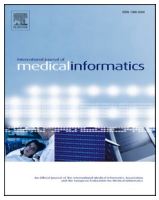
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Journal of Medical Informatics 160 (2022) 104688

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Can we reliably automate clinical prognostic modelling? A retrospective cohort study for ICU triage prediction of in-hospital mortality of COVID-19 patients in the Netherlands

I. Vagliano [a], S. Brinkman [b], A. Abu-Hanna [a], M.S Arbous [c], D.A. Dongelmans [d], P.W.G. Elbers [e], D.W. de Lange [f], M. van der Schaar [g], N.F. de Keizer [b], M.C. Schut [a,*], on behalf of The Dutch COVID-19 Research Consortium [1]

[a] *Department of Medical Informatics, Amsterdam University Medical Centers, Amsterdam Public Health research institute, Meibergdreef 9, 1105 AZ, Amsterdam, the Netherlands*
[b] *Department of Medical Informatics, Amsterdam University Medical Centers, Amsterdam Public Health research institute and National Intensive Care Evaluation (NICE) foundation, Meibergdreef 9, 1105 AZ, Amsterdam, the Netherlands*
[c] *Department of Intensive Care, Leiden University Medical Center, Leiden, the Netherlands*
[d] *Department of Intensive Care Medicine, Amsterdam University Medical Centers, Location AMC, Meibergdreef 9, 1105 AZ, Amsterdam, the Netherlands*
[e] *Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence (LCCCI), Amsterdam Medical Data Science (AMDS), Amsterdam UMC, De Boelelaan 1117, 1081 HV, Amsterdam, the Netherlands*
[f] *Department of Intensive Care Medicine and Dutch Poisons Information Center (DPIC), University Medical Center Utrecht, University Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands*
[g] *The Alan Turing Institute, University of California and University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom*

A B S T R A C T

Background: Building Machine Learning (ML) models in healthcare may suffer from time-consuming and potentially biased pre-selection of predictors by hand that can result in limited or trivial selection of suitable models. We aimed to assess the predictive performance of automating the process of building ML models (AutoML) in-hospital mortality prediction modelling of triage COVID-19 patients at ICU admission versus expert-based predictor pre-selection followed by logistic regression.
Methods: We conducted an observational study of all COVID-19 patients admitted to Dutch ICUs between February and July 2020. We included 2,690 COVID-19 patients from 70 ICUs participating in the Dutch National Intensive Care Evaluation (NICE) registry. The main outcome measure was in-hospital mortality. We asessed model performance (at admission and after 24h, respectively) of AutoML compared to the more traditional approach of predictor pre-selection and logistic regression.
Findings: Predictive performance of the autoML models with variables available at admission shows fair discrimination (average AUROC = 0·75-0·76 (sdev = 0·03), PPV = 0·70-0·76 (sdev = 0·1) at cut-off = 0·3 (the observed mortality rate), and good calibration. This performance is on par with a logistic regression model with selection of patient variables by three experts (average AUROC = 0·78 (sdev = 0·03) and PPV = 0·79 (sdev = 0·2)). Extending the models with variables that are available at 24h after admission resulted in models with higher predictive performance (average AUROC = 0·77-0·79 (sdev = 0·03) and PPV = 0·79-0·80 (sdev = 0·10-0·17)).
Conclusions: AutoML delivers prediction models with fair discriminatory performance, and good calibration and accuracy, which is as good as regression models with expert-based predictor pre-selection. In the context of the restricted availability of data in an ICU quality registry, extending the models with variables that are available at 24h after admission showed small (but significantly) performance increase.

Abbreviations: APACHE, Acute Physiology and Chronic Health Evaluation; AutoML, Automated machine learning; AUPRC, Area under the Precision-Recall Curve; AUROC, Area under the Receiver Operator Characteristic; CT, Computed tomography; CV, Cross validation; GCS, Glasgow coma scale; LDA, Linear discriminant analysis; ML, Machine learning; NPV, Negative predictive value; PPV, Positive predictive value.

* Corresponding author.
E-mail address: m.c.schut@amsterdamumc.nl (M.C. Schut).

[1] Collaborators of the Dutch COVID-19 ICU Research Consortium: D.P. Verbiest, L.F. te Velde, E.M. van Driel, T. Rijpstra, P.H.J. Elbers, A.P.I Houwink, L. Georgieva, E. Verweij, R.M. de Jong, F.M. van Iersel, T.J.J. Koning, E. Rengers, N. Kusadasi, M.L. Erkamp, R. van den Berg, C.J.M.G. Jacobs, J.L. Epker, A.A. Rijkeboer, M.T. de Bruin, P. Spronk, A. Draisma, D.J. Versluis, A.E. van den Berg, M. Vrolijk-de Mos, J.A. Lens, R.V. Pruijsten, H. Kieft, J. Rozendaal, F. Nooteboom, D.P. Boer , I.T.A. Janssen, L. van Gulik, M.P. Koetsier, V.M. Silderhuis, R.M. Schnabel, I. Drogt, W. de Ruijter, R.J. Bosman, T. Frenzel, L.C. Urlings-Strop, A. Dijkhuizen, I. Z. Hené, A.R. de Meijer, J.W.M. Holtkamp, N. Postma, A.J.G.H. Bindels, R.M.J. Wesselink, E.R. van Slobbe-Bijlsma, P.H.J. van der Voort, B.J.W. Eikemans, D.J. Mehagnoul-Schipper, D. Gommers, J.G. Lutisan, M. Hoeksema, M.G.W. Barnas, B. Festen-Spanjer, M. van Lieshout, N.C. Gritters, M. van Tellingen, G.B. Brunnekreef, J. Vandeputte, T.P.J. Dormans, M.E. Hoogendoorn, M. de Graaff, D. Moolenaar, A.C. Reidinga, J.J. Spijkstra, R. de Waal.

https://doi.org/10.1016/j.ijmedinf.2022.104688
Received 15 November 2021; Received in revised form 28 December 2021; Accepted 11 January 2022
Available online 22 January 2022

1. Introduction

The prevalent approach to clinical prediction modeling often involves the manual selection of potentially relevant variables by experts, followed by regression analysis. Recent advancements in Machine Learning (ML) render this classical approach restrictive (uses only one model type), inefficient (labor-intensive manual selection) and potentially biased (predictor pre-selection). Automated Machine Learning (AutoML) is the automation of the ML design process which includes, among others, automatic model and variable selection and hyperparameter tuning [1]. The promise of AutoML is to remove or lessen the burden of manual ML design tasks. In this study, we assess the predictive performance of AutoML for clinical prognosis modeling by comparing classical modeling (manual variable selection followed by regression) and AutoML modeling approaches. In particular, we assess the performance of AutoPrognosis [2]for the prediction of in-hospital mortality of COVID-19 patients that were admitted to the ICU. AutoPrognosis is an AutoML tool developed for clinical prognostic modeling that learns 20 ML models (e.g., regression, neural networks, and linear discriminant analysis) simultaneously. The case study is particularly relevant for challenging the classical model approach, because (1) the largest proportion of prediction models for diagnosis and prognosis of COVID-19 were developed in the classical way (dd. July 2021: 89 out of 238 models used regression); [3] and (2) efficient automated approaches might be part of a rapid response strategy in a crisis situation.

The classical approach to develop prediction models based on expert-based predictor preselection followed by logistic regression can be time and labor intensive and may be biased. In case of new and yet unknown diseases, such predictor selection is not even possible. New and highly infectious diseases with high chances of leading to pandemic outbreaks, like COVID-19, require a rapid response in order to obtain and disseminate new information about the disease. It is unclear whether automated clinical prognostic modelling approaches based on different machine learning algorithms, which are more rapid and less labor-intensive, are able to reliably predict in-hospital mortality for COVID-19 patients [2].

The aim of this study is twofold. First, to assess the performance of prognostic models to predict in-hospital mortality of COVID-19 patients admitted to Dutch ICUs using automated clinical prognostic modelling versus using the more traditional approach with expert-based predictor preselection followed by logistic regression. Second, to assess the performance of these models based on data available at ICU admission versus data available after 24h of ICU admission.

