scientific reports

OPEN

Boosting the accuracy of existing models by updating and extending: using a multicenter COVID-19 ICU cohort as a proxy

Daniek A. M. Meijs [1,2,3✉], Laure Wynants [4,5,6], Sander M. J. van Kuijk [7], Clarissa I. E. Scheeren[8], Anisa Hana[2,9], Jannet Mehagnoul-Schipper [10], Björn Stessel [11,12], Margot Vander Laenen[13], Eline G. M. Cox [1,14], Jan-Willem E. M. Sels [1,15], Luc J. M. Smits [4], Johannes Bickenbach [16], Dieter Mesotten [12,13], Iwan C. C. van der Horst [1,3], Gernot Marx [16], Bas C. T. van Bussel [1,3,4] & CoDaP Investigators*

Most published prediction models for Coronavirus Disease 2019 (COVID-19) were poorly reported, at high risk of bias, and heterogeneous in model performance. To tackle methodological challenges faced in previous prediction studies, we investigated whether model updating and extending improves mortality prediction, using the Intensive Care Unit (ICU) as a proxy. All COVID-19 patients admitted to seven ICUs in the Euregio-Meuse Rhine during the first pandemic wave were included. The 4C Mortality and SEIMC scores were selected as promising prognostic models from an external validation study. Five predictors could be estimated based on cohort size. TRIPOD guidelines were followed and logistic regression analyses with the linear predictor, APACHE II score, and country were performed. Bootstrapping with backward selection was applied to select variables for the final model. Additionally, shrinkage was performed. Model discrimination was displayed as optimism-corrected areas under the ROC curve and calibration by calibration slopes and plots. The mortality rate of the 551 included patients was 36%. Discrimination of the 4C Mortality and SEIMC scores increased from 0.70 to 0.74 and 0.70 to 0.73 and calibration plots improved compared to the original models after updating and extending. Mortality prediction can be improved after updating and extending of promising models.

Since the Coronavirus Disease 2019 (COVID-19) pandemic, there has been a high demand for prediction models to support clinical decision-making. Especially in intensive care, capacity was severely strained, and the hope was that prediction models could assist in decision-making[1,2]. Many prediction models for the diagnosis and prognosis of COVID-19 patients have been developed. All published diagnostic and prognostic models for COVID-19 were reviewed and were mostly at high risk of bias and poorly reported[3]. These methodological

[1]Department of Intensive Care Medicine, Maastricht University Medical Center + (Maastricht UMC+), P. Debyelaan 25, 6229 HX Maastricht, the Netherlands. [2]Department of Intensive Care Medicine, Laurentius Ziekenhuis, Roermond, the Netherlands. [3]Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, the Netherlands. [4]Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, the Netherlands. [5]Department of Development and Regeneration, KULeuven, Leuven, Belgium. [6]Epi-Centre, KULeuven, Leuven, Belgium. [7]Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht UMC+, Maastricht, the Netherlands. [8]Department of Intensive Care Medicine, Zuyderland Medisch Centrum, Heerlen/Sittard, the Netherlands. [9]Department of Intensive Care Medicine, University Hospital of Zurich, Zurich, Switzerland. [10]Department of Intensive Care Medicine, VieCuri Medisch Centrum, Venlo, the Netherlands. [11]Department of Intensive Care Medicine, Jessa Hospital, Hasselt, Belgium. [12]Faculty of Medicine and Life Sciences, UHasselt, Diepenbeek, Belgium. [13]Department of Intensive Care Medicine, Ziekenhuis Oost-Limburg (ZOL), Genk, Belgium. [14]Department of Intensive Care Medicine, University Medical Center Groningen (UMCG), Groningen, the Netherlands. [15]Department of Cardiology, Maastricht UMC+, Maastricht, the Netherlands. [16]Department of Intensive Care Medicine, University Hospital Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Aachen, Germany. *A list of authors and their affiliations appears at the end of paper. ✉email: daniek.meijs@mumc.nl

shortcomings could lead to overestimated model performance and, consequently, unreliable predictions that could cause harm when decisions are based on these numbers. A large individual participant data (IPD) meta-analysis that externally validated promising prognostic models showed that model performance varied strongly and there was substantial between cluster heterogeneity[4]. These methodological challenges have been identified accompanied by recommendations for future prediction research, among which model updating and extending of available prediction models using data from multiple countries and healthcare systems[3,4]. Updating and extending, also often referred to as recalibration and redesign, mean that the linear predictor (LP) of the original model is recalibrated (other regression coefficient) and that additional variables can be added to the score, e.g. a country variable[5–7].

Prior research on prediction modeling mainly focused on developing new models in each individual cohort rather than updating and extending promising existing models. Existing models tended to be rejected because of poor model quality or performance, so new ones were developed. This typically occurred in small datasets, particularly early in the pandemic[3]. Unfortunately, the same pattern then reoccurs, leading to many prediction models with limited generalizability that should be discouraged from being applied in daily patient care. However, all those rejected models contain predictive information that is valuable to a certain degree. For that reason, it seems beneficial to update and extend existing models resulting in models based on more evidence from more studies and, thus, more individuals. Hence, models are less overfitted, and performance in a new setting could be improved, while requiring less data[5–7]. Appropriate validation of these models will lead to more accurate predictions and better comparison between studies. Eventually, this could help healthcare providers with decision-making in daily patient care and healthcare policy, and ultimately improve patient outcomes.

