Journal of Clinical Epidemiology 152 (2022) 257−268

Journal of Clinical Epidemiology

COVID-19 SERIES

Predicting COVID-19 prognosis in the ICU remained challenging: external validation in a multinational regional cohort

Daniek A.M. Meijs[a,b,c,*], Sander M.J. van Kuijk[d], Laure Wynants[d,e,f], Björn Stessel[g,h], Jannet Mehagnoul-Schipper[i], Anisa Hana[b,j], Clarissa I.E. Scheeren[k], Dennis C.J.J. Bergmans[a,l], Johannes Bickenbach[m], Margot Vander Laenen[n], Luc J.M. Smits[d], Iwan C.C. van der Horst[a,c], Gernot Marx[m], Dieter Mesotten[h,n], Bas C.T. van Bussel[a,c,d], CoDaP Investigators

[a]Department of Intensive Care Medicine, Maastricht University Medical Centre (Maastricht UMC+), Maastricht, The Netherlands
[b]Department of Intensive Care Medicine, Laurentius Ziekenhuis, Roermond, The Netherlands
[c]Cardiovascular Research Institute Maastricht (CARIM), Maastricht, The Netherlands
[d]Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands
[e]Department of Development and Regeneration, KULeuven, Leuven, Belgium
[f]Epi-centre, KULeuven, Leuven, Belgium
[g]Department of Intensive Care Medicine, Jessa Hospital, Hasselt, Belgium
[h]Faculty of Medicine and Life Sciences, UHasselt, Diepenbeek, Belgium
[i]Department of Intensive Care Medicine, VieCuri Medisch Centrum, Venlo, The Netherlands
[j]Department of Intensive Care Medicine, University Hospital of Zurich, Zurich, Switzerland
[k]Department of Intensive Care Medicine, Zuyderland Medisch Centrum, Heerlen/Sittard, The Netherlands
[l]School of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University, Maastricht, The Netherlands
[m]Department of Intensive Care Medicine, University Hospital Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Aachen, Germany
[n]Department of Intensive Care Medicine, Ziekenhuis Oost-Limburg, Genk, Belgium

Accepted 19 October 2022; Published online 27 October 2022

Abstract

Objectives: Many prediction models for coronavirus disease 2019 (COVID-19) have been developed. External validation is mandatory before implementation in the intensive care unit (ICU). We selected and validated prognostic models in the Euregio Intensive Care COVID (EICC) cohort.

Author contributions: All authors contributed substantially to this study. Conceptualization and study design: SMJvK, LW, BS, JM-S, AH, CIES, JB, LJMS, ICCvdH, GM, DM, BCTvB. Methodology: DAMM, SMJvK, LW, LJMS, ICCvdH, BCTvB. Software: DAMM, SMJvK, LW, LJMS, ICCvdH, BCTvB. Validation: DAMM, BCTvB. Formal analysis: DAMM, SMJvK, LW, LJMS, ICCvdH, BCTvB. Investigation: DAMM, BS, JM-S, AH, CIES, JB, DM, BCTvB. Resources: DAMM, SMJvK, LW, BS, JM-S, AH, CIES, DCJJB, JB, MvL, LJMS, ICCvdH, GM, DM, BCTvB, CoDaP investigators. Data curation: DAMM, SMJvK, LW, LJMS, BCTvB, CoDaP investigators. Writing−original draft: DAMM, ICCvdH, BCTvB. Writing−review and editing: DAMM, SMJvK, LW, BS, JM-S, AH, CIES, DCJJB, JB, MvL, LJMS, ICCvdH, GM, DM, BCTvB, CoDaP investigators. Visualization: DAMM, SMJvK, LW, LJMS, ICCvdH, BCTvB. Supervision: DAMM, ICCvdH, BCTvB. Project administration: DAMM, BCTvB. Funding acquisition: BS, ICCvdH, GM, DM, BCTvB.

Declaration of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent: Informed consent was waived by the METC of Maastricht UMC+. However, each of the participating hospitals had its own policy and approach. For example, in Maastricht UMC+, the board of directors adopted a policy to inform patients and ask their consent to use collected data. Our manuscript does not contain individual person's data. Therefore, consent to publish does not apply.

Data availability: The datasets generated and/or analyzed during the current study are not publicly available due to data-sharing agreements of the participating hospitals. Individual patient data and the pseudo-anonymized dataset will not be made available to others.

Funding: This work was supported by the "Interreg Euregio Meuse-Rhine" [Covid Data Platform (CoDaP) grant: Interreg-EMR 187]. The funding source was not involved in the study design, data collection, data analysis, data interpretation, writing process, and decision to submit for publication.

* Corresponding author: Maastricht UMC+, Department of Intensive Care Medicine, P. Debyelaan 25, 6229 HX Maastricht, The Netherlands. Tel.: +31620126764; fax: +31433874330.

E-mail address: daniek.meijs@mumc.nl (D.A.M. Meijs).

https://doi.org/10.1016/j.jclinepi.2022.10.015

Study Design and Setting: In this multinational cohort study, routine data from COVID-19 patients admitted to ICUs within the Euregio Meuse-Rhine were collected from March to August 2020. COVID-19 models were selected based on model type, predictors, outcomes, and reporting. Furthermore, general ICU scores were assessed. Discrimination was assessed by area under the receiver operating characteristic curves (AUCs) and calibration by calibration-in-the-large and calibration plots. A random-effects meta-analysis was used to pool results.

Results: 551 patients were admitted. Mean age was 65.4 ± 11.2 years, 29% were female, and ICU mortality was 36%. Nine out of 238 published models were externally validated. Pooled AUCs were between 0.53 and 0.70 and calibration-in-the-large between −9% and 6%. Calibration plots showed generally poor but, for the 4C Mortality score and Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC) score, moderate calibration.

Conclusion: Of the nine prognostic models that were externally validated in the EICC cohort, only two showed reasonable discrimination and moderate calibration. For future pandemics, better models based on routine data are needed to support admission decision-making.