2. Methods

2.1. Data

This study used prospectively collected data on all patients admitted between February 15th and July 1st 2020 with confirmed COVID-19 to a Dutch ICU extracted from the Dutch National Intensive Care Evaluation (NICE) registry. This NICE dataset contains, amongst other items, demographic data, minimum and maximum values of physiological data in the first 24h of ICU admission, diagnoses (reason for admission as well as comorbidities), ICU as well as in-hospital mortality data and length of stay [4]. This data collection takes place in a standardized manner according to strict definitions and stringent data quality checks to ensure high data quality [5].

Patients were considered to have COVID-19 when the RT-PCR of their respiratory secretions was positive for SARS-CoV-2 or when their CT-scan was consistent with COVID-19 (i.e. a CO-RADS score of ≥ 4 in combination with the absence of an alternative diagnosis) [6]. All analyses were performed on two variants of the NICE dataset: (1) when including only variables available at ICU admission (0h) and (2) when including all variables available after the first 24h of ICU admission (24h).

2.2. Outcome measurements

The primary outcome of this study was in-hospital mortality. During the peak of COVID-19 there was a shortage of ICU beds in some hospitals and many patients were transferred to other ICUs. For transferred patients we could follow their transfers through the Netherlands (because all Dutch ICUs participate in the used registry) and used the survival status of the last hospital the patient was admitted to during one and the same COVID-19 episode.

2.3. Analyses

We applied AutoPrognosis to build prognostic models for prediction of in-hospital mortality using an automated machine learning (AutoML) process [2]. Supplementary Section 1 provides a brief technical overview of how AutoPrognosis works.

Comparative design – In our study, we compared three different approaches (see Table 1) to develop a prognostic model to predict the in-hospital mortality of confirmed COVID-19 patients. Additionally, as a reference, we applied a recalibrated version of the Acute Physiology and Chronic Health Evaluation IV (APACHE IV) regression model, [7] which is one of the most common prognostic model used in intensive care, on our COVID-19 patient population. Such a reference enabled us to verify if developing an ad-hoc model makes sense at all (independently from the used approach).

Statistical Analysis – All the analyses were performed using Python v3.6 and R version 3.5.1 x64 with publicly available software packages[2]. For the reporting of this study, we followed the TRIPOD statement (https://www.tripod-statement.org) and the IJMEDI checklist for assessment of medical AI (https://zenodo.org/record/4835800) [8]. The file is available in an Open Science Foundation (osf.io) repository (https://osf.io/d68cr/).

2.4. Data processing

Table 2 includes an overview of the processing operations that were performed. For the expert-selection approach, three intensivists (DD, DdL, SA) independently preselected predictors from a list of available variables in the NICE registry. Discrepancies were resolved by discussion and based on consensus. The APACHE III acute physiology score [9] and the overall Glasgow Coma Scale (GCS) [10] score were included, and the raw predictors that these scores take into account were excluded (we tried adding the raw predictors but this did not improve results). A further selection on the predictors was done with a backward stepwise AIC selection model.

Table 1
Model approaches.

Approach	Description
Fully-automated	We performed an AutoPrognosis analysis on all available patient variables and these variables were not processed, i.e., selected or transformed.
Semi-automated	We performed an AutoPrognosis analysis on patient variables that were selected by means of stepwise regression and subsequently transformed (capped and normalized) - see the Section Table 1 for details.
Expert-selection	We performed a more traditional logistic regression analysis on patient variables that were selected based on experts' opinions (i.e. intensivists) and by means of stepwise regression.

[2] We used autoprognosis (https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/master/alg/autoprognosis/; included R packages on https://cran.r-project.org/) and scikit (https://scikit-learn.org/) software packages (date of last acces October 31, 2021).

Table 2
Overview of data processing operations.

Model approach	Concerns	Operation
All approaches	**Missings**	Missing values for numerical variables were imputed by using fast k-nearest neighbour (kNN)[27] and mode imputation for categorical variables. Multiple imputation by chained equations (MICE)[28] yielded similar results.
	Derived variables	In addition to the original patient variables as collected and described above, we included a derived variable for the body mass index (BMI) based on weight divided by squared length.
Semi-automated approach	**Variable selection**	Variables were selected with a backward stepwise AIC (Akaike information criterion) selection model[29] before application of AutoPrognosis.
	Extreme values	Extreme values were removed by capping numerical variables (below 1th percentile and above 99th percentile).
	Rescaling	All variables were rescaled to the range [0,1] by min–max normalization: $x' = \frac{x - min(x)}{max(x) - min(x)}$, where $\times$ is the original value and x' is the normalized value.

2.5. Model performance

We measured (1) *discrimination*: Area Under the Receiver Operating Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), sensitivity, Positive predictive value (PPV), Negative predictive value (NPV), Brier score (i.e., the mean squared error of the prediction); (2) *calibration*: calibration curves; and (3) *interpretation*: model coefficients. AUROC and AUPRC were provided by AutoPrognosis; we computed separately the other required measurements. For PPV, NPV and sensitivity, the decision threshold was set to 0.3, which is the average mortality rate in this patient population, corresponding to outcome prevalence [11]. For some models built by AutoPrognosis, e.g., neural networks, interpretation was not readily available (but involves more elaborative techniques like SHAP [12] or LIME [13]), this was not measured.

2.6. Validation

The model performance was evaluated as the average performance over a five-fold cross validation (this is the default validation in AutoPrognosis). For all three approaches, the folds were kept identical to enable fair comparison. The original APACHE IV model as a baseline was first-level recalibrated with the same five folds to achieve a better fit with our specific population, and was then also evaluated with the same five folds. Following Moreno and Apolone, [14] recalibration was done by computing a new intercept α_{new} and the overall calibration slope β_{new} by fitting a logistic regression model with the APACHE IV probability (lp_{APIV}) as the only covariate: $lp_{APIVrecal} = \alpha_{new} + \beta_{new} lp_{APIV}$.

2.7. Approach comparison

Performance measures for discrimination and calibration were assessed by averaging the mean predicted values and the fraction of positives of the best models per fold. To determine the best model per fold, we perform a model comparison within AutoPrognosis. The best model is the one which achieved the highest average AUROC over the five folds. We used the 5x2 cross validation (CV) F-test statistical test for determining the best model [15,16]. For interpretation, we provided feature importance results for the best performing model within each approach. The interpretation results were judged on clinical relevance by intensivists (DdL, SA, DD).

3. Results

3.1. Study population

In total 2,706 confirmed COVID-19 patients of 70 ICUs were included, of which 2,690 (99·4%) could be followed up until hospital discharge; 796 patients (29·6%) died during their hospital stay. Table 3 (data at admission) and supplementary Table 1 (data at 24h) show the descriptive summary statistics of the patient population stratified by hospital survival state.

We observe that survivors were significantly younger (60·8 vs 68·6 years), were more often woman (30·5 vs 22·1%), were less often admitted from the emergency room (23·2 vs 30·9%), and were less often on mechanical ventilation at ICU admission (45·4 vs 55·5%).

3.2. Models' performance

Discrimination – Tables 4a (models with data at admission; referred to as 0h models onwards) and 4b (models with data after 24h; referred to as 24h models onwards) show the AUROC, AUPRC, PPV, NPV, and Brier scores of the three approaches. The obtained 0h and 24h models have fair discriminatory performance (AUROC = 0·75-0·78). For both the 0h and 24h models, there is a significant difference in discriminatory performance in terms of AUROC, AUPRC and Brier score between the fully- and semi-automated approaches (AUROC 0h: $p < 0{\cdot}05$, AUROC 24h $p < 0{\cdot}01$, AUPRC and Brier score both 0h and 24h: $p < 0{\cdot}01$, for 5x2 CV F-test). Additionally, for the 24h models the results of the APACHE IV model are significantly different to all other models for all measures but NPV ($p < 0{\cdot}01$ for 5x2 CV F-test). The best 0h and 24h models obtained by the fully-automated approach were linear discriminant analysis (LDA) models. The best 0h models of the semi-automated approach was LDA; the best 24h was a logistic regression (logR) model. The PPV in the context of triage is most important as one does not want to falsely identify non-survivors and abstain them from ICU care. The 0h model PPVs range between 0·70 (fully-automated) and 0·79 (expert-selection); there is no significant difference in PPV between the three approaches ($p > 0{\cdot}05$ for 5x2 CV F-test).

