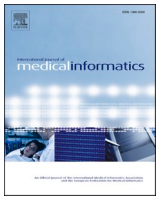
International Journal of Medical Informatics 167 (2022) 104863

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Assess and validate predictive performance of models for in-hospital mortality in COVID-19 patients: A retrospective cohort study in the Netherlands comparing the value of registry data with high-granular electronic health records

Iacopo Vagliano [a,*], Martijn C. Schut [a], Ameen Abu-Hanna [a], Dave A. Dongelmans [b,c], Dylan W. de Lange [b,d], Diederik Gommers [e], Olaf L. Cremer [f], Rob J. Bosman [g], Sander Rigter [h], Evert-Jan Wils [i], Tim Frenzel [j], Remko de Jong [k], Marco A.A. Peters [l], Marlijn J.A. Kamps [m], Dharmanand Ramnarain [n], Ralph Nowitzky [o], Fleur G.C.A. Nooteboom [p], Wouter de Ruijter [q], Louise C. Urlings-Strop [r], Ellen G.M. Smit [s], D. Jannet Mehagnoul-Schipper [t], Tom Dormans [u], Cornelis P.C. de Jager [v], Stefaan H.A. Hendriks [w], Sefanja Achterberg [x], Evelien Oostdijk [y], Auke C. Reidinga [z], Barbara Festen-Spanjer [aa], Gert B. Brunnekreef [ab], Alexander D. Cornet [ac], Walter van den Tempel [ad], Age D. Boelens [ae], Peter Koetsier [af], Judith Lens [ag], Harald J. Faber [ah], A. Karakus [ai], Robert Entjes [aj], Paul de Jong [ak], Thijs C.D. Rettig [al], M.C. Reuland [c], Sesmu Arbous [am], Lucas M. Fleuren [an], Tariq A. Dam [an], Patrick J. Thoral [an], Robbert C.A. Lalisang [ao], Michele Tonutti [ao], Daan P. de Bruin [ao], Paul W.G. Elbers [an], Nicolette F. de Keizer [a,b], on behalf of the Dutch COVID-19 Research Consortium, the Dutch ICU Data Sharing Against COVID-19 Collaborators[1]

[a] *Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands*
[b] *National Intensive Care Evaluation (NICE) foundation, Amsterdam, The Netherlands*
[c] *Department of Intensive Care Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands*
[d] *Department of Intensive Care Medicine, University Medical Center Utrecht, University Utrecht, Utrecht, The Netherlands*
[e] *Department of Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands*
[f] *Intensive Care, UMC Utrecht, Utrecht, The Netherlands*
[g] *ICU, OLVG, Amsterdam, The Netherlands*
[h] *Department of Anesthesiology and Intensive Care, St. Antonius Hospital, Nieuwegein, The Netherlands*
[i] *Department of Intensive Care, Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands*
[j] *Department of Intensive Care Medicine, Radboud University Medical Center, Nijmegen, The Netherlands*
[k] *Intensive Care, Bovenij Ziekenhuis, Amsterdam, The Netherlands*
[l] *Intensive Care, Canisius Wilhelmina Ziekenhuis, Nijmegen, The Netherlands*
[m] *Intensive Care, Catharina Ziekenhuis Eindhoven, Eindhoven, The Netherlands*
[n] *Department of Intensive Care, ETZ Tilburg, Tilburg, The Netherlands*
[o] *Intensive Care, Haga Ziekenhuis, Den Haag, The Netherlands*
[p] *Intensive Care, Laurentius Ziekenhuis, Roermond, The Netherlands*
[q] *Department of Intensive Care Medicine, Northwest Clinics, Alkmaar, The Netherlands*
[r] *Intensive Care, Reinier de Graaf Gasthuis, Delft, The Netherlands*
[s] *Intensive Care, Spaarne Gasthuis, Haarlem en Hoofddorp, The Netherlands*
[t] *Intensive Care, VieCuri Medisch Centrum, Venlo, The Netherlands*
[u] *Intensive care, Zuyderland MC, Heerlen, The Netherlands*
[v] *Department of Intensive Care, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands*
[w] *Intensive Care, Albert Schweitzerziekenhuis, Dordrecht, The Netherlands*
[x] *ICU, Haaglanden Medisch Centrum, Den Haag, The Netherlands*
[y] *ICU, Maasstad Ziekenhuis Rotterdam, Rotterdam, The Netherlands*
[z] *ICU, SEH, BWC, Martiniziekenhuis, Groningen, The Netherlands*