Keywords: COVID-19; SARS-CoV-2; Critical care; Intensive care unit; Prediction; Prognosis

1. Introduction

During the coronavirus disease 2019 (COVID-19) pandemic, many prediction models were developed for diagnostic and prognostic purposes. The accurate prediction was paramount to support clinical decision-making, particularly during the early phase of the pandemic when little was known about the manifestations of the disease caused by the new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Furthermore, prediction of patient outcomes can improve effective management of bed availability in times of a pandemic where knowledge and capacity are under pressure. This was especially the case in the intensive care unit (ICU), as many patients with severe SARS-CoV-2 infection required organ support there [1,2].

A prediction model needs to meet several criteria to be useful in daily clinical practice. In the third update of the living systematic review by Wynants et al. [3], 238 prediction models for prognosis and diagnosis in COVID-19 have been identified and assessed for risk of bias. The risk of bias of all included models was evaluated as being high or, at best, unclear. For a model to perform well, both discrimination and calibration are important. In addition, model predictors must be routinely available. Furthermore, models need to be applicable to the population and settings requiring prediction, such as prognosis in the ICU, particularly during scarce bed availability. However, external validation of prediction models, which means testing the model in another sample of patients than it has been developed in, is often omitted, particularly in the ICU [4]. External validation is essential to generalize results to future patients and should precede the implementation of models in daily clinical practice [5,6]. Several external validation studies of prediction models for COVID-19 patients have been conducted. However, these studies focused mostly on patients admitted to the hospital ward instead of the ICU [7−9]. There is still a lack of ICU-specific prediction models, and applicability of general models to the ICU population is likely possible for some models only [3,10].

Therefore, we aimed to evaluate the predictive performance of published prediction models by selecting promising prognostic prediction models with clinically available predictors for external validation in our multinational COVID-19 cohort consisting of patients admitted to the ICUs within the Euregio Meuse-Rhine. As the majority of the 238 evaluated prediction models were developed at the beginning of the pandemic, we used data from the first pandemic wave for external validation.

2. Materials and methods

The paper is reported according to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosisclustered data reporting guideline [11−14]. Every section of the Materials and Methods is detailed in Appendix A.2.

2.1. Model selection

Prognostic prediction models for COVID-19 patients in the ICU were identified and extracted from https://www.covprecise.org/, the international Precise Risk Estimation to optimise COVID-19 Care for Infected or Suspected patients in diverse sEttings (COVID-PRECISE) group, in collaboration with the Cochrane Prognosis Methods Group according to the living systematic review of Wynants et al. (Fig. 1) [3]. Inclusion and exclusion criteria are described in Appendix A.2.1 and the selection process is shown in Fig. 1.

What is new?

Key findings

- Of 238 reviewed prognostic prediction models, nine were externally validated in the ICU.
- Only two out of these nine models showed reasonable discrimination and moderate calibration.

What this adds to what was known?

- External validation of prediction models is often omitted in the ICU.
- Despite great efforts have been made to develop prediction models early in the pandemic, their clinical value to support decision-making in the ICU is, overall, poor.

What is the implication and what should change now?

- For future pandemics, better prediction models based on routine data are needed to support admission decision-making.

2.2. External validation cohort

All patients with polymerase chain reaction and/or chest computed tomography scan confirmed COVID-19 and respiratory failure admitted to ICUs of any of the seven participating Euregio hospitals were consecutively included between March 2, 2020, and August 12, 2020 (Fig. 2) [17]. Hence, the study sample size was determined pragmatically. An extensive description of our methods and cohort has been described in Appendix A.2.2 and elsewhere [16,18].

2.3. Predictors

Using a predefined study protocol [16,18], predictor data up to 24 hours of ICU admission were acquired from electronic medical records and manually or electronically collected depending on the center. The collected variables used as predictors and outcomes are described in A.2.3. and Table A.1 of the Appendix [19]. Unknown, inappropriate, and inapplicable data were considered missing at random since missingness of data were related to other variables in the dataset and unlikely to be related to the true value itself [20–22].

2.4. Outcomes

Follow-up ended when patients were either discharged from the ICU or died in the ICU and was determined as ICU discharge or death. Patients whose outcome status after transportation could not be retrieved after recontacting the hospital were censored (Appendix A.2.4). Sensitivity analyses were performed without censored patients.

2.5. Description of included prediction models

The study characteristics of included prediction models and risk of bias are described in more detail in Appendix A.2.5 [23–30]. The risk of bias of the individual studies was scored by Wynants et al. [3] using the Prediction model study Risk Of Bias Assessment Tool (PROBAST) [15].

2.6. Ethics approval

Ethical approval was obtained from the medical ethics committee (Medisch Ethische Toetsingscommissie 2020-1565/3 00 523) of Maastricht UMC+.

2.7. Statistical analyses

IBM SPSS Statistics version 25 (IBM corporation, NY, USA) and R version 4.0.4 were used for all analyses. Microsoft PowerPoint version 16.59 was used to create figures. Data are presented as mean ± SD, median [IQR], or percentages. Descriptive statistics were performed for the whole cohort as well as for the individual Euregio countries. We included all patients in the analyses. In addition, sensitivity analyses were performed without censored transferred patients who, in the main analysis, contribute to the survived group. Missing data were imputed using multiple imputation if $<50\%$ of values on a variable were missing. Variables with more missings were omitted from the analysis. The number of imputations was based on the percentage of patients with missing data [31]. Continuous and categorical predictors were appropriately handled using the same definitions and cutoff values as the development study. The prognostic index was calculated for each patient by the sum of the models' regression coefficients, reported in the development studies, multiplied by the individual patient values. The prognostic index was transformed into a probability score when a model intercept was reported. For the Sequential Organ Failure Assessment (SOFA) score and the Acute Physiology And Chronic Health Evaluation II (APACHE II) score, risk scores instead of separate variables were already available for all patients and therefore directly assessed. The performance of the models was assessed by both discrimination and calibration measures. Model discrimination, the ability to separate patients who died in the ICU from those who are discharged, was determined as the area under the receiver operating characteristic (ROC) curve (AUC). An AUC of 0.5 implies inability to distinguish between those who die in the ICU and those who are discharged, whereas one means perfect discrimination. Model calibration refers