Calibration – Fig. 1a (data at admission: 0h) and 1b (data at 24h: 24h) show the calibration curves of the three approaches. The 0h and 24h models were well calibrated (calibration curves closely follow the 45° line) and the 24h models outperformed the calibration of the APACHE IV model.

Interpretation of the models – Fig. 2a (data at admission: 0h) and Fig. 2b (data at 24h: 24h) show the coefficients of the best performing models of the fully-automated approach (Linear Discriminant Analysis). Supplementary Table 2 includes the 0h-model description for the fully-automated approach. Supplementary Table 3 includes the 0h-model description for the semi-automated approach. For both the LDA models, the major harmful risk factor for mortality was the patient's age and the major protective risk factor for mortality was the date at which the patient was admitted to the ICU (later date lower mortality risk). Fig. 3a (data at admission) and 3b (data at 24h) show the coefficients of the best performing models of the semi-automated approach (best 0h

Table 3
Descriptive summary statistics stratified by in-hospital mortality for the variables available at admission. The variables used for the first 24h from admission are available in Supplementary Table 1.

		Overall	Survivor	Non-survivor	P-Value	Missing
Number of patients		2,690	1,894	796		
Age, mean (SD)		63·1 (11·2)	60·8 (11·3)	68·6 (9·1)	<0·001	
Gender female, n (%)		754 (28·0)	578 (30·5)	176 (22·1)	<0·001	
Body mass index, mean (SD)		28·7 (5·0)	28·9 (5·1)	28·3 (4·9)	0·016	
Origin of admission, n (%)	General ward same hospital	1,834 (68·3)	1,334 (70·6)	500 (62·8)	0·009	5
	Emergency room same hospital	685 (25·5)	439 (23·2)	246 (30·9)		
	CCU/IC of another hospital	78 (2·9)	54 (2·9)	24 (3·0)		
	CCU/IC of the same hospital	23 (0·9)	14 (0·7)	9 (1·1)		
	Special/Medium care of the same hospital	19 (0·7)	15 (0·8)	4 (0·5)		
	Others	46 (1·6)	33 (1·9)	13 (1·7)		
Readmission to the ICU, n (%)		11 (0·4)	6 (0·3)	5 (0·6)	0·319	
Referring specialty, n (%)	pulmonary diseases	1,703 (64·1)	1,191 (63·9)	512 (64·6)	0·173	33
	internal medicine	822 (30·9)	588 (31·5)	234 (29·5)		
	cardiology	32 (1·2)	16 (0·9)	16 (2·0)		
	surgery	21 (0·8)	17 (0·9)	4 (0·5)		
	other specialism	17 (0·6)	11 (0·6)	6 (0·8)		
	others	62 (2·2)	42 (2·6)	20 (2·5)		
Planned admission, n (%)		46 (1·7)	36 (1·9)	10 (1·3)	0·311	
Hospital length of stay prior to ICU admission, mean (SD)		2·2 (2·8)	2·3 (2·9)	1·9 (2·4)	<0·001	
Comorbidities						
Confirmed infection, n (%)		2,143 (79·7)	1,509 (79·7)	634 (79·6)	0·97	
Acute renal failure, n (%)		243 (9·0)	114 (6·0)	129 (16·2)	<0·001	
Gastro intestinal bleeding, n (%)		4(0·1)	4 (0·2)		0·326	
Aids or Immunological insufficiency, n (%)		196 (7·2)	124 (6·6)	72 (9·0)	0·025	
Chronic cardiovascular insufficiency, n (%)		30 (1·1)	12 (0·6)	18 (2·3)	0·001	
Chronic renal insufficiency, n (%)		76 (2·8)	34 (1·8)	42 (5·3)	<0·001	
Cirrhosis, n (%)		3 (0·1)	2 (0·1)	1 (0·1)	1	
Chronic Obstructive Pulmonary Disease, n (%)		222 (8·3)	130 (6·9)	92 (11·6)	<0·001	
Diabetes, n (%)		516 (19·2)	319 (16·8)	197 (24·7)	<0·001	
Malignancy, n (%)		63 (2·3)	31 (1·7)	32 (4·0)	0·296	
Chronic respiratory insufficiency, n (%)		104 (3·9)	65 (3·4)	39 (4·9)	0·091	
APACHE IV reason for admission, N (%)	Pneumonia, viral	2,508 (93·4)	1,785 (94·4)	723 (90·9)	<0·001	5
	Pneumonia, other	20 (0·7)	11 (0·6)	9 (1·1)		
	Cardiac arrest	20 (0·7)	2 (0·1)	18 (2·3)		
	Pneumonia, bacterial	17 (0·6)	15 (0·8)	2 (0·3)		
	ARDS-adult respiratory distress syndrome, non-cardiogenic pulmonary edema	11 (0·4)	8 (0·4)	3 (0·4)		
	Others	109 (2·4)	69 (5·4)	40 (4·9)		
Interventions						
Cardio Pulmonary Resuscitation before or at ICU admission, n (%)		25(0·9)	2 (0·1)	23 (2·9)	<0·001	
Mechanical ventilation at ICU admission, n (%)		1,301 (48·4)	859 (45·4)	442 (55·5)	<0·001	
Outcome						
In-hospital mortality, n (%)		796 (29·6)		796 (100·0)	<0·001	

Table 4a
Comparison of the automated, semi-automated, and expert-selection approaches using data available on admission (0h). We outline the average results for the five-fold cross validation with the standard deviation in between brackets and considering the best model per fold. For both PPV and NPV, the decision threshold was set to 0·3.

Approach	AUROC	AUPRC	PPV	NPV	Sensitivity	Brier
Fully-automated	0·753 (0·028)	0·565 (0·029)	0·695 (0·109)	0·720 (0·008)	0·092 (0·050)	0·181 (0·011)
Semi-automated	**0·771 (0·022)**	**0·600 (0·029)**	**0·816 (0·124)**	**0·721 (0·013)**	**0·090 (0·082)**	**0·187 (0·018)**
Expert-selection	0·762 (0·027)	0·579 (0·032)	0·762 (0·122)	0·717 (0·007)	0·070 (0·045)	0·179 (0·012)

model: LDA, best 24h model: logR). Again, age (harmful) and ICU admission date (protective) were found as most important risk factors. Fig. 4a (data at admission) and 4b (data at 24h) show the coefficients of the logR models of the expert-selection approach. Most important factors were again age (harmful) and ICU admission (protective). Supplementary Table 4 includes the model description of the 0h logR model.

Variable selection – Supplementary Table 5 shows the selections of variables. For the 0h models, the semi-automated approach selects the least number of variables (16 versus 25 selected by the experts); and there is a major overlap (13 out of 16 variables) in the variable selections in the semi-automated and expert-selection approaches. For the 24h models, the semi-automated approach selected more variables (34) than the experts did (30), but the overlap of variables (13) is the same as in the 0h models.

Table 4b
Comparison of the automated, semi-automated and expert-selection approaches and the APACHE IV baseline using data from the first 24h after admission (24h). We outline the average results for the five-fold cross validation with the standard deviation in between brackets and considering the best model per fold. For both PPV and NPV, the decision threshold was set to 0·3.