Can outcome prediction in an Intensive Care Unit (ICU) population be improved taking all the important methodological considerations[3] and the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)[8] guideline into account? In this study, we therefore aim to show a complete and comprehensive method on how updating and extending of existing models can be performed. We choose a moderate-size multinational cohort of ICU patients early during the pandemic to show how to deal with moderate cohort size and a different setting than the developmental model cohorts, and used early pandemic data, as many new models were developed then, while updating was rare[3]. We hypothesize that model performance after updating and extending increases in the ICU setting. The multiregional ICU cohort will be used as an example, since there continues to be a lack of high-quality prognostic prediction models for the ICU setting nowadays[9]. The objective is to investigate whether mortality prediction by the 4C Mortality score and Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC), two models that showed reasonable model performance in a previous external validation study, can be further improved by model updating and extending[9–11].

Results
Cohort
In total, 551 patients with COVID-19 were admitted to seven ICUs within the Euregio Meuse-Rhine from March 2 to August 12, 2020 (Fig. 1). The median age of the cohort was 67 [15] years, and 29% were female (Table 1). Demographic and clinical characteristics, comorbidities, risk scores, and vital signs are shown in Table 1.

Outcomes
In the full cohort, 196 (36%) patients were deceased in the ICU (Table 1). From 27 (5%) patients, survival status remained unknown after contacting the centers where patients had been transferred to.

Predictors
Definitions, methods of measurement and missing values of all included predictors are described in Supplementary Table S1. The percentage of missing values for the predictors included in the updated and extended 4C Mortality score and SEIMC score varied from 0 to 18.1%.

The 4C Mortality score updating and extending
Logistic regression analyses were performed to update and extend the 4C Mortality score (Table 2). No predictors were excluded from the final model after bootstrapping with backward selection. The RCS function was thus
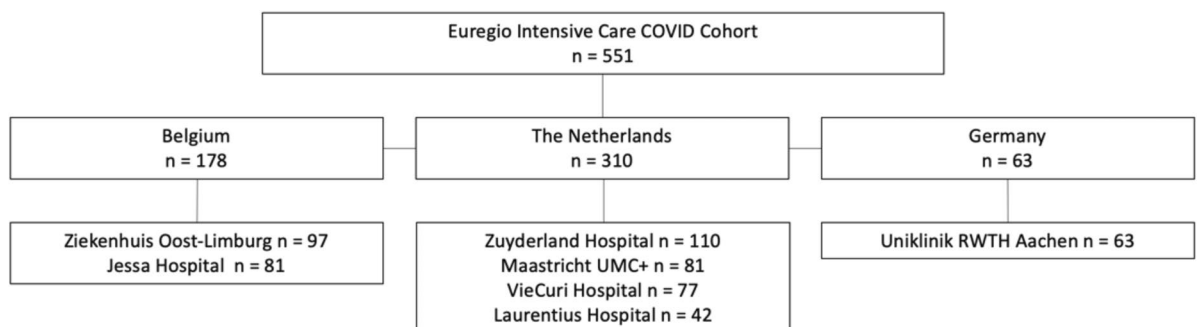

Figure 1. Flowchart EICC cohort[9,14,19].

Characteristics	Full cohort n = 551
Age, year	67.0 [15.0]
Female, n (%)	159 (29)
Body mass index, kg/m^2	28.1 [5.9]
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)
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)
Malignancy, n (%)	63 (11)
APACHE II score	16.0 [6]
Admission location	
Emergency department, n (%)	184 (33)
Hospital ward, n (%)	277 (50)
Other ICU, n (%)	90 (16)
GCS at admission	15 [0]
Respiratory rate at admission, /min	24.0 [8.0]
SpO$_2$ at admission, %	92.7 [7.0]
Lowest MAP at admission, mmHg	64.0 [15.0]
Heart rate at admission, bpm	92.0 [24.0]
Vasopressor use at admission, %	360 (65)
Temperature at admission, °C	37.7 [1.4]
Creatinine at admission, μmol/l	79.0 [48.0]
CRP at admission, mg/l	172.6 [155.0]
Neutrophils at admission, *10^9/l	7.2 [4.6]
Lymphocytes at admission, *10^9/l	0.7 [0.5]
Transfer to another ICU, n (%)	79 (14)
Invasive mechanical ventilation during ICU stay, n (%)	434 (79)
ICU mortality, n (%)	196 (36)
Length of ICU stay, days	15.2 [6.0–29.9]

Table 1. Baseline characteristics and primary outcomes of the EICC cohort. Data are presented as median [IQR] or absolute numbers (percentages). *HIV* human immunodeficiency virus, *APACHE II* Acute Physiology And Chronic Health Evaluation II, *ICU* Intensive Care Unit, *GCS* Glasgow Coma Scale, *SpO2* peripheral capillary oxygen saturation, *MAP* mean arterial pressure, *CRP* C-reactive protein.

included, meaning that the LP has a non-linear effect on survival (on the logit scale) in the ICU. Additionally, after accounting for the 4C Mortality score, a higher Acute Physiology and Chronic Health Evaluation (APACHE) II and being in a Dutch or German hospital increase mortality risks (Table 2).

SEIMC score updating and extending
Again, bootstrapping with backward selection did not result in exclusion of predictors from the final model indicating that the RCS function should be included in the final model. An increased SEIMC score leads to a higher mortality risk (Table 2). After adjustment for the SEIMC score, higher APACHE II scores and being admitted to a Dutch or German hospital also increase mortality risk.

