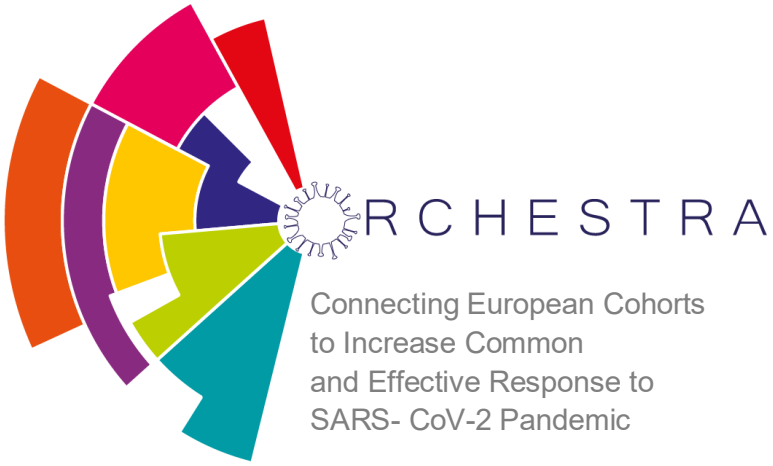


WP2 Deliverable 2.3

Task 2.5 Development of predictive models to inform therapeutic algorithms

University Hospital Cologne (UHC)

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Project Classification

Project Acronym:	ORCHESTRA
Project Title:	Connecting European Cohorts to Increase Common and Effective Response to SARS- CoV-2 Pandemic
Coordinator:	UNIVR
Grant Agreement Number:	101016167
Funding Scheme:	Horizon 2020
Start:	1st December 2020
Duration:	48 months
Website:	www.orchestra-cohort.eu
Email:	info@orchestra.eu

Document Classification

WP No:	WP2
Deliverable No:	D2.3
Title:	Development of predictive models to inform therapeutic algorithms
Lead Beneficiary:	UHC
Other Involved Beneficiaries:	All WP2 members, UNIVR, HMGU
Nature:	Report
Dissemination Level:	Public
Due Delivery Date:	31.05.2023
Submission Date:	31.07.2023
Justification of delay:	To increase accuracy of the analysis, to perform sensitivity analyses and include medical and statistical expertise from other beneficiaries.
Status:	Final
Version:	1.0
Author(s):	Katharina S. Appel, Sina M. Hopff, Ramsia Geisler, Anna Gorska, Elisa Gentilotti, Lorenzo M. Canziani, Evelina Tacconelli, Jörg J. Vehreschild

History of Changes

Version	Date	Created/Modified by
0.1	24.7.2023	Katharina S. Appel, Sina M. Hopff, Ramsia Geisler, Jörg J. Vehreschild
0.2	8.8.2023	Lorenzo M. Canziani
1.0	28.9.2023	Katharina S. Appel, Lorenzo M. Canziani, Jörg J. Vehreschild

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Table of contents

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Executive summary

A multitude of prognostic indices have been disseminated to facilitate risk categorization pertaining to the SARS-CoV-2 infection, commonly known as Coronavirus disease 2019 (COVID-19). For Task 2.5 / Deliverable 2.3 *Development of predictive models to inform therapeutic algorithms, w*e performed a systematic review to identify and assess clinical scores for confirmed or clinically assumed COVID-19 cases and performed a subsequent external validation using ORCHESTRA's Work Package 2 (WP2) long-term sequalae data.

An exhaustive evaluation and risk of bias (ROB) analysis were carried out, employing the *Prediction model Risk Of Bias ASsessment Tool* (PROBAST), for those scores which met predetermined criteria. From the 1,522 studies retrieved from the MEDLINE/Web of Science databases (as of 20th February 2023), a total of 242 scores were identified for prognostication of COVID-19 outcomes, including mortality (109 scores), disease severity (116 scores), hospitalization (14 scores), and long-term sequelae (3 scores). Examination of the predictors revealed a predilection towards the use of laboratory data and sociodemographic information in constructing the mortality and severity scores. It was observed that most of the scores were developed using retrospective cohorts (75.2%) or single-center cohorts (57.1%). Forty-nine scores were considered for the comprehensive analysis. The analysis yielded a diverse range of quality and predictor selection, with only five scores exhibiting a low risk of bias. Most of the scores raised some concern regarding the ROB and/or lacked robust validation, emphasizing the need for further refinement to prevent suboptimal performance and misclassification.

For external validation on ORCHESTRA's WP2 long-term sequelae data, we applied a mass validation design combining possible subpopulations, timings of predictor measurements, and outcomes. The ORCHESTRA WP2 cohort consists of both outpatients and inpatients diagnosed with SARS-CoV-2 infection. The performance was analyzed using discrimination and classification measures. External validation was possible for 39 (79.6%) of the scores that have been assessed in the in-depth analysis. Certain scores could not be validated due to the unavailability of specific information in the dataset or unusual predictor choices in the original studies. The dataset exhibited a significant number of missing values for the acute infection of patients (due to a possible enrolment after primary infection), particularly for specific biochemistry or vital sign assessments, and this, combined with the time variation of predictor assessment, substantially reduced the patient pool for score calculation. Sample sizes varied considerably, with larger sizes for scores encompassing demographic information, comorbidities, and vital signs, and smaller sizes for scores that included specific laboratory information or social and functional patient assessments. We provide a comprehensive analysis and tools to enable a selection of scores for different stratification demands, care settings and predictor availability.

Up to now, none of the scores has found entry into COVID-19 treatment guidelines. This analysis uncovers a gap in reliable COVID-19 predictive scoring systems may contribute to the application of scores across settings and regions.

Future pandemic preparedness could be enhanced through collaborative data sharing, unified score development concepts, and strict adherence to guidelines such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist, all of which would improve the transparency, reliability, and practical utility of predictive scores.

Dissemination level: Public.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Abbreviations

4C	4C Mortality Score
AFEM	African Federation for Emergency Medicine
AU(RO)C	Area under the (receiver operating characteristic) curve
BIPAP	Biphasic Positive Airway Pressure
CCEDRRN	Canadian Community Epidemiology and Drug Response Network COVID-19
CHARMS	CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies
CI	Confidence interval
CINECA	Consorzio Interuniversitario
COPS	COVID-19 prognosis score
COVID-19	Coronavirus disease 2019
CPAP	Continuous positive airway pressure
CRP	C-reactive protein
ED	Emergency department
EPV	Events-per-variable
FU	Follow-up
ICU	Intensive Care Unit
IDSA	Infectious Diseases Society of America
INSERM	Institut National de la Santé et de la Recherche Médicale
ISARIC	International Severe Acute Respiratory and emerging Infection Consortium's
IQR	Interquartile range
L1	Level 1 according to selection of scores for in-depth analysis (those not fulfilling L2 criteria)
L2	Level 2 according to selection of scores for in-depth analysis
LMIC-PRIEST	Low- and middle-income country Pandemic Respiratory Infection Emergency System Triage
LTS	Long-term sequelae
MV	Mechanical ventilation
NEWS	National Early Warning Score
NPV	Negative predictive value
OURMAPCN	Acronym using the predictor components of the score: Oxygen saturation, blood Urea nitrogen, Respiratory rate, admission before the date the national Maximum number of daily new cases was reached, Age, Procalcitonin, C-reactive protein (CRP), and absolute Neutrophil counts
PCC	Post-COVID condition
PPV	Positive predictive value
PRIEST	Pandemic Respiratory Infection Emergency System Triage
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PROBAST	Prediction model Risk Of Bias ASsessment Tool
qCSI	Quick COVID severity index
ROB	Risk of bias
SARS2	Acronym using the predictor components of the score: Sex, Age, Race, Socioeconomics status, Smoking status
SARS-CoV-2	syndrome coronavirus type 2
SAS	Andalusian Health Service
SD	Standard deviation
SEIMC	Spanish Society for Infectious Diseases and Clinical Microbiology

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
UHC	University Hospital Cologne
UNIBO	University of Bologna
UNIVR	University of Verona
WP2	Work package 2
WHO	World Health Organization

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has precipitated a state of emergency in healthcare systems worldwide. Hospital facilities were inundated with patients, necessitating expedited management decisions, while concurrently resource limitations obstructed the provision of sufficient therapies for all patients.[1] As of 2023, the COVID-19 pandemic has transitioned from an epidemic to an endemic state.[2,3] During this phase, the persistent and dynamic evolution of severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) variants, as well as the impact on immunity, vaccines, and therapeutic interventions, continue to be relevant factors.[4,5] Despite the progress made, some individuals, particularly those who are older and have comorbidities, still develop severe disease.[6]

During the initial surge of the pandemic, both scientific and clinical professionals rapidly accelerated their endeavours to bolster decision-making processes pertaining to the administration and therapeutic interventions for infected patients. This often involved the establishment of criteria delineating symptom severity or assessment scores. Such clinical prognostic scores originate from models which compute the likelihood of a specific condition for an individual by integrating multiple predictive factors, typically in a user-friendly format. The trade-off between information and precision incurred during the transition from a model to a score is counterbalanced by the enhanced practical utility. The rationale behind employing scores in general lies in their ability to expedite the acquisition of predictive skills, particularly for inexperienced clinicians, while also facilitating standardized communication among medical professionals and uniform estimation of risks in scientific endeavours. Clinical scores are routinely used as "clinical prediction rules", with the overall goal of enhancing patient care and mitigating severe consequences through adjusting therapeutic strategies based on the identified risks.[7] The design of these scores can lead to a variety of scenarios, such as predicting in-hospital mortality on hospital admission or hospitalization at the point of diagnosis, thereby rendering their application pertinent across diverse clinical settings.

While an abundance of predictive models for COVID-19 have been published[8,9] none of the scoring systems have demonstrated both applicability and reliability sufficient for universal implementation in routine clinical care and treatment protocols. The current Infectious Diseases Society of America (IDSA) guidelines (as of 05/2023) do not endorse a specific tool for outcome prognosis.[10] Similarly, the World Health Organization's (WHO) guidelines on Therapeutics and COVID-19 (as of 01/2023)[11] articulate the need for reliable tools, especially pertaining to the usage of available medication. Although it references the International Severe Acute Respiratory and emerging Infection Consortium's (ISARIC) 4C Mortality Score (4C),[12] the guidelines underscore the "need for better evidence on prognosis" and the imperative to validate prediction models in localized settings. The WHO's Living guidance for clinical management of COVID-19 (as of 01/2023)[13] also advocates for "clinical judgment […] rather than currently available prediction models for prognosis". However, it does recommend employing the National Early Warning Score (NEWS) 2 in screening for deterioration in COVID-19 pneumonia.[13] In summation, while the exigency for reliable stratification tools is underscored in COVID-19 guidelines, the evidence base for prognostic scores is insubstantial and their translation into clinical practice remains elusive.

For the identification of effective scores in the abundance of current COVID-19 literature,[8,9] thorough reviews and external validations are important to provide effective overviews for decision makers. In this report regarding task 2.5 *Development of predictive models to inform*

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

therapeutic algorithms, we identify, characterize, and validate a set of scoring systems for COVID-19 in the large, prospective ORCHESTRA dataset. Our project focuses on a critical evaluation of predictors and the transferability of clinical scores across settings or regions. This analysis may facilitate implementation in routine care, may guide therapeutic decisions and pave the way for enhanced preparedness for potential future pandemics.

Methods

Systematic review

Systematic review question, inclusion, and exclusion criteria

This systematic review identified prognostic clinical scores for COVID-19 developed from the inception of the pandemic in late 2020 until February 2023. The review incorporated original scores that were either designed or modified for the management or treatment of COVID-19, based on individual patient data from clinically presumed or confirmed COVID-19 cases. Parameters such as the level of patient care, timing of predictor measurement, prediction interval, predictor types, or COVID-19-related outcomes were not pre-selected. We excluded models based on regression or other predictive techniques that were not specifically designed for clinical scoring, as well as single predictors based on single observations. Scores constructed for distinct subpopulations (e.g., those with comorbidities, participants in pharmaceutical trials), and mathematical virus transmission simulations were also omitted (refer to **Table S1** for additional information).

In the initial stage, we extracted pertinent information from all identified studies that met the primary inclusion criteria (henceforth referred to as "all scores"). In the second stage, we selected scores for a more comprehensive analysis (Level 2, L2) based on pre-defined criteria: (I) a reported area under the (receiver operating characteristic) curve (AUC) of ≥ 0.75; (II) the report of a separate validation cohort as the minimal validation procedure; (III) the development in a multi-center setting (≥ 2 centers); (IV) a points-based application (for further details refer to **Table S1**). Scores that did not meet the primary inclusion criteria (Level 1, L1) were not subjected to further evaluation. Scores that satisfied these criteria (L2) were subsequently examined in detail and appraised for their risk of bias (ROB).

Data sources, search strategy and data extraction

We performed repeated searches in PubMed/MEDLINE and Web of Science on 14 April 2022 and 20 February 2023 employing a predetermined search strategy that integrated search blocks related to the terms "COVID-19", "Prediction", "Scoring", and "Validation metrics" (for additional details, refer to **Table S2**). We utilized the PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*)[14] guidelines, as well as a modified version of the *CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies* (CHARMS) checklist[15] (refer to **Supplementary text S1** for extracted information). Two reviewers (KA, RG) independently carried out each step, encompassing screening, data extraction, and assessment of the risk of bias (ROB). In cases of disagreement, consensus was achieved through discussion.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Unless otherwise specified, the unit of analysis was a single score per outcome and predictor set for the ROB analysis. As the design for the external validation (see “External Validation”) itself aimed for multiple outcome testing, the unit of analysis for the external validation is a single score.

The extracted AUCs are conveyed either as a range or as a median with an interquartile range (IQR); categorical data is reported as absolute numbers and percentages (n (%) respectively). The sample size was evaluated using the (estimated) events-per-variable (EPV), with lower EPVs indicating a heightened risk of overfitting (for further details, refer to **Supplementary text S1**).

Outcomes and categorization of scores

The examined publications applied the following outcomes: fatal outcomes (in-hospital mortality, death within specified time intervals), disease severity (characterized as composite outcomes such as the necessity for mechanical ventilation, intensive care unit (ICU) admission, or death), hospitalization, and the Post-COVID condition (PCC). We categorized the scores based on the type of outcome and the timing of predictor measurement (as displayed in **Table 1**).

Table 1: Categories by timing of predictor measurement and outcome

No.	Category
1	First/early contact to health care facility → Death
2	First/early contact to health care facility → Deterioration (severity, ICU admission, need for mechanical ventilation, respiratory complication, specific organ failures or death as composite endpoint etc.)
3	Severe disease or ICU admission → Deterioration or death
4	First diagnosis and contact to out-patient health care facility → Hospitalization
5	Acute infection → PCC

Contact to health care also includes hospital or emergency department admission or the first diagnosis by Sars-CoV-2 Testing. Intensive care unit (ICU); Post-COVID-19 condition (PCC).

Risk of bias assessment

Deficiencies in the design, execution, or analysis methodologies of a study can induce systematic errors, or bias, in effect estimates. The *Prediction model Risk Of Bias ASsessment Tool* (PROBAST)[16] provides guidance for assessing the adequacy of methods addressing potential biases during the development of a clinical prediction rule. It assesses and classifies the risk of bias within its four subdomains: "participants", "predictors", "outcome", and "analysis". The ratings "low", "unclear", or "high" evaluate the validity of the study and consolidate into an overall risk of bias. If at least one question or subdomain receives a "high" rating, the overall risk of bias is classified as "high", in accordance with the rules of the PROBAST guideline.[16]

External validation

Population and data collection

ORCHESTRA (Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic), is a project financed under the Horizon 2020 (H2020) initiative. The

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

primary goal of this project is to counteract the challenges engendered by the COVID-19 pandemic through the establishment of an international large-scale cohort study, intended to generate profound and harmonized scientific evidence pertaining to the prevention and treatment of SARS-CoV-2 infection. Work Package 2 (WP2) under ORCHESTRA is dedicated to developing a multi-country prospective observational cohort that undergoes regular follow-up assessments from the point of SARS-CoV-2 diagnosis. An initial inclusion during FU with a retrospective documentation of the acute infection is possible. Clinical, virological, biochemical, and immunological data were systematically documented using a pre-defined and harmonized format in the REDCap (Research Electronic Data CAPpture) tool, a specialized electronic data capture instrument, hosted at Consorzio Interuniversitario (CINECA).

For the presented analysis, we used patients from ORCHESTRA's WP2 with polymerase-chain reaction confirmed SARS-COV-2 infection recruited from 2020/02/07 to 2023/05/04 in five multi-centred cohorts (UNIVR (University Hospital of Verona, Italy), UNIBO (University Hospital of Bologna, Italy), INSERM (National Institute of Health and Research, France), CovidHOME (University of Groningen, the Netherlands), SAS (Andalusian Health Service, Spain). While most of the included patients were hospitalized during the acute infection, COVID HOME comprises only of non-hospitalized COVID-19 cases. For the external validation procedures, we used outcomes within a time interval from primary infection up to Follow-Up (FU) at month 12 of these patients.