Approach	AUROC	AUPRC	PPV	NPV	Sensitivity	Brier
Fully-automated	0·764 (0·030)	0·594 (0·032)	0·736 (0·080)	0·740 (0·014)	0·191 (0·082)	0·177 (0·017)
Semi-automated	**0·785 (0·027)**	**0·629 (0·037)**	**0·818 (0·107)**	**0·740 (0·017)**	**0·183 (0·099)**	**0·170 (0·019)**
Expert-selection	0·778 (0·029)	0·608 (0·042)	0·785 (0·165)	0·727 (0·019)	0·117 (0·099)	0·173 (0·014)
APACHE IV(original)	0·706 (0·028)	0·521 (0·045)	0·697 (0·105)	0·724 (0·010)	0·107 (0·040)	0·186 (0·008)
APACHE IV(recalibrated)	..	..	..	..	..	0·186 (0·006)

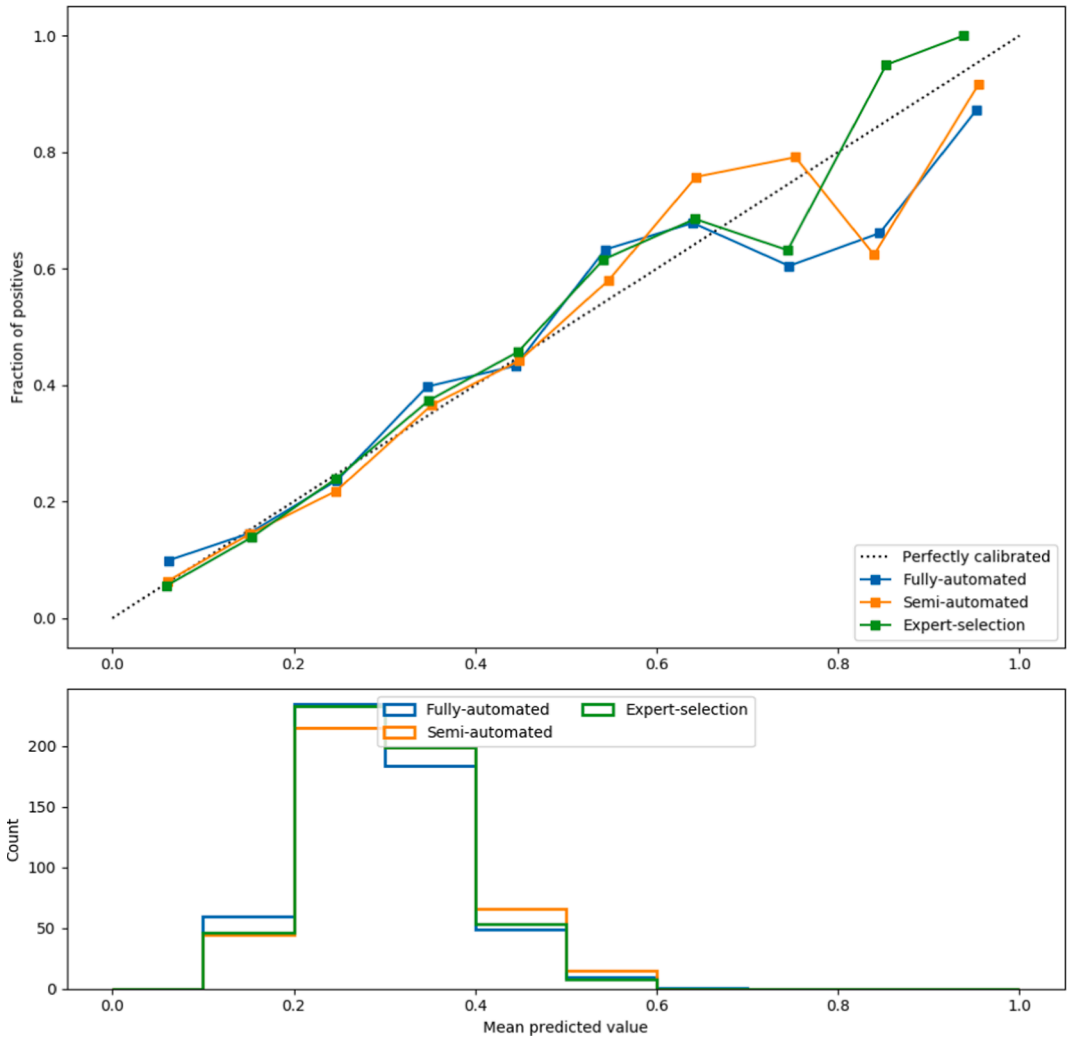


Fig. 1a. Calibration curves of the fully-automated, semi-automated, and expert-selection approaches using data available on admission. Below the distribution of predicted values is shown.

4. Discussion

In this study, we assessed the predictive performance of automated clinical prognostic modelling (AutoML) for in-hospital mortality of ICU-admitted confirmed COVID-19 patients by comparing two automated modelling approaches using (fully-automated and semi-automated) AutoML and one expert-selection approach where intensivists selected potentially relevant variables and a logistic regression analysis was performed. In addition, we compared predictive performance of models that had access to only variables available at admission (0h) with models that had access to variables available at 24h after ICU admission (24h). Overall, predictive performance in terms of discrimination (AUROC) was fair (0·7-0·8).

For the 0h models, there was no significant difference for discrimination (AUROC) between the automated and manual approaches. The semi-automated constructed LDA model (best model of the semi-automated approach) did significantly outperform the fully automated constructed LDA model (best model of the fully-automated approach), but the difference was too small to be clinically relevant. There was no significant difference in PPV between the three approaches.

The 24h models performed similarly in terms of discrimination (AUROC), PPV, and calibration. The selected best model for the semi-automated approach was different for 0h and 24h (0h: LDA, 24h: logR), for the fully-automated approach the best 0h and 24 models were the same (both LDA).

The 24h models were found to perform significantly better than the

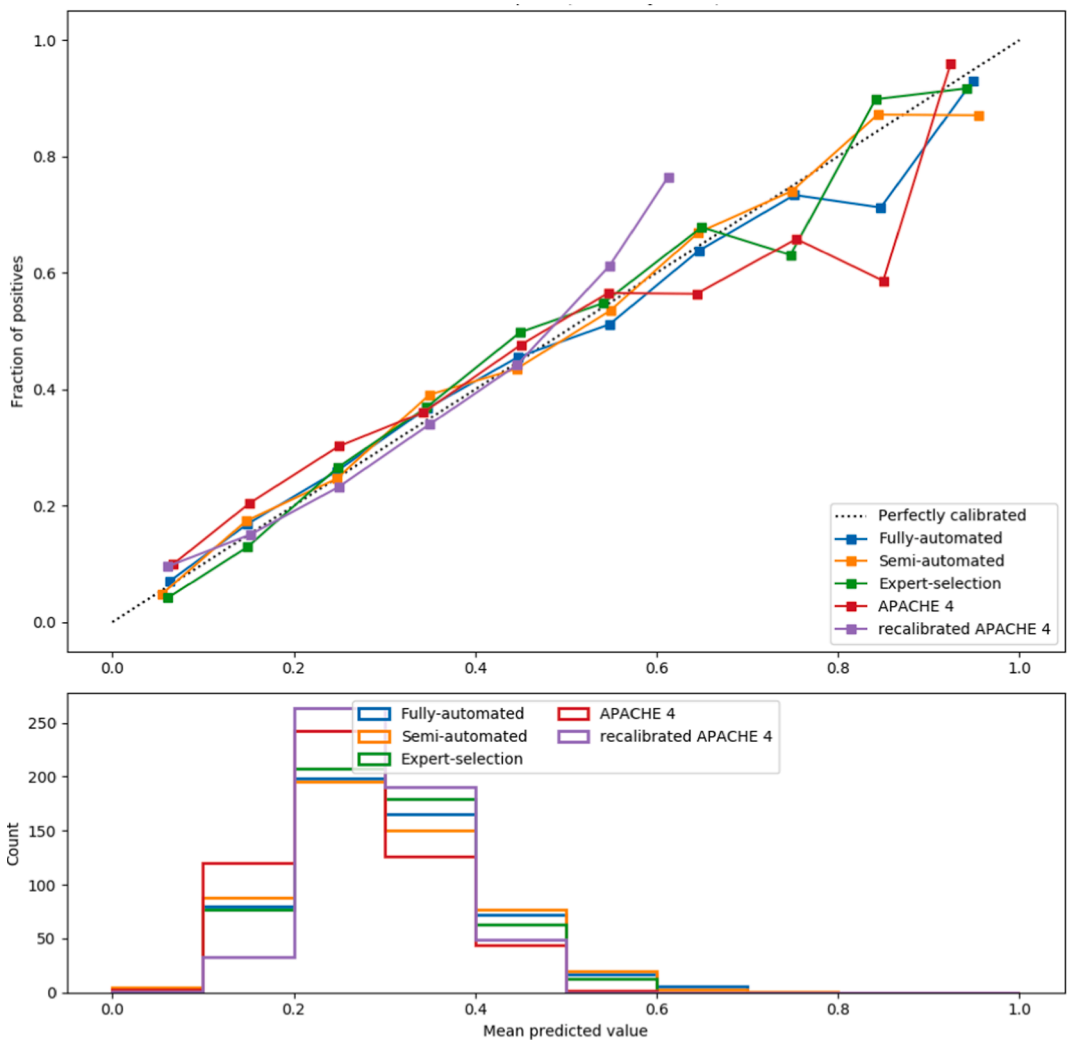


Fig. 1b. Calibration curves for the fully-automated, semi-automated, and expert-selection approaches using data from the first 24h from admission. Below the distribution of predicted values is shown.