Performance
Internal validation using bootstrapping yielded optimism-corrected areas under the receiver operating characteristic (ROC) curves of 0.74 and 0.73 for the 4C Mortality score and SEIMC score, respectively (Table 3).

Pooled calibration slopes for the updated and extended 4C Mortality score and SEIMC score were 0.97 and 0.96, respectively (Table 3). For both scores, the line in the flexible calibration plots closely approximates the diagonal with only slight under- and overestimation (Fig. 2).

Model	Predictors	Pooled coefficient (95% CI)
4C Mortality score	Intercept	− 1.06 (− 2.26 to 0.15)
	Linear predictor	1.24 (0.65 to 1.83)
	Linear predictor'	− 0.53 (− 1.09 to 0.02)
	APACHE II	0.06 (0.02 to 0.10)
	Dutch category	1.30 (0.80 to 1.79)
	German category	1.03 (0.38 to 1.68)
SEIMC score	Intercept	− 1.46 (− 2.61 to − 0.30)
	Linear predictor	0.83 (0.30 to 1.36)
	Linear predictor'	− 0.24 (− 0.80 to 0.31)
	APACHE II	0.06 (0.02 to 0.10)
	Dutch category	1.01 (0.55 to 1.46)
	German category	1.17 (0.51 to 1.83)

Table 2. Logistic regression coefficients of the updated and extended 4C Mortality score and SEIMC score. Data are presented as pooled coefficient estimate and 95% confidence interval per predictor included in the updated and extended model. *CI* confidence interval, *APACHE II* Acute Physiology And Chronic Health Evaluation II, *SEIMC* Spanish Society of Infectious Diseases and Clinical Microbiology.

Updated and extended model	Optimism-corrected area under ROC curve	Optimism-corrected calibration slope
4C Mortality score	0.74	0.97
SEIMC score	0.73	0.96

Table 3. Bootstrapping results of the updated and extended 4C Mortality score and SEIMC score. Data are presented as pooled optimism-corrected areas under the ROC curve and pooled optimism-corrected calibration slopes. *ROC* receiver operating characteristic, *SEIMC* Spanish Society of Infectious Diseases and Clinical Microbiology.

Shrinkage
To correct for overfitting, the beta regression coefficients reported in Table 2 were shrunken. Definitive beta regression coefficients of the updated and extended 4C Mortality score and SEIMC score are shown in Supplementary Table S2.

Comparison updated and extended models with original models
To compare the performance of the updated and extended 4C Mortality score and SEIMC score with the original 'crude' models, the areas under the ROC curve and calibration plots of this study were compared with external validation results in the EICC cohort[9]. The areas under the ROC curve from the 4C Mortality score increased from 0.70 to 0.74 after updating and extending (Supplementary Table S3). The area under the ROC curve rose from 0.70 to 0.73 after updating and extending the SEIMC score. Flexible calibration curves of the updated and extended 4C Mortality and SEIMC scores approached the 45 degrees diagonal more than the original scores (Fig. 2). For both scores, less underestimation in the lower predicted risks and less overestimation in the higher predicted risks was observed compared to the original model (Fig. 2).

Discussion
In this study, we were able to improve model discrimination and calibration by means of updating and extending of two promising prediction models in the ICU. While previous studies continuously focused on developing new models resulting in poor external validation results, we are the first that demonstrate that updating and extending of available models improve mortality prediction. Eventually, this approach could lead to better prediction models that bring us one step closer towards clinical implementation.

