart failure, peripheral vascular disease, hypertension, atrial fibrillation or coronary artery disease. An additional variable was also calculated for neurological comorbidities, defined as any of the following reported for the patient: hemiplegia, dementia, cerebrovascular disease or stroke, multiple sclerosis, myasthenia gravis, neuromyelitis optica spectrum disorder (NMOSD), movement disorder (e.g., Parkinson's disease, dystonia, ataxia, and tremor), motor neurone diseases (e.g., amyotrophic lateral sclerosis, and spinal muscular atrophy), other neurological autoimmune

diseases and other prior neurological diagnoses. We defined a predictor for any malignant neoplastic disease as any of the following: H/O lymphoma, leukemia, solid tumor, solid metastasized tumor, and stem cell transplantation.

2.2 | Statistical analysis

All analyses were performed using R (version 3.6.3). Random forest (RF) analyses (including missing value imputations and individual Boruta stability selection steps) were calculated using the "randomForestSRC" package by Ishwaran and Kogalur.⁷

Among the available baseline variables of the LEOSS data set (≈170 predictors), we selected those with <50% missing values among the combined derivation and validation data set ($n = 1946$ patients, Figure 2), with the exception of troponin T (52% missing) and pancreatic lipase (56% missing). This resulted in a total of 104 predictors (Tables 1 and S1). The time-to-event data in the anonymized LEOSS cohort was grouped for patients experiencing an event at ≥8 days after study inclusion, the time variable was coded accordingly as 1–7 days and ≥8 days, resulting in eight bins for the time variable (Table S1). These were used for the time-to-event approaches: random survival forest and Cox models, and for C-index calculation. Continuous predictors were binned as value ranges in the LEOSS cohort due to anonymization, and the ranges were coded as consecutively increasing integers.

We performed RF missing value imputation using multivariate unsupervised splitting as described by Tang and Ishwaran⁸ and two iterations per imputation. An RF approach has been previously shown to be the method of choice for ordinal variables,⁹ which are the main target of imputation in our data set (because continuous variables were categorized). The imputations were performed either for the data of the combined derivation and validation data set ($n = 1946$ patients) or, separately, the full test set ($n = 682$ patients, Figure 2), while withholding the outcome variables. Twenty imputed data sets were thus generated for each cohort.

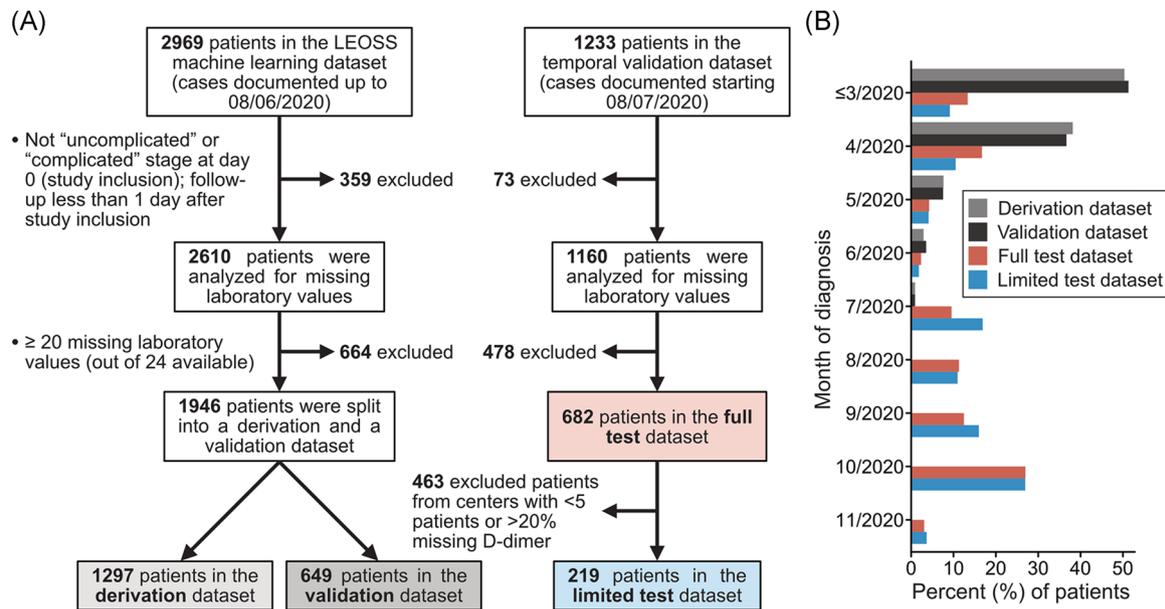


FIGURE 2 Patient flow diagram (A) and months of COVID-19 diagnosis (B) for the different data sets

We performed a split into a derivation and validation cohort with similar characteristics based on the following predefined potential confounders: age, sex, presence of dyspnea, neutrophil count, lymphocyte count, lactate dehydrogenase (LDH), bilirubin, CRP, procalcitonin (PCT), D-dimer, H/O malignant neoplasia, presence of CV comorbidity (as defined above) and the number of events. We performed 1000 random splits at 2/3 and 1/3 ratios and calculated the standardized mean difference for each split, selecting the split with the smallest maximal standardized mean difference between these predictors.