0h models (improved AUROC of 0·02), but since it is only a small improvement, it may not be clinically relevant.

4.1. Related work

The studies that are most closely related to our work focus on the development and assessment of prognostic models of mortality among COVID-19 infected patients[17,18] and the identification of prognostic factors for severity and mortality in patients infected with COVID-19. [19–23]

As for development of prognostic models, reported predictive performance varies from fair (AUROC 0·7-0·8) to very good (AUROC > 0·9), other performance measures than AUROC are rarely assessed (e.g., calibration), the studies show an high risk of bias and concern sample sizes up to a maximum of 577 (Table 1 in Wynants et al.[3]).

As for finding strong prognostic factors, similar to other studies we found age, sex and patient history (comorbidities) to be predictors of mortality among COVID-19 patients.

Additionally other indicative predictors were found in other studies such as body temperature, disease signs and symptoms (such as shortness of breath and headache), blood pressure, features derived from CT images, oxygen saturation on room air, hypoxia, diverse laboratory test abnormalities, biomarkers of end-organ dysfunction. [17,18,20,21,23] Most of these other predictors were not included in our dataset (mainly because the used registry data did not include detailed individual patient information). For some of the included comorbidities, we have no explanation why these were not selected as predictors in our models, other than that it is a result from dependences and correlations that are specific for our set of predictors. Our best performing models included CPR, gastro intestinal bleedings and neoplasm, which were not mentioned before in other studies. This may be because these data items are not systematically recorded in other datasets, or that the combination of COVID-19 with another important reason for ICU admission cannot be identified in other studies. A bad prognosis of ICU patients with cancer and after CPR, even independent of COVID-19, is expected and known.[24,25]

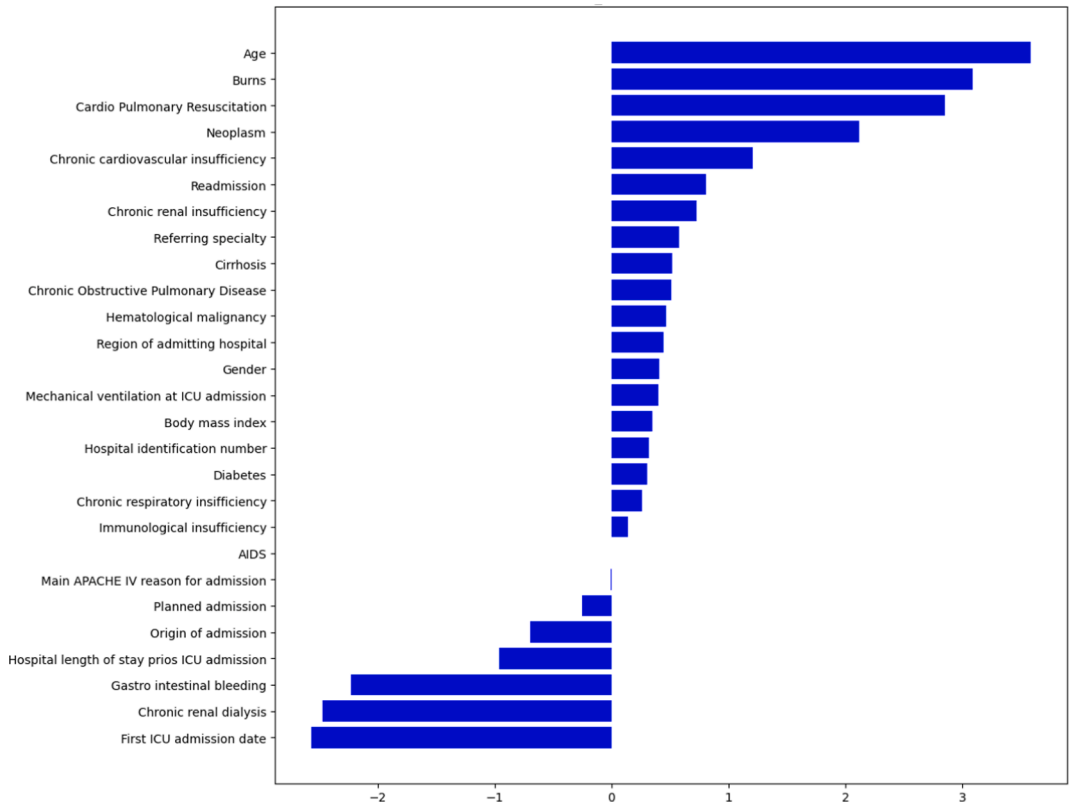


Fig. 2a. Coefficients of the *linear discriminant analysis* model of the fully-automated approach using data available at admission.

4.2. Strengths

The sample size of our study is large (i.e., contains many confirmed COVID-19 patients), and the dataset is comprehensive (i.e., contains many features per patient). As for the analysis, our evaluation is rigorous in that we use multiple performance measures. In general, our approach enables the rapid development of prediction models in case of the COVID-19 epidemic crisis since the registry data that we use are readily available and we use an autormated machine learning approach.

4.3. Limitations

Regarding the model development, we enabled the logistic regression model to perform better to some degree (e.g., with/without variable selection, inclusion of either aggregate (APACHE, GCS) scores or the raw predictors) but this was not done exhaustively. Boosting logR performance is still possible, for example by allowing it to use the best form of predictors (i.e., transformation with for example restricted cubic splines [26]). We found further model tweaking to be out of scope, because we primarily compare (automated versus traditional) approaches and not models.

As for data, the used NICE registration data does not include all laboratory or other individual patient variables, but a specific selection and sometimes an aggregation of routinely collected data. As other studies do include more and different individual patient information such as time series of laboratory values and features derived from CT images that may explain their higher predictive performance.

4.4. Implications

Our study shows the value of automated modelling. After further development and extensive validation, these models are of great importance to assist medical staff in making decisions on ICU admittance and treatment, thereby supporting the use of ICU capacity as efficiently as possible.

Since we do not find clinically relevant differences between models using data at admission time compared to after 24h, this may affect the triage process itself as well: when considering predicted mortality under high pressure on ICU capacity, it may not be effective to admit patients only to see how they develop in the first 24h. However, in case limited ICU capacity is not the main pressure for triage one might say that 24h is not long enough to accurately estimate individuals' survival chances.

4.5. Future work

The models achieve fair (AUROC 0·7-0·8) but not good (AUROC > 0·8) predictive performance. The addition of more individual patient information such as other and more detailed laboratory values (instead of min/max values that we included) and findings of CT images obtained from the electronic patient record may increase the performance since other COVID-19 models including those predictors show better performance than we do, and this is thus worthwhile to investigate.