To date, many prognostic prediction models have been developed for COVID-19, but almost none focused on model updating and extending. A study in Japan with 160 non-ICU patients investigated the external validity of four prediction models that predicted respiratory support and death[12]. They found that these prediction models overestimated the presence of the outcome event, which improved after recalibration. The model extended with non-routinely available urinary β2-microglobulin (β2MG) did not improve model performance. We used available ICU data that are clinically relevant to accomplish models with the best implementation in clinical practice. Elmoheen et al. externally validated the CURB-65 and pneumonia severity index (PSI) in 1181 COVID-19 patients admitted to a hospital in Qatar[13]. They demonstrated improved discrimination for the PSI model and better calibration for both scores after recalibration. Recently, a large international IPD meta-analysis that externally validated the most promising COVID-19 prediction models showed that performance of prognostic
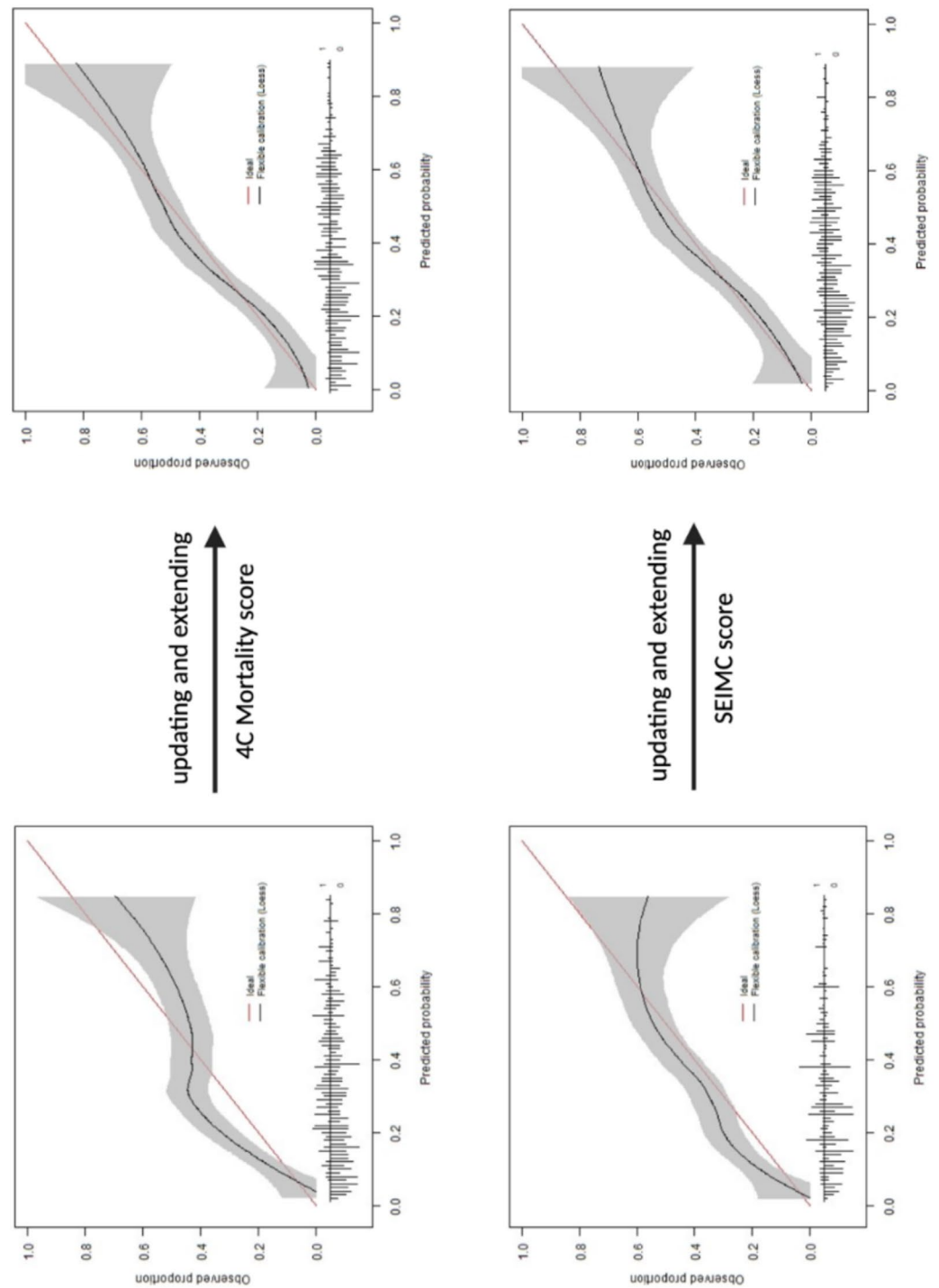
Figure 2. Calibration plots of the original compared to the updated and extended scores. The predicted probability of mortality is reported on the x-axis and the actual (observed) probability on the y-axis. The diagonal reflects optimal calibration and above the x-axis, histograms of predicted risk are illustrated for patients who died and survived in the ICU. In addition, 95% confidence boundaries are shown by the grey area.

models was heterogeneous across countries and that local and dynamic adjustments of promising models are required before clinical implementation[4]. We attempted to show how this could be achieved.

As recommended, it is important to compare the target setting and population to those from the original model to reveal possible sources of heterogeneity[3]. The 4C Mortality score development and validation cohort were considerably larger than the EICC cohort and comprised a population that presented at hospital admission[10]. Patients were included during the same pandemic wave. Baseline characteristics differed moderately, whereas mortality rates were comparable[9,14]. Similarly, the SEIMC cohort comprised first pandemic wave patients[11]. The cohort size was larger than the EICC cohort. Baseline characteristics varied, and mortality rates were lower in the SEIMC cohort compared to the EICC cohort[9,14]. Both scores were primarily developed in hospitalized patients instead of ICU patients, which indicates that patients in the EICC cohort were at more advanced disease stages or had more severe illnesses, indicating another setting. Additionally, patient selection likely plays a role in the ICU, especially in a pandemic when resources are limited. The EICC cohort reflects a different, more homogeneous case-mix than the general ward population resulting in worse discrimination in the former setting[6,7]. To adjust for this case-mix difference, the APACHE II score, an important disease severity prediction score for the ICU population, and country variables to correct for data complexity were added to the original models[14].

The main objective of the present study was not the deliver a valid COVID-19 prediction model for clinical practice in the ICU. Instead, we proposed a step-by-step approach to show model updating and extending of existing models according to the highest methodological standards taking a moderate cohort size into account[15]. Although, the COVID-19 ICU population was used as a proxy, this method could be applied to any other patient population following the TRIPOD guideline and recommendations by experts accurately[3,8]. The EICC dataset is representative for the ICU setting and includes patients from various healthcare systems and countries, improving understanding of generalizability and implementation across different settings and populations[3]. We show that model updating and extending can take the complexity in data structure into account by appropriately adding countries. We consider this essential in our heterogeneous EICC cohort[14]. Importantly, the research proposal and analyses were performed by a multinational and multidisciplinary team. With regard to the analyses, multiple imputation to appropriately handle missing data, pooling of parameter estimates and performance measures, adding restricted cubic splines to examine non-linearity, bootstrapping, backward selection and multiple imputation in each bootstrap sample to assess optimism, and shrinkage to correct for overfitting were performed. Finally, not only discrimination, but also calibration was shown by flexible calibration curves and calibration slopes, as appropriate.