* Corresponding author at: Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health research institute, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands.
E-mail address: i.vagliano@amsterdamumc.nl (I. Vagliano).

https://doi.org/10.1016/j.ijmedinf.2022.104863
Received 12 April 2022; Received in revised form 19 August 2022; Accepted 3 September 2022
Available online 22 September 2022

[aa] Intensive Care, Ziekenhuis Gelderse Vallei, Ede, The Netherlands
[ab] Department of Intensive Care, Ziekenhuisgroep Twente, Almelo, The Netherlands
[ac] Department of Intensive Care, Medisch Spectrum Twente, Enschede, The Netherlands
[ad] Department of Intensive Care, Ikazia Ziekenhuis Rotterdam, Rotterdam, The Netherlands
[ae] Anesthesiology, Antonius Ziekenhuis Sneek, Sneek, The Netherlands
[af] Intensive Care, Medisch Centrum Leeuwarden, Leeuwarden, The Netherlands
[ag] ICU, IJsselland Ziekenhuis, Capelle aan den IJssel, The Netherlands
[ah] ICU, WZA, Assen, The Netherlands
[ai] Department of Intensive Care, Diakonessenhuis Hospital, Utrecht, The Netherlands
[aj] Department of Intensive Care, Adrz, Goes, The Netherlands
[ak] Department of Anesthesia and Intensive Care, Slingeland Ziekenhuis, Doetinchem, The Netherlands
[al] Department of Anesthesiology, Intensive Care and Pain Medicine, Amphia Ziekenhuis, Breda, The Netherlands
[am] Intensivist, LUMC, Leiden, The Netherlands
[an] Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands
[ao] Pacmed, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Covid-19 [C01.748.610.763.500]
Critical care [E02.760.190]
In-hospital mortality
[E05.318.308.985.550.400]
Prognosis [E01.789]
Machine learning [G17.035.250.500]
Electronic Health Record
[E05.318.308.940.968.625.500]

ABSTRACT

Purpose: To assess, validate and compare the predictive performance of models for in-hospital mortality of COVID-19 patients admitted to the intensive care unit (ICU) over two different waves of infections. Our models were built with high-granular Electronic Health Records (EHR) data versus less-granular registry data.
Methods: Observational study of all COVID-19 patients admitted to 19 Dutch ICUs participating in both the national quality registry National Intensive Care Evaluation (NICE) and the EHR-based Dutch Data Warehouse (hereafter EHR). Multiple models were developed on data from the first 24 h of ICU admissions from February to June 2020 (first COVID-19 wave) and validated on prospective patients admitted to the same ICUs between July and December 2020 (second COVID-19 wave). We assessed model discrimination, calibration, and the degree of relatedness between development and validation population. Coefficients were used to identify relevant risk factors.
Results: A total of 1533 patients from the EHR and 1563 from the registry were included. With high granular EHR data, the average AUROC was 0.69 (standard deviation of 0.05) for the internal validation, and the AUROC was 0.75 for the temporal validation. The registry model achieved an average AUROC of 0.76 (standard deviation of