Variable selection was performed using the Boruta algorithm¹⁰ at 100 iterations using equal proportions of the 20 imputed derivation data sets and a p value of 0.01 for selection. For the classification RFs, we used the presence of an event (critical phase or COVID-19-related death) within 7 days of diagnosis as the outcome of interest during Boruta selection. We used the balanced method by Chen et al.¹¹ both during Boruta selection and modeling with the selected variables. We used survival random forest (RSF) as described by Ishwaran et al.,¹² during Boruta selection, and during the final modeling of time-to-event data. As RSFs take time to event into account, events occurring beyond 7 days after diagnosis were also included. Variable importance was calculated using permutation. For Cox and logistic (binomial) regression models, we performed ridge (L2) penalization optimized using 20 \times fold cross-validation on the imputed derivation data sets. Score values were calculated from the ridge penalized binomial regression coefficients of the model containing the five selected predictors on the derivation data set with missing values replaced with the most common value of the 20 imputed data sets for this patient, and predictor and event within 7 days as the outcome. Finally, the regression coefficients were divided by the smallest value and rounded to the next integer.

Two-sided p values for the binomial ridge penalized coefficients were obtained as suggested by Cule et al.,¹³ by repeating the ridge regression procedure on a data set with randomly permuted outcomes 1000 times (using equal numbers of the 20 imputed data sets).

The area under the receiver operating characteristics curve (AUC) and Harrell's C-indices were calculated using linear predictors from the binomial and Cox ridge-penalized regression models or out-of-bag predictor estimates for the RF approaches. The 95% confidence intervals (CIs) for AUC and C-indices were calculated using 1000 bootstraps of patients' scores using equal contributions from the imputed data sets.

3 | RESULTS

3.1 | Patient population

Important characteristics of the LEOSS cohort were previously described.⁵ More diagnosed SARS-CoV-2 cases were available for the current analysis compared with the previous report (2969 in the first data set, patients from the first wave of the pandemic, and 1233 patients in the second test set; Figure 2).⁵ Based on the predefined disease phase (Figure 1) and the availability of laboratory values, a total of 1946 patients were included in the first round of analysis and assigned to derivation and validation groups with similar characteristics (Figure 2). Important characteristics are summarized in Table 1, with a summary of the remaining predictors provided in Table S1.

The age distribution in the first data set was centered, with approximately equal contributions of patients aged ≤ 65 and >65 years. There were more men than women (55%–59% vs. 41%–45%). At least 56% presented with known CV comorbidity. The incidence of the combined endpoint, critical phase or COVID-19-related death

TABLE 2 Summary of the predictive performances of the analyzed models

Selection	Model	N pr.	AUC, 7d (imp. range)		AUC, all (imp. range)	
			Derivation	Validation	Derivation	Validation
All pr.	RF	104	0.83 (0.82–0.83)	0.83 (0.82–0.83)	0.83 (0.82–0.83)	0.83 (0.82–0.83)
	Binomial ridge	104	0.88 (0.87–0.89)	0.81 (0.80–0.81)	0.86 (0.86–0.87)	0.81 (0.80–0.82)
RF Boruta	RF	5	0.74 (0.72–0.75)	0.73 (0.71–0.76)	0.73 (0.72–0.75)	0.74 (0.73–0.77)
	Binomial ridge	5	0.80 (0.80–0.80)	0.81 (0.81–0.81)	0.80 (0.80–0.80)	0.81 (0.81–0.81)
	Score	5	0.80 (0.80–0.80)	0.81 (0.81–0.81)	0.80 (0.80–0.80)	0.81 (0.81–0.81)
			95% CI, 0.77–0.83	95% CI, 0.77–0.85	95%CI, 0.77–0.83	95% CI, 0.77–0.85
Validation on the full test set				0.83 (0.82–0.83)	0.82 (0.82–0.83)	
				95% CI, 0.78–0.87	95% CI, 0.78–0.86	
Validation on the limited test set				0.82 (0.82–0.82)	0.83 (0.83–0.84)	
				95% CI, 0.73–0.90	95% CI, 0.76–0.90	

Note: Initial derivation and validation analyses were performed on the respective data sets ($n = 1297$ and 649 , respectively) as summarized in Figure 2. As indicated, the final score was additionally independently validated on the full and the limited test sets ($n = 682$ and 219 , as described in Figure 2). Indicated are the median values and the full range for the imputed data sets (in brackets). AUC values were calculated for an event within 7 days of diagnosis ("7d") and for all time points ("all"). 95% confidence intervals (95% CI) were calculated for score predictions using bootstrapping with equal contributions of the imputed data sets. Results of the performance of the final score (median AUC and 95% CI) in the respective validation and test datasets are highlighted in bold.