We were limited by cohort size as our moderate cohort was sufficient for the prediction of only five predictors. Consequently, the individual predictors of the 4C Mortality score and SEIMC score could not be re-estimated separately. Nevertheless, we show that a LP, which was recalibrated as a reflection of the individual model predictors, leaves the opportunity to estimate two additional predictors. It is likely that model performance after updating and extending has improved by adding country as a factor due to heterogeneity within the EICC cohort. However, this heterogeneity highlights the importance of updating models to a new setting. The outcome status of 27 patients could not be retrieved after transport. Since these patients were classified as survivors, ICU mortality could have been underestimated. No external validation dataset was currently available to validate the updated and extended models. As a result, these findings could not be generalized to patients admitted later during the pandemic and patients admitted in the future. If our intention was to improve this model for application in clinical practice, external validation, updating of these adjusted models in other pandemic waves, and impact studies are essential additional steps before clinical implementation would be possible. Figure 3 shows the framework of the prediction process, demonstrating how updating and extending are integrated into this process (Fig. 3, Box). This requires more real time data for updating and extending during a pandemic, as the pandemic evolved faster than updating and extending were possible. In fact, this study has been overtaken by time as current virus variants differ from those in 2020 and stress on healthcare systems is considerably less than during the first pandemic wave. However, this study aimed to make a case against a focus on developing new prediction models on separate datasets as was done during the pandemic[3].

This approach of prediction model updating and extending is beneficial and could be applied to any available or future risk score in any setting. It provides the opportunity to increase the potential of a model that has originally been developed for as specific patient group and period, as it can be continuously updated across different settings and over time and enriched with new variables or simplified instead. This leads to more reliable and sustainable prediction models, that have increased potential for clinical implementation. Updating and extending increase efficiency because predictive information from several studies is combined, predictive performance is improved, generalizability is increased, and bias is reduced. In future pandemics or when the prediction of patients' outcomes is vital, better prediction of patients' outcomes could be realized by reviewing available literature, external validation of available prediction models, and updating and extending of these models, taking these steps into account. The data density in the ICU, combined with complicated patients' diseases, set the stage for more advanced regression and machine-learning techniques that consider dynamic and temporal predictor trends[16].

Box
Figure 3 shows the framework of the prediction process, demonstrating how updating and extending are integrated into this process. First, one should consider the goal, study design, and modeling strategy and compute a sample size before developing the model. After the model has been fit, it should be validated, preferably internally and externally, to evaluate its generalization. Updating and extending, as described in this study, will be the next stage. When a model shows adequate performance, an impact study can be conducted to assess its effectiveness in clinical practice. If effectivity has been proven, the model can be implemented in guidelines, training programs, software, and eventually clinical practice.
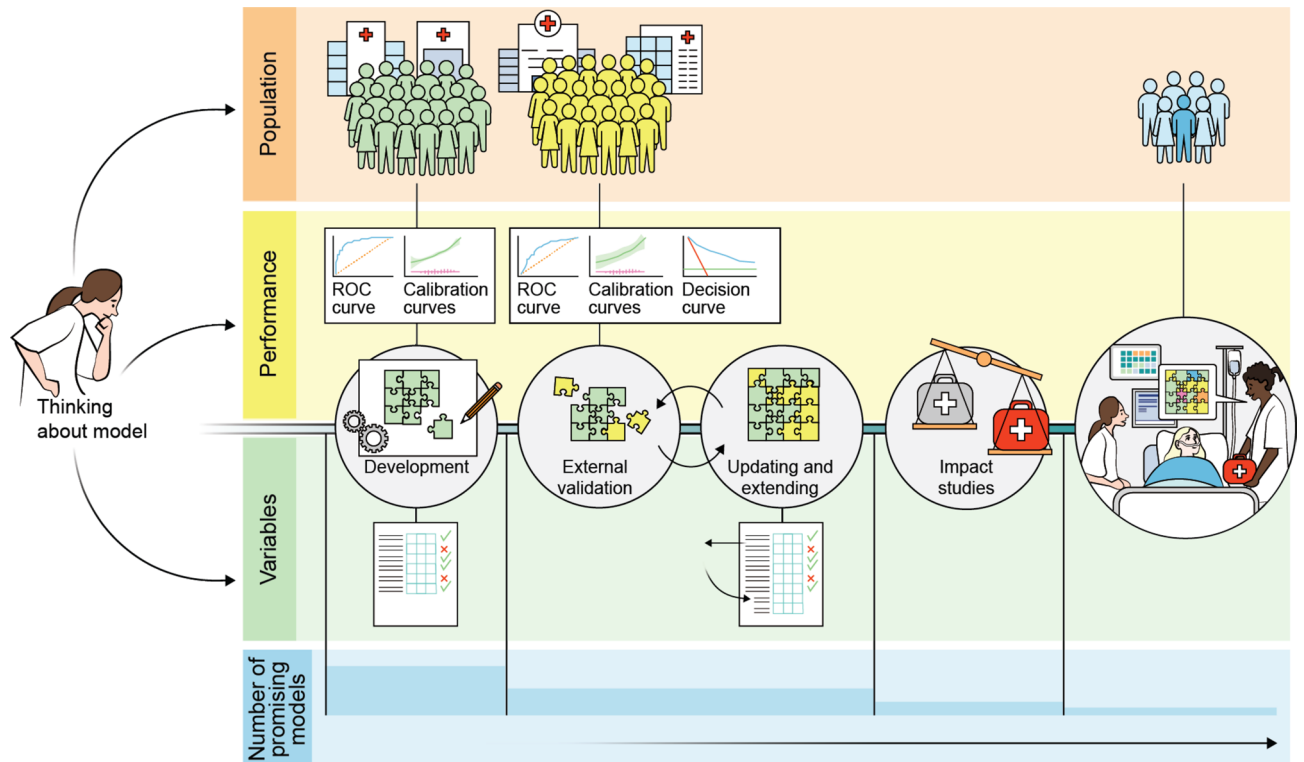