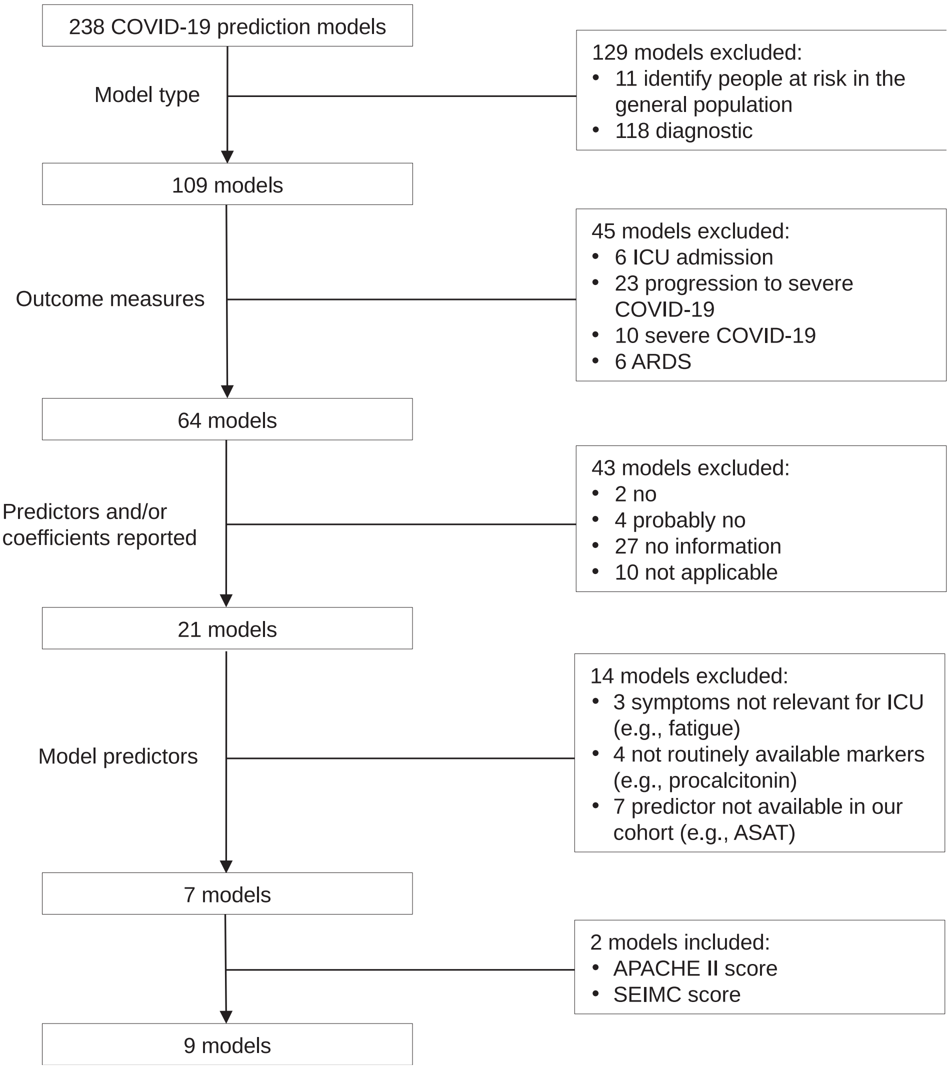

Fig. 1. Flowchart identifying prediction models. COVID-19, coronavirus disease 2019; ICU, intensive care unit; ARDS, acute respiratory distress syndrome; ASAT, aspartate aminotransferase. Legend: models for diagnosis and identifying people at risk in the general population were excluded. The remaining models were mainly prognostic, and further selection was based on outcome measures. As our cohort was composed of ICU patients only, in whom severe COVID-19 infection can be assumed, the outcome ICU admission, as well as progression to severe COVID-19, severe COVID-19, and ARDS, were excluded. Outcome measures length of hospital stay, in-hospital mortality, and in-hospital or out-of-hospital mortality were used. Since reporting of predictors and coefficients are necessary in order to validate prediction models as specifically assessed in step 4.9 in PROBAST [15], a tool to assess the risk of bias and applicability of prediction model studies, models which did not report or probably did not report this, or were machine learning or artificial intelligence studies, were excluded. Finally, predictors included in one of the final 21 prediction models were evaluated. Again, as we only included ICU patients and our goal was to validate models containing routinely available data, models including symptoms not relevant for ICU patients, not routinely available data, or data that were not available in the EICC cohort (e.g., $\geq$50% missing data) were excluded. Additionally, two promising models, which were not available in the COVID-PRECISE, were added. *Abbreviations:* PROBAST, Prediction model study Risk Of Bias Assessment Tool; EICC, Euregio Intensive Care COVID; COVID-PRECISE, Precise Risk Estimation to optimise COVID-19 Care for Infected or Suspected patients in diverse settings.

to the agreement between observed risk and the predicted risk [32,33]. Calibration was assessed by calibration-in-the-large (i.e., the difference between the predicted and observed probability of mortality) and by visual inspection of the calibration plot. Calibration could only be assessed in models that reported an intercept to calculate a probability instead of a unitless risk score only. The cohort was divided into deciles according to the estimated probability score,

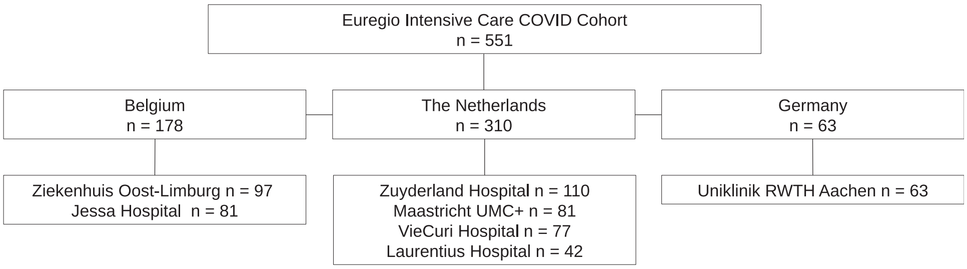
Euregio Intensive Care COVID Cohort
n = 551
Belgium
n = 178
The Netherlands
n = 310
Germany
n = 63
Ziekenhuis Oost-Limburg n = 97
Jessa Hospital n = 81
Zuyderland Hospital n = 110
Maastricht UMC+ n = 81
VieCuri Hospital n = 77
Laurentius Hospital n = 42
Uniklinik RWTH Aachen n = 63