Abbreviations: AUC, area under the receiver operating characteristic (ROC) curve; imp., imputation; N pr., number of predictors in the model; pr., predictors; RF, random forest for classification.

within 7 days was 14%–16%, and 20% when including any time point during the follow-up period (Table 1).

From the second test set (patients entered into the registry after the first data export for score derivation), 682 patients fulfilled the selection criteria. This set largely consisted of patients diagnosed after June 2020 (Figure 2). Compared with the derivation/validation cohorts, the patients were younger (60% ≤ 65 years) and more were diagnosed during an uncomplicated phase (72% vs. 64%–68%). Consequently, the event rate was lower, with only 10% experiencing an event within 7 days of diagnosis and 12% during follow-up (Table 1). Both the derivation and validation data sets consisted almost exclusively of patients receiving inpatient care.

3.2 | Predictor selection

We performed Boruta variable stability selection using RF for classification, resulting in the selection of 5 (out of 104) predictors (Table 2). These were CRP, disease phase, age, serum urea, and D-dimer levels (Figure S1A). Interestingly, including only these five predictors in a logistic regression model achieved results almost on par with the full set of variables (Table 2, "RF Boruta," binomial ridge, median AUC = 0.81 in the validation cohort).

We additionally performed a Boruta stability selection using an RSF approach. Twenty-four predictors were retained, with the five predictors from RF Boruta among the most important variables (Figure S1B). Increasing the number of predictors from 5 to 24 had a minor impact on the model's performance in the validation data set as

TABLE 3 Results of the ridge-penalized binomial regression on the five variables selected by RF Boruta

Predictor	Ridge β	p value	Weight
Age	0.07	0.024	1
Disease phase	0.40	0.003	5
Urea	0.26	0.013	3
CRP	0.14	0.002	2
D-dimer	0.09	0.041	1

Note: Indicated are β coefficients from binomial ridge regression (outcome: event within 7 days) and the resulting weights per step increase in the respective predictor group (all groups are listed in Table 4). p values were calculated using ridge regression on the derivation data set with permutations of the outcome.

Abbreviation: CRP, C-reactive protein.

measured by Harrell's C-index (median C-index: 0.76 vs. 0.77, respectively; Table S2).

3.3 | Derivation and validation of a simplified predictive score

Based on the encouraging results and simple interpretability, we used the coefficients obtained in the binomial ridge regression model with five predictors (Table 3) to derive an additive score to predict COVID-19 progression to the critical phase or death. The score is listed in Table 4. It exhibited similar performance as the binomial model in

TABLE 4 Calculation of the CAPS-D score

Predictor	Score	Predictor	Score
Age (year)		CRP (mg/L)	
≤25	-	<3	-
26–35	+1	3–29	+2
36–45	+2	30–69	+4
46–55	+3	70–119	+6
56–65	+4	120–179	+8
66–75	+5	180–249	+10
76–85	+6	>249	+12
>85	+7	Disease phase	
D-dimer (LN)		Uncomplicated	-
Normal	-	Complicated	+5
>1x, ≤2x	+1	Urea (LN)	
>2x, ≤5x	+2	<Normal	-
>5x, ≤10x	+3	Normal	+3
>10x, ≤20x	+4	>1x, ≤2x	+6
>20x	+5	>2x	+9
		Maximum score	38

Abbreviations: CRP, C-reactive protein; LN, laboratory normal range, “x” indicates multiples of the upper limit of the normal range.

both the derivation and validation data sets (median AUC in validation data set for events within 7 days of diagnosis: 0.81, 95% CI: 0.77–0.85; for all events, 0.81, 95% CI: 0.77–0.85; Table 2). Interestingly, the simplified score yielded a similar performance as a Cox regression or an RSF approach with both 5 and 24 predictors as measured by Harrell's C-index (median C-index of 0.76, 95% CI: 0.73–0.80 in the validation cohort; Table S2).

We used the second test set of patients whose data were entered into the registry after the initial data export ($n = 682$ patients, “full test set” in Figure 2) as an independent prospective validation group. To further reduce the impact of missing values on the estimation of score performance, we additionally removed patients from centers with >20% missing values for D-dimer, the variable with the most missing values (42%–47% missing). Centers that enrolled <5 patients were also excluded, which produced an additional “limited test set” ($n = 219$ patients; Figure 2). This data set had few missing values (CRP, 1%; serum urea, 2%; D-dimer, 7% missing; Table 2).