Figure 3. The prediction process. This visual representation conveys the sequential steps involved in the prediction process.

Methods

Guidelines

The TRIPOD guideline was followed (Supplementary File S4)[8].

Research population

As it was recommended to perform updating in individual patient data from multiple countries and healthcare systems[3], we used the Euregio Intensive Care Covid (EICC) cohort to address our research question. This retrospective cohort is part of the Interreg Covid Data Platform (CoDaP) project and includes seven ICU departments in the Euregio Meuse-Rhine that collaborated on COVID-19 during the first pandemic wave. The seven participating departments include the Intensive Care Medicine departments of Maastricht University Medical Center + (MUMC +, Maastricht, the Netherlands), Zuyderland Hospital (Heerlen/Sittard, the Netherlands), VieCuri Hospital (Venlo, the Netherlands), Laurentius Hospital (Roermond, the Netherlands), Ziekenhuis Oost-Limburg (Genk, Belgium), Jessa Hospital (Hasselt, Belgium), and University Hospital Rheinisch-Westfälische Hochschule (RWTH) Aachen (Aachen, Germany). All patients with confirmed COVID-19 and respiratory failure admitted to the ICU of any of the abovementioned hospitals were consecutively included between March 2 and August 12, 2020. COVID-19 diagnosis was based on either virus detection with polymerase chain reaction or a chest CT scan of 4–5 based on the COVID-19 Reporting and Data System (CO-RADS) score[17]. No exclusion criteria were set. Patients were admitted to the ICU via the emergency department, hospital ward, or transportation from other ICUs within or outside the Euregio because of tertiary care requirements or limited bed availability[18]. More detailed information on the cohort can be found in previous publications[9,14,19].

Sample size calculation

Updating and extending can be done with varying degrees of complexity. The more complex the updating or extending strategy, the more data needed to execute it. The sample size for the EICC cohort was determined pragmatically. All patients were included since there was a desperate need for COVID-19 research during the first pandemic wave. For this research question, we calculated the number of predictors that could be estimated for model updating and extending purposes based on the available sample size. Unfortunately, little research has been conducted to determine adequate sample sizes for prediction model updating. Several rules of thumb are often used to estimate sample sizes for prediction model development studies[6,7,15,20–24]. However, Riley et al.[15] stated that these are too simplistic and advocated that a more scientific approach that tailors sample size to the setting is required. Therefore, they developed a step-by-step guideline for sample size calculation in prediction models with binary outcomes, consisting of four separate steps, which were applied in this study. The lowest of the four sample size calculations per step was used to set the absolute number of predictors. Details of the sample size calculation are described in Supplementary File S5. To conclude, the pragmatic sample size of 551 patients

was sufficient to estimate a model with a maximum of five predictors. As the available cohort size was insufficient to re-estimate all predictors included in the two original scores, the original LP was included as a predictor in the updated model, and one regression coefficient was estimated for the LP.

Model selection
Candidate prognostic models for model updating and extending were selected from a previous external validation study of COVID-19 models and established ICU prediction models in the EICC cohort[9]. Of all nine included and externally validated models, the 4C Mortality and SEIMC scores demonstrated the best discrimination and calibration and were selected for model updating and extending[10,11]. As recommended[3], setting and model characteristics of the 4C Mortality and SEIMC scores have been detailed in Supplementary File S5.

Predictors
Early March 2020, a study protocol with certain demographic, anthropometric, vital, laboratory, and clinical variables was written and shared among the participating hospitals to construct the EICC cohort. In addition, numerous routinely available variables from admission to discharge were collected, among which the predictors included in the 4C Mortality score and SEIMC score. More information on collected predictors is outlined in Supplementary File S5.

Candidate predictors for model updating and extending
Target predictors for model updating and extending are predictors presumed to discriminate better in the target population and setting compared to the original model. ICU patients generally presented with more severe COVID-19 than ward patients. The established APACHE II[25] score discriminates between severe and mild illness in ICU patients and was, therefore, one of the chosen predictors to enrich the 4C Mortality and SEIMC score. Furthermore, country was added as a categorical predictor to consider the multinational nature of the cohort since heterogeneity in the EICC cohort was observed previously[14].

Outcomes
Patients were followed until the outcome occurred, either ICU death or ICU discharge to another hospital or the general hospital ward. These centers were contacted to retrieve the outcome status if patients were transported to other ICUs. If the outcome status remained unknown, patients were classified as survivors in the primary analyses with the potential risk of mortality underestimation.