Fig. 2. Flowchart Euregio Intensive Care COVID cohort [16].

displayed by points in the calibration plot. Perfect calibration is shown by the diagonal reference line, indicating agreement between predicted and observed probabilities over the range of predictions. Dots located above the reference line indicate underestimation by the model, while overestimation is reflected by the points below the reference line. Pooled AUCs and calibration-in-the-large were calculated for the three Euregio country parts using random-effects meta-analysis and 95% confidence intervals (CIs) were computed [12,13].

3. Results

3.1. Model selection

A total of 238 prediction models for COVID-19 were identified by COVID-PRECISE. Firstly, 129 models were excluded because they were diagnostic or not applicable to the ICU population (Fig. 1). Subsequently, 45 models were excluded due to unusable outcome measures such as ICU admission or severe COVID-19 pneumonia. Forty-three models were excluded as full information on predictors, intercepts, and coefficients was not present in the original article or supplement. Of the 21 potential prognostic models, three were not applicable since some predictors were not relevant for the ICU (e.g., cough, fatigue), four models included predictors that were not routinely available in Euregio ICUs (e.g., interleukin 6 or procalcitonin), and seven were excluded because they contained predictors that were more than 50% missing in our cohort. The APACHE II model [26] is widely used in the ICU and was added as prognostic model. The SOFA [30] and Confusion, Urea >7 mmol/l, Respiratory rate ≥ 30/min, low systolic (<90 mmHg) or diastolic (≤ 60 mmHg) Blood pressure, age ≥ 65 years (CURB-65) score [29], models that are also broadly implemented, were already included in the models selected via COVID-PRECISE. Furthermore, the Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC) score [27], which applied to the Euregio Intensive Care COVID (EICC) cohort, but was not available in COVID-PRECISE, was investigated. Thus, nine potential prognostic prediction models were selected for external validation. One model had an unclear risk of bias, five had a high risk of bias, and three models comprised already established prediction scores (Fig. 1 and Table 1).

3.2. External validation cohort

From March 2, 2020, to August 12, 2020, 551 patients with COVID-19 pneumonia were admitted to seven ICUs across the Netherlands, Belgium, and Germany (Fig. 2). Demographic and clinical characteristics and outcome measures are reported in Table 2 for the full EICC cohort and in Table A.2 (Appendix) for the individual country parts. The mean age of the cohort was 65.4 ± 11.2 years, the mean body mass index was 29.0 ± 5.3 kg/m^2, and 29% were female. At ICU admission, disease severity, as defined by APACHE II and SOFA scores, was 16.1 ± 5.5 and 6.2 ± 3.0.

3.3. Predictors

In our dataset, 309 (56%) of the patients had at least one missing value on any of the variables from the full set of predictors. Therefore, the number of imputations of the multiple imputation model was set to 56.

3.4. Outcomes

The ICU mortality rate was 36%, and the median [IQR] length of stay was 15.2 [6.0-29.9] days (Table 2). From 27 (5%) transported patients, survival status could not be retrieved after re-contacting individual hospitals and was therefore censored.

Table 1. Model characteristics of included prognostic prediction models

Study	Model	Derivation and validation cohort	Setting development study	Patients/disease
Unclear risk of bias prognostic model for COVID-19				
Knight et al. [23]	4C Mortality score	n = 35,463 (derivation) n = 22,361 (validation)	General hospital ward	Adults with COVID-19
High risk of bias prognostic models for COVID-19				
Zhang et al. [24]	DL-death	n = 775 (derivation) n = 226 (validation)	General hospital ward	Adults with RT-PCR confirmed COVID-19
Zhang et al. [24]	DCSL-death	n = 775 (derivation) n = 226 (validation)	General hospital ward	Adults with RT-PCR confirmed COVID-19
Wang et al. [25]	Clinical model	n = 286 (derivation) n = 44 (validation)	General hospital ward	RT-PCR/genetic testing confirmed, and imaging suspected COVID-19 cases
Bello-Chavolla et al. [28]	Mechanistic COVID-19 lethality score	n = 41,307 (derivation) n = 10,326 (validation)	Outpatients and general hospital ward	Suspected, confirmed and negative COVID-19 cases
Berenguer et al. [27]	SEIMC	n = 3,358 (derivation) n = 1,269 (validation)	General hospital ward	RT-PCR confirmed COVID-19 cases
Established prognostic models				
Lim et al. [29]	CURB-65 score	n = 718 (derivation) n = 214 (validation)	General hospital ward	CAP patients
Knaus et al. [26]	APACHE II score	n = 5,815 (validation)	ICU	Patients admitted to ICU
Vincent et al. [30]	SOFA score	n = 1,643 (derivation)	ICU	ICU patients (without short stay and postoperative patients)

Abbreviations: ICU, intensive care unit; COVID-19, coronavirus disease 2019; RT-PCR, reverse transcription-polymerase chain reaction; CAP, community-acquired pneumonia; CRP, C-reactive protein; SpO_2, peripheral capillary oxygen saturation; eGFR, estimated glomerular filtration rate; CKD-EPI, Chronic Kidney Disease Epidemiology Collaboration; FiO_2, fraction of inspired oxygen; PaO_2/FiO_2 ratio, the ratio of partial pressure of oxygen in arterial blood divided by the fraction of inspired oxygen.

[a] We only included models having mortality as outcome.

[b] One point was scored if systolic blood pressure was <90 mmHg or diastolic blood pressure was ≤ 60 mmHg.