Inclusion criteria were age ≥18 years and the availability of information about the primary infection of the patient.

Feasibility and mapping of scores

For most scores, a 1:1 mapping of required information to the ORCHESTRA dataset was possible. Some scores could not be mapped, others were only feasible when applying minor assumptions (for details see Table **S3**).

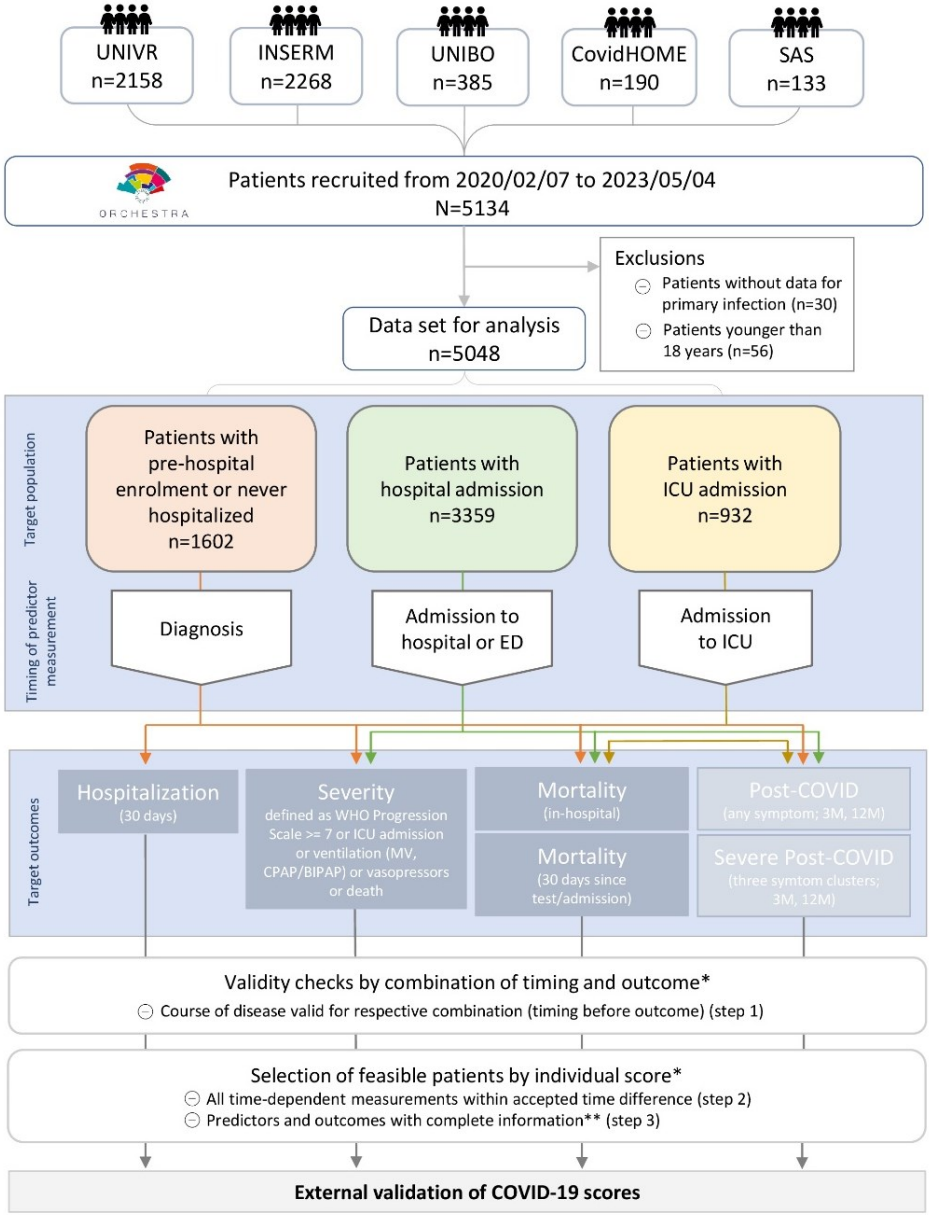
Outcomes, mass validation design, and work flow

To generate comparable results, we designed synthesized outcomes and did not predict the exact outcome defined within a publication. The synthesized outcomes were 30-days hospitalization, severity (defined as WHO Progression Scale ≥ 7 or ICU admission or ventilation (MV, CPAP/BIPAP) or vasopressors or death) and fatal outcomes (in-hospital death; survival 30 days after testing or hospital admission). Particularly "severity" scores used heterogenous definitions and composite outcomes (including e.g. ICU admission or mechanical ventilation) to describe a severe course of disease.

A mass validation design was established to create a comprehensive environment for multi-design testing (**Figure 1**). Starting from the distinct sub-populations that can be distinguished within the WP2 cohort, there are three related timings of predictor measurement that apply, respectively: (i) diagnosis or testing, (ii) admission to hospital or emergency department (ED), and (iii) admission to ICU. Depending on the subpopulation and the respective time in the course of disease, there are different outcomes that offer to be predicted: hospitalization, severity, fatal outcome or PCC, respectively. A score was tested for all combinations of predictor measurement and outcomes that make sense in terms of context, irrespective of the original design of the score.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

While the focus of the analysis are short-term outcomes, we also tested whether the scores designed to predict adverse outcomes for the primary infection of COVID-19 could predict the PCC at month three and 12, as severe disease was shown to be risk factor of PCC[17]. PCC was defined as in a previous ORCHESTRA study[18], with the presence of at least one COVID-19



related symptom cluster as “PCC” and the presence of three clusters (respiratory, chronic pain, chronic fatigue) as “severe PCC”, both at month three and 12 after primary infection. For PCC prediction, we only included patients with a documented FU at month three and 12.

Figure 1: Mass validation design.

Note that the subpopulations are not necessarily distinct (e.g., might be enrolled in an outpatient setting a must be hospitalized later). The time-variation of predictor assessment is then accounted for in the “timing of predictor measurement” section of the workflow. After iterating all technically possible combinations, we reduced the resulting table to those courses of disease that make sense in terms of content. *Individual sample sizes for each score and scenario due to selection steps and missing values per predictor.

To ensure a clean analysis, we introduced filter steps selecting (i) only patients where the timing of predictor measurement was at least one day before the outcome occurred and (ii) all time-

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

dependent measurements within the accepted time interval. The default time interval of predictor measurement for the respective point in time was 48 hours (see sensitivity analyses). If a measurement was performed outside the time interval, the value for the scenario was set as missing and the patient was not included in the respective score validation.

The calculation of a score is dependent on the completeness of all components. Therefore, the available sample size applicable to a particular score and scenario within the mass validation design depends on the number, time-dependence, and missing status of each variable. For instance, if a score primarily uses sociodemographic information and comorbidities for outcome prediction (not time-dependent, good documentation rate), the number of patients the score can be validated on is higher than for a score with multiple laboratory or vital signs (time-dependent, mixed documentation rate) incorporated.

Data and statistical analysis

Summary statistics are presented with mean and standard deviation (SD) for continuous and with number and percent (n (%)) for categorial variables.

The ensuing discrimination measures serve as the key metrics for inter-score comparison.[7] The AUC is the most used (discriminatory) performance indicator and describes the likelihood that, when presented with one individual who has experienced the outcome and one who hasn't, the model will allocate a superior predictive probability to the individual with the outcome as opposed to the one without. An AUC of 0.5 is equivalent to arbitrary estimations, whereas an AUC of 1 is perfect discrimination. With an AUC of 0.75 an higher, the prediction is useful and reliable.[19]

We identified a data-based threshold for each score using the Youden Index and calculated sensitivity, specificity, accuracy and positive (PPV) and negative predictive values (PPV). If the original publication suggested a single threshold, resulting in a binary risk classification, it is tested (in contrast to multiple risk groups).

Additionally, the overall performance can be assessed using measures quantifying the distance between observed and predicted outcomes, such as the Brier score,[7] where 0 represents a perfect match between predicted probabilities and observed outcomes, and 1 represents a bad match.

The statistical analysis was performed using R 4.2.2.

Handling of missing values

Data imputation was executed exclusively in instances where data could be inferred from known information or to reduce missing information based on the requirements of the study protocol, such as missing biochemistry units based on the cohort's usual use.

Sensitivity analyses

Sensitivity analyses were performed using an accepted time interval of seven days for predictor measurements. By reducing the selection criteria, both the sample size per combination and the heterogeneity increased.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Results

Systematic review

The evaluation procedure is illustrated by the PRISMA flow chart (**Figure 2**). Out of 1,522 studies procured from the database, 242 original COVID-19 scores satisfied the primary inclusion criteria (L1), and 49 met the criteria (L2) (refer to **Table S4** for details on all scores and **Table S5** for details on L2). Reasons for exclusion were AUC ≤ 0.75 or AUC missing (17.4%), separate validation cohort missing (42.6%), a single-center setting for model development (56.6%), use of another approach than a points-based application (e.g., formula) (27.7%) (Note that multiple reasons for exclusion per score are possible). Comparative summary statistics correlating to this section are displayed in **Table 2**.

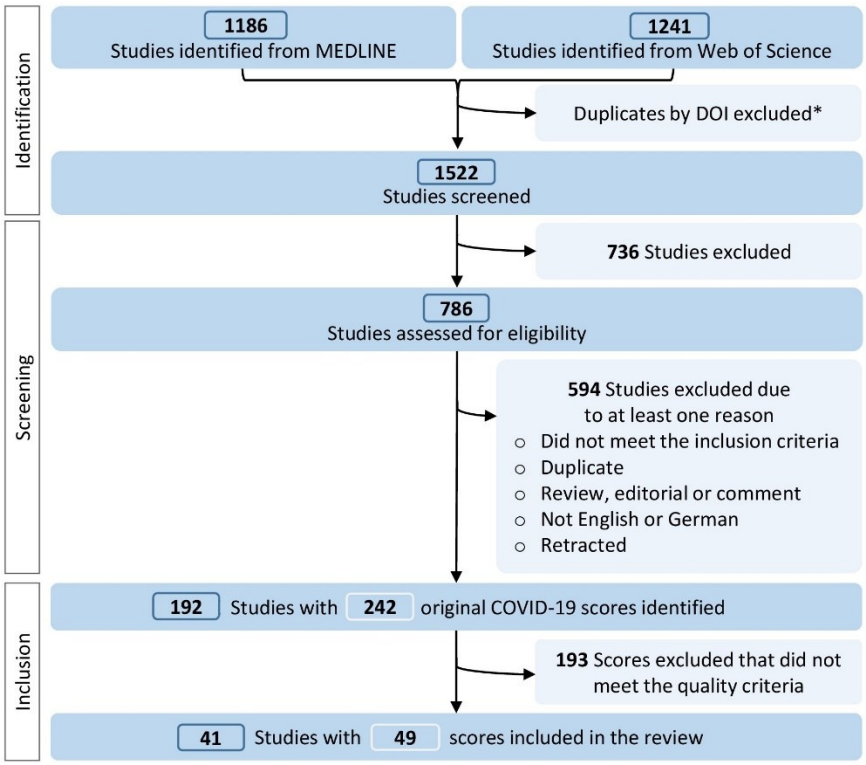


Figure 2: PRISMA flow chart.

The criteria for score selection were as follows: (I) a reported area under the (receiver operating characteristic) curve (AUC) of ≥ 0.75; (II) the report of a separate validation cohort as the minimal validation procedure; (III) the development in a multi-center setting (≥ 2 centers); (IV) a points-based application (for further details refer to **Table S1**).

Data basis and general study characteristics (for all scores)

All studies were published within the timeframe of 2020 to 2023. Most of the scores were developed based on cohorts with n<1,000 participants (64.0%) in a retrospective (75.2%) and/or single-center (57.1%) design. Fifty-seven percent of the models underwent validation in a

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

separate cohort, which included random splits, temporal, and/or geographical (external) validation. The median AUC was 0.83, with an IQR of [0.77, 0.90].

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Table 2: Characteristics of the included scores.

Characteristics	All	Level	
	N = 242 n (%)	Level 1, N = 193 n (%)	Level 2, N = 49 n (%)
Category			
1 First/early contact to health care facility → Death	100 (41.3%)	79 (40.9%)	21 (42.9%)
2 First/early contact to health care facility → Deterioration	112 (46.3%)	94 (48.7%)	18 (36.7%)
3 Severe disease or ICU admission → Deterioration or death	13 (5.4%)	12 (6.2%)	1 (2.0%)
4 First diagnosis and contact to out-patient health care facility → Hospitalization	14 (5.8%)	5 (2.6%)	9 (18.4%)
5 Acute infection → PCC	3 (1.2%)	3 (1.6%)	0 (0.0%)
Study design			
Prospective	33 (13.6%)	30 (15.5%)	3 (6.1%)
Retro- and prospective	12 (5.0%)	3 (1.6%)	9 (18.4%)
Retrospective	182 (75.2%)	150 (77.7%)	32 (65.3%)
Unknown	15 (6.2%)	10 (5.2%)	5 (10.2%)
Multi-center design			
≥ 2 centers	103 (42.9%)	54 (28.3%)	49 (100.0%)
Samples size			
Cumulative number of participants ≥ 1,000	87 (36.0%)	47 (24.4%)	40 (81.6%)
Estimated events per variable[a] (Median, IQR)	-	-	15.6 (6.6, 267.3)
Health sector			
Hospitals/emergency department	216 (89.6%)	182 (94.8%)	34 (69.4%)
In- or outpatient sites	16 (6.6%)	3 (1.6%)	13 (26.5%)
Outpatient sites	7 (2.9%)	5 (2.6%)	2 (4.1%)
Other	2 (0.8%)	2 (1.0%)	0 (0.0%)
Population			
Patients in emergency department	25 (10.3%)	18 (9.3%)	7 (14.3%)
In-patients with severe disease	37 (15.3%)	35 (18.1%)	2 (4.1%)
In-patients without restriction to specific conditions[b]	158 (65.3%)	132 (68.4%)	26 (53.1%)
Inhabitants of one region	1 (0.4%)	1 (0.5%)	0 (0.0%)
Out- and inpatients	11 (4.5%)	2 (1.0%)	9 (18.4%)
Outpatients	10 (4.1%)	5 (2.6%)	5 (10.2%)
Study/recruitment time			
2020	-	-	38 (77.6%)
2020-2021	-	-	4 (8.2%)
2020-2022	-	-	7 (14.3%)
Country			
China	45 (18.6%)	39 (20.2%)	6 (12.2%)
Italy	25 (10.3%)	24 (12.4%)	1 (2.0%)
USA	33 (13.6%)	18 (9.3%)	15 (30.6%)
Other	139 (57.4%)	112 (58.0%)	27 (55.1%)
Timing of predictor measurement			
Admission to hospital or ED	190 (79.8%)	159 (84.1%)	31 (63.3%)
Admission to ICU	7 (2.9%)	6 (3.2%)	1 (2.0%)
SARS-CoV2-Testing/diagnosis	13 (5.5%)	12 (6.3%)	1 (2.0%)
Other	28 (11.8%)	12 (6.3%)	16 (32.7%)
Outcomes			
Deterioration or death	112 (46.3%)	94 (48.7%)	18 (36.7%)
Death (single endpoint)	113 (46.7%)	91 (47.2%)	22 (44.9%)
Hospitalization	14 (5.8%)	5 (2.6%)	9 (18.4%)
Post-COVID condition	3 (1.2%)	3 (1.6%)	0 (0.0%)
Handling of missing values			
Any imputation method applied	-	-	19 (38.8%)
Multiple Imputation	-	-	11 (22.4%)
Modelling technique			
Cox, (Bayesian) Logistic, or LASSO Regression	-	-	41 (83.7%)
Machine learning	-	-	2 (4.1%)
Mixed methods or other	-	-	6 (12.2%)
Validation[c]			
Separate cohort present	138 (57.0%)	89 (46.1%)	49 (100%)
Geographical validation	-	-	10 (20.4%)
Temporal validation	-	-	17 (34.7%)

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Temporal and geographical validation	-	-	7 (14.3%)
Random split	-	-	13 (26.5%)
Validation with different population characteristics	-	-	1 (2.0%)
Independent external validation	-	-	2 (4.1%)
Discrimination			
AUC of the strongest validation ≥ 0.75	190 (78.5%)	141 (73.1%)	49 (100.0%)
AUC (Median, IQR)	0.83 (0.77, 0.90)	0.84 (0.77, 0.91)	0.81 (0.80, 0.85)
Calibration[c]			
Any method applied	-	-	30 (61.2%)
Calibration plot or table	-	-	23 (46.9%)
Hosmer-Lemeshow	-	-	12 (24.5%)
Application			
Formula	65 (26.9%)	65 (33.7%)	0 (0.0%)
Points-based and formula	172 (71.1%)	123 (63.7%)	49 (100.0%)
Formula	3 (1.2%)	3 (1.6%)	0 (0.0%)
Other	2 (0.8%)	2 (1.0%)	0 (0.0%)

We present n (%) for categorical information and the median (interquartile range) for continuous information. The column "All" includes all scores fulfilling the *a priori* inclusion criteria, whereas Level 1 merely includes scores that did not fulfill the selection criteria and Level 2 only includes the scores fulfilling the respective criteria (see Methods section). Resulting from two granularity levels of data extraction, some information is only available for Level 2 scores.