In both full and limited test sets, we confirmed the similar performance of the developed scoring system, with a trend toward higher AUC and C-index values compared with the validation data set (full test set, median AUC for 7 days: 0.83, 95% CI: 0.78–0.87; all events: AUC 0.82, 95% CI: 0.78–0.86; limited test set, median AUC for 7 days: 0.82, 95% CI, 0.73–0.90; all events: AUC 0.83, 95% CI, 0.76–0.90; Table 2; median C-index for full test set: 0.80, 95% CI, 0.76–0.84; limited test set: 0.81, 95% CI: 0.74–0.87; Table S2).

Depending on the clinical application, different cut-off values may be considered. Therefore, we provide the predictive metrics of the score, such as sensitivity, specificity, and positive and negative predictive values (PPV and NPV) versus the cut-off (Figure 3), as well as the absolute event risks for specific score values (Figure S2).

Apart from the discriminative performance, we observed good calibration with a slope ranging from 0.949 to 1.113 in the different validation/test data sets (Figure S3). Interestingly, the Brier score was tendentially smaller in the “full test” compared to the validation data set (0.076–0.091 vs. 0.106–0.124, respectively; Figure S3), mirroring the tendency toward better discriminative performance in this data set (Tables 2 and S2). Calibration-in-the-large for the “full test” set, which yielded a lower event per case rate, was similar to that in the validation set for an event within 7 days (intercept: -0.160 vs. -0.174, respectively), but lower for all events (intercept: -0.314 vs. 0.010, respectively, potentially reflecting the differences in event rates between the cohorts).

One method for selecting a cut-off is to optimize the modified Youden's J .¹⁴ For the proposed score, the optimal J in the combined validation and full test data set was at a cut-off of ≥ 17 , both for predictions at 7 days after diagnosis and for all events. Applying this cut-off, on average, 69% of patients were predicted not to progress to critical illness (Table 5, combined validation/test data set) at an NPV of 95% for 7 days after diagnosis and an NPV of 94% for full follow-up. Patients with scores at or above this threshold had ~3-fold increased odds of experiencing an event, whereas patients below this threshold had ~3-fold decreased odds as measured by the respective likelihood ratios (Table 5).

4 | DISCUSSION

We describe the derivation and validation of a COVID-19 risk score for the prediction of the combined endpoint of critical disease or COVID-19-related death using five predictors. We derive the score in an untargeted manner by selecting the most stable predictors among over 100 available at baseline in the LEOSS registry in an RF approach and using regularized regression to calculate the coefficients.

A number of approaches for COVID-19 risk stratification have been reported (reviewed by Wynants et al.¹⁵); several had a similar aim of predicting critical disease, as indicated by admission to the ICU^{16–19} or death.^{17,18,20,21}

The availability of factors such as hospital or ICU beds was limited during the height of the pandemic with the resulting strain on healthcare systems. Thus, difficulties in generalizing outcome predictions obtained under these constraints in the currently available scores may arise. Some important limiting factors must be considered. If hospital beds are limited, the study population for inpatient analyses may overrepresent patients with symptoms of exceptional severity and high-risk groups, which may limit generalizability. Similarly, if ICU resources are limited, the indications for admission may be more conservative; a patient may be identified as having a favorable outcome (not admitted to ICU) despite having fulfilled the clinical criteria at some point.