3.5. Model performance

3.5.1. Unclear risk of bias prognostic model for COVID-19

The 4C Mortality score [23] had a pooled AUC of 0.70 (95% CI 0.64-0.76) for the full cohort (Table 3). Pooled calibration-in-the-large was −1% (95% CI −19% to 17%) (Table 3). The calibration plot is shown in Fig. 3. Sensitivity analyses (Table A.3 and Fig. A.1, Appendix) and country-specific analyses (Table A.4, Appendix) showed highly comparable discrimination. Calibration-in-the-large, however, varied between the three Euregio country parts (Table A.4, Appendix).

3.5.2. High risk of bias prognostic models for COVID-19

The DL-death and DCSL-death model [24] had a pooled AUC of 0.53 (95% CI 0.43-0.64) and 0.53 (95% CI 0.42-0.63), respectively. The pooled AUC of the clinical model [25] was 0.70 (95% CI 0.65-0.74), the mechanistic COVID-19 lethality score [28] 0.67 (95% CI 0.62-0.72), and the SEIMC [27] 0.70 (95% CI 0.65-0.74) (Table 3).

Pooled calibration-in-the-large were −2% (95% CI −14% to 10%) for the DL-death model, 6% (95% CI −6% to 18%) for the DCSL-death model, and −5% (95% CI −20% to 11%) for the SEIMC model (Table 3). Fig. 3 shows calibration plots for the DL-death, DCSL-death, and SEIMC models. Similar results were observed in sensitivity analyses (Table A.3 and Fig. A.1, Appendix). Minor differences in model discrimination existed between the three Euregio country parts, with the DL-death and DCSL-death having the lowest AUC in the Belgian part, whereas for the clinical model, mechanistic COVID-19 mortality score and SEIMC lowest AUCs were observed in the German part (Table A.4, Appendix).

Year, country	Predictors	Outcome
Unclear risk of bias prognostic model for COVID-19		
2020, England, Scotland, and Wales	Age, sex, number of comorbidities, respiratory rate, peripheral oxygen saturation, Glasgow Coma Scale, urea, CRP	Mortality
High risk of bias prognostic models for COVID-19		
2020, China and the United Kingdom	Age, sex, neutrophil count, lymphocyte count, platelet count, CRP, creatinine	Mortality (and poor outcome)[a]
2020, China and the United Kingdom	Age, sex, chronic lung disease, diabetes mellitus, malignancy, cough, dyspnea, neutrophil count, lymphocyte count, platelet count, CRP, creatinine	Mortality (and poor outcome)[a]
2020, China	Age, history of hypertension, history of coronary heart disease	Mortality
2020, Mexico	Age, diabetes, diabetes*age, obesity, pneumonia, chronic kidney disease, chronic obstructive pulmonary disease, immunosuppression	Mortality
2020, Spain	Age, low age-adjusted SaO_2, neutrophil-to-lymphocyte ratio, eGFR (CKD-EPI [19]), dyspnea, sex	Mortality
Established prognostic models		
2003, United Kingdom, New Zealand, and the Netherlands	Confusion, urea, respiratory rate, systolic or diastolic blood pressure[b], age	Mortality
1985, United States	Age, history of severe organ failure or immunocompromise, temperature, mean arterial pressure, pH, heart rate or pulse, respiratory rate, sodium, potassium, creatinine, acute kidney failure, hematocrit, white blood cell count, Glasgow Coma Scale, FiO_2	Mortality
1996, Europe and the United States	PaO_2/FiO_2, platelets, Glasgow Coma Scale, bilirubin, mean arterial pressure or vasoactive agents, creatinine	Mortality

Calibration-in-the-large, however, varied largely between the individual countries (Table A.4, Appendix).

3.5.3. Established prognostic models to predict mortality for acute respiratory illness and ICU patients

The pooled AUC was 0.68 (95% CI 0.64-0.73) for the CURB-65 score [29], 0.65 (95% CI 0.60-0.69) for the APACHE II score [26], and 0.62 (95% CI 0.56-0.68) for the SOFA score [30] (Table 3).

Pooled calibration-in-the-large was −9% (95% CI −21% to 3%) for the APACHE II score, and the calibration plot is shown in Fig. 3. Similar model performance was observed in sensitivity analyses (Table A.3 and Fig. A.1, Appendix). However, the German part had a lower AUC than the Belgian and Dutch Euregio parts, whereas calibration-in-the-large was best in the Belgian part (Table A.4, Appendix).

4. Discussion

In this study, we reviewed 238 prognostic prediction models for COVID-19 and externally validated nine using routinely available data in a multinational cohort of COVID-19 patients admitted to seven ICUs in Belgium, the Netherlands, and Germany during the first pandemic wave. In addition, established ICU prediction models were added for external validation in COVID-19 patients. Most studied models, among which prediction models for