[a] Events per variable (EPV) were estimated using the absolute number of candidate predictors. Some studies did not precisely name the number of candidate predictors. To generate assumptions regarding the sample size, we counted predictors indicated as candidate in tables or texts (signed by "~" in table S5), even though we acknowledge the fact that it is more precise to use the number of regression coefficients instead[16].

[b] Regarding population characteristics, "severe disease" includes ICU patients and patients with respiratory complication, pneumonia, intubation or other severe conditions.

[c] Multiple options possible.

The study populations incorporated hospitalized cases without limitations to specific conditions (65.3%), focused on patients with severe disease (15.3%), or patients admitted to the ED (10.3%). The principal point in time for prediction occurred during admission to the hospital or ED (79.8%). Predicted outcomes for all scores can be categorized as mortality (only) (45.0%), severity (47.9%), hospitalization (5.8%), or the PCC (1.2%).

Among the 188 distinct predictors (extracted from all evaluated scores), age (68.2%) emerged as the most included, followed by C-reactive protein (CRP) (29.8%). This trend was also observed in scores predicting mortality or severity, where the significance of laboratory data, demographics, and physiological information was evident. Hospitalization scores frequently incorporated age (87.8%) and dyspnea (57.1%). The most common comorbidities were diabetes mellitus type 2 and hypertension. The number of predictors per score varied from two to 29. **Figure 3** illustrates the relationship between predictor frequency and AUC. We also visually represent the predictor classification by score, category, and inclusion level (**Figure 4**, **Figure S1**, **Figure S2**). We did not detect a significant shift in predictor composition when comparing L1 and L2 scores.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

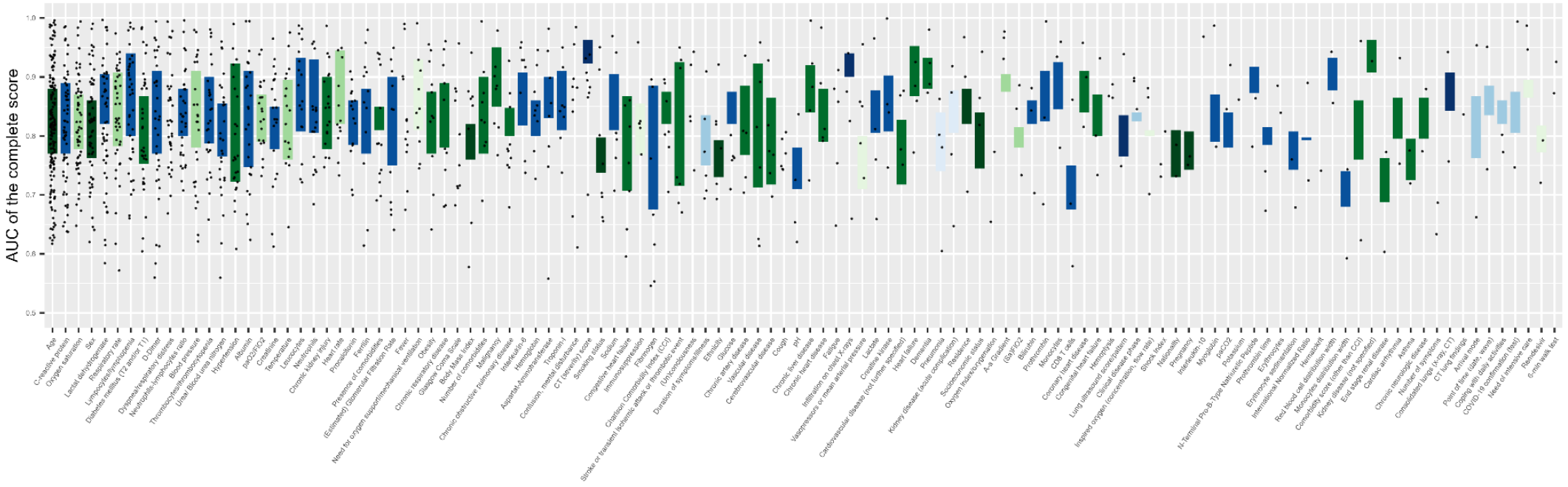


Figure 3: Relationship between predictor frequency within all scores (irrespective of category) and area under the curve (AUC). The predictors were grouped by predictor classification. Legend on the right.

Please note that some predictors may be included twice or multiple times. However, if the AUC is missing, only one point is displayed in the figure. Predictors that were included only once are not shown in the figure: Alcohol consumption, Employment, Tidal volume, Partial pressure (p50c), Interleukin-8, Cytokinemia, Endocan, Proteinuria, Serum protein, Bicarbonate, Triglycerides, Alanin-Aminotransferase, Alkaline phosphatase, Antithrombin, Partial thromboplastin time, Protein C, Blood group, No. of cytopenia, Macrophage activation, Corpuscular volume, Immature granulocytes-to-lymphocyte ratio, Delta-Hemoglobin-Equivalent, Immunoglobulin M, CD3+CD4+, CD3+CD8+, CD4, CD24+CD38lo/- B cells, Naïve CD4+ T cells, CD16+/CD56+ NK cells, Fluorescence of CD57 in CD8+ T cells, Cycle threshold (PCR), Acute kidney injury, Myocardial infarction, Hypotension, Myocardial injury, Chronic cardiopulmonary disease, Epilepsy, Sensory polyneuropathy, Hematological dysfunction, Coagulopathy, Wheezing, Tachypnoea, Sore throat, Rhinorrhea, Respiratory wrestling signs, Chest pain, Diarrhea, Nausea, Lung parenchymal involvement, Chest radiography abnormality, RALE score, Need for hospitalization, Prolonged colonization, Number of past ICU days, Vaccination, Central venous catheter, Statines, ACE inhibitors, Tocilizumab, Corticosteroids, Need for intubation, Hemodialysis, Pneumothorax, Ischemia, Multifocal colonization, Haemophagocytosis, Hepatomegaly/splenomegaly, Clinical Frailty Scale, Modified Medical Research Council (mMRC), Braden scale, Norton scale, DIC-ISTH score.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

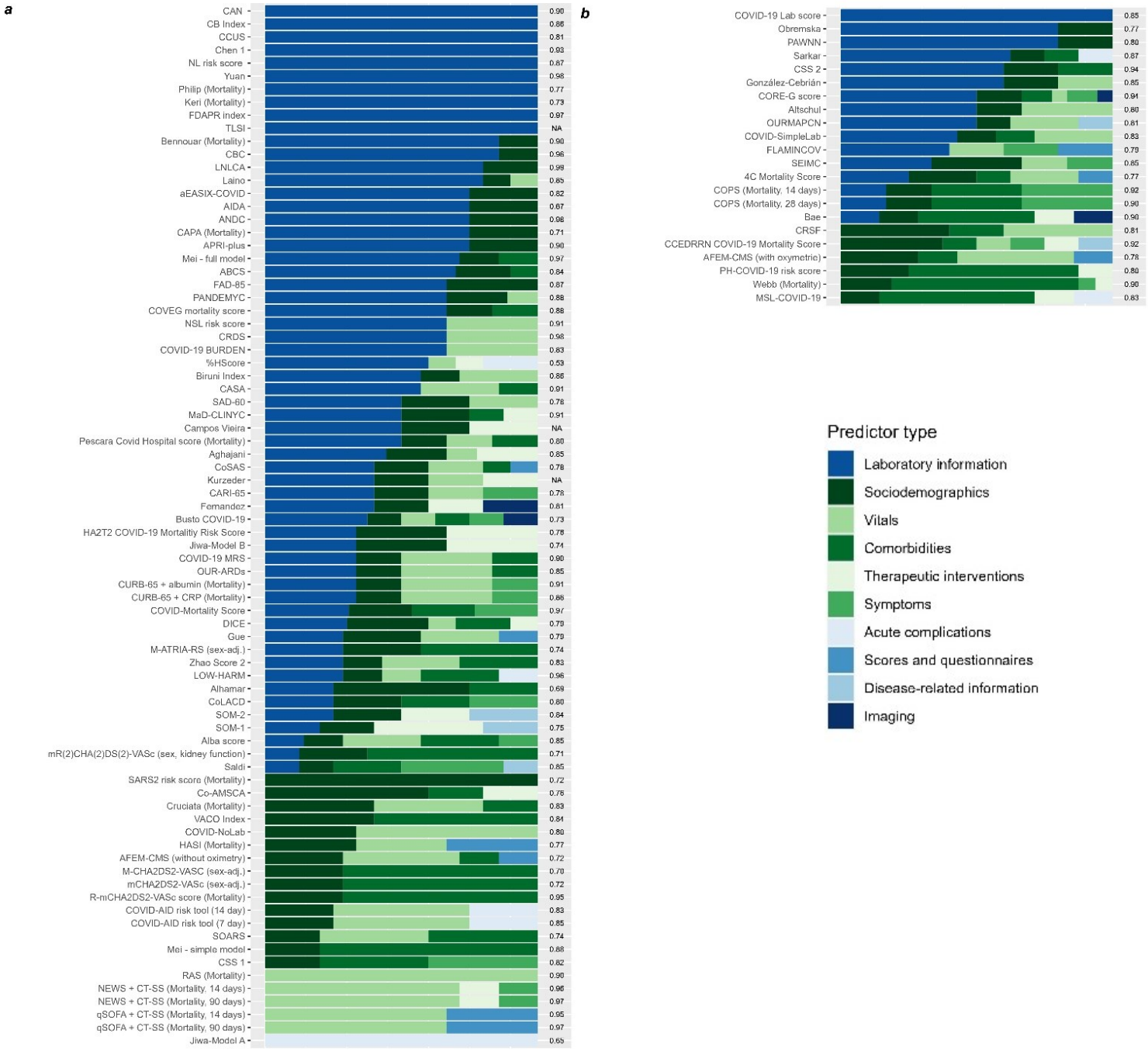


Figure 4: Predictor composition aggregated by predictor type for all scores assigned to category 1 stratified by the level of selection.

(a) Level 1, (b) Level 2. The AUC is displayed on the right. The sorting of the scores is determined by (I) the absolute number of categories and (II) the relative proportion across all scores. The color gradient from green to blue indicates the availability of the category, although in case of doubt this also depends on the level of care. Similar presentations of scores assigned to category 2 to 4 are displayed in the supplementary material S1 and S2.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Characteristics of selected scores (Level 2)

The most prevalent outcome was (in-hospital) mortality as a standalone outcome (mortality outcomes within categories 1 and 3 as per Table 1: 44.9%). Thirty-seven percent of scores predicted a composite outcome of "severity" (category 2 and 3). Among the scores exclusively predicting mortality and severity, between 0.4% to 51.2% and 3.7% to 51.6% of the patients within the development cohorts reached the respective outcomes. The estimated EPV ranged from 0.9 to 709.8, with 47.5% of the scores having an eEPV less than 10. These scores achieved a median AUC of 0.81 (IQR=[0.80, 0.87]).

Nine scores predicting hospitalization (category 4: 18.4%) met the L2 criteria with 4.0 to 38.9% of patients experiencing this outcome. These scores benefited from a large sample size with EPVs ranging from 15.6 to 120.7. The median AUC was 0.84 (IQR=[0.80, 0.85]).

The scores identified for predicting PCC (category 5) mostly utilized symptom information. However, none of them met the L2 criteria of AUC ≥ 0.75 and were therefore not further investigated.

Risk of bias

Numerous studies failed to comply with established guidelines,[15,16,20] resulting in the exclusion of critical information necessary for proper evaluation. Most scores elicited at least one issue across the four domains (participants, predictors, outcomes, analysis) within PROBAST, culminating in an overall high risk of bias (low 10.2%; unclear 6.1%; high 83.7%) (refer to **Figure S3**, **Table S6**). The primary source of concern predominantly pertained to the "analysis" domain. In particular, the concerns pertained to the absence of calibration measures[21], failure to adjust for over-optimism,[16,22] inappropriate management of missing values, and insufficient validation methodologies.[22]

Table 3 provides an overview of scores with an overall risk of bias (ROB) rating classified as "low" or "unclear". Five scores were evaluated as having an overall "low" ROB: The ISARIC's 4C Mortality Score (4C),[12] the Canadian Community Epidemiology and Drug Response Network COVID-19 (CCEDRRN)'s Mortality Score,[23] and the Spanish Society for Infectious Diseases and Clinical Microbiology (SEIMC)[24] score for mortality prediction, the (Pandemic Respiratory Infection Emergency System Triage study)'s PRIEST score[25] and the Low- and Middle-Income Countries PRIEST (LMIC-PRIEST)[26] for severity prediction.

Additionally, three scores received an overall "unclear" ROB rating (African Federation for Emergency Medicine (AFEM)[27], OURMAPCN[28], and SARS2[29] scores). These scores share certain features: their aggregate sample size was relatively large (except for AFEM), they utilized missing value imputation (except for SEIMC) and calibration measures (except for AFEM). During the analysis, potential overfitting and data complexities were addressed. All scores incorporated age, most including gender, and indicators of respiratory function as predictors. Among laboratory indicators, markers of kidney function were predominantly included.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Table 3: Characteristics of scores with low or unclear *Risk of Bias* rating.

Score	Reference	Study design	Population	Outcome	Year(s)[a]	Sample size (DC)	Sample size (VC(s))	No. of outcomes[b]	No. of final predictors	Predictors (description)	Strongest type of validation reported	Performance of strongest validation reported AUC (95 %-CI)	Overall ROB
4C Mortality Score	Knight et al. 2020[12]	Prospective observational cohort	Inpatients with "high likelihood" of COVID-19	Mortality (in-hospital)	2020	35,463	22,361	11,426	8	age, sex, number of comorbidities, RR, SaO2, GCS, urea level, CRP	Temporal validation with geographic subsetting	0.77 (0.76-0.77)	low
AFEM-CMS - with SaO2	Pigoga et al. 2021[27]	Retrospective observational cohort	Inpatients with suspected, probable, or confirmed COVID-19	Mortality (in-hospital)	2020	374	93	239*	7	sex, age, number of comorbidities, GCS, systolic BP, RR, SaO2	Random split of with cross-validation	0.78 (0.74-0.81)	unclear
CCEDRRN COVID-19 Mortality Score	Hohl et al. 2022[23]	Retrospective observational cohort	ED patients with confirmed or suspected COVID-19	Mortality (in-hospital/ ED)	2020-2021	6,758	2,054	471	8	age, sex, type of residence, arrival mode, chest pain, severe liver disease, RR, level of respiratory support	Geographical validation (same country, different centers)	0.92 (0.90-0.93)	low
LMIC-PRIEST	Marincowitz et al. 2022[26]	Observational cohort study	ED patients with suspected or confirmed COVID	Mortality (in-hospital), intubation, NIV or ICU adm. (30 days)	2020-2022	305,564	140,520+ 20,698	12,610	11	RR, SaO2, heart rate, systolic BP, temperature, alertness, inspired oxygen, sex, age, diabetes, heart disease	Geographical validation (other country)	0.79 (0.79-0.80)	low
OURMAPCN	Chen et al. 2021[28]	Retrospective observational cohort	Inpatients with confirmed COVID-19	Mortality (in-hospital)	2020	6,415	6,351+ 2169+ 553	462	8	CRP, SpO2, admission date, age, BUN, RR, procalcitonin, neutrophils	Geographical validation (different country)	0.81 (0.76-0.86)	unclear
PRIEST	Goodacre et al. 2021[25]	Retrospective observational cohort	Inpatients with suspected covid-19	Death or organ support (30 days)	2020	11,773	9,118	2,421	9	age, sex, RR, systolic BP, SaO2/inspired oxygen ratio, performance status, consciousness, renal impairment, respiratory distress	Geographical validation (same country, different centers)	0.80 (0.79-0.81)	low
SARS2 risk score	Dashti et al. 2021[29]	Retrospective observational cohort	Out- and inpatients with confirmed COVID-19	Hospitalization (30 days)	2020	10,496	1,851	3,197	5	age, sex, race, socioeconomic status, smoking	Validation with different population (medical staff)	0.77 (0.73-0.80)	unclear
SEIMC	Berenguer et al. 2021[24]	Retrospective observational cohort	Inpatients with confirmed COVID-19	Mortality (30 days)	2020	4,035	2,126	1,047	6	age, SaO2, NLR, GFR, dyspnea, sex	Temporal and geographical validation (same country, different centers)	0.85 (0.82-0.87)	low

Further information on the selected set of scores and for all scores assessed in Level 2 are presented in supplementary table S4. Area under the receiver operating characteristic curve (AUC), blood pressure (BP), blood urea nitrogen (BUN), C-reactive protein (CRP), development cohort (DC), emergency department (ED), Glasgow Coma Scale (GCS), glomerular filtration rate (GFR), respiratory rate (RR), Non-invasive ventilation (NIV), neutrophils-lymphocytes-ratio (NLR), oxygen saturation (SaO2), validation cohort (VC). [a] recruitment year; [b] Cumulative sample size consists of development cohort plus validation cohort(s). Outcomes in the development cohort or in the whole cohort (*) if not otherwise stated.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

External validation

Population characteristics

During the included study period (February 2020 to May 2023), the cohorts COVIDHome (n=190), INSERM (n=2268), SAS (n=133), UNIBO (n=385), and UNIVR (n=2158) contributed to an overall samples size of 5,134 patients with confirmed COVID-19 infection. The analysis was performed on 5,048 adults with documented primary infection (**Figure 1**).