4.6. Conclusions

This study shows that automated clinical prognostic modelling (AutoML) delivers prediction models with fair predictive performance in

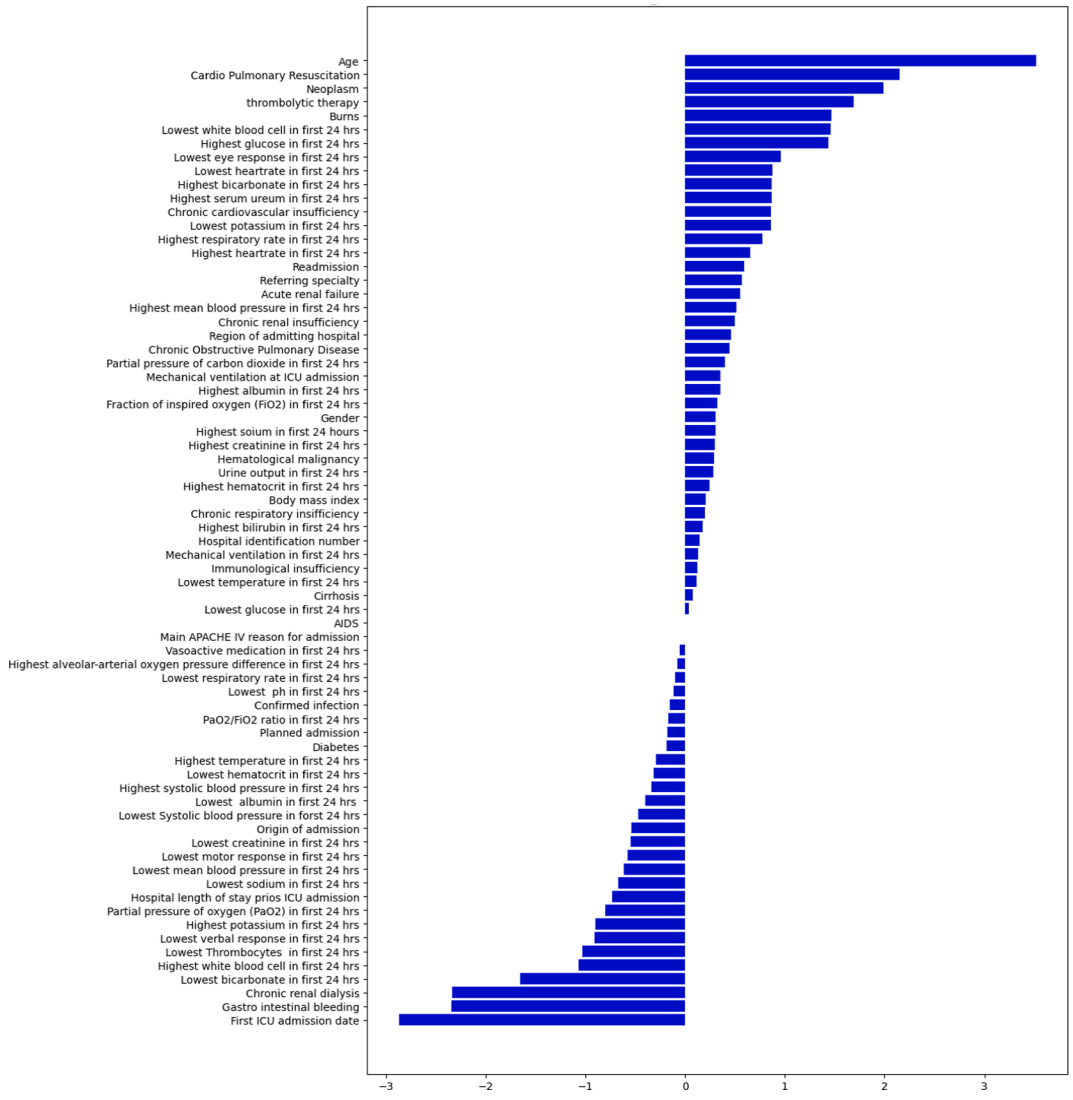


Fig. 2b. Coefficients of the *logistic regression* model of the fully-automated approach using data available at 24h after admission.

terms of discrimination, calibration, and accuracy. The model performance is as good as models that were developed using the more time-consuming regression analysis with expert-based predictor preselection. Models including data from the first 24h of ICU admission did significantly outperform models based on admission data, but the clinical relevance is small. These results pave the way to serve as a baseline for rapid automated model development in times of pandemics or other enduring crises that affect ICU capacity and hence increase the need for patient triage.

Other declarations

The investigators were independent from the funders; IV, SB and MCS had full access to the data, have verified the data, and take responsibility for the integrity of the data and the accuracy of the data analysis; the lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Summary Table

What was already known on the topic:

- Classical prediction models (i.e., regression models with manual predictor selection) yield good performance for clinical diagnosis and prognosis, but the modeling process is potentially biased and limited.
- Automated prognostic modelling (AutoML) facilitates automatic model and variable selection and hyperparameter tuning, and can lessen the burden of carrying out manual design tasks for prediction modeling.

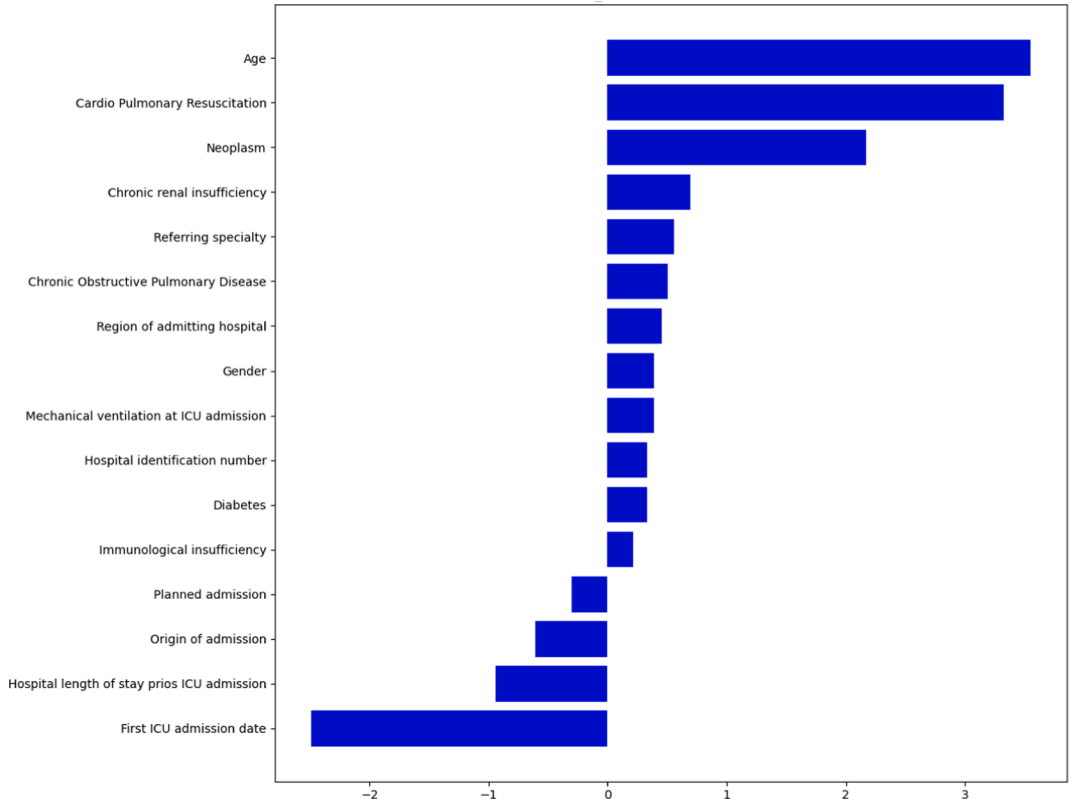


Fig. 3a. Coefficients of the *linear discriminant analysis* model of the semi-automated approach using data available at admission.