Table 2. Characteristics for the full Euregio Intensive Care COVID cohort

Characteristics	**Full cohort** ***n* = 551**
Age, y	65.4 ± 11.2
Female, *n* (%)	159 (29)
Height, m	1.73 ± 0.1
Weight, kg	87.3 ± 17.1
Body mass index, kg/m^2	29.0 ± 5.3
Obesity, *n* (%)	175 (32)
Dyslipidemia, *n* (%)	149 (27)
Diabetes mellitus, *n* (%)	141 (26)
Hypertension, *n* (%)	260 (47)
Smoking, *n* (%)	112 (20)
Chronic liver disease, *n* (%)	4 (1)
Chronic lung disease, *n* (%)	101 (18)
Chronic kidney disease, *n* (%)	68 (12)
Myocardial infarction, *n* (%)	13 (2)
Chronic cardiac disease, *n* (%)	118 (21)
Dementia, *n* (%)	4 (1)
Neurological conditions, *n* (%)	64 (12)
Connective tissue disease, *n* (%)	11 (2)
HIV/aids, *n* (%)	0 (0)
Immunosuppression, *n* (%)	21 (4)
Malignancy, *n* (%)	63 (11)
APACHE II score	16.1 ± 5.5
SOFA score	6.2 ± 3.0
Admission location	
Emergency department, *n* (%)	184 (33)
Hospital ward, *n* (%)	277 (50)
Other ICU, *n* (%)	90 (16)
Glasgow Coma Scale at admission	14.7 ± 1.1
Respiratory rate at admission, /min	24.6 ± 7.1
SpO$_2$ at admission, %	91.4 ± 6.8
pH at admission	7.4 ± 0.1
Lowest PaO$_2$/FiO$_2$ ratio at admission	15.4 ± 10.6
Highest FiO$_2$ at admission, %	71.2 ± 21.5
Lowest MAP at admission, mmHg	68.5 ± 18.8
Heart rate at admission, bpm	93.1 ± 18.9
Vasopressor use at admission, *n* (%)	360 (65)
Creatinine at admission, µmol/L	101.2 ± 82.4
Urea at admission, mmol/L	11.6 ± 11.1
Dialysis at admission, *n* (%)	37 (7)
Bilirubin at admission, µg/L	10.0 ± 8.6
Thrombocytes at admission, *10^9/L	248.7 ± 105.7
Temperature at admission, ° Celsius	37.6 ± 1.2
CRP at admission, mg/L	184.8 ± 98.0
Neutrophils at admission, *10^9/L	8.3 ± 6.0
Lymphocytes at admission, *10^9/L	0.89 ± 11.6
Invasive mechanical ventilation during ICU stay, *n* (%)	434 (79)
Reintubation, *n* (%)	44 (8)
Duration of invasive mechanical ventilation, d	11.4 [2.3−23.0]
Mechanical circulatory support, *n* (%)	32 (6)
Kidney replacement therapy, *n* (%)	112 (20)
ICU mortality, *n* (%)	196 (36)
Length of ICU stay, d	15.2 [6.0−29.9]

Data are presented as mean ± SD, median [IQR], or percentages. HIV, human immunodeficiency virus; APACHE II, Acute Physiology And Chronic Health Evaluation II; SOFA, Sequential Organ Failure Assessment; ICU, intensive care unit; SpO$_2$, peripheral capillary oxygen saturation; PaO$_2$/FiO$_2$ ratio, the ratio of partial pressure of oxygen in arterial blood divided by the fraction of inspired oxygen; FiO$_2$, the fraction of inspired oxygen; MAP, mean arterial pressure; CRP, C-reactive protein.

COVID-19 rated as high risk of bias and established ICU scores, revealed poor performance regarding both discrimination and calibration. However, the 4C Mortality score and SEIMC showed reasonable model performance after external validation in an ICU cohort. Taken together, this shows that, despite the huge effort to develop many models early in the pandemic, their clinical value to support decision-making is, overall, poor. This highlights that data infrastructure for high-quality studies on model development, external validation, and implementation are required to improve data-driven decision support in future pandemics [34].

A direct comparison of model performance is hampered as case-mix differences exist between the model development population and the EICC cohort. These case-mix differences as well as possible explanations for the observed model performance, are extensively described in A.4 of the Appendix. Except for the APACHE II score and SOFA score, the included models were developed and/or validated in hospitalized patients or outpatients, with none of them or only a small subset of the cohort being admitted to the ICU. All patients included in the EICC cohort, on the contrary, were admitted to the ICU, indicating more severe illness and/or advanced disease course. Furthermore, in the ICU, patient selection likely played a role as patients with a high age and burden of comorbidities were often excluded from ICU admission. The EICC cohort reflects a case-mix with a relatively homogeneous population compared to model development studies on the hospital ward or general

Table 3. External validation of prognostic prediction models in the Euregio Intensive Care COVID cohort

Study	Model	Discrimination[a]	Calibration-in-the-large[b]
Unclear risk of bias prognostic model for COVID-19			
Knight et al. [23]	4C Mortality score	0.70 (95% CI 0.64–0.76)	−1% (95% CI −19% to 17%)
High risk of bias prognostic models for COVID-19			
Zhang et al. [24]	DL-death	0.53 (95% CI 0.43–0.64)	−2% (95% CI −14% to 10%)
Zhang et al. [24]	DCSL-death	0.53 (95% CI 0.42–0.63)	6% (95% CI −6% to 18%)
Wang et al. [25]	Clinical model	0.70 (95% CI 0.65–0.74)	-[c]
Bello-Chavolla et al. [28]	Mechanistic COVID-19 lethality score	0.67 (95% CI 0.62–0.72)	-[c]
Berenguer et al. [27]	SEIMC	0.70 (95% CI 0.65–0.74)	−5% (95% CI −20% to 11%)
Established prognostic models			
Lim et al. [29]	CURB-65 score	0.68 (95% CI 0.64–0.73)	-[c]
Knaus et al. [26]	APACHE II score	0.65 (95% CI 0.60–0.69)	−9% (95% CI −21% to 3%)
Vincent et al. [30]	SOFA score	0.62 (95% CI 0.56–0.68)	-[c]

Abbreviations: COVID-19, coronavirus disease 2019; SEIMC, Spanish Society of Infectious Diseases and Clinical Microbiology; APACHE II, Acute Physiology And Chronic Health Evaluation II; SOFA, Sequential Organ Failure Assessment; CI, confidence interval.

[a] Discrimination is reported as the pooled area under the ROC curve with 95% CI for all 56 imputed sets using random-effects meta-analysis. ROC, receiver operating characteristic; CI, confidence interval.

[b] Calibration-in-the-large is reported as the pooled difference between the predicted and observed mortality risk with 95% CI for all 56 imputed sets using random-effects meta-analysis. Positive values suggest overestimation, whereas negative values suggest underestimation. CI, confidence interval.