[1] Collaborators of the Dutch COVID-19 ICU Research Consortium: D.P. Verbiest, L.F. te Velde, E.M. van Driel, T. Rijpstra, P.W.G. 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, 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, D. Ramnarai, The Dutch ICU Data Sharing Against COVID-19 Collaborators. *From collaborating hospitals having shared data:* Julia Koeter, Roger van Rietschote, Laura van Manen, Leon Montenij, Jasper van Bommel, Roy van den Berg, Ellen van Geest, Anisa Hana, B. van den Bogaard, Prof. Peter Pickkers, Pim van der Heiden, Claudia (C.W.) van Gemeren, Arend Jan Meinders, Martha de Bruin, Emma Rademaker, Frits H.M. van Osch, Martijn de Kruif, Nicolas Schroten, Klaas Sierk Arnold, J.W. Fijen, Jacomar J.M. van Koesveld, Koen S. Simons, Joost Labout, Bart van de Gaauw, Michael Kuiper, Albertus Beishuizen, Dennis Geutjes, Johan Lutisan, Bart P. Grady, Remko van den Akker, Tom A. Rijpstra, W. Boersma, *From collaborating hospitals having signed the data sharing agreement:* D. Pretorius, Menno Beukema, Bram Simons, A.A. Rijkeboer, Marcel Aries, Niels C. Gritters van den Oever, Martijn van Tellingen, MD, EDIC, Annemieke Dijkstra, Rutger van Raalte, *From the Laboratory for Critical Care Computational Intelligence:* Mark Hoogendoorn, Armand R.J. Girbes, Luca Roggeveen, Fuda van Diggelen, Ali el Hassouni, David Romero Guzman, Sandjai Bhulai, Dagmar M. Ouweneel, Ronald Driessen, Jan Peppink, Harm-Jan de Grooth, G.J. Zijlstra, A.J. van Tienhoven, Evelien van der Heiden, Jan Jaap Spijkstra, Hans van der Spoel, Angelique M.E. de Man, Thomas Klausch, Heder J. de Vries, *From Pacmed:* Sebastiaan J.J. Vonk, Mattia Fornasa, Tomas Machado, Michael de Neree tot Babberich, Olivier Thijssens, Lot Wagemakers, Hilde G.A. van der Pol, Tom Hendriks, Julie Berend, Virginia Ceni Silva, Robert F.J. Kullberg, Taco Houwert, Hidde Hovenkamp, Roberto Noorduijn Londono, Davide Quintarelli, Aletta A. de Beer, Giovanni Cino, Willem E. Herter, Adam Izdebski, *From RCCnet:* Leo Heunks, Nicole Juffermans, Arjen J.C. Slooter, *From other collaborating partners:* Martijn Beude.

0.05) in the internal validation and 0.77 in the temporal validation. In the EHR data, age, and respiratory-system related variables were the most important risk factors identified. In the NICE registry data, age and chronic respiratory insufficiency were the most important risk factors.
Conclusion: In our study, prognostic models built on less-granular but readily-available registry data had similar performance to models built on high-granular EHR data and showed similar transportability to a prospective COVID-19 population. Future research is needed to verify whether this finding can be confirmed for upcoming waves.

1. Introduction

The coronavirus disease 2019 (COVID-19) has challenged global health and society at large. Most countries have experienced multiple COVID-waves in the last years. Models that estimate the risk of in-hospital mortality of COVID-19 patients in the intensive care unit (ICU) could be valuable for decision making on treatment (intensify or withhold) and capacity planning. Many prognostic models have been developed, often using data purposely collected from electronic health records (EHR) [1]. Various existing ICU data registries or specifically developed COVID-19 data collections improved our understanding of the relation between patient characteristics and disease progress at the ICU [2–4].

EHR data typically have high granularity (multiple variables and measurements over time). They potentially support the application of advanced methods, but combining these data from multiple centers, each using different data models and coding lists, requires a considerable effort and time. In a sudden pandemic or crisis situation, a rapid response is needed. Therefore, waiting to collect, curate and aggregate EHR data might not be possible. In contrast, high-quality registry data are already collected, more uniform, quality-controlled, and thereby readily-usable. Thus, they may enable a faster response, although have a lower granularity and a possibly-delayed availability. A comparison of the value of registry data with high-granular electronic health records for building prognostic models is still lacking.

This study aims to assess, validate over successive waves of infections, and compare the predictive performance of models for in-hospital mortality of COVID-19 patients admitted to the intensive care unit (ICU), when such models are developed with high-granular ICU data collected from various hospitals' EHR or low-granular registry data.

2. Methods

2.1. Study design and population

This was a multi-center observational study on prospectively collected EHR data on patients from 19 ICUs participating in the Dutch ICU Data Warehouse [5] as well as Dutch National Intensive Care Evaluation (NICE) registry [6,7]. We included all patients that were 18 years and older and were admitted between February 15th, 2020 and January 1st, 2021 with confirmed COVID-19. Thirteen of those 19 ICUs uploaded EHR data during the second wave. For the NICE registry, we selected data from the same 19 ICUs.