Statistical analysis
IBM SPSS Statistics version 25 (IBM corporation, NY, USA) and R version 4.0.4 were used for the analyses. Data are presented as median [IQR] or percentages. Descriptive statistics were performed for the whole cohort. All patients were included in the analyses. Missing data were handled by multiple imputation if < 50% of values on a variable were missing; in other cases, variables were omitted from the analysis. Missing values were multiply imputed as documented elsewhere[9,14,19,26]. Continuous and categorical predictors were handled using the same definitions and cut-off values defined in the development study. For each patient, the LP (i.e., it summarizes the developed prediction model under investigation) was calculated by the intercept and sum of the models' regression coefficients, reported in the 4C Mortality score and SEIMC score development studies, multiplied by the individual patient values[10,11]. The LP was then transformed into a probability score using the inverse logit transformation.

Different methods for model updating and extending exist[5–7]. As the sample size was sufficient for estimating a maximum of five predictors for each prognostic prediction model, a combination of model recalibration and extending was performed instead of re-estimating all predictors in the model. Therefore, for each model, the model intercept and estimated slope of the LP were updated. Additionally, the APACHE II score and a country factor were added to extend the updated models. The Belgian country part functioned as a reference group. A logistic regression model was fitted for the 4C Mortality score and SEIMC score with the LP, APACHE II score, Dutch country category and German country category included as model predictors, and ICU mortality as the outcome. Finally, parameter estimates of the individual imputed sets were pooled using the total covariance matrix pooling[27].

In order to examine non-linear effects in the linear predictor, a restricted cubic spline (RCS) function with three knots for the LP was added to the model. Bootstrapping with backward selection using the Prediction Model Pooling, Selection and Performance Evaluation Across Multiply Imputed Datasets (psfmi) package was performed to determine whether the RCS function should be included in the model. A p-value less than 0.1 was considered to be statistically significant, advocating the inclusion of the RCS function in the final model[27]. After the updated and extended models had been fitted, bootstrapping was performed 200 times to validate the updated and extended models internally. To assess the optimism, multiple imputation and backward selection were repeated in each bootstrap sample with a p-value of 1.0[27]. Model performance was examined by discrimination and calibration[6–8,28,29]. Model discrimination implies the ability of a prediction model to distinguish between patients who develop the outcome of interest and those who do not. This is reflected by the optimism-corrected area under the ROC curve obtained after bootstrapping. An area under the ROC curve of 1 represents ideal discrimination, whereas 0.5 represents no discrimination[6,7]. The second parameter, model calibration, refers to the correspondence between the observed outcome proportions and predicted outcome risk. This is illustrated by the calibration intercept, the calibration slope, and flexible calibration curves[6,7,30]. Ideally, the calibration intercept is 0 and the calibration slope is 1. The optimism-corrected calibration slope was retrieved after bootstrapping[27]. Flexible calibration curves using local regression were constructed for each updated and

extended model and the original models[6,7,10,11,23]. To create these curves with a multiple imputed dataset, the mean LP of all multiple imputed sets was computed and used. The final step was the shrinkage of the regression coefficients towards zero to prevent overfitting. The beta regression coefficients from the updated and extended models were multiplied by the shrinkage factor based on the optimism-corrected calibration slope, resulting in shrunken regression coefficients[6,7]. Afterwards, the model intercept was re-estimated by logistic regression with the new LP as an offset variable.

Ethics approval

The medical ethics committee (Medisch Ethische Toetsingscommissie 2020–1565/3 00 523) of Maastricht UMC + provided ethical approval[31]. The study was conducted in accordance with the General Data Protection Regulation (GDPR) and national data privacy laws. Data sharing agreements were composed by legal officers of Maastricht UMC + and Clinical Trial Center Maastricht (CTCM), judged by the legal department of each hospital, tailored to each center, and then signed to ensure adequate and safe data sharing.

Conclusions

This study demonstrated a stepwise approach to prediction model updating and extending and showed that updating and extending of two promising prognostic COVID-19 prediction models lead to improved mortality prediction in the ICU. Instead of developing new models on separate datasets done during the pandemic, this study makes a case towards clinical implementation of prediction models that requires various steps, from literature reviewing of developed models, extensive data collection, and external validation, to updating and extending the most promising models.

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. Only data for the full cohort or a particular subcohort will be published and shared after the provision of a research proposal and signed data access agreement of each participating hospital. Please contact the corresponding author (Daniek.meijs@mumc.nl) to request these data.