Patients had a mean age of 59.44 (SD: 20.01); 2,193 (43.4%) were female. Among all patients, 1,602 had an out-patient status at enrollment, and 3,359 and 932 were hospitalized or admitted to ICU in the course of the disease, respectively. The most frequent comorbidities were hypertension (35%), diabetes mellitus (13%), and asthma (8%). Further details on the patient characteristics stratified by subpopulation are provided in **Table S7**.

Feasibility of scores

The mapping, thus the external validation, was possible for 39 (79.6%) of the scores described in the systematic review part of the analysis. The remaining scores were not feasible due to specific information not available in the ORCHESTRA dataset: Arrival mode (Self; ambulance or police), hemoptysis, activities of daily living (ADL) scale, Norton scale (a performance status measurement), haemocytometric parameters (e.g., reactive lymphocytes), congenital heart disease, mean corpuscular volume, and comorbidity cardiac arrhythmia. All scores including urea or blood urea nitrogen (BUN) were only feasible due to additional data sharing of INSERM.

The datasets contained a considerable number of missing values, especially for (specific) biochemistry or vital sign assessments. The missingness was dependent on the cohorts contributing to ORCHESTRA's WP2. This – in addition to the time variation of predictor assessment – significantly reduced the number of patients on which the score calculation was possible. Even though the number of patients in each sub population was considerable, the final sample size ranged from n=0 to n=3,358 patients, depending on the included predictor availability. The highest samples sizes were yielded for those scores that merely included demographic information, comorbidities, and vital signs. Low samples sizes were observed for scores including specific (such as Interleukin-6 or activated Partial Thromboplastin Time) or heterogeneously documented laboratory information (such as urea) or when including information regarding the social status (such as income), constitution or functional assessment of patients (such as the Glasgow Coma Scale, other gradings of consciousness, or functional scales).

Performance of scores on short-term and long-term outcomes

A graphical overview of relevant outcomes for hospitalized patients at timing of hospital admission is presented in **Figures 5 to 11.** A table of predictions for all combination of subpopulation, predictor measurement and outcome scores is reported in **Table S8**, where interested researchers can filter and make choices for specific combinations of timing and outcome, predictor availability and importance of performance measures according to the respective research question or stratification demand.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

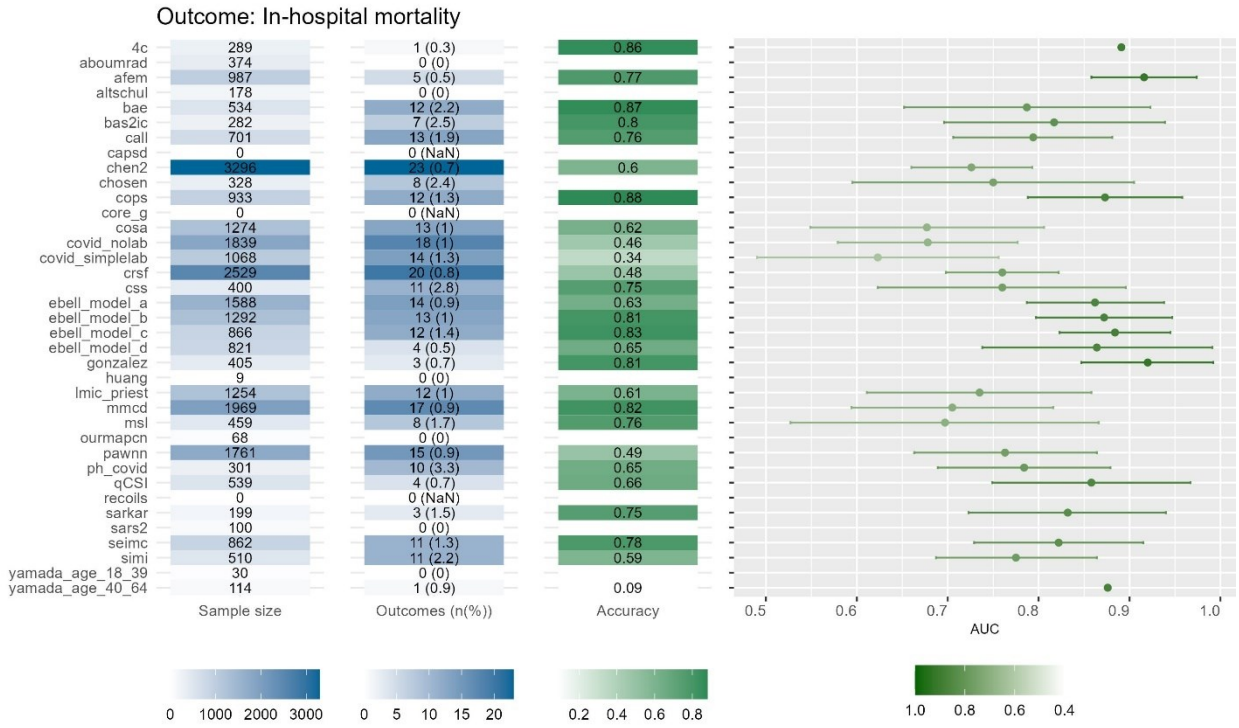


Figure 5: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome in-hospital mortality.

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

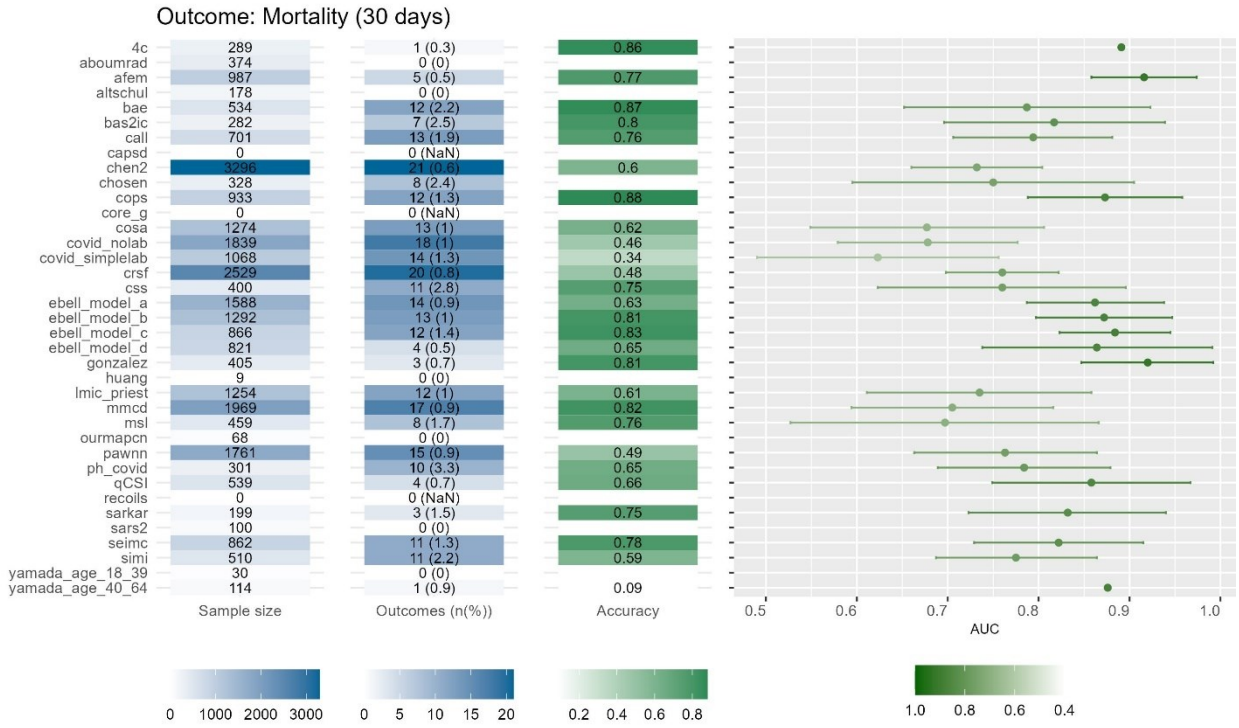


Figure 6: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome mortality (within 30 days of admission).

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

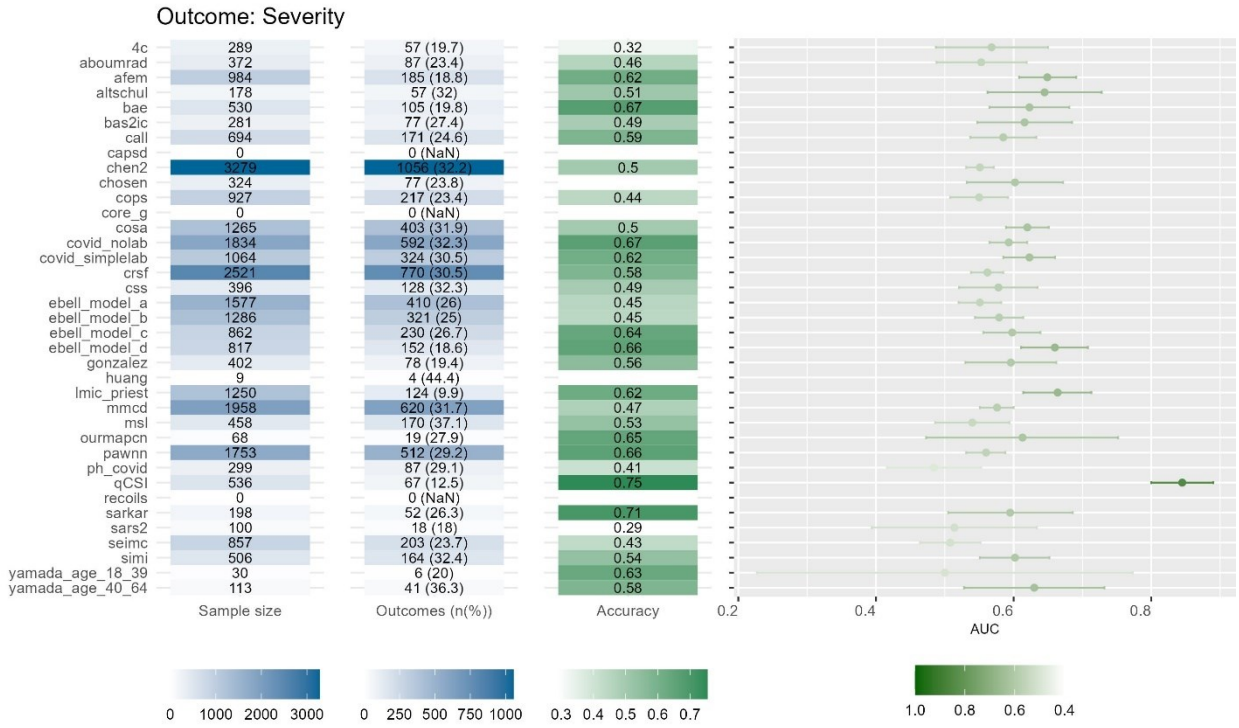


Figure 7: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome severity.

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

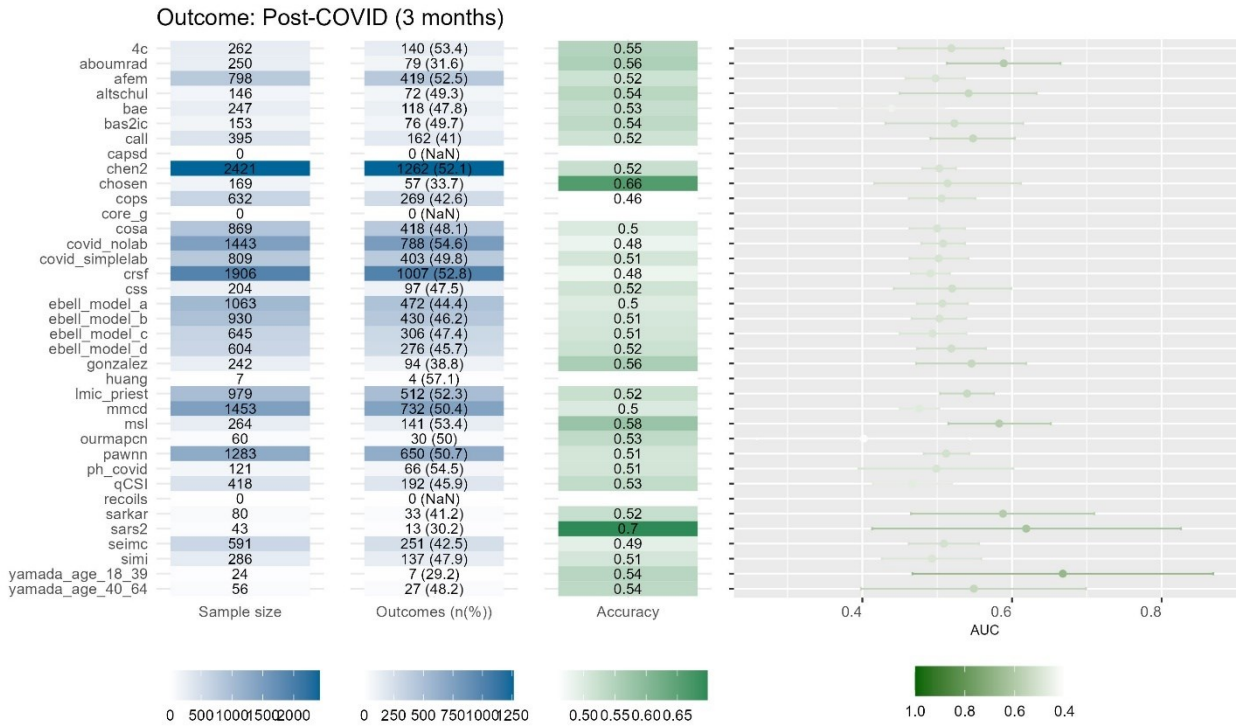


Figure 8: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome Post-COVID (at month 3).

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

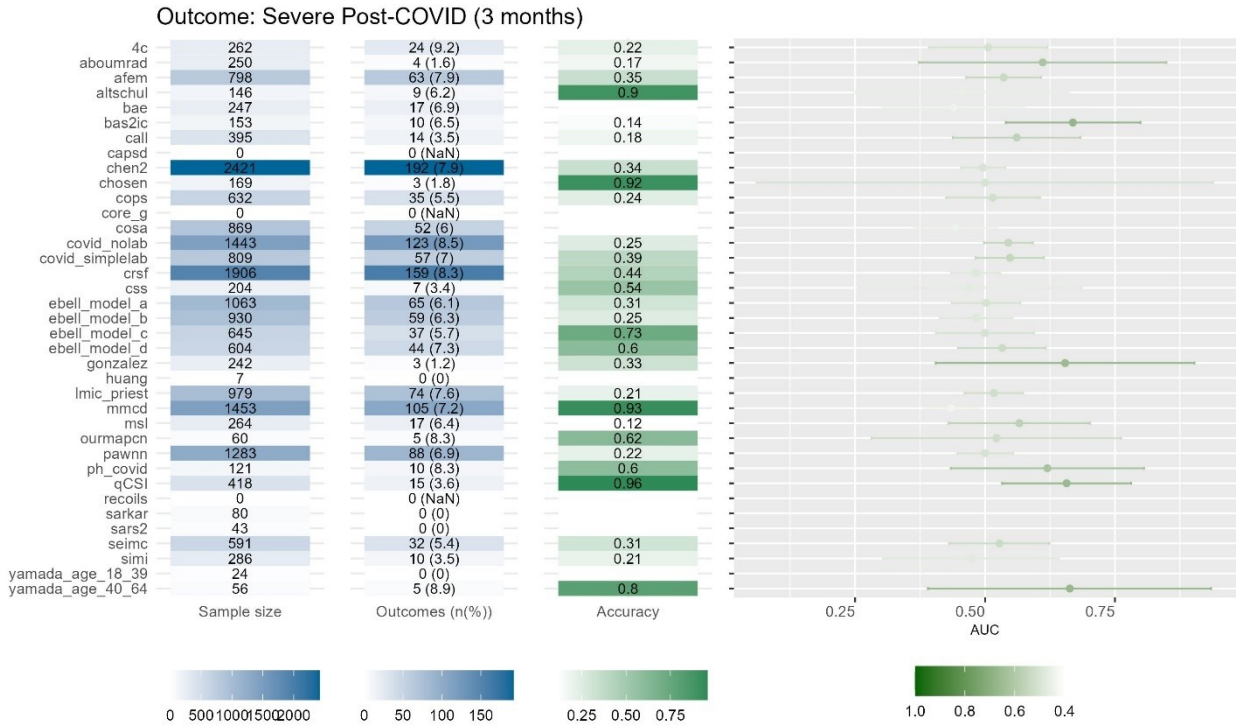


Figure 9: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome Severe Post-COVID (at month 3).