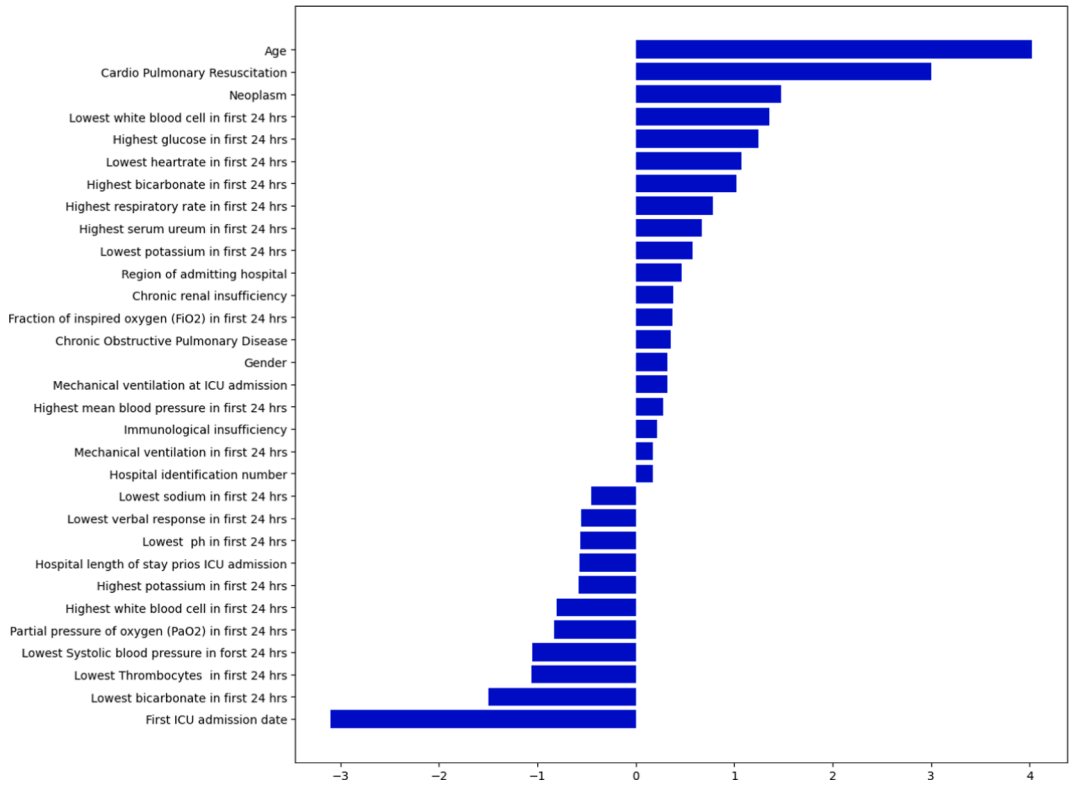


Fig. 3b. Coefficients of the *linear discriminant analysis* model of semi-automated approach using data available at 24h after admission.

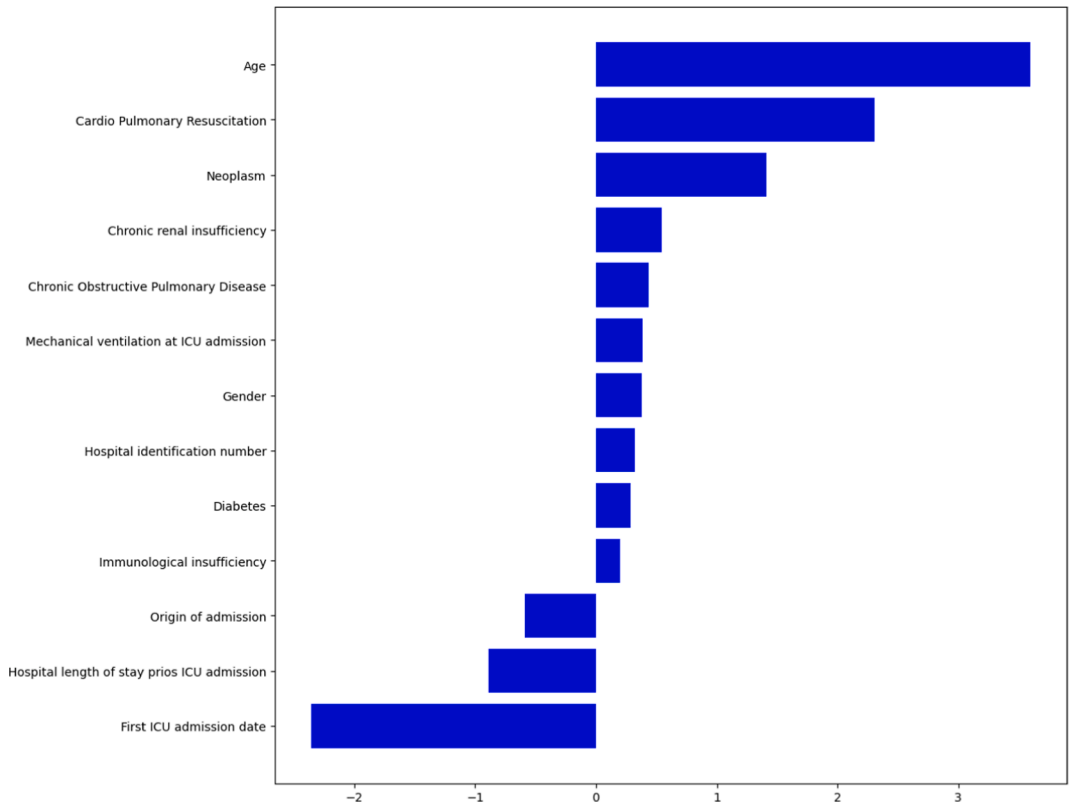


Fig. 4a. Coefficients of the *logistic regression* model of the expert-selection approach using data available at admission.

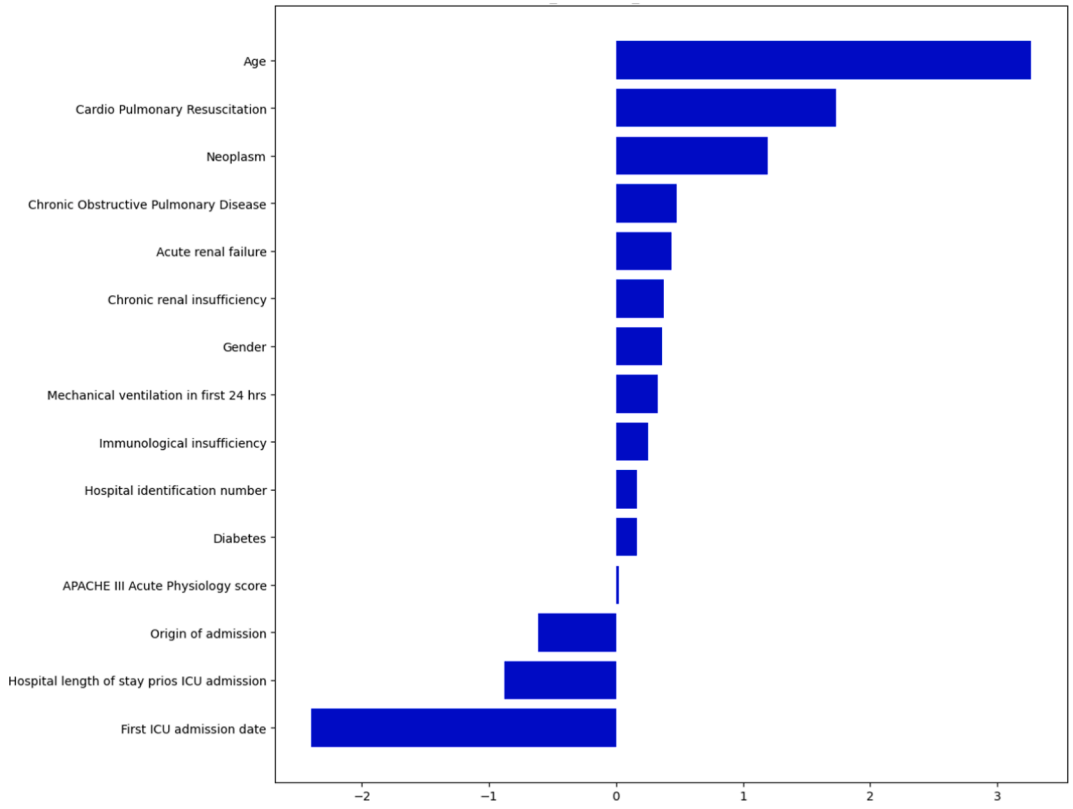


Fig. 4b. Coefficients of the *logistic regression model* of the expert-selection approach using data available at 24h after admission.

- The largest proportion of prediction models for diagnosis and prognosis of COVID-19 were developed in the classical way (regression with manual predictor selection).

What this study added to our knowledge:

- Automated modeling can deliver clinical prediction models that perform on par with more classical models (regression models with manual predictor selection).
- Automated modelling can assist decision-making on ICU admittance and treatment, and can support efficient use of ICU capacity.
- Admitting of COVID-19 patients to the ICU to see how they develop in the first 24 hours may not be effective.