References

1. Ma, X. & Vervoort, D. Critical care capacity during the COVID-19 pandemic: Global availability of intensive care beds. *J. Crit. Care* **58**, 96–97. https://doi.org/10.1016/j.jcrc.2020.04.012 (2020).
2. Douin, D. J. *et al.* ICU bed utilization during the Coronavirus Disease 2019 pandemic in a multistate analysis-March to June 2020. *Crit. Care Explor.* **3**, e0361. https://doi.org/10.1097/CCE.0000000000000361 (2021).
3. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. *BMJ* **369**, m1328. https://doi.org/10.1136/bmj.m1328 (2020).
4. de Jong, V. M. T. *et al.* Clinical prediction models for mortality in patients with covid-19: External validation and individual participant data meta-analysis. *BMJ* **378**, e069881. https://doi.org/10.1136/bmj-2021-069881 (2022).
5. Moons, K. G. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698. https://doi.org/10.1136/heartjnl-2011-301247 (2012).
6. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (Springer, 2019).
7. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (Springer, 2009).
8. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63. https://doi.org/10.7326/M14-0697 (2015).
9. Meijs, D. A. M. *et al.* Predicting COVID-19 prognosis in the ICU remained challenging: External validation in a multinational regional cohort. *J. Clin. Epidemiol.* **152**, 257–268. https://doi.org/10.1016/j.jclinepi.2022.10.015 (2022).
10. Knight, S. 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* **370**, m3339. https://doi.org/10.1136/bmj.m3339 (2020).
11. Berenguer, J. *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* **76**, 920–929. https://doi.org/10.1136/thoraxjnl-2020-216001 (2021).
12. Yamada, G. *et al.* External validation and update of prediction models for unfavorable outcomes in hospitalized patients with COVID-19 in Japan. *J. Infect. Chemother.* **27**, 1043–1050. https://doi.org/10.1016/j.jiac.2021.04.008 (2021).
13. Elmoheen, A. *et al.* External validation and recalibration of the CURB-65 and PSI for predicting 30-Day mortality and critical care intervention in multiethnic patients with COVID-19. *Int. J. Infect. Dis.* **111**, 108–116. https://doi.org/10.1016/j.ijid.2021.08.027 (2021).
14. Mesotten, D. *et al.* Differences and similarities among Coronavirus Disease 2019 patients treated in seven ICUs in three countries within one region: An observational cohort study. *Crit. Care Med.* **50**, 595–606. https://doi.org/10.1097/CCM.0000000000005314 (2022).
15. Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441. https://doi.org/10.1136/bmj.m441 (2020).
16. Xie, J. *et al.* A review of modeling methods for predicting in-hospital mortality of patients in intensive care unit. *J. Emerg. Crit. Care Med.* https://doi.org/10.21037/JECCM.2017.08.03 (2017).
17. Prokop, M. *et al.* CO-RADS: A categorical CT assessment scheme for patients suspected of having COVID-19-definition and evaluation. *Radiology* **296**, E97–E104. https://doi.org/10.1148/radiol.2020201473 (2020).
18. Strauch, U. *et al.* QUIT EMR trial: A prospective, observational, multicentre study to evaluate quality and 24 hours post-transport morbidity of interhospital transportation of critically ill patients: Study protocol. *BMJ Open* **7**, e012861. https://doi.org/10.1136/bmjopen-2016-012861 (2017).
19. Meijs, D. A. M. *et al.* Better COVID-19 Intensive Care Unit survival in females, independent of age, disease severity, comorbidities, and treatment. *Sci. Rep.* **12**, 734. https://doi.org/10.1038/s41598-021-04531-x (2022).

20. Collins, G. S., Ogundimu, E. O. & Altman, D. G. Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. *Stat. Med.* **35**, 214–226. https://doi.org/10.1002/sim.6787 (2016).
21. Harrell, F. E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (Springer, 2015).
22. Snell, K. I. E. *et al.* External validation of clinical prediction models: Simulation-based sample size calculations were more reliable than rules-of-thumb. *J. Clin. Epidemiol.* **135**, 79–89. https://doi.org/10.1016/j.jclinepi.2021.02.011 (2021).
23. Van Calster, B. *et al.* A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005 (2016).
24. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. & Habbema, J. D. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* **58**, 475–483. https://doi.org/10.1016/j.jclinepi.2004.06.017 (2005).
25. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: A severity of disease classification system. *Crit. Care Med.* **13**, 818–829 (1985).
26. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30**, 377–399. https://doi.org/10.1002/sim.4067 (2011).
27. Heymans, M. & Eekhout, I. *psfmi: Prediction Model Pooling, Selection and Performance Evaluation Across Multiply Imputed Datasets.* https://cran.r-project.org/web/packages/psfmi/index.html. Accessed July 2022.
28. Debray, T. P. *et al.* Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: Guidance on their use. *PLoS Med.* **12**, e1001886. https://doi.org/10.1371/journal.pmed.1001886 (2015).
29. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, i3140. https://doi.org/10.1136/bmj.i3140 (2016).
30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 230. https://doi.org/10.1186/s12916-019-1466-7 (2019).
31. Tas, J. *et al.* Serial measurements in COVID-19-induced acute respiratory disease to unravel heterogeneity of the disease course: Design of the Maastricht Intensive Care COVID cohort (MaastrICCht). *BMJ Open.* **10**, e040175. https://doi.org/10.1136/bmjopen-2020-040175 (2020).
32. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612. https://doi.org/10.7326/0003-4819-150-9-200905050-00006 (2009).
33. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383. https://doi.org/10.1016/0021-9681(87)90171-8 (1987).

Author contributions

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70333-6.

Correspondence and requests for materials should be addressed to D.A.M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CoDaP Investigators

Nanon F. L. Heijnen[1], Mark M. G. Mulder[1], Marcel Koelmann[1], Julia L. M. Bels[1], Nick Wilmes[1,3,15], Charlotte W. E. Hendriks[1], Emma B. N. J. Janssen[15], Micheline C. D. M. Florack[1,8], Chahinda Ghossein-Doha[1,3,15], Meta C. E. van der Woude[8], Laura Bormans-Russell[8], Noëlla Pierlet[13], Ben Goethuys[13], Jonas Bruggen[13], Gilles Vermeiren[13], Hendrik Vervloessem[13] & Willem Boer[13]