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

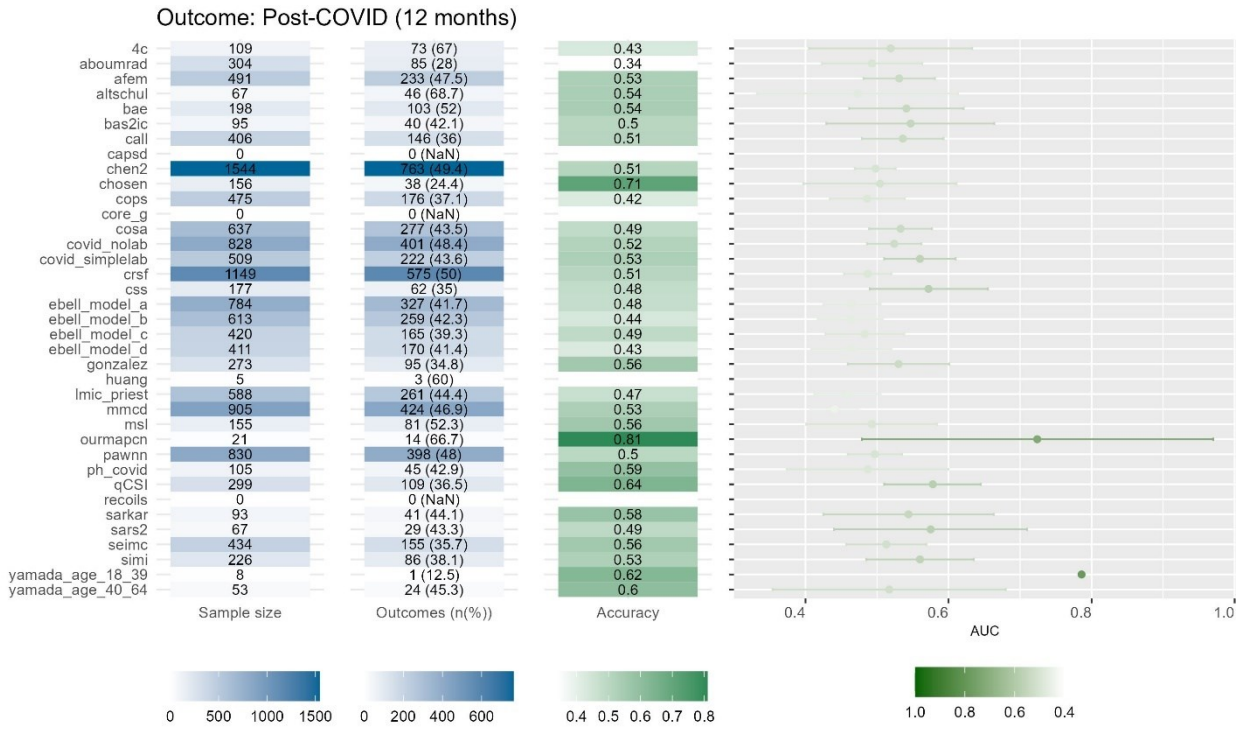


Figure 10: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome Post-COVID (at month 12).

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

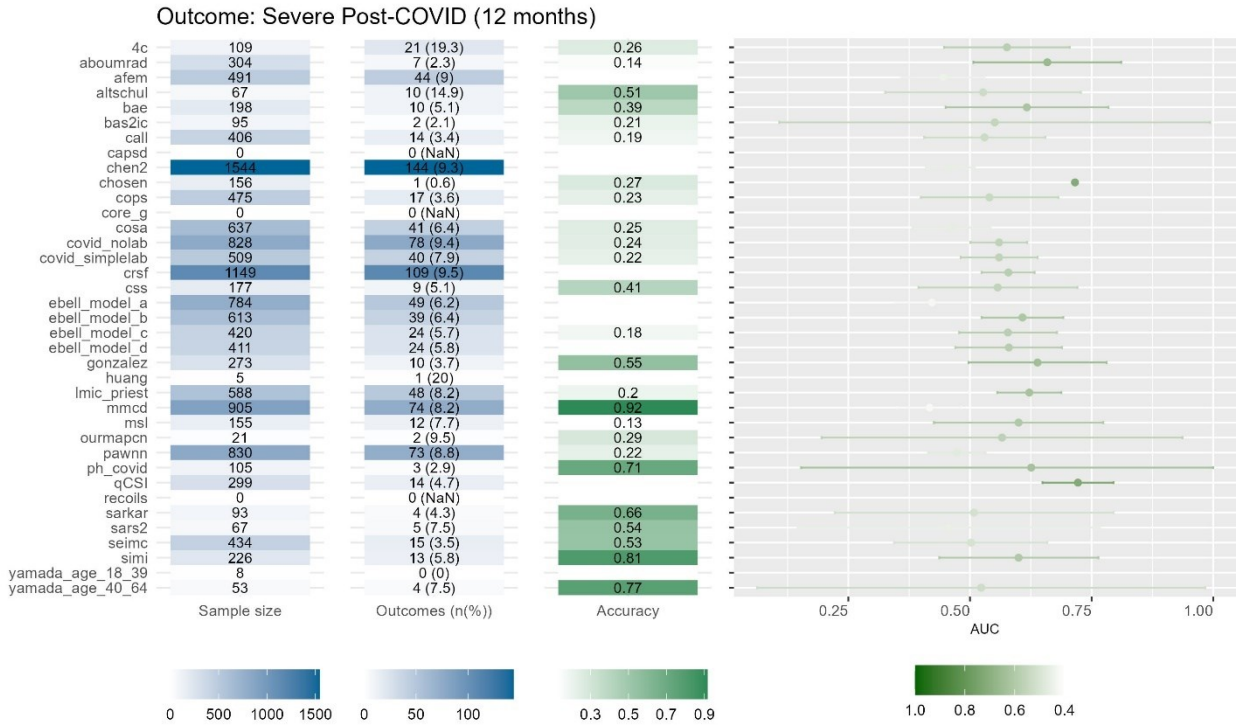


Figure 11: Overview of accuracy and AUC in the context of sample size and number of outcomes for the outcome Severe Post-COVID (at month 12).

Area under the (receiver operating characteristic) curve (AUC) is displayed with 95% confidence interval.

For the outcome severity, the performance of most scores was moderate. The quick COVID-Severity Index (qCSI) by Haimovich et al.,[30] which was designed for the prediction of respiratory failure or death (within 24 hours of admission) presented with an outstanding AUC of 0.85 (95%-CI: 0.80-0.89) and a comparably good accuracy (0.75) and relatively high number of outcome events ($n_{outcomes}$=67).

For in-hospital mortality predictions, there were few outcomes (ranging from 0 to 23). With special regard to the low number of outcome events and the associated limitations, the AFEM score by Pigoga et al.[27] (n_{events}=5; AUC 0.92 (95%-CI: 0.86-0.97), accuracy: 0.77) and the score by González-Cebrián et al.[31] (n_{events}=3; AUC 0.92 (95%-CI: 0.85-0.99), accuracy: 0.81) presented with a good predictive performance. Additionally, there were several scores with AUC above 0.80 along with outcome events above n=10 such as the Ebell Models A to C[32] or the COPS by Cho et al.[33]. The 4C by Knight et al. also showed good discrimination and accuracy (AUC 0.89, 95%-CI: N/A, accuracy: 0.86), but was only validated using a single outcome event. The results regarding the prediction of mortality within 30 days of admission were comparable, as most in-hospital deaths occurred within a 30 days interval.

The experimental approach to test the predictive ability of the included scores for long-term outcomes showed overall poor AUCs, especially in the context of a sufficient number of outcomes (n≥100): high AUCs came along with limited sample size and power. For instance, the prediction of PCC (3 months) at hospital admission resulted in AUCs ranging from 0.40 to 0.67. The best trade-off between accuracy and AUC was yielded using the SARS2 by Dasthi et al. (AUC: 0.62, (95%-CI: 0.41-0.83), accuracy: 0.70), but the sample size and number of outcomes was low ($n_{sample\ size}$=43, $n_{outcomes}$=13). For a detailed overview of PCC predictions regarding sample sizes, accuracy and AUC see **Figures 8 to 11**.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

The sensitivity analysis using an accepted time interval of seven days around the target timing of predictor measurement is presented in **S9**.

The comparison of the AUCs reported in the original studies and the AUC resulting from the ORCHESTRA's WP2 validation did not reveal significant patterns (**Figure 12**). Some scores performed better in the original study than in the validation (e.g., Sarkar, CSS or CALL) while others even performed better in the WP2 cohorts than in the original study (e.g., AFEM, 4C Mortality Score). We would like to emphasize that we might not have created the exact scenario for which the score was designed (e.g., another composite definition of "severity").

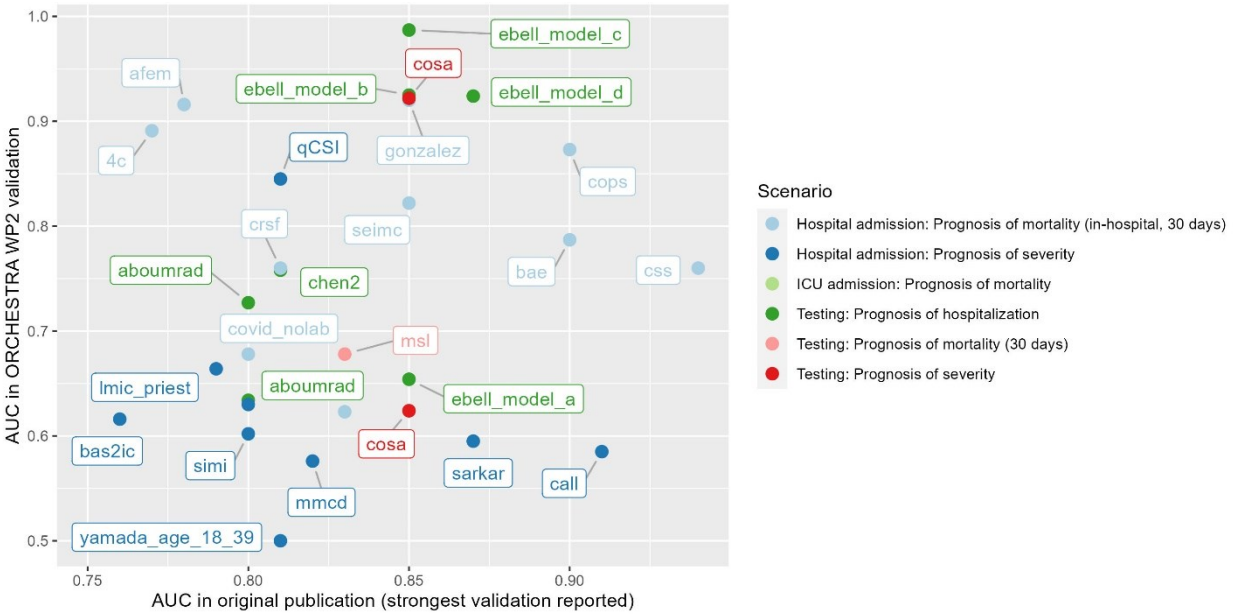


Figure 12: Comparison of AUCs of original studies and the AUCs calculated in the ORCHESTRA's WP2 validation.

We chose the combination of population, timing and outcome that is closest to the original target scenario. If an original study included "in- and outpatients" we display the comparison twice, one using outpatients and one using inpatients, if applicable. Note that some of these combinations did not reach sample sizes or number of outcomes above 1, thus the AUC is missing in these cases.

Discussion

In this analysis, we investigated both the quantity and quality of published clinical scores purposed for predicting various types of COVID-19 outcomes and tested their performances on the ORCHESTRA WP2 dataset. Despite the development of numerous scores explicitly for this aim, none were incorporated into COVID-19 therapeutic guidelines as a standard component of the clinical routine.[10,11,13] Our examination revealed that numerous scores inadequately complied with the requisite quality criteria necessary to ascertain their validity, reliability, and credibility. We subsequently delve into a discussion concerning the shortcomings and prospects inherent in score development, focusing specifically on aspects such as risk of bias (ROB) and predictor selection.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Scores identified with low or unclear ROB

A significant number of scores (n=41) were found to carry a significant ROB, attributable to factors such as the study design, inadequate sample sizes, the reliance on single-center data, or the absence of validation procedures. Notably, only five scores presented a low ROB, and three scores exhibited an uncertain ROB.

The ISARIC's 4C score, designed for predicting mortality, is grounded on a comprehensive, prospective cohort and has demonstrated commendable performance across various external validation efforts.[12] The CCEDRRN COVID-19 Mortality Score, formulated using a retrospective cohort, interestingly incorporates the "arrival mode" (e.g., by ambulance) of patients[23] - a piece of information that is often not readily available, thus complicating the process of external validation. The SEIMC score was also constructed using a substantial, retrospective cohort and has undergone both temporal and geographical validation.[24] The PRIEST score[25] and the LMIC-PREST,[26] which were both formulated based on extensive cohorts, were identified as being appropriate for severity prediction.

Scores demonstrating an ambiguous risk of bias (ROB) were associated with specific constraints: The AFEM score was derived from a relatively small retrospective patient cohort from two centers; it lacked external validation and calibration processes[27]. Regarding the OURMAPCN score,[28] we determined a low effective eEPV. Nevertheless, this score exhibited solid performance in three external validation studies and considered the potential issues of overfitting. The SARS2 score, intended for predicting hospitalization, employed a sizeable sample for its development, and the validation process was conducted using a cohort of medical staff.[29]

Predictor selection, applicability, and complexities

A vast array of predictors, emanating from diverse domains, were included in the scores (**Figure 2** and **Figure 3**). We can hypothesize that the selection of predictors is often more influenced by the availability of data than by the adherence to optimal practice guidelines for score development.[7,34] Various discrepancies across studies, such as the range of data sources employed, criteria for inclusion of analyzed cohorts (notably their baseline status), slight differences in endpoints and definitions, and the statistical methodologies used, likely account for this observed variety in choices. This heterogeneity may have been further amplified by the emergence of different strains, and the changing landscape of vaccination statuses. Nonetheless, this review reveals the presence of a common set of predictors utilized across numerous scores (such as age and CRP), whereas some predictors were included sparingly (e.g., nausea or hypotension). In L2, most scores incorporated age (87.8%) and respiratory measurements (like oxygen saturation and respiratory rate), symptoms (such as dyspnea and cough), comorbidities (like chronic obstructive pulmonary disease), lung imaging, or indicators of acute respiratory complications (73.5%).

Pre-test probability, also known as prior probability, is the initial estimate of the likelihood that a person has a particular condition before any diagnostic tests are conducted. It guides decisions regarding which tests to perform and how to interpret their results in medical settings. In different healthcare settings, such as prehospital care versus the emergency room, pre-test probabilities can vary due to differences in patient populations and available information. Additionally, the importance of specific variables in predicting outcomes may differ based on the setting and the availability of relevant data and resources. In this context, our findings suggest

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

that scores predicting COVID-19 mortality or severity should incorporate factors such as age, respiratory conditions, laboratory data, and comorbidities to reliably predict the respective outcomes.[35,36] Pre-hospital scores, such as those predicting hospital admission, primarily utilize information on comorbidities and socio-demographics, which can be applied without the need for extensive diagnostic infrastructure. In general, symptoms, and imaging appear to play a minor role in these prediction models.