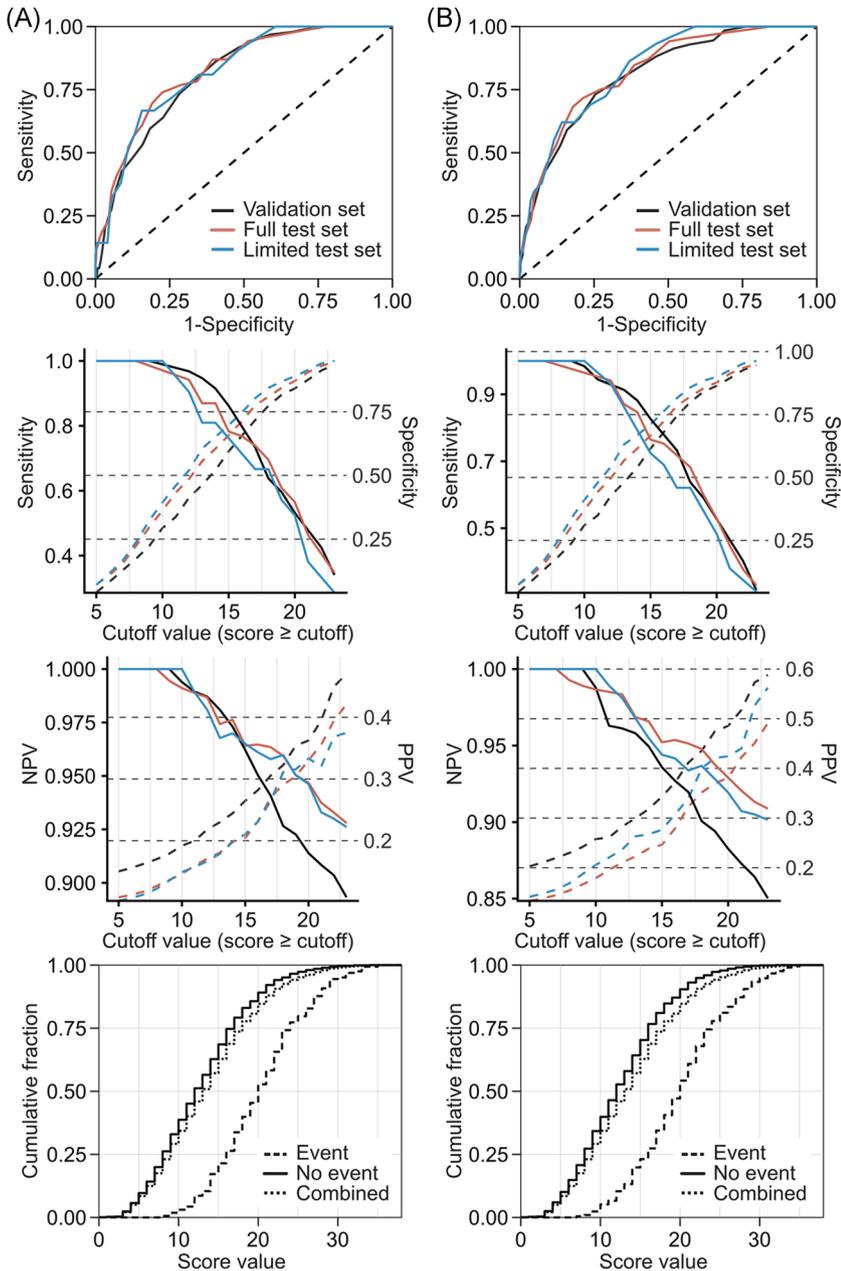


FIGURE 3 Summary of key characteristics of the score for predicting the combined endpoint of critical phase or COVID-19-related death (A) within 7 days of the diagnosis or (B) at any time point during follow-up in the validation and test cohorts. Color codes distinguish the different data sets as indicated. Sensitivity and NPV are indicated by continuous lines and the corresponding y axis scaling on the left, while specificity and PPV are indicated by dashed lines and y axis scaling on the right side of the respective panels. Bottom panels show cumulative fractions of patients meeting respective score cut-offs for the combined validation and full test set (combined $n = 1331$). For all panels, the median score (rounded to the next whole integer) of the imputations was calculated for patients with missing values

Another important consideration is the generalizability of mortality as an outcome in patient stratification. Case fatality rates differ widely across countries,²² perhaps partly attributed to country-specific differences in the clinical management of COVID-19 patients and to resource availability during the first wave of the pandemic.⁴ This may limit generalizability and potentially require an update to existing scores for mortality prediction,²³ as care providers gain experience with COVID-19 management and the strain on hospitals is reduced.

A previous review on COVID-19 prognosis scores came to an overall negative assessment of the potential bias of these scores, which discouraged their use.¹⁵ To our knowledge, a combination of characteristics sets our approach apart from those available at the time of writing making it potentially more generalizable for clinical

application: (a) the outcome was not defined in terms of a specific treatment (or lack thereof, i.e., admission to the ICU), but based on clinical features (a predefined “critical phase”); (b) the inclusion was based on predefined clinical criteria (“uncomplicated” or “complicated” phase), and (c) the use of a stability selection approach to reduce the number of predictors, as discussed below. Additionally, the majority (>90%) of patients enrolled in the LEOSS cohort were from Germany,⁵ where the capacity of the healthcare system was not generally exceeded during the first wave of the pandemic.²⁴

To address bias in predictor selection, we used an untargeted approach and resampling techniques (stability selection and cross-validated ridge regression) to internally test the predictions on the derivation data set and then validate them on a withheld validation cohort. Stability selection aids in ensuring the internal validity and

TABLE 5 Score characteristics at the selected cut-off of ≥ 17

	Validation set (7d/all)	Full test set (7d/all)	Combined (7d/all)
Sensitivity	0.73/0.73	0.74/0.72	0.74/0.73
Specificity	0.72/0.75	0.77/0.79	0.75/0.77
PPV	0.31/0.41	0.27/0.32	0.29/0.37
NPV	0.94/0.92	0.96/0.95	0.95/0.94
LR+	2.6/2.9	3.3/3.3	2.9/3.1
LR-	0.37/0.36	0.34/0.36	0.35/0.36
%score < cut-off	65%	72%	69%

Abbreviations: 7d, event (critical disease or COVID-19-related death) within 7d of diagnosis; all, all events during follow-up; LR+/-, positive/negative likelihood ratio; NPV, negative predictive value; PPV, positive predictive value; %score < cut-off, percentage of patients with scores below the cut-off value (≤ 16).

adequate sample size for the derivation data set; too small a sample will typically reduce variable stability and lead to fewer variables being selected. Ridge regression shrinks the regression coefficients to achieve improved predictions in a binomial model with internal (cross-) validation in the derivation data set.