[c] Intercept not reported or risk score.

population, as patients at highest risk, who are not accepted for ICU admission, and lowest risk, not requiring intensive organ support were likely not included. However, considerable heterogeneity was observed in the EICC cohort [16], also illustrated by differences in model performance between the Euregio country parts. Since the discriminatory performance depends on case-mix variability, models developed or validated in hospitalized or outpatient populations showed lower AUCs after external validation in our relatively homogeneous ICU cohort [32,33]. Previous validation studies evaluating prediction models in other cohorts often included general populations, explaining why higher AUCs are observed compared to the EICC cohort. Therefore, it is inappropriate to directly compare AUC from validation studies in a general population to the ICU population. Nevertheless, high-quality prediction models could support a multifactorial decision when stress on ICU bed availability increases during a pandemic, particularly when driven by an intervening national healthcare policy [16,35].

4.1. Limitations

We evaluated nine prognostic models, including only one model at unclear risk of bias, five models at high risk of bias, and three established models with moderate to poor performance, which indicates that there is still a lack of well-performing and valid prediction models for the ICU population. However, we could not evaluate all high risk of bias prediction models as data on certain variables were missing, excluding these prediction models. Our analyses cannot provide evidence that other high risk of bias models should be discouraged, although as a proof of concept, our study may warrant caution, at the very least. Furthermore, we externally validated the APACHE II score instead of the more recent and advanced APACHE IV score [36] as data for the APACHE II score were more complete. Another limitation was the lack of information after transport to another ICU for 25 patients. However, we performed sensitivity analyses without these patients that showed comparable results. In addition, the original article of certain models did not report an intercept, and calibration could therefore not be assessed. The included COVID-19 prediction models were developed in the early phase of the pandemic and externally validated using patient data from the first pandemic wave. The dynamic development of the virus was not considered and, therefore, our results could not be generalized to ICU patients admitted later in the pandemic and suffering from other SARS-CoV-2 variants. However, the first pandemic wave data were used since the stress on healthcare systems and the accompanying need for prediction was highest during that period. As considerable heterogeneity is observed between SARS-CoV-2 variants and pandemic waves, models should be externally validated or updated in other pandemic wave cohorts [37,38]. Model updating and extension could further improve model performance which has not been performed yet [32,33]. Our study, therefore, sets the stage for model updating and extension of the promising 4C Mortality score and SEIMC model.

5. Conclusions

In this study, nine out of 238 available COVID-19 prognostic models were externally validated in the EICC cohort

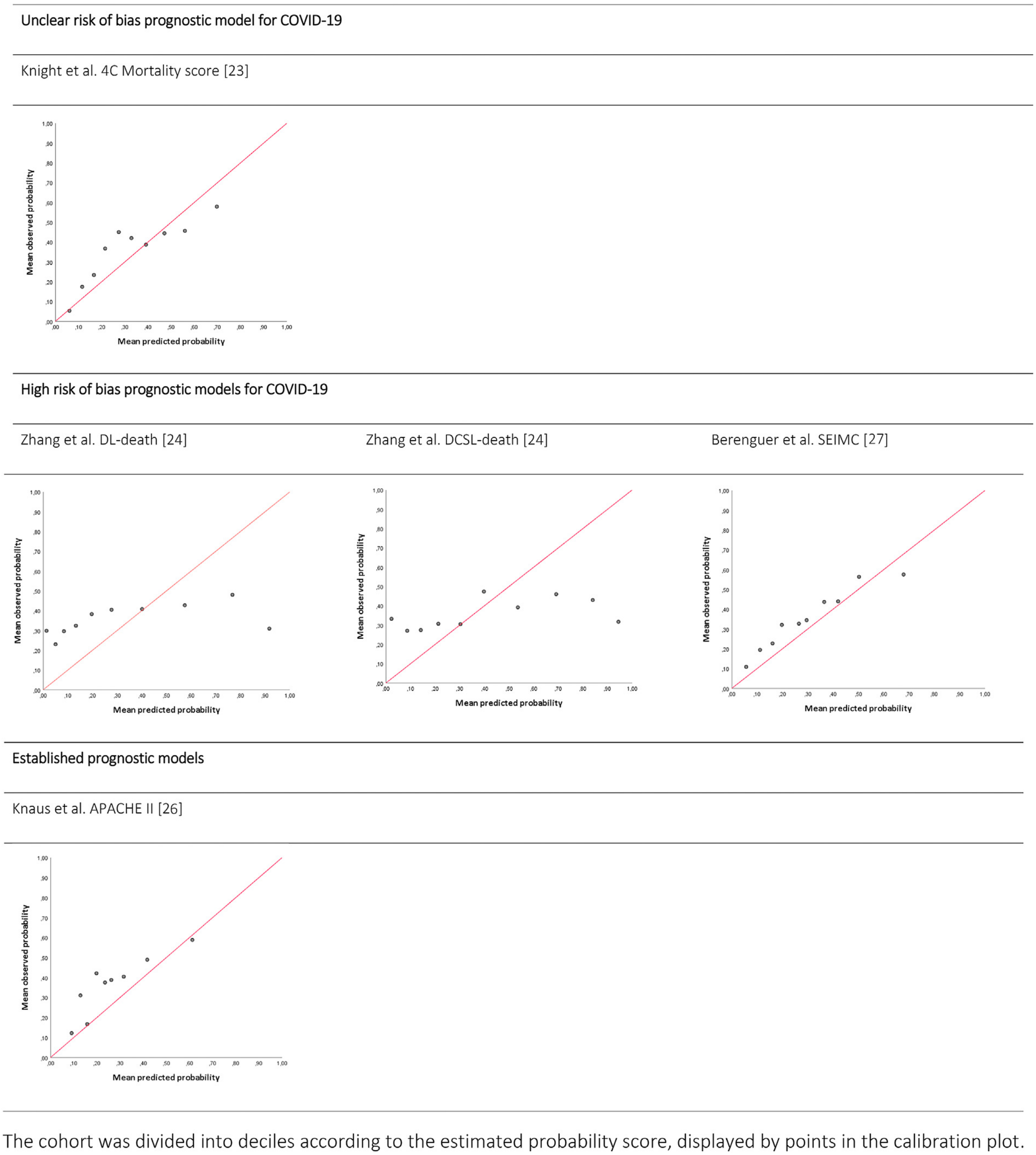

Fig. 3. Calibration plots prediction models. The cohort was divided into deciles according to the estimated probability score, displayed by points in the calibration plot.

based on routinely collected data. Only two of these nine models, the 4C Mortality score and the SEIMC, showed reasonable discrimination and moderate calibration. For future pandemics, better prediction models based on routine data are essential to improve data-driven decision support. Therefore, data infrastructure for high-quality studies on model development and external validation are required.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.10.015.