The predictor set of frequently utilized variables (Top 20), consists of six elements (age, sex, diabetes mellitus, hypertension, blood urea nitrogen, creatinine) are components of baseline assessment for (organ-related) infection outcomes or differential diagnoses. Ten elements (CRP, lactate dehydrogenase, oxygen saturation, respiratory rate, neutrophils-lymphocyte ratio, lymphopenia, dyspnea, thrombocytopenia, blood pressure, paO2/FiO2, temperature, leukocytes) are acknowledged proxies for clinical severity. Interestingly, only two of the predictors (D-dimer, albumin) may not enjoy universal recognition as components integral to the preliminary assessment of moderate to severe respiratory infections. While albumin is not COVID-specific and is usually used for detection of malnutrition, abnormal D-dimer levels have been shown to be a characteristic of COVID-19 with the progress of research findings (but may not have been at study initiation). A partial explanation for these resultsmight lie in the role of pre-existing knowledge as a central factor in shaping the selection of datasets, which were later incorporated into analytical processes. This interpretation suggests that widely used criteria are generally accessible inpatient care scenarios, especially in medium to high resource settings. Yet, it also implies that these scoring systems might have limited value in augmenting our existing understanding of respiratory infection outcomes, and their specificity to COVID-19 could be questionable.

Non-routine laboratory indicators, such as D-dimer and Interleukin-6, confine the utility of the scoring systems to environments with substantial resources. Notwithstanding, considering that D-dimer emerged as one of the top 10 single predictors in our analysis (**Figure 3**), subsequent research endeavours are necessitated to delineate the supplementary value these parameters contribute to the efficacious management of patients. Overall, laboratory tests for scores, e.g., indicators of renal function (such as urea, blood urea nitrogen, creatinine, glomerular filtration rate) or inflammation markers like CRP versus leukocytes may be constrained in specific resource and management contexts. For instance, most primary healthcare providers and outpatient departments cannot conduct exhaustive blood examinations predicated on moderate respiratory symptoms.[37] The potential for bias stemming from the connection between the availability of data and the care environment could hinder the application of results across different settings. Unconventional or time-independent predictors, like the date of admission,[28] present limited generalizability for implementations that are irrespective of time, geographical region, and setting. It is noteworthy that only a single score incorporated the variable of vaccination status, a factor primarily attributed to the development of most scores using data from the early stages of the pandemic. Clinical trial outcomes suggest that the current vaccination status could be one of the most significant prognostic indicators.[38,39] Additionally, the multifaceted nature of scores, incorporating numerous predictors across varied domains and multiple cutoff levels, poses challenges for healthcare professionals when applying them at the bedside,[7] underlining that clinical judgment remains irreplaceable.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Limitations of studies included in the review and comparison to other reviews

Variations in the design of score development could contribute to disparities in performance.[16] Notably, we identified a substantial variability in sample sizes and settings, as well as differences in case definitions and severity specifications. Population characteristics such as age distributions,[36,40] ethnicity,[41] and immune status are recognized to affect COVID-19 outcomes and these characteristics demonstrated variability across studies. Additionally, we also observed differences in prerequisites for specific treatments and/or hospital admissions among various countries.[42] Furthermore, the comparability of composite outcomes is constrained due to the variability in their composition selected by different study groups.

High performance metrics paired with small sample sizes or inconsistent reporting of both discrimination and calibration measures suggest an increased risk for overfitting.[16] High AUCs were frequently identified (78.5% ≥ 0.75) even though most of the scores were developed utilizing small sample sizes (64.0% ≤ 1,000 patients). Conversely, it is important to consider that a smaller AUC may arise from a more heterogeneous patient population. Additionally, a considerable number of scores have only undergone inadequate internal validation using a random split of patients (26.5%).[22] Consequently, the application of these scores in clinical practice should be postponed until validations conducted in different countries exhibit consistent performance across varying patient characteristic distributions.[19,43]

Given the profusion of published models and scores, it is evidently challenging to pinpoint "all" pertinent items. Thus, a range of complementary strategies is required. We identified a handful of reviews on COVID-19 predictions, all of which focused on divergent approaches and resulted in a (marginally) different set of models.[8,9,44–46] The 4C Mortality Score,[12] the PRIEST score,[25] and the well-established NEWS2 were repeatedly highlighted as commendable prognostic models among others. Furthermore, we focused on easy-to-use scores and excluded non-simplified (regression or machine learning) models. With the loss of information during the transition to a score, the precision of prediction might be reduced. We acknowledge the necessity of a careful trade-off between high-precision prediction, practicability and utility, depending on the respective setting and medical problem.

High-performing scores in the external validation

Considering both the ROB analysis and the validation results (AUC vs. accuracy), the AFEM seem reasonable choices for predicting mortality outcomes, respectively. This score was validated with a comparably high number of outcome events, showed good AUCs, and does not require laboratory testing. The 4C could not be validated reliably (n=1 outcomes) due to high missing rates in its components and few fatal outcomes, but is discussed as promising in various settings, including Omicron cohorts.[46–49]

Regarding severity predictions, our analysis highlighted the qCSI by Haimovich et al.[30] Notably, the score raised concerns in the PROBAST-based ROB assessment. Also, the mapping of the required oxygen flow rate (L/min) was not possible in a 1:1 manner, as this information is documented categorically in ORCHESTRA, using other cutoffs than required in the score (assumptions see **Table S3**). In contrast, the other scores had only a low to moderate performance.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

The prediction of PCC using scores developed for short-term prognosis presented with low performance, thus this approach is not helpful facing the currently most significant health risk, PCC.

Limitations of the review process and the external validation

We limited the ROB analysis and comprehensive data extraction to a subset of scores that met predefined selection criteria. To increase the feasibility, we applied a set of selection criteria ((I) a reported area under the curve (AUC) of ≥ 0.75; (II) the report of a separate validation cohort as the minimal validation procedure; (III) the development in a multi-center setting (≥ 2 centers); (IV) a points-based application) that also aim to pre-select studies in terms of quality characteristics. While an AUC of ≥ 0.75 is described to be a 'clearly useful discrimination'[20], heterogenous population characteristics – which are in fact favorable – might reduce the AUC. Also, a very well-designed single-center study can, in individual cases, potentially be of higher quality than a poorly conducted multi-center study. However, we wanted to highlight that multi-centre studies increase the heterogeneity of the study population and reduce the effect of individual clinical (treatment) routines, thus enhancing transferability of results. We acknowledge that, due to the strict determination of cut-offs, we might have missed single high-quality studies that could not fulfil these criteria for some reason. A more expansive approach, encompassing additional sources and thorough analysis of more or all scores, may have uncovered more pertinent studies. We did not reach out to the authors of the primary studies for supplemental or missing data and utilized a constrained CHARMS checklist (see **Supplementary text S1**), focusing on aspects most relevant to our research question. Well-established early warning scores fell outside our research scope, however, they are reported to exhibit robust predictive performance in external validation studies.[50]

With the heterogeneity of multiple studies from five European countries within ORCHESTRA's WP2, the validation power was higher when restricting the predictors to demographics, vital signs and comorbidities. With a complete dataset, other scores might have had sufficient sample size to highlight its predictive capabilities. As patients can be enrolled in the WP2 cohorts after the primary infection, the amount of missing might have been lower with prospective documentation of the primary infection. In contrast, as ORCHESTRA already is a cohort with in-depth documentation, transferability to low-resource or outpatient settings is only possible when avoiding predictor classes requiring high resources. Thus, this analysis is a real-world check on the applicability of scores on routine data.

Furthermore, the focus of the ORCHESTRA's WP2 cohort is long-term follow-up (with few deaths) which comes along with changes in public health priorities. The decision to validate scores for the acute phase using the prospective cohorts of WP2 raises concerns. The primary focus of the WP2 prospective cohort is to characterize post-acute symptoms, and individuals can be enrolled after the acute infection stage. This introduces a survival bias, since those with better survival prospects are included, leading to the observed low mortality rates. The inclusion of additional retrospective data sources might balance the described complexities, but also brings disadvantages such as harmonization issues or sparse information. In consequence, validations for these scenarios have reduced reliability.

The validation was limited due to a considerable number of missing values and the variation of timing of predictor measurement for the documentation of time-dependent information at the primary infection, both reducing representativeness and reliability. Particularly for the prediction of low-prevalence outcomes, the number of outcome events recommended to generate

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

meaningful results was often not reached.[16] Furthermore, for some scores a 1:1 mapping was not possible, or assumptions were necessary to derive the score. Other scores could not be mapped due to missing information in the dataset. These limitations can be attributed to both the dataset itself (more granular documentation is always possible) and the suitability of predictor choices in the development studies. Additionally, in the context of the outcome synthesis approach, we acknowledge the possibility that the scores might have performed better when predicting the exact outcome it was designed for.

Conclusion

This analysis reveals a comprehensive analysis of COVID-19 scores with respect to predictor evaluation and applicability and a connected external validation of identified potential high-quality scores. None of the many scores that have been developed have been devised and subsequently garnered robust guideline endorsements on a global scale. Up to now, the prevailing consensus suggests that while predictive tools are useful, they should serve to supplement, not supplant, the judgement of the practicing clinician.

With three years of COVID-19 investigations at the time of data collection, we also acknowledge the lack of dependable scoring systems for predicting PCC. As health outcomes have progressively ameliorated since the pandemic's initial surge, PCC, commonly referred to as Long COVID, has now superseded severe illness and mortality as a health risk for COVID-19 patients. Dependable predictors of adverse long-term outcomes could serve as a beneficial tool for clinical decision-making and contribute to the conceptualization of forthcoming clinical trials.

Most scores we reviewed exhibited a marked ROB and lacked external validation. We identified a set of scores with low or unclear risk of bias (ROB) and critically examined the recurring challenges inherent in the development of prognostic scoring systems. Most of these scores were validated in the ORCHESTRA WP2 dataset. The recommendations of specific scores should be based on the desired stratification demand, the predictors available in the respective setting, and a careful tradeoff of performance measures. In the absence of such validation, there exists a substantial risk of suboptimal performance and inaccurate categorization within diverse populations and settings.

In the event of future pandemics, fostering data and resource sharing, coupled with the harmonization of score development concepts, would enhance the quality and visibility of these scores. This advancement could facilitate their improved implementation, ultimately serving to optimize patient management, therapeutic decisions, and outcomes.

In conclusion, we stress the need for a delicate balance of validity and practicality of predictors in the development of scores. In particular, enhancements in the reporting of methodological approaches and adherence to the *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD) checklist[20] would improve the lucidity and trustworthiness of the formulated scores.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

References

1 Filip R, Puscaselu RG, Anchidin-Norocel L, Dimian M, Savage WK. Global challenges to public health care systems during the COVID-19 pandemic: A review of pandemic measures and problems. *J Pers Med* 2022;12:1295.

2 Dyer O. Covid-19: Europe could be headed for pandemic "endgame," says WHO region chief. *BMJ* 2022;376:o205.

3 Biancolella M, Colona VL, Mehrian-Shai R *et al.* COVID-19 2022 update: transition of the pandemic to the endemic phase. *Hum Genomics* 2022;16:19.

4 Shrestha LB, Foster C, Rawlinson W, Tedla N, Bull RA. Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. *Rev Med Virol* 2022;32:e2381.

5 Moghadas SM, Vilches TN, Zhang K *et al.* The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States. *Clinical Infectious Diseases* 2021;73:2257–64.

6 Long B, Carius BM, Chavez S *et al.* Clinical update on COVID-19 for the emergency clinician: Presentation and evaluation. *Am J Emerg Med* 2022;54:46–57.

7 Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagn Progn Res* 2019;3:16.

8 Wynants et al. *Living review of Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal*. Available online at https://www.covprecise.org/living-review/ [Accessed 24 February 2023].

9 Miller JL, Tada M, Goto M *et al.* Prediction models for severe manifestations and mortality due to COVID-19: A rapid systematic review. *Acad Emerg Med* 2022;29:206–16.

10 Infectious Diseases Society of America. *Guidelines on the Treatment and Management of Patients with COVID-19: May 2023*. Available online at https://www.idsociety.org/COVID19guidelines [Accessed 21 September 2023].

11 World Health Organization. Therapeutics and COVID-19: Living Guideline: January 2023.

12 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.

13 World Health Organization. Clinical management of COVID-19: living guideline: January 2023.

14 Page MJ, McKenzie JE, Bossuyt PM *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.

15 Moons KGm, Groot JAH de, Bouwmeester W *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744.

16 Moons KGM, Wolff RF, Riley RD *et al.* PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med* 2019;170:W1-W33.

17 Xie Y, Bowe B, Al-Aly Z. Burdens of post-acute sequelae of COVID-19 by severity of acute infection, demographics and health status. *Nature Communications* 2021;12:6571.

18 Gentilotti E, Gorska A, Tami A *et al.* Clinical phenotypes and quality of life to define post-COVID-19 syndrome: a cluster analysis of the multinational, prospective ORCHESTRA cohort. *EClinicalMedicine* 2023;62:102107.

19 Alba AC, Agoritsas T, Walsh M *et al.* Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA* 2017;318:1377–84.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

20 Collins GS, Reitsma JB, Altman DG, Moons KGm. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.
21 van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.
22 Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021;14:49–58.
23 Hohl CM, Rosychuk RJ, Archambault PM *et al.* The CCEDRRN COVID-19 Mortality Score to predict death among nonpalliative patients with COVID-19 presenting to emergency departments: a derivation and validation study. *CMAJ Open* 2022;10:E90-E99.
24 Berenguer J, Borobia AM, Ryan P *et al.* Development and validation of a prediction model for 30-day mortality in hospitalised patients with COVID-19: the COVID-19 SEIMC score. *Thorax* 2021;76:920–9.
25 Goodacre S, Thomas B, Sutton L *et al.* Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: The PRIEST observational cohort study. *PLoS One* 2021;16:e0245840.
26 Marincowitz C, Hodkinson P, McAlpine D *et al.* LMIC-PRIEST: Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19 in a middle-income setting. *PLoS One* 2023;18:e0287091.
27 Pigoga JL, Omer YO, Wallis LA. Derivation of a contextually-appropriate COVID-19 mortality scale for low-resource settings. *Ann Glob Health* 2021;87:1–15.
28 Chen Z, Chen J, Zhou J *et al.* A risk score based on baseline risk factors for predicting mortality in COVID-19 patients. *Curr Med Res Opin* 2021;37:917–27.
29 Dashti H, Roche EC, Bates DW, Mora S, Demler O. SARS2 simplified scores to estimate risk of hospitalization and death among patients with COVID-19. *Sci Rep* 2021;11:4945.
30 Haimovich AD. Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation. *Annals of Emergency Medicine* 2020;76:442–53.
31 González-Cebrián A, Borràs-Ferrís J, Ordovás-Baines JP *et al.* Machine-learning-derived predictive score for early estimation of COVID-19 mortality risk in hospitalized patients. *PLoS One* 2022;17:e0274171.
32 Ebell M, Hamadani R, Kieber-Emmons A. Development and Validation of Simple Risk Scores to Predict Hospitalization in Outpatients with COVID-19 Including Omicron. *JABFM*;35:1058–64.
33 Cho S-Y, Park S-S, Song M-K *et al.* Prognosis Score System to Predict Survival for COVID-19 Cases: a Korean Nationwide Cohort Study. *J Med Internet Res* 2021;23:e26257.
34 Schummers L, Himes KP, Bodnar LM, Hutcheon JA. Predictor characteristics necessary for building a clinically useful risk prediction model: a simulation study. *BMC Med Res Methodol* 2016;16:123.
35 Chu K, Alharahsheh B, Garg N, Guha P. Evaluating risk stratification scoring systems to predict mortality in patients with COVID-19. *BMJ Health Care Inform* 2021;28:e100389.
36 Marin BG, Aghagoli G, Lavine K *et al.* Predictors of COVID-19 severity: A literature review. *Rev Med Virol* 2021;31:1–10.
37 Ebell MH, Cai X, Lennon R *et al.* Development and validation of the COVID-NoLab and COVID-SimpleLab risk scores for prognosis in 6 US Health Systems. *J Am Board Fam Med* 2021;34:127-135.
38 Polack FP, Thomas SJ, Kitchin N *et al.* Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;383:2603–15.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

39 El Sahly HM, Baden LR, Essink B *et al.* Efficacy of the mRNA-1273 SARS-CoV-2 vaccine at completion of blinded phase. *N Engl J Med* 2021;385:1774–85.
40 Bellou V, Tzoulaki I, van Smeden M *et al.* Prognostic factors for adverse outcomes in patients with COVID-19: a field-wide systematic review and meta-analysis. *Eur Respir J* 2022;59:2002964.
41 Mackey K, Ayers CK, Kondo KK *et al.* Racial and ethnic disparities in COVID-19–related infections, hospitalizations, and deaths: A Systematic Review. *Ann Intern Med* 2020;174:362–73.
42 Bentivegna M, Hulme C, Ebell MH. Primary care relevant risk factors for adverse outcomes in patients with COVID-19 infection: A systematic review. *JABFM* 2021;34:113–26.
43 Shamsoddin E. Can medical practitioners rely on prediction models for COVID-19? A systematic review. *Evid Based Dent* 2020;21:84–6.
44 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.
45 Buttia C, Llanaj E, Raeisi-Dehkordi H *et al.* Prognostic models in COVID-19 infection that predict severity: a systematic review. *Eur J Epidemiol* 2023;38:355–72.
46 Lombardi Y, Azoyan L, Szychowiak P *et al.* External validation of prognostic scores for COVID-19: a multicenter cohort study of patients hospitalized in Greater Paris University Hospitals. *Intensive Care Med* 2021;47:1426–39.
47 Crocker-Buque T, Myles J, Brentnall A *et al.* Using ISARIC 4C mortality score to predict dynamic changes in mortality risk in COVID-19 patients during hospital admission. *PLoS One* 2022;17:e0274158.
48 Ngiam JN, Chew NWS, Tham SM *et al.* Utility of conventional clinical risk scores in a low-risk COVID-19 cohort. *BMC Infect Dis* 2021;21:1094.
49 Vito A de, Colpani A, Saderi L *et al.* Is the 4C Score Still a Valid Item to Predict In-Hospital Mortality in People with SARS-CoV-2 Infections in the Omicron Variant Era? *Life* 2023;13:183.
50 Gupta RK, Marks M, Samuels THA *et al.* Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. *Eur Respir J* 2020;56:2003498.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Supplementary Results

Table S1

Inclusion and exclusion criteria for the systematic review of prognostic scoring systems.