COVID-19 was defined as a positive real-time reverse transcriptase polymerase chain reaction (RT-PCR) assay for SARS-CoV-2 or, in the early phase of the pandemic, as a CT-scan consistent with COVID-19, i.e. a COVID-19 Reporting and Data System (CO-RADS) score of ≥ 4 in combination with the absence of an alternative diagnosis) [8].

2.2. Data collection

NICE is a quality registry with national coverage since 2016. ICUs extract for all their patients a predefined dataset from the routinely collected data from their EHR and upload this dataset each month after manual validation and completion. This predefined dataset includes demographics, minimum and maximum values of physiological data in the first 24 h of ICU admission, diagnoses, ICU and in-hospital mortality and length of stay [6]. Data collection is standardized with strict definitions and stringent data-quality checks [7]. Hereafter we call this data source REG.

The Dutch ICU Data Warehouse includes high-granular data of critically-ill patients with COVID-19 in the Netherlands. The raw data were extracted from the participating hospitals' EHR. Parameters were mapped to a common nomenclature by a team of clinicians, data entry errors were filtered, and derived parameters were added (e.g., body-mass index) when not directly provided [9]. Data included demographics, administrative variables, comorbidities, and physiological data and information regarding the patient positioning and ventilation characteristics. Hereafter we call this data source EHR. EHR patients transferred to another ICU were linked when data from the referring and receiving hospital were available, otherwise excluded as their final in-hospital outcome was unknown.

2.3. Outcome and predictors

The outcome of this study was in-hospital mortality. The variables available in the two data sources in the first 24 h were included as predictors. The model is intended to be used at the first 24 h from admission. The predictors finally included are provided in Tables 1 and 2.

2.4. Data preprocessing

The data preprocessing was the same for both datasets, unless differently specified. We removed administrative variables which did not have clinical value (such as identifiers), variables regarding discharge (date, destination), and variables that have zero variance. In REG, which had less missing data than EHR, we removed variables with over 45 % of missing data, in EHR, variables with over 85 % of missing. For the remaining numerical variables, missing values were imputed by using the multiple imputation by chained equations (MICE) [10]. Mode imputation was used for the remaining categorical variables. Backward stepwise variable selection was used with the Akaike information criterion [11]. Numerical variables were capped below the 1st percentile and above the 99th percentile. All variables were rescaled to the range [0,1] with min–max scaling. In EHR, the average, minimum value, maximum value, difference between the last and first measurement, and slope were computed based on the repeated measurements of each numerical variable available in the first 24 h. For categorical variables, the mode was selected.

2.5. Analyses

We developed several prognostic models to predict in-hospital mortality with each of the two datasets. We used AutoPrognosis, an automated machine learning process [12,13]. The best model per dataset was chosen based on predictive performance, variability of performance, and interpretability, and it was then internally and temporally validated.

2.6. Performance measures, internal validation, and temporal validation

We measured discrimination with the area under the receiver operating characteristics (AUROC), the area under the precision-recall curve (AUPRC), positive predictive values (PPV), negative predictive values (NPV), and the Brier score. We also assessed model calibration with calibration curves and provided model coefficients to interpret the models. Brier score is used to assess discrimination due to its known limitations to assess calibration [14]. A calibration curve gives better insight into risk prediction areas with larger deviation between predicted and true risk. For both PPV and NPV, the decision threshold was set to 0.3, the average mortality rate in this patient population.

Model performance was internally validated by the average performance over a fivefold cross validation on all COVID-19 patients admitted to Dutch ICUs between February 15th and June 30th, 2020 (first wave).