Ethics approval and consent to participate

The study protocol was reviewed by the Medical Ethics Committee of the Amsterdam Medical Center, the Netherlands. This committee provided a waiver from formal approval (W20_273 # 20.308) and informed consent since this trial does not fall within the scope of the Dutch Medical Research (Human Subjects) Act.

CRediT authorship contribution statement

I. Vagliano: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **S. Brinkman:** Conceptualization, Methodology, Investigation, Writing – original draft. **A. Abu-Hanna:** Methodology, Writing – review & editing. **M.S Arbous:** Methodology, Writing – review & editing. **D.A. Dongelmans:** Methodology, Writing – review & editing. **P.W.G. Elbers:** Writing – review & editing. **D.W. de Lange:** Methodology, Writing – review & editing. **M. van der Schaar:** Methodology, Writing – review & editing. **N.F. de Keizer:** Conceptualization, Methodology, Investigation, Writing – original draft, Supervision, Project administration. **M.C. Schut:** Conceptualization, Writing – original draft, Methodology, Investigation, Supervision, Project administration.

Declaration of Competing Interest

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.

Data availability

Data is available under stringent conditions as described on the NICE website https://www.stichting-nice.nl/extractieverzoek_procedure.jsp (in Dutch).

Acknowledgements

We thank all participating ICUs for making this study possible.

Funding

This research was funded by The Netherlands Organisation for Health Research and Development (ZonMw) COVID-19 Programme in the bottom-up focus area 1 “Predictive diagnostics and treatment” for theme 3 “Risk analysis and prognostics” (project number 10430 01 201 0011: IRIS). The funder had no role in the design of the study or writing the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104688.

References

[1] R. Elshawi, M. Maher, S. Sakr, Automated machine learning: State-of-the-art and open challenges. *arXiv preprint arXiv:190602287* 2019.
[2] A.M. Alaa, M. van der Schaar, AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning. 2018.
[3] L. Wynants, B. Van Calster, G.S. Collins, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369: m1328. doi: 10.1136/bmj.m1328 [published Online First: 2020/04/09].
[4] N. van de Klundert, R. Holman, D.A. Dongelmans, et al. Data Resource Profile: the Dutch National Intensive Care Evaluation (NICE) Registry of Admissions to Adult Intensive Care Units. *Int J Epidemiol* 2015;44(6):1850-50h. doi: 10.1093/ije/dyv291 [published Online First: 2015/11/29].
[5] D.G. Arts, N.F. De Keizer, G.J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9(6):600-11. doi: 10.1197/jamia.m1087 [published Online First: 2002/10/19].
[6] M. Prokop, W. van Everdingen, T. van Rees Vellinga, H. Quarles van Ufford, L. Stöger, L. Beenen, B. Geurts, H. Gietema, J. Krdzalic, C. Schaefer-Prokop, B. van Ginneken, M. Brink, CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19-Definition and Evaluation, Radiology 296 (2) (2020) E97–E104.
[7] J.E. Zimmerman, A.A. Kramer, D.S. McNair, F.M. Malila, Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today’s critically ill patients, Critical Care Medicine 34 (5) (2006) 1297–1310, https://doi.org/10.1097/01.CCM.0000215112.84523.F0.
[8] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies, Int J Med Inform 153 (2021) 104510, https://doi.org/10.1016/j.ijmedinf.2021.104510.
[9] J.E. Zimmerman, D.P. Wagner, E.A. Draper, L. Wright, C. Alzola, W.A. Knaus, Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database, Crit Care Med 26 (8) (1998) 1317–1326.
[10] G. Teasdale, B. Jennett, Assessment of coma and impaired consciousness. A practical scale, Lancet 2 (7872) (1974) 81–84, https://doi.org/10.1016/s0140-6736(74)91639-0 [published Online First: 1974/07/13].
[11] E.A. Freeman, G.G. Moisen, A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa, Ecol Model 217 (1–2) (2008) 48–58, https://doi.org/10.1016/j.ecolmodel.2008.05.015.
[12] A unified approach to interpreting model predictions. Advances in neural information processing systems; 2017.
[13] “ Why should I trust you?” Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
[14] R. Moreno, G. Apolone, Impact of different customization strategies in the performance of a general severity score, Crit Care Med 25 (12) (1997) 2001–2008, https://doi.org/10.1097/00003246-199712000-00017 [published Online First: 1997/12/24].
[15] T.G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Comput 10 (7) (1998) 1895–1923, https://doi.org/10.1162/089976698300017197 [published Online First: 1998/09/23].
[16] E. Alpaydin, Combined 5 x 2 cv F test for comparing supervised classification learning algorithms, Neural Comput 11 (8) (1999) 1885–1892, https://doi.org/10.1162/089976699300016007 [published Online First: 1999/12/01].
[17] R.K. Gupta, M. Marks, T.H.A. Samuels, A. Luintel, T. Rampling, H. Chowdhury, M. Quartagno, A. Nair, M. Lipman, I. Abubakar, M. van Smeden, W.K. Wong, B. Williams, M. Noursadeghi, Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study, Eur Respir J 56 (6) (2020) 2003498, https://doi.org/10.1183/13993003.03498-2020 [published Online First: 2020/09/27].
[18] L. Wynants, B. Van Calster, G.S. Collins, et al., Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal, BMJ 369 (2020), m1328, https://doi.org/10.1136/bmj.m1328 [published Online First: 2020/04/09].
[19] X. Fang, S. Li, H. Yu, P. Wang, Y. Zhang, Z. Chen, Y. Li, L. Cheng, W. Li, H. Jia, X. Ma, Epidemiological, comorbidity factors with severity and prognosis of COVID-19: a systematic review and meta-analysis, Aging (Albany NY) 12 (13) (2020) 12493–12503.
[20] B. Gallo Marin, G. Aghagoli, K. Lavine, et al. Predictors of COVID-19 severity: A literature review. *Rev Med Virol* 2020:e2146. doi: 10.1002/rmv.2146 [published Online First: 2020/08/28].
[21] A. Izcovich, M.A. Ragusa, F. Tortosa, M.A. Lavena Marzio, C. Agnoletti, A. Bengolea, A. Ceirano, F. Espinosa, E. Saavedra, V. Sanguine, A. Tassara, C. Cid, H.N. Catalano, A. Agarwal, F. Foroutan, G. Rada, C. Lazzeri, Prognostic factors for severity and mortality in patients infected with COVID-19: A systematic review, PLOS ONE 15 (11) (2020) e0241955, https://doi.org/10.1371/journal.pone.0241955 [published Online First: 2020/11/18].
[22] X. Lai, J. Liu, T. Zhang, L. Feng, P. Jiang, L. Kang, Q. Liu, Y. Gao, Clinical, laboratory and imaging predictors for critical illness and mortality in patients with COVID-19: protocol for a systematic review and meta-analysis, BMJ Open 10 (12) (2020) e039813, https://doi.org/10.1136/bmjopen-2020-039813.
[23] J.A. Siordia, Epidemiology and clinical features of COVID-19: A review of current literature, J Clin Virol 127 (2020) 104357, https://doi.org/10.1016/j.jcv.2020.104357.

[24] M.M.E.M. Bos, N.F. de Keizer, I.A. Meynaar, F. Bakhshi-Raiez, E. de Jonge, Outcomes of cancer patients after unplanned admission to general intensive care units, Acta Oncol 51 (7) (2012) 897–905.
[25] L. Mandigers, F. Termorshuizen, N.F. de Keizer, D. Gommers, D. dos Reis Miranda, W.J.R. Rietdijk, C.A. den Uil, A nationwide overview of 1-year mortality in cardiac arrest patients admitted to intensive care units in the Netherlands between 2010 and 2016, Resuscitation 147 (2020) 88–94.
[26] S. Durrleman, R. Simon, Flexible regression models with cubic splines, Statistics in medicine 8 (5) (1989) 551–561.
[27] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525.
[28] S.v. Buuren, K. Groothuis-Oudshoorn, MICE: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software 45 (3) (2011), https://doi.org/10.18637/jss.v045.i03.
[29] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen E, Tanabe K, Kitagawa G, eds. Selected Papers of Hirotugu Akaike. New York, NY: Springer New York 1998:199-213.