Inclusion criteria	Exclusion criteria
• Original publication of a new or modified score developed in the context of COVID-19 • In- and outpatients with clinically assumed or confirmed COVID-19 (laboratory tests (PCR, serological test), or clinical assumption or diagnosis) • Transparent and reproducible algorithm presented • Combination of at least two predictors • English or German language	• Models built on of specific sub populations (e.g. comorbidities, specific pharmaceutical interventions, pregnancy) • Models without scoring character or pure presentation as web calculators or nomograms • Models of single predictors[a], genetic factors or psychological outcomes (burnout, stress, resilience) • (Quantitative) Radiologic scores with image processing[a]

[a] A radiological score without further combination with other clinical predictors was considered as single predictor.

Reasons for the selection criteria:

Choosing the Area under the (receiver operating characteristic) curve (AUC) as the primary performance metric is a common approach in studies involving predictive models. Although the AUC alone doesn't provide a comprehensive measure of a model's performance, it does offer an effective way to assess the model's discriminatory ability - that is, its capacity to distinguish between different outcomes. It's generally accepted that an AUC of 0.75 or higher indicates a model with 'clearly useful discrimination'[19]. We reported the performance measure estimated from the strongest validation reported (external validation (different centers, countries, populations) > temporal validation > random split > development and validation in the exact same dataset)[22,44]. A validation cohort is a fundamental quality standard for generating clinical predictive models because typically a model shows improved performance on the cohort from which it was originally derived[7]. Additionally, models developed and validated on more diverse, multi-center data are generally more likely to be transferrable to different settings[19]. To ensure that a model is applicable in a real-world clinical environment and doesn't require excessive resources to use, we added point-based calculation as a criterion and excluded models that require sophisticated calculations or formulas. This also enhances their usability in the hectic and resource-limited environments often found in clinical settings.

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Table S2

Search strategy in PubMed/MEDLINE and Web of Science

#1 AND #2 AND #3 AND #4

#1 COVID-19

PubMed	Web of Science
("COVID-19"[Mesh] OR "SARS-CoV-2"[Mesh] OR “COVID-19”[tiab] OR “COVID19”[tiab] OR “SARS-CoV-2 Infectio*”[tiab] OR “2019 Novel Coronavirus”[tiab] OR “2019-nCoV Diseas*”[tiab] OR “2019-nCoV Infectio*”[tiab] OR “COVID-19 Virus Diseas*”[tiab] OR “COVID-19 Virus Infectio*”[tiab] OR “Coronavirus Disease 2019”[tiab] OR “Severe Acute Respiratory Syndrome Coronavirus 2 Infectio*”[tiab] OR (("coronavirus"[MeSH Terms] OR "coronavirus"[tiab] OR "COV"[tiab]) AND 2019/11/01[PDAT] : 2023/12/31[PDAT]))	TS=(“COVID-19” OR COVID19 OR “SARS-CoV-2 Infectio*” OR “2019 Novel Coronavirus” OR “2019-nCoV Diseas*” OR “2019-nCoV Infectio*” OR “COVID-19 Virus Diseas*” OR “COVID-19 Virus Infectio*” OR “Coronavirus Disease 2019” OR “Severe Acute Respiratory Syndrome Coronavirus 2 Infectio*”) OR (TS=(coronavirus OR COV) AND (DOP=(2019-11-01/2023-12-31)))

#2 Prediction

PubMed	Web of Science
("forecas*"[tiab] OR “stratif*”[tiab] OR “prognos*”[tiab] OR "predic*"[tiab] OR "risk assessment"[tiab])	TS=(forecas* OR stratif* OR prognos* OR predic* OR “risk assessment”)

#3 Scoring

PubMed	Web of Science
("Clinical Decision Rules"[Mesh] OR "scor*"[title] OR "index"[title] OR "indices"[title] OR "scal*"[title] OR "clinical decision rul*"[title] OR "tool*"[title] OR "algorith*"[title] OR “clinical signature”[title])	TI=(scor* OR index OR indices OR scal* OR “clinical decision rul*” OR tool* OR algorith* OR “clinical signature”)

#4 Validation Metrics

PubMed	Web of Science
("Area Under Curve"[Mesh] OR "ROC Curve"[Mesh] OR "Sensitivity and Specificity"[Mesh] OR "Data Accuracy"[Mesh] OR “validat*”[tiab] OR “discriminat*”[tiab] OR “calibrat*”[tiab] OR “AUC”[tiab] OR “AUCs”[tiab] OR “AUROC”[tiab] OR “ROC”[tiab] OR “area under the”[tiab] OR “receiver operating characteristi*”[tiab] OR “sensitivity”[tiab] OR “specificity”[tiab] OR “coefficient of determination”[tiab] OR “incidence”[tiab] OR “accuracy”[tiab] OR “negative predictive value”[tiab] OR “positive predictive value”[tiab] OR “NPV”[tiab] OR “PPV”[tiab] OR “c-statisti*”[tiab] OR “false-positive”[tiab] OR “false-negative”[tiab] OR “false discovery rate”[tiab])	TS=(validat* OR discriminat* OR calibrat* OR AUC OR AUCs OR AUROC OR ROC OR “area under the” OR “receiver operating characteristi*” OR sensitivity OR specificity OR “coefficient of determination” OR incidence OR accuracy OR “negative predictive value” OR “positive predictive value” OR NPV OR PPV OR “c-statisti*” OR “false-positive” OR “false-negative” OR “false discovery rate”)

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

The results of both data bases were merged using the Digital Object Identifier (DOI).

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Supplementary text S1

Data extraction items

The following data items were extracted:

Name of the score, first author, year, title, study design, number of participating centres, health care level, sample size (development cohort, validation cohort), population, age, country of derivation, timing of predictor measurement, primary outcome(s), outcome events (in the development cohort if stated), number of predictors, predictors, type of combination of predictors, separated validation cohort and the area under the receiver operating characteristic curve (AUC/AUROC). We reported on performance statistics (AUC) as stated from the strongest form of validation available (see S1).

Additional data for the selected set of scores (Level 2) were extracted for the following items using elements of the CHARMS checklist[15] and PROBAST guidelines[16]:

Study design (additional information), inclusion and exclusion criteria, study dates, definition and method for measurement of outcome, time of outcome occurrence or summary of duration of follow-up, number of candidate predictors (self-counted if not precisely stated; indicated by "~"), handling of continuous variables, events-per-variable*, number of participants with any missing value, handling of missing values, modelling method, selection method of final predictors, shrinkage of predictor weights or regression coefficients/account for overfitting and optimism, complexities in the data, calibration measures, classification measures, method used for testing model performance.

*The absolute sample size of a study is an initial indicator of its generalizability, but it's not sufficient on its own for determining the study's power. The events-per-variable (EPV) criterion is a better approach to measure the reliability of a model, especially when it comes to complex predictive models that are based on numerous predictors. The EPV ratio is calculated by dividing the number of outcome events (such as deaths in a mortality study) by the number of candidate predictors used in the model. This method is used to avoid overfitting, which can occur when a model is too closely fit to the specific dataset used for its creation, limiting its applicability to other datasets. Even though the number of regression coefficients would be more precise[16], we derived the absolute number of candidate predictors used during model development as approximation. The rule of thumb often used in many predictive modeling studies is that EPV should not be below 10, and even better if it's 20 or above[16].

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Table S3

Assumptions for score mapping.

Information	Assumption
Infiltration on chest X-ray	Presence of ground-glass opacity or consolidation on "radiographic imaging" or documentation of any degree of "involvement" on right or left lung
BMI/Obesity	BMI >=30 -> obesity; left as missing if NA
(Chronic) heart disease	Coronary heart disease (I25.1) or Congestive heart failure (I50.0)
Chronic kidney disease (stages 1-3, stages 4-6)	By eGFR using creatinine: GFR < 60 ml/min ~ stages 4-6; GFR > 59 ml/min ~ stages 1-3 (no reference reported for staging in Aboumrad et al.)
Chronic renal failure	Chronic kidney disease with and without dialysis or "Yes, not specified (N18.9)"
Diabetes	Includes diabetes type 1 (E10), diabetes type 2 (E11), type not specified (E14) or other type of diabetes (E12-E14) if not further specified.
No mental disturbance or confusion	"currently altered consciousness and/or confusion"(sct_40917007) = No or if Glasgow coma scale = 15 (GCS derived as reported)
Alertness (alert - reacts to voice, confused or reacts to pain, unresponsive)	Alert = 15, Reacts to voice = GCS 14, Confused or reacts to pain = GCS 4-13, unresponsive = 3
Disease severity in CAPS-D (Werfel et al.)	According to WHO criteria: complicated phase = WHO progression scale > 5 (not LEOSS criteria)
eGFR	According to CKD-EPI-formula
Fever	Symptom documented or body temperature > 38.0°C at respective point in time
Glasgow coma scale (missing)	All patients at the normal ward (at the respective time point) or those never hospitalized have GCS 15
Household income (< $60 K, $60-$80 K, >= $80 K) for SARS2 risk score (Hospitalization)	Adjusted by Purchasing power parities (PPP) (https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm; The last year with information for the "Euro Area (19 countries)" (included cohorts in Italy, France, Netherlands, Spain) is 2022 (USA to EU): 0.685) and adjusted for net income (after taxes) (Taxes on labour as % of GDP - Income from employment in the EU in https://taxation-customs.ec.europa.eu/taxation-1/economic-analysis-taxation/data-taxation-trends_en - Table 45: EU-27 in 2021: 20,9%). The authors (Dashti et al.) do not report whether it is gross or net income, but considering the height of the incomes, we assume the gross income. Income information was imputed from later points in time, if it was not available at primary infection.
No. of comorbidities	Standardized to set included in Orchestra (s. Medical history from; is ~ CCI) plus obesity; not individual set defined by score
Presence of (any) comorbidities	Standardized to set included in Orchestra (s. Medical history from; is ~ CCI) plus obesity; not individual set defined by score (any=yes)
Oxygen flow rate cut offs (qCSI by Haimovich et al.)	Other cut offs in WP2 dataset than proposed by the score. The following derivation displays the proposed cut offs in the qCSI (Haimovich et al.) in relation (~) to the best fitting assumptions in the WP2 dataset (according to oxygen support used, oxygen therapy maximum flow volume, FiO2): Flowrate ≤ 2 L/min ~ "no oxygen support", Flowrate 3-4 L/min ~ "Flowrate 1-5 L/min", Flowrate 5-6 L/min ~ "Flowrate ≥ 6 L/min OR ((high flow nasal cannula OR NIV (CPAP/BIPAP) OR MV))
Oxygen saturation	Assumed to be applied on room air, if not otherwise stated in the publication
Vasopressors	Used irrespective of dosis (not documented)

Presence of/admitted with pneumonia	Infiltration on lung x-ray and fever
Smoking	Only active smokers included if not further specified
Upper limit of the normal range - BUN	Laborlexikon.de: 6-25 mg/dl
Upper limit of the normal range - ddimer	Herold: <0.55 µg/ml
Upper limit of the normal range - procalcitonin	Laborlexikon.de and Herold: 0,5 ng/ml (0,5µg/l)
Upper limit of the normal range - urea	Herold: 12-50 mg/dl; Laborlexikon.de: age and sex dependent
Upper limit of the normal range - CRP	Herold: < 5 mg/L
Unit of laboratory result missing	Imputed by cohort, if usage is predominated by one unit and value range is plausible (not if a centre used multiple options at random for existing units)

References: Gerd Herold: Innere Medizin 2021, Köln. Laborlexikon.de (https://flexikon.doccheck.com/de**).**

Table S4

Characteristics of all included scores.

Download the Excel file here: https://cloud.idcohorts.net/s/5EdSoR8rZX68Sty

Table S5

Characteristics of the set of scores included in Level 2.

Download the Excel file here: https://cloud.idcohorts.net/s/MXxxcsiPwMfykyT

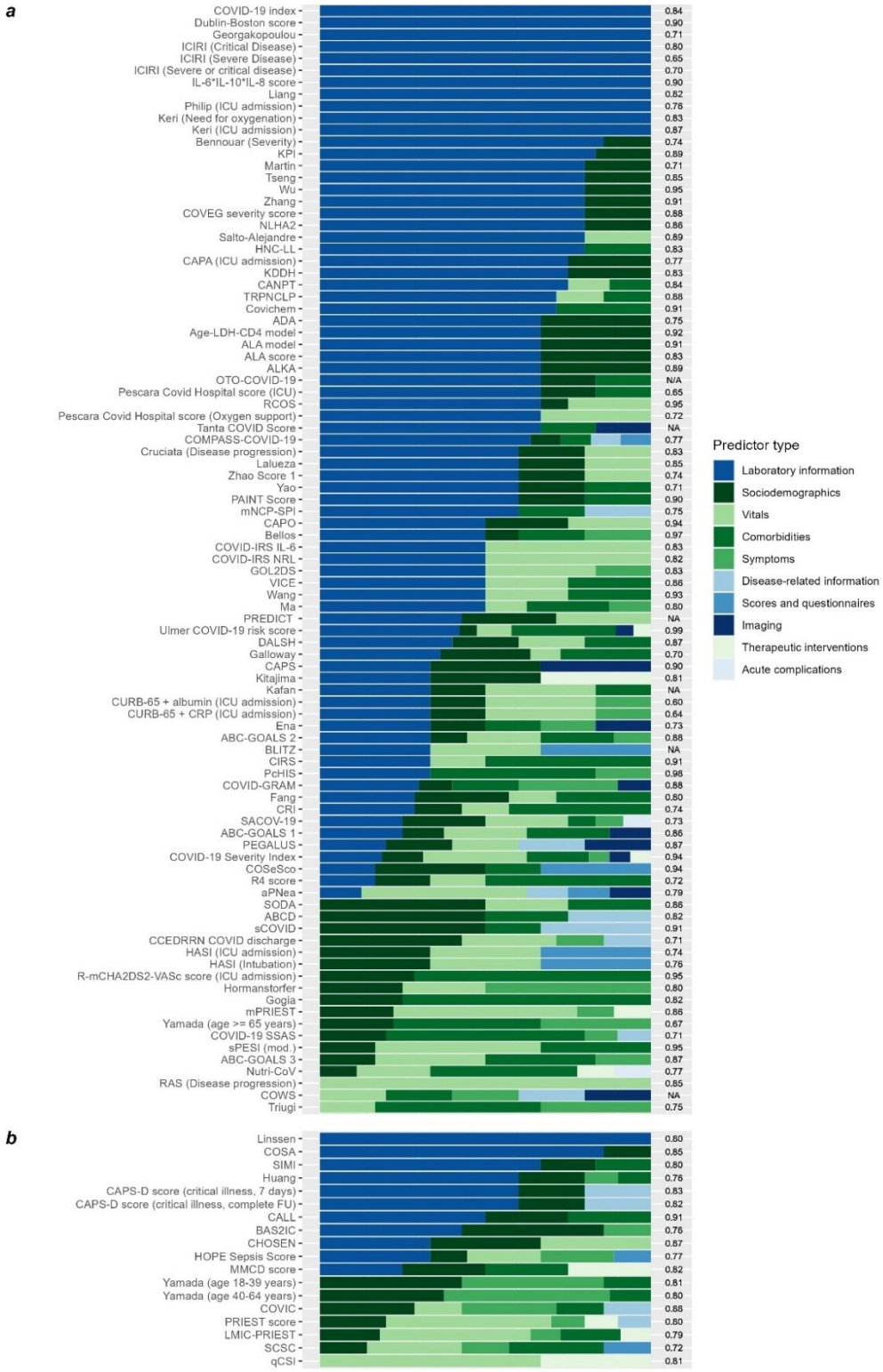


Figure S1

Predictor composition aggregated by predictor type for scores assigned to category 2.