Various factors (virus mutations, treatment strategies, etc.) may impact model performance over time. To validate our models over time [15], we validated on a prospective dataset of all COVID-19 patients admitted to 13 of the same 19 Dutch ICUs between July 1st, 2020 and January 1st, 2021 (second wave) as the other 6 ICUs did not provide data in this time period. Following Debray et al. [16], we assessed the degree of relatedness between development and validation population to understand whether temporal validation was estimating the model reproducibility or transportability. Model reproducibility means that a model performs sufficiently accurate across new samples from the same target population. Transportability is the ability of a model to perform well across samples from different but related populations. To assess the degree of relatedness between development and validation population, we evaluated their corresponding case-mix differences: We built a logistic-regression membership model that uses the same predictors used by the in-hospital mortality models plus the in-hospital mortality outcome. The outcome of the membership model was the predicted probability of a patient to belong to the development or validation population. When such a model performed poorly, it meant development and validation population had similar case-mix and therefore the temporal validation assessed the reproducibility of the model. When such a model performed well, development and validation population had different case-mix and therefore the temporal validation tested the transportability of the model. The membership model performance was assessed with the AUROC and interpreted according to Hosmer and Lemeshow [17].

Statistical difference among the performance results for the models built on EHR and REG was assessed with a paired Student's-*t* test for dependent samples after bootstrapping each measure over 300 iterations.

Table 1
Descriptive summary statistics of the EHR patient population used in the development (and internal validation) as well as temporal validation, stratified by in-hospital mortality. We only show the variables selected after the variable selection with backward elimination. APTT stands for activated partial thromboplastin time, FiO_2 for fraction of inspired oxygen, $PaCO_2$ for partial pressure of carbon dioxide, PaO_2 for partial pressure of oxygen.

Variable		Development population (first wave)					Temporal validation population (second wave)				
		Overall	Survivor	Non-survivor	Missing	P-value	Overall	Survivor	Non-survivor	Missing	P-value
n		992	676	316			541	360	181		
Age, mean (SD)		63.7 (11.8)	61.5 (12.1)	68.6 (9.5)	0	<0.001	64.5 (11.8)	61.5 (12.2)	70.5 (8.1)	0	<0.001
Physiological and blood values											
Average estimated glomerular filtration rate in first 24 hrs, mean (SD)		63.1 (21.1)	65.4 (20.3)	59.1 (22.0)	459	0.001	61.1 (22.0)	65.6 (19.9)	54.6 (23.2)	213	<0.001
Lowest estimated glomerular filtration rate in first 24 hrs, mean (SD)		58.2 (22.8)	61.0 (22.1)	53.3 (23.2)	459	<0.001	56.4 (22.9)	61.1 (21.3)	49.4 (23.5)	213	<0.001
Average FiO_2 in first 24 hrs, mean (SD)		55.5 (16.0)	53.2 (15.1)	60.1 (16.9)	82	<0.001	61.9 (16.1)	60.2 (16.3)	64.9 (15.2)	35	0.001
Highest erythrocytes in first 24 hrs, mean (SD)		11.2 (148.0)	13.8 (173.7)	4.2 (0.7)	518	0.309	5.4 (15.7)	4.5 (1.0)	7.3 (28.4)	268	0.374
Highest glucose in first 24 hrs, mean (SD)		9.9 (4.5)	9.4 (4.3)	10.9 (4.7)	134	<0.001	12.3 (5.4)	11.6 (5.1)	13.6 (5.8)	5	<0.001
Highest prothrombin time in first 24 hrs, mean (SD)		9.4 (8.0)	9.6 (6.5)	9.2 (10.5)	380	0.608	12.6 (7.1)	12.3 (5.6)	13.2 (9.3)	100	0.316
Highest neutrophils in first 24 hrs, mean (SD)		7.4 (4.0)	7.3 (3.9)	7.6 (4.1)	599	0.480	9.0 (4.5)	8.5 (4.0)	10.3 (5.4)	280	0.007
Lowest measured respiratory rate in first 24 hrs, mean (SD)		14.8 (5.3)	14.5 (5.1)	15.4 (5.8)	0	0.026	14.2 (5.1)	14.1 (4.6)	14.4 (5.8)	0	0.441
Highest verbal response in first 24 hrs, n (%)	1	250 (38.8)	157 (36.4)	93 (43.5)	347	0.267	85 (21.2)	53 (19.3)	32 (25.4)	140	0.452
	2	2 (0.3)	2 (0.5)				1 (0.2)	1 (0.4)			
	3	3 (0.5)	2 (0.5)	1 (0.5)			1 (0.2)	1 (0.4)			
	4	8 (1.2)	4 (0.9)	4 (1.9)			9 (2.2)	5 (1.8)	4 (3.2)		
	5	382 (59.2)	266 (61.7)	116 (54.2)			305 (76.1)	215 (78.2)	90 (71.4)		
Average verbal response in first 24 hrs, mean (SD)		3.1 (1.9)	3.2 (1.9)	2.9 (1.8)	347	0.016	3.9 (1.7)	4.0 (1.6)	3.7 (1.7)	140	0.137
Lowest verbal response in first 24 hrs, n (%)	1	346 (53.6)	215 (49.9)	131 (61.2)	347	0.041	133 (33.2)	84 (30.5)	49 (38.9)	140	0.172
	2	3 (0.5)	1 (0.2)	2 (0.9)			2 (0.5)	1 (0.4)	1 (0.8)		
	3	5 (0.8)	4 (0.9)	1 (0.5)			4 (1.0)	4 (1.5)			
	4	11 (1.7)	7 (1.6)	4 (1.9)			16 (4.0)	9 (3.3)	7 (5.6)		
	5	280 (43.4)	204 (47.3)	76 (35.5)			246 (61.3)	177 (64.4)	69 (54.8)		
Interventions											
Averaged measured P0.1 in first 24 hrs, mean (SD)		2.2 (2.3)	2.1 (1.5)	2.4 (3.8)	841	0.531	1.6 (1.3)	1.6 (1.3)	1.7 (1.3)	402	0.762