We successfully confirmed the performance of our score in an independent test set, consisting of the majority of COVID-19 cases diagnosed after the first wave of the pandemic.

An important contributor to the predictive performance of the final score was the predefined clinical phase ("complicated" vs. "uncomplicated"), which summarizes the presence of manifest organ involvement of the lungs, heart, or liver. Of note, some parameters of the complicated phase, such as arterial partial pressure of oxygen (PaO_2) and pericardial effusion, were acquired by indication (e.g., if an echocardiography or arterial blood gas analysis was performed, but not routinely). Therefore, for phase assignment, these do not have to be taken into account in the absence of an indication for the respective measurement.

Serum urea, likely as a measure of kidney involvement, was an important predictor and outperformed creatinine previously for mortality.^{20,25} This predictor potentially summarizes both pre-existing chronic kidney disease as a risk factor (Williamson et al.²⁶; Figure S1B) and acute kidney injury (AKI) due to COVID-19 as organ involvement (also stable in RSF Boruta; Figure S1B). Different mechanisms of AKI in COVID-19 patients have been observed, including indirect involvement, such as cardiorenal syndrome, direct virus-induced injury, and immunological causes such as complement activation.^{27,28} Differentiating the type of acute kidney involvement in COVID-19 patients may provide further insights and refine risk stratification in future analyses.

Overall, our score, despite being limited to five predictors and applying a point system, compared favorably to more complex prediction models.^{17,20} We suggested a threshold for patients with an increased risk of critical disease at ≥ 17 points, based on the modified Youden's *J*. At this threshold, we obtained a positive likelihood ratio

of threefold while retaining a good negative predictive value of 94%–95%. Different cut-offs may be considered based on the application and local circumstances (e.g., different local ratios of critical disease per case, and travel time to the next hospital in case of deterioration in an outpatient setting). The graphs provided in Figure 3 for sensitivity/specificity and PPV/NPV (based on the prevalence in the validation and test data sets) as well as in Figure S2 for absolute risk prediction may assist in determining such thresholds.

4.1 | Limitations

Our study had several limitations. The LEOS registry is anonymized, and continuous parameters are categorized, thus potentially reducing the predictive performance of laboratory measures. As a real-world data set, given the heterogeneity of clinical procedures across centers, our analysis had to compensate for missing values. This typically reduces the predictive performance of the respective variables and the probability that they pass the stability selection criteria. Therefore, some predictors may have been underestimated or missed.

Our analysis was limited to predicting disease progression with information obtained at the time of the first positive SARS-CoV-2 testing (typically occurring during presentation at the medical facility), without considering the dynamics of the predictors. The days since the onset of symptoms (uncomplicated phase) to the diagnosis were included as a variable; however, the stability criteria were not met. In addition, there were differences between the validation and the test data set, with the latter having a higher proportion of patients diagnosed in the uncomplicated phase (suggesting earlier diagnosis, possibly due to expanded testing capacities after the first wave). Nevertheless, the score exhibited similar or better performance in the test set. This indirect evidence suggests that the application of our score may be valid for time points after diagnosis (or initial presentation), such as if the patient's condition or laboratory values deteriorate. Further studies are required to evaluate its suitability in such settings.

No information on patient race/ethnicity was available for this analysis, and it may be assumed that the distribution follows that in the German population and represents largely Caucasians, which may limit generalizability. External validation in different patient populations is therefore required, also with regard to socioeconomic factors and local standards of care.

Extensive information on comorbid conditions for study participants was available. Although some passed the criteria in RSF stability selection, none passed the RF stability criteria. However, more predictors (24 vs. 5) did not improve the overall predictive performance. This suggests that the increased risk due to these comorbidities may already be reflected by the remaining five predictors (collinearity), thereby relieving the need for inclusion in the score. However, this may not hold true for less common comorbidities, as the overall prediction improvement will be low for low prevalence predictors, even if they strongly affect the patients suffering from these comorbidities. A score based on the total population, as

presented here, may underestimate high-risk constellations due to rare comorbidities, such as specific cancers or autoimmune diseases/immunosuppressive treatments. To our knowledge, this limitation applies to most, if not all, available COVID-19 prognosis scores derived from the total population. Unfortunately, these patients may deteriorate rapidly. It is therefore important to establish the additional risk for such rare conditions in addition to the score used in future studies.