References

[1] Ma X, Vervoort D. Critical care capacity during the COVID-19 pandemic: global availability of intensive care beds. J Crit Care 2020;58:96–7.

[2] Douin DJ, Ward MJ, Lindsell CJ, Howell MP, Hough CL, Exline MC, et al. ICU bed utilization during the Coronavirus disease 2019 pandemic in a multistate analysis-March to June 2020. Crit Care Explor 2021;3(3):e0361.

[3] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. BMJ 2020;369:m1328.

[4] Keuning BE, Kaufmann T, Wiersema R, Granholm A, Pettila V, Moller MH, et al. Mortality prediction models in the adult critically ill: a scoping review. Acta Anaesthesiol Scand 2020;64(4):424−42.

[5] Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

[6] Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

[7] van Dam P, Zelis N, van Kuijk SMJ, Linkens A, Bruggemann RAG, Spaetgens B, et al. Performance of prediction models for short-term outcome in COVID-19 patients in the emergency department: a retrospective study. Ann Med 2021;53(1):402−9.

[8] Gupta RK, Marks M, Samuels THA, Luintel A, Rampling T, Chowdhury H, 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(6):2003498.

[9] Luo M, Liu J, Jiang W, Yue S, Liu H, Wei S. IL-6 and CD8+ T cell counts combined are an early predictor of in-hospital mortality of patients with COVID-19. JCI Insight 2020;5(13):e139024.

[10] de Jong VMT, Rousset RZ, Antonio-Villa NE, Buenen AG, Van Calster B, Bello-Chavolla OY, et al. Clinical prediction models for mortality in patients with COVID-19: external validation and individual participant data meta-analysis. BMJ 2022;378:e069881.

[11] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med 2015;162:735−6.

[12] Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG, Cochrane IPDM-aMg. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. Plos Med 2015;12(10):e1001886.

[13] Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.

[14] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.

[15] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1−33.

[16] Mesotten D, Meijs DAM, van Bussel BCT, Stessel B, Mehagnoul-Schipper J, Hana A, et al. Differences and similarities among COVID-19 patients treated in seven ICUs in three countries within one region: an observational cohort study. Crit Care Med 2022;50(4):595−606.

[17] Prokop M, van Everdingen W, van Rees Vellinga T, Quarles van Ufford H, Stoger L, Beenen L, et al. CO-RADS: a categorical CT assessment scheme for patients suspected of having COVID-19-definition and evaluation. Radiology 2020;296:E97−104. https://doi.org/10.1148/radiol.2020201473.

[18] Meijs DAM, van Bussel BCT, Stessel B, Mehagnoul-Schipper J, Hana A, Scheeren CIE, et al. Better COVID-19 Intensive Care Unit survival in females, independent of age, disease severity, comorbidities, and treatment. Sci Rep 2022;12(1):734.

[19] Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009;150:604−12.

[20] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.

[21] Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: the Treatment and Reporting of Missing data in Observational Studies framework. J Clin Epidemiol 2021;134:79−88.

[22] Perkins NJ, Cole SR, Harel O, Tchetgen Tchetgen EJ, Sun B, Mitchell EM, et al. Principled approaches to missing data in epidemiologic studies. Am J Epidemiol 2018;187:568−75.

[23] Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, 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.

[24] Zhang H, Shi T, Wu X, Zhang X, Wang K, Bean D, et al. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. medRxiv 2020. https://doi.org/10.1101/2020.04.28.20082222. [preprint: not peer reviewed].

[25] Wang K, Zuo P, Liu Y, Zhang M, Zhao X, Xie S, et al. Clinical and laboratory predictors of in-hospital mortality in patients with Coronavirus disease-2019: a cohort study in Wuhan, China. Clin Infect Dis 2020;71:2079−88.

[26] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13(10):818−29.

[27] Berenguer J, Borobia AM, Ryan P, Rodriguez-Bano J, Bellon JM, Jarrin I, 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(9):920−9.

[28] Bello-Chavolla OY, Bahena-Lopez JP, Antonio-Villa NE, Vargas-Vazquez A, Gonzalez-Diaz A, Marquez-Salinas A, et al. Predicting mortality due to SARS-CoV-2: a mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico. J Clin Endocrinol Metab 2020;105:dgaa346.

[29] Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 2003;58(5):377−82.

[30] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European society of intensive care medicine. Intensive Care Med 1996;22(7):707−10.

[31] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377−99.

[32] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.

[33] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Cham, Switzerland: Springer; 2019.

[34] Sperrin M, Grant SW, Peek N. Prediction models for diagnosis and prognosis in Covid-19. BMJ 2020;369:m1464.

[35] Bauer PR. Influence of geopolitics on severity and outcome in COVID-19. Crit Care Med 2022;50(4):700−2. Bauer PR. Influence of Geopolitics on Severity and Outcome in COVID-19. Crit Care Med 2022;50(4):700-702.

[36] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (Apache) IV: hospital mortality

assessment for today's critically ill patients. Crit Care Med 2006; 34(5):1297−310.

[37] Lambermont B, Rousseau AF, Seidel L, Thys M, Cavalleri J, Delanaye P, et al. Outcome improvement between the first two waves of the Coronavirus disease 2019 pandemic in a single tertiary-care hospital in Belgium. Crit Care Explor 2021;3(5):e0438.

[38] Carbonell R, Urgeles S, Rodriguez A, Bodi M, Martin-Loeches I, Sole-Violan J, et al. Mortality comparison between the first and second/third waves among 3,795 critical COVID-19 patients with pneumonia admitted to the ICU: a multicentre retrospective cohort study. Lancet Reg Health Eur 2021;11:100243.