Table 2
Descriptive summary statistics of the REG patient population used in the development (and internal validation) as well as temporal validation, stratified by in-hospital mortality. We only show the variables selected after the variable selection with backward elimination.

Variable		Development population (first wave)					Temporal validation population (second wave)				
		Overall	Survivor	Non-survivor	Missing	P-Value	Overall	Survivor	Non-survivor	Missing	P-Value
n		972	650	322			591	410	181		
Age, mean (SD)		63.4 (11.3)	61.0 (11.3)	68.2 (9.6)	0	<0.001	63.9 (12.1)	61.2 (12.3)	70.0 (9.2)	0	<0.001
Gender, n (%)	Male	704 (72.4)	454 (69.8)	250 (77.6)	0	0.013	418 (70.7)	294 (71.7)	124 (68.5)	0	0.490
Hospital identification number, n (%)	1	40 (4.1)	22 (3.4)	18 (5.6)	0	<0.001	59 (10.0)	39 (9.5)	20 (11.0)	0	0.795
	2	18 (1.9)	8 (1.2)	10 (3.1)			1 (0.2)	1 (0.2)			
	3	63 (6.5)	33 (5.1)	30 (9.3)			67 (11.3)	47 (11.5)	20 (11.0)		
	4	44 (4.5)	25 (3.8)	19 (5.9)			34 (5.8)	23 (5.6)	11 (6.1)		
	5	37 (3.8)	23 (3.5)	14 (4.3)			3 (0.5)	1 (0.2)	2 (1.1)		
	6	125 (12.9)	93 (14.3)	32 (9.9)			2 (0.3)	2 (0.5)			
	7	101 (10.4)	84 (12.9)	17 (5.2)							
	8	60 (6.2)	42 (6.5)	18 (5.6)							
	9	39 (4.0)	18 (2.8)	21 (6.5)							
	10	11 (1.1)	10 (1.5)	1 (0.3)			33 (5.6)	25 (6.1)	8 (4.4)		
	11	47 (4.8)	30 (4.6)	17 (5.3)							
	12	84 (8.6)	61 (9.4)	23 (7.1)			180 (30.5)	130 (31.7)	50 (27.6)		
	13	37 (3.8)	29 (4.5)	8 (2.5)			36 (6.1)	25 (6.1)	11 (6.1)		
	14	42 (4.3)	37 (5.7)	5 (1.6)							
	15						1 (0.2)		1 (0.6)		
	16	48 (4.9)	33 (5.1)	15 (4.7)							
	17	70 (7.2)	36 (5.5)	34 (10.6)			102 (17.3)	69 (16.8)	33 (18.2)		
	18	75 (7.7)	41 (6.3)	34 (10.6)			4 (0.7)	3 (0.7)	1 (0.6)		
	19	31 (3.2)	25 (3.8)	6 (1.9)			69 (11.7)	45 (11.0)	24 (13.3)		
Comorbidities											
Acute renal failure, n (%)		92 (9.5)	41 (6.3)	51 (15.8)	0	<0.001	55 (9.3)	26 (6.3)	29 (16.0)	0	<0.001
Chronic Obstructive Pulmonary Disease, n (%)		92 (9.5)	48 (7.4)	44 (13.7)	0	0.002	54 (9.1)	35 (8.5)	19 (10.5)	0	0.543
Chronic respiratory insufficiency, n (%)		31 (3.2)	12 (1.8)	19 (5.9)	0	0.001	12 (2.0)	7 (1.7)	5 (2.8)	0	0.527
Diabetes, n (%)		209 (21.5)	117 (18.0)	92 (28.6)	0	<0.001	159 (26.9)	102 (24.9)	57 (31.5)	0	0.116
Main APACHE IV reason for admission, n (%)	Pneumonia, viral	922 (94.9)	624 (96.0)	298 (92.5)	0	0.032	549 (92.9)	382 (93.2)	167 (92.3)	0	0.230