ACKNOWLEDGMENTS

The LEOSS registry was supported by the German Centre for Infection Research and the Willy Robert Pitzer Foundation. We express our deep gratitude to all study teams supporting the LEOSS study. The LEOSS study groups who contributed at least 5 per mille to the analyses of this study: University Hospital Freiburg (Siegbert Rieg), Hospital Ingolstadt (Stefan Borgmann), Technical University of Munich (Christoph Spinner), Klinikum Dortmund (Martin Hower), Johannes Wesling Hospital Minden (Kai Wille), Hospital Passau (Martina Haselberger), Hospital Ernst von Bergmann (Lukas Tometten), University Hospital Jena (Maria Madeleine Rührich), University Hospital Essen (Sebastian Dolff), University Hospital Frankfurt (Maria Vehreschild), University Hospital Heidelberg (Uta Merle), Hospital Bremen-Center (Christiane Piepel), University Hospital Regensburg (Frank Hanses), University Hospital Ulm (Beate Grüner), University Hospital Munich/LMU (Michael von Bergwelt-Baildon), University Hospital Cologne (Norma Jung), University Hospital Erlangen (Richard Strauß), Hacettepe University (Murat Akova), Bundeswehr Hospital Koblenz (Dominic Rauschnig), Hospital Leverkusen (Lukas Eberwein), Hospital Maria Hilf GmbH Moenchengladbach (Juergen vom Dahl), University Hospital Würzburg (Nora Isberner), Tropical Clinic Paul-Lechler Hospital Tübingen (Claudia Raichle), St. Josef-Hospital—Catholic Hospital Bochum (Kerstin Hellwig), University Hospital Tübingen (Siri Goepel), Municipal Hospital Karlsruhe (Christian Degenhardt), University Hospital Schleswig-Holstein—Kiel (Anette Friedrichs), Hospital Kreuznacher Diakonie Hunsrueck (Wolfgang Rimili), University Hospital Dresden (Katja de With), University Hospital Düsseldorf (Björn Jensen), Clinic Munich (Wolfgang Guggemos), Petrus Hospital Wuppertal (Sven Stieglitz), University Hospital Saarland (Robert Bals), Marien Hospital Herne, University Hospital Bochum (Timm Westhoff), Robert-Bosch-Hospital Stuttgart (Katja Rothfuss), University Hospital Bonn (Jacob Nattermann), Justus-Liebig-University Giessen (Janina Trauth), Hospital St. Joseph-Stift Dresden (Lorenz Walter), Sophien and Hufeland Clinic Weimar (Jessica Rüdell), Pamukkale University School of Medicine (Hüseyin Turgut), Hospital Universitari Arnau de Vilanova (Juan Antonio Schoenenberger-Arnaiz), Robert-Koch-Institute (Thomas Kratz), Hospital South-Eastern Bavaria AG Trostberg (Thomas Glück), Malteser Hospital St. Franziskus-Hospital Flensburg (Milena Milovanovic), Hospital Fulda (Philipp Markart), Oberlausitz-Hospital (Maximilian Worm), University Hospital Hamburg-Eppendorf (Sabine Jordan), Agaplesion Diaconia Hospital Rotenburg (David Heigener), St. Josef Hospital Kupferdreh (Ingo Voigt).

THE LEOSS STUDY INFRASTRUCTURE GROUP: Jörg Janne Vehreschild (Goethe University Frankfurt), Lisa Pilgram (Goethe University Frankfurt), Carolin E. M. Jakob (University Hospital of Cologne), Melanie Stecher (University Hospital of Cologne), Maximilian Schons (University Hospital of Cologne), Susana Nunes de Miranda (University Hospital of Cologne), Nick Schulze (University Hospital of Cologne), Sandra Fuhrmann (University Hospital of Cologne), Annika Claßen (University Hospital of Cologne), Bernd Franke (University Hospital of Cologne), Fabian Praßer (Charité, Universitätsmedizin Berlin) und Martin Lablans (University Medical Center Mannheim).