Subfigure a and b correspond to scores within category 2 assigned to Level 1 and 2, respectively. The sorting of the scores is determined by (I) the absolute number of categories and (II) the relative proportion across all scores. The color gradient from green to blue indicates the availability of the category, although in case of doubt this also depends on the level of care.

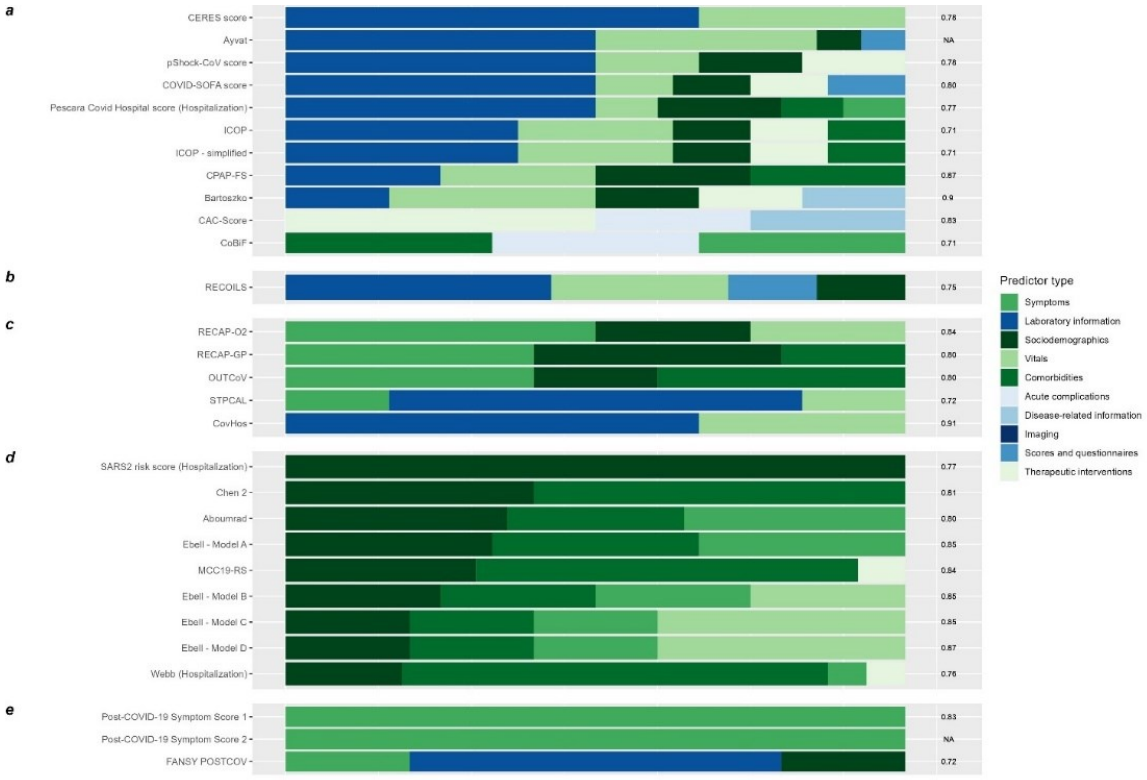


Figure S2

Predictor composition aggregated by predictor type for scores assigned to category 3, 4 and 5.

Subfigures a, b, c, d and e correspond to category 3 (L1), 3 (L2), 4 (L1), 4 (L2), and 5, respectively. The sorting of the scores is determined by (I) the absolute number of categories and (II) the relative proportion across all scores. The color gradient from green to blue indicates the availability of the category, although in case of doubt this also depends on the level of care.

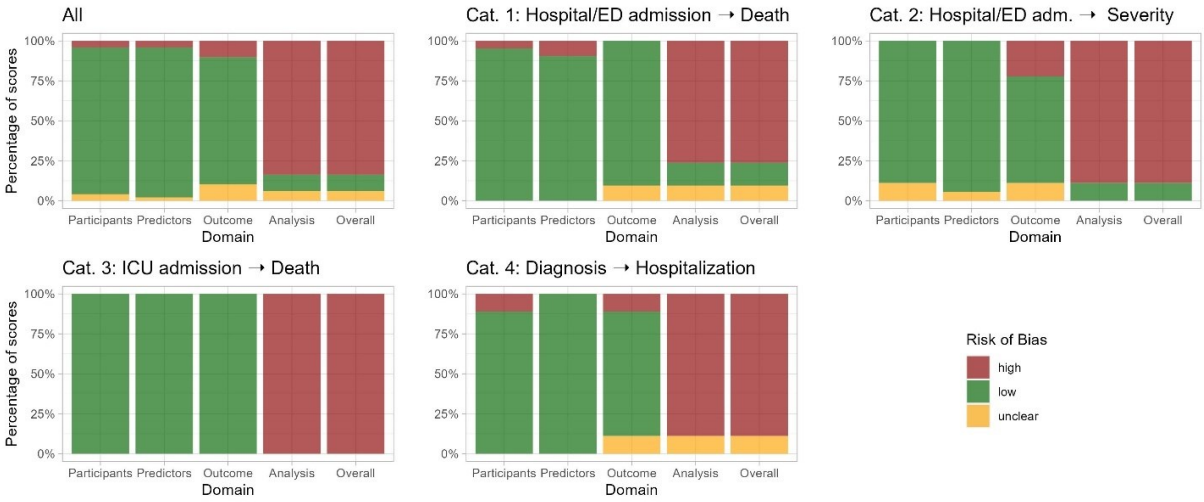


Figure S3

Assessment of the risk of bias evaluation utilizing PROBAST. Emergency Department (ED), Intensive Care Unit (ICU).

Table S6

PROBAST results by score and domain.

Name Score	Participants	Predictors[a]	Outcome	Analysis	Overall
4C Mortality Score	low	low	low	low	low
Aboumrad	low	low	low	high	high
AFEM COVID-19 Mortality Scale (AFEM-CMS) - with oxymetrie	low	low	low	unclear	unclear
Altschul	low	low	low	high	high
Bae	low	low	low	high	high
BAS^{2}IC Score	low	low	low	high	high
CALL (Comorbidity, Age, Lymphocyte, LDH)	low	low	high	high	high
CAPS-D score (critical illness, 7 days)	low	low	low	high	high
CAPS-D score (critical illness, complete FU)	low	low	low	high	high
CCEDRRN COVID-19 Mortality Score	low	low	low	low	low
Chen 2	low	low	high	high	high
CHOSEN (COVID Home Safely Now)	low	unclear	high	high	high
COPS (COVID-19 Prognosis Score) - Mortality (14 days)	low	low	low	high	high
COPS (COVID-19 Prognosis Score) - Mortality (28 days)	low	low	low	high	high
CORE-G score	low	low	low	high	high
COSA	low	low	low	high	high
COVIC	low	low	low	high	high
COVID-NoLab	low	high	low	high	high
COVID-SimpleLab	low	high	low	high	high
CRSF (COVID-19 Risk-Score in Fars Province)	high	low	low	high	high
CSS (Scoring System of COVID-19)	low	low	low	high	high
Ebell - Model A	low	low	low	high	high
Ebell - Model B	low	low	low	high	high
Ebell - Model C	low	low	low	high	high
Ebell - Model D	low	low	low	high	high
FLAMINCOV score	low	low	low	high	high
González-Cebrián	low	low	low	high	high
HOPE Sepsis Score	low	low	unclear	high	high
Huang	low	low	low	high	high
Linssen	low	low	low	high	high
LMIC-PRIEST	low	low	low	low	low
MCC19-RS (Mayo Clinic COVID-19 risk score)	high	low	low	high	high
MMCD score	low	low	high	high	high
MSL-COVID-19	low	low	unclear	high	high
Obremska	low	low	low	high	high
OURMAPCN-score	low	low	low	unclear	unclear
PAWNN score	low	low	low	high	high
PH-(Patient History) COVID-19 risk score	low	low	low	high	high
PRIEST	low	low	low	low	low
qCSI (quick Covid-19 Severity Index)	low	low	high	high	high
RECOILS (Rapid Evaluation of Coronavirus Illness Severity score)	low	low	low	high	high
Sarkar	low	low	unclear	high	high
SARS2 risk score (Hospitalization)	low	low	low	unclear	unclear
SEIMC	low	low	low	low	low
SIMI score	low	low	low	high	high
Webb (Hospitalization)	low	low	unclear	high	high
Webb (Mortality)	low	low	unclear	high	high
Yamada (age 18-39 years)	unclear	low	low	high	high
Yamada (age 40-64 years)	unclear	low	low	high	high

We performed the analysis according to PROBAST guidelines[16]. "No information" if no information was provided. If we rated 2.2 as “no information” we allocated a low overall rating for the predictor domain, to acknowledge the little ROB for objective predictors, where applicable.

Table S7

Patient characteristics of the ORCHESTRA WP2 cohort stratified by subpopulation.

Characteristic	All	Hospitalized patients	ICU patients	Outpatients
	N = 5,048[1]	N = 3,359[2]	N = 932[2]	N = 1,602[2]
Cohort				
COVIDHome	158/5,048 (3%)	7/3,359 (0%)	0/932 (0%)	157/1,602 (10%)
INSERM	2,263/5,048 (45%)	2,262/3,359 (67%)	737/932 (79%)	22/1,602 (1%)
SAS	133/5,048 (3%)	105/3,359 (3%)	5/932 (1%)	28/1,602 (2%)
UNIBO	382/5,048 (8%)	367/3,359 (11%)	48/932 (5%)	8/1,602 (0%)
UNIVR	2,112/5,048 (42%)	618/3,359 (18%)	142/932 (15%)	1,387/1,602 (87%)
Age	59.4 (20.0)	60.6 (14.8)	60.7 (12.0)	56.6 (27.8)
Sex				
Female	2,193 (43.5%)	1,247/3,359 (37%)	239/932 (26%)	895/1,601 (56%)
Male	2,846 (56.5%)	2,112/3,359 (63%)	693/932 (74%)	706/1,601 (44%)
Body mass index	28.8 (12.0)	28.6 (7.2)	29.5 (5.7)	29.1 (17.9)
Smoking history				
Yes	310 (6.4%)	162/3,349 (5%)	40/931 (4%)	149/1,570 (9%)
Former smoker	1,186 (24.3%)	812/3,349 (24%)	242/931 (26%)	386/1,570 (25%)
Non-smoker	2,601 (53.4%)	1,789/3,349 (53%)	518/931 (56%)	834/1,570 (53%)
Unknown if ever smoked	778 (16.0%)	586/3,349 (17%)	131/931 (14%)	201/1,570 (13%)
Course of disease				
WHO clinical progression scale	3.4 (1.8)	5.3 (1.3)	7.0 (1.4)	2.2 (0.7)
Hospitalization				55/1,602 (3%)
Severity[3]	1,112 (22.0%)	1,112/3,359 (33%)	932/932 (100%)	20/1,602 (1%)
Mortality (ever)	28 (0.6%)	24/3,359 (1%)	14/932 (2%)	9/1,602 (1%)
Vital signs				
Respiratory rate (breaths/min)	22.7 (8.6)	23.6 (8.0)	26.7 (8.8)	17.9 (9.5)
Oxygen saturation (%)	94.8 (5.3)	94.4 (5.6)	92.8 (6.8)	96.6 (4.5)
Heart rate (beats/min)	86.1 (17.1)	87.8 (17.3)	90.9 (18.3)	77.0 (12.4)
Systolic blood pressure (mmHg)	129.7 (20.2)	129.7 (20.6)	130.3 (21.3)	130.1 (17.4)
Biochemistry				
Leukocytes (10^9/L)	14.3 (210.1)	14.6 (213.7)	20.7 (316.7)	6.5 (3.5)
Lymphocytes (10^9/L)	4.1 (57.5)	4.3 (60.1)	2.4 (24.3)	1.7 (1.5)
Neutrophils (10^9/L)	14.9 (239.3)	16.3 (251.4)	7.2 (8.6)	2.3 (2.2)
Thrombocytes (10^9/L)	577.2 (9,785.2)	590.1 (9,952.8)	387.7 (4,274.8)	203.3 (71.2)
Neutrophils-to-lymphocytes ratio	372.7 (14,746.7)	408.2 (15,435.0)	0.2 (0.4)	1.2 (0.9)
C-reactive protein (mg/L)	85.7 (93.3)	93.2 (94.5)	131.0 (113.5)	21.1 (41.8)
Interleukine 6 (pg/mL)	25.0 (14.3)	25.0 (14.3)	26.5 (15.8)	NA (NA)
Lactate dehydrogenase (U/L)	326.8 (201.6)	348.1 (211.7)	463.0 (341.8)	221.1 (65.7)
Creatinine (mol/L)	0.1 (3.6)	0.2 (3.7)	0.0 (0.2)	0.0 (0.0)
Urea (mmol/L)	7.1 (8.4)	7.1 (8.4)	8.1 (11.7)	7.8 (3.5)
D-dimer (mg/L)	6.6 (64.7)	8.1 (71.7)	7.9 (70.9)	0.7 (2.5)

Comorbidities				
At least one comorbidity	3,080/5,048 (61%)	2,148/3,359 (64%)	630/932 (68%)	960/1,602 (60%)
Chronic liver disease	121/4,855 (2%)	96/3,348 (3%)	18/928 (2%)	29/1,552 (2%)
Chronic kidney disease				
Chronic kidney disease with dialysis	18/4,855 (0%)	12/3,345 (0%)	2/928 (0%)	5/1,554 (0%)
Chronic kidney disease without dialysis	283/4,855 (6%)	227/3,345 (7%)	69/928 (7%)	59/1,554 (4%)
Yes, not specified (N18.9)	3/4,855 (0%)	1/3,345 (0%)	0/928 (0%)	2/1,554 (0%)
No chronic kidney disease present	4,497/4,855 (93%)	3,092/3,345 (92%)	857/928 (92%)	1,445/1,554 (93%)
Unknown if chronic kidney disease present	629/5,048 (12%)	536/3,359 (16%)	162/932 (17%)	98/1,602 (6%)
Chronic heart disease[4]	3,266/5,048 (65%)	2,069/3,359 (62%)	532/932 (57%)	1,090/1,602 (68%)
Hypertension				
Diabetes mellitus	33/4,819 (1%)	24/3,297 (1%)	5/894 (1%)	10/1,565 (1%)
Yes, diabetes type 1 (E10)	446/4,819 (9%)	320/3,297 (10%)	90/894 (10%)	128/1,565 (8%)
Yes, diabetes type 2 (E11)	166/4,819 (3%)	163/3,297 (5%)	56/894 (6%)	7/1,565 (0%)
Yes, not specified (E14)	8/4,819 (0%)	4/3,297 (0%)	0/894 (0%)	5/1,565 (0%)
Yes, other type of diabetes (E12-E14)	4,166/4,819 (86%)	2,786/3,297 (85%)	743/894 (83%)	1,415/1,565 (90%)
No	381/5,048 (8%)	251/3,359 (7%)	59/932 (6%)	135/1,602 (8%)
Asthma (J45.9)	113/5,048 (2%)	64/3,359 (2%)	3/932 (0%)	49/1,602 (3%)
Chronic obstructive pulmonary disease (COPD, J44.9)	8/5,048 (0%)	3/3,359 (0%)	1/932 (0%)	5/1,602 (0%)
Pulmonary hypertension (I27.0)	11/5,048 (0%)	3/3,359 (0%)	0/932 (0%)	8/1,602 (0%)
Restrictive lung disease (J98.4)	158/5,048 (3%)	7/3,359 (0%)	0/932 (0%)	157/1,602 (10%)

[1] n (%); Median (IQR); [2] n/N (%); Mean (SD); [3] defined as WHO Progression Scale ≥ 7 or ICU admission or ventilation (MV, CPAP/BIPAP) or vasopressors or death); [4] Comprises of Congestive heart failure (I50.0) and Coronary heart disease (I25.1).

Table S8

External validation results by score, population, timing, and outcome.

Download the Excel file here: https://cloud.idcohorts.net/s/btNXdoTedLzPwyH

The ending _data refers to the data-based cutoff calculated using the Youden Index, the ending _publ refers to the cutoff suggested in the original publication.

Table S9

Sensitivity analysis using 7 days of deviation from targeted timing of predictor measurement.

Download the Excel file here: https://cloud.idcohorts.net/s/yJ65HDDfpLx6Cjz

The ending _data refers to the data-based cutoff calculated using the Youden Index, the ending _publ refers to the cutoff suggested in the original publication.