	Cardiac arrest	8 (0.8)		8 (2.5)			7 (1.2)	2 (0.5)	5 (2.8)		
	Cerebrovascular accident	4 (0.4)	1 (0.2)	3 (0.9)			6 (1.0)	5 (1.2)	1 (0.6)		
	Pneumonia, bacterial	4 (0.4)	3 (0.5)	1 (0.3)			2 (0.3)	1 (0.2)	1 (0.6)		
	Pneumonia, other	4 (0.4)	2 (0.3)	2 (0.6)			2 (0.3)	2 (0.5)			
	Respiratory-medical, other	4 (0.4)	3 (0.5)	1 (0.3)			3 (0.5)		3 (1.7)		
	Others	26 (2.7)	17 (3.4)	8 (2.7)			22 (3.7)	17 (3.4)	4 (2.4)		
Physiological and blood values											
Highest albumin in first 24 hrs, mean (SD)		87.9 (766.4)	96.7 (818.6)	71.0 (656.1)	298	0.659	71.2 (642.9)	29.8 (4.8)	159.6 (1136.1)	111	0.159
Lowest albumin in first 24 hrs, mean (SD)		26.8 (11.1)	27.1 (11.8)	26.1 (9.8)	298	0.226	27.8 (11.3)	28.9 (5.0)	25.4 (18.3)	110	0.022
Highest bicarbonate in first 24 hrs, mean (SD)		36.6 (325.0)	42.2 (397.3)	25.2 (4.0)	30	0.285	42.6 (416.7)	25.4 (3.4)	81.7 (754.0)	18	0.324
Highest creatinine in first 24 hrs, mean (SD)		111.2 (333.8)	109.6 (403.6)	114.5 (82.5)	36	0.767	149.7 (600.8)	102.9 (131.6)	256.9 (1066.6)	21	0.060
Fraction of inspired oxygen (FiO_2) in first 24 hrs, mean (SD)		65.8 (25.1)	63.4 (24.9)	70.7 (24.8)	36	<0.001	73.7 (26.2)	72.4 (23.7)	76.6 (30.9)	15	0.113
Lowest glucose in first 24 hrs, mean (SD)		6.3 (3.9)	6.1 (4.5)	6.8 (2.2)	22	0.004	6.8 (4.9)	7.0 (1.7)	6.4 (8.5)	6	0.384
Highest heartrate in first 24 hrs, mean (SD)		105.0 (21.8)	102.4 (18.9)	110.4 (25.9)	14	<0.001	104.7 (26.9)	103.1 (28.2)	108.3 (23.4)	2	0.020
Hospital length of stay priors ICU admission, mean (SD)		2.5 (3.0)	2.6 (3.3)	2.2 (2.4)	2	0.028	2.6 (4.5)	2.3 (4.1)	3.3 (5.3)	1	0.017

(*continued on next page*)

3. Study population

In EHR, the development population included 992 confirmed COVID-19 patients admitted to 19 ICUs, which could be followed until hospital discharge. In total, 316 patients (31.9 %) died during their hospital stay. Survivors were significantly younger (61.5 vs 68.6 years) and less often males (71.0 % vs 78.8 %) than non-survivors. For the temporal-validation population, 541 confirmed COVID-19 patients of 13 ICUs were included; 181 patients (33.5 %) died during their hospital stay. As in the development population, survivors were significantly younger (61.5 vs 70.5 years) and were less often males (70.6 % vs 74.0 %) than non-survivors. Table 1 shows the descriptive summary statistics of each patient population.