CONFLICT OF INTERESTS

Dr. Spinner reports grants, personal fees, and nonfinancial support from Gilead Sciences, grants and personal fees from Janssen-Cilag, personal fees from Formycon, other from Aperion, other from Eli Lilly, during the conduct of the study; personal fees from AbbVie, personal fees from MSD, grants and personal fees from GSK/ViiV Healthcare outside the submitted work. Dr. Rührich reports grants from the IZKF outside the submitted work. Dr. Vehreschild reports personal fees from Merck/MSD, Gilead, Pfizer, Astellas Pharma, Basilea, German Centre for Infection Research (DZIF), University Hospital Freiburg/Congress and Communication, Academy for Infectious Medicine, University Manchester, German Society for Infectious Diseases (DGI), Ärztekammer Nordrhein, University Hospital Aachen, Back Bay Strategies, German Society for Internal Medicine (DGIM), and grants from Merck/MSD, Gilead, Pfizer, Astellas Pharma, Basilea, German Centre for Infection Research (DZIF), German Federal Ministry of Education and Research (BMBF), (PJ-T: DLR), University of Bristol, Rigshospitalet Copenhagen. The remaining authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Jörg J. Vehreschild: Initiation and leading of LEOSS. **Jörg J. Vehreschild, Carolin E. M. Jakob, and Maximilian Schons:** Developing and maintaining LEOSS. **Stanislas Werfel, Christoph Schmaderer, and Christoph Spinner:** Conception of this study and critical data interpretation. **Stanislas Werfel:** Machine learning and statistical analyses, generation of tables and figures and manuscript preparation. **Carolin E. M. Jakob:** Data management, extraction, and additional statistical analyses. **Uwe Heemann and Jochen Schneider:** Data interpretation and critical revision of the manuscript. **Stefan Borgmann, Jochen Schneider, Martin Hower, Kai Wille, Martina Haselberger, Hanno Heuzeroth, Maria M. Rührich, Sebastian Dolff, Johanna Kessel, and Siegbert Rieg:** Acquisition of data; all authors revised and approved the final version of the manuscript.

ORCID

Stanislas Werfel  <http://orcid.org/0000-0001-9130-0057>

Christoph Spinner  <https://orcid.org/0000-0002-3875-5367>

Sebastian Dolff  <http://orcid.org/0000-0001-5549-1706>

REFERENCES

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382:727-733.
2. Pitlik SD. Covid-19 compared to other pandemic diseases. *Rambam Maimonides Med J*. 2020;11:e0027.
3. Short KR, Kedzierska K, van de Sandt CE. Back to the future: lessons learned from the 1918 influenza pandemic. *Front Cell Infect Microbiol*. 2018;8:343.
4. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob Heal*. 2020;8:e480.
5. Jakob CEM, Borgmann S, Duygu F, et al. First results of the "Lean European Open Survey on SARS-CoV-2-Infected Patients (LEOSS)". *Infection*. 2020;49:63-73.
6. Jakob CEM, Kohlmayer F, Meurers T, et al. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. *Sci Data*. 2020;7:435. <https://doi.org/10.1038/s41597-020-00773-y>
7. Ishwaran H, Kogalur UB. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package, version 2.9.3. 2020. <https://cran.r-project.org/package=randomForestSRC>
8. Tang F, Ishwaran H. Random forest missing data algorithms. *Stat Anal Data Min*. 2017;10:363-377.
9. Liao SG, Lin Y, Kang DD, et al. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*. 2014;15:346.
10. Kursu MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36:1-13.
11. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. Univ California Berkeley; 2004. 110, p. 24.
12. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841-860.
13. Cule E, Vineis P, De, Iorio M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*. 2011;12:372.
14. Perkins NJ, Schisterman EF. The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol*. 2006;163:670-675.
15. 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;369:369.
16. Liang W, Liang H, Ou L, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med*. 2020;180:1081-1089.
17. Vaid A, Somani S, Russak AJ, et al. Machine learning to predict mortality and critical events in COVID-19 positive New York City patients: a cohort study. *J Med Internet Res*. 2020;22:e24018.
18. Zhao Z, Chen A, Hou W, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PLOS One*. 2020;15:e0236618.
19. Galloway JB, Norton S, Barker RD, et al. A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: an observational cohort study. *J Infect*. 2020;81:282-288.
20. 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;370:22.
21. Wang X, Yang J, Wei C, et al. Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *Int J Epidemiol*. 2020;49:1918-1929.
22. Mortality Analyses—Johns Hopkins Coronavirus Resource Center. Accessed November 20, 2020. <https://coronavirus.jhu.edu/data/mortality>
23. Sperrin M, Mcmillan B. Prediction models for covid-19 outcomes. *BMJ*. 2020;371:3777.
24. Han E, Tan MMJ, Turk E, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. *Lancet*. 2020;396:1525-1534.
25. Razavian N, Major VJ, Sudarshan M, et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. *npj Digit Med*. 2020;3:130.
26. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584:430-436.
27. Ronco C, Reis T, Husain-Syed F. Management of acute kidney injury in patients with COVID-19. *Lancet Respir Med*. 2020;8:738-742.
28. Nadim MK, Forni LG, Mehta RL, et al. COVID-19-associated acute kidney injury: consensus report of the 25th Acute Disease Quality Initiative (ADQI) Workgroup. *Nat Rev Nephrol*. 2020;16:747-764.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Werfel S, Jakob CEM, Borgmann S, et al. Development and validation of a simplified risk score for the prediction of critical COVID-19 illness in newly diagnosed patients. *J Med Virol*. 2021;93:6703-6713. <https://doi.org/10.1002/jmv.27252>