In REG, 972 patients admitted to the same 19 ICUs as EHR were included in the development population. In total, 322 patients (33.1 %) died during their hospital stay. Survivors were significantly younger (61.0 vs 68.2 years) and were less often males (69.8 % vs 77.6 %) than non-survivors. For the temporal-validation population, 591 confirmed COVID-19 patients of the same 13 ICUs as EHR were included; 181 patients (30.6 %) died during their hospital stay. As in the development population, survivors were significantly younger (61.2 vs 70.0 years), but were more often males (71.7 % vs 68.5 %) than non-survivor. Table 2 shows the descriptive statistics of each population.

4. Results

Among the 20 models built with Autoprognosis, the best model on both datasets was logistic regression. Table 3 shows the discrimination of the EHR and REG models, in terms of AUROC, AUPRC, PPV, NPV, and Brier scores. Both models have fair discriminatory performance in the internal validation (AUROC = 0.69 vs 0.74). On all measures, the best model developed on REG data performed significantly better ($p < 0.01$) than the best model developed on EHR data. Fig. 1 shows the calibration curves of the models for the internal and temporal validation. In the internal validation, the REG model and the EHR model are similarly calibrated.

Fig. 2 shows the coefficients of the EHR model. Age, fraction of inspired oxygen and glucose were the most important risk factors. The estimated globular filtration rate and erythrocytes were other important risk factors. The EHR membership model showed acceptable

Table 3
Discrimination of the EHR and REG model. For internal validation, we outline the average results for the fivefold cross validation with the standard deviation in brackets. For temporal validation the performance on the new population is reported. For PPV and NPV, the decision threshold was 0.3.

Data	Validation	AUROC	AUPRC	PPV	NPV	Brier score
EHR	Internal	0.693 (0.047)	0.506 (0.039)	0.630 (0.111)	**0.708 (0.016)**	0.205 (0.015)
REG	Internal	**0.737 (0.050)**	**0.574 (0.055)**	**0.732 (0.079)**	0.703 (0.009)	**0.195 (0.017)**
EHR	Temporal	0.746	0.562	0.706	0.677	0.188
REG	Temporal	**0.754**	**0.571**	**0.761**	**0.732**	**0.177**

Table 2 (*continued*)

Variable		Development population (first wave)					Temporal validation population (second wave)				
		Overall	Survivor	Non-survivor	Missing	P-Value	Overall	Survivor	Non-survivor	Missing	P-Value
Lowest mean blood pressure in first 24 hrs, mean (SD)		61.5 (14.3)	63.6 (12.4)	57.1 (16.7)	5	<0.001	62.3 (17.9)	63.8 (16.4)	58.9 (20.6)	4	0.005
Highest alveolar-arterial oxygen pressure difference in first 24 hrs, mean (SD)		359.8 (170.3)	345.9 (167.9)	387.2 (171.9)	95	0.001	434.4 (149.5)	422.3 (150.2)	461.4 (144.7)	46	0.004
Partial pressure of oxygen (PaO_2) in first 24 hrs, mean (SD)		80.4 (35.6)	82.4 (37.0)	76.5 (32.4)	164	0.019	75.9 (28.5)	77.5 (25.9)	72.3 (33.5)	23	0.073
Highest respiratory rate in first 24 hrs, mean (SD)		33.0 (8.6)	32.5 (8.5)	33.9 (8.8)	9	0.024	34.3 (8.7)	33.6 (8.2)	36.1 (9.4)	3	0.002
Lowest respiratory rate in first 24 hrs, mean (SD)							15.0 (4.8)	14.7 (4.5)	15.8 (5.3)	9	0.015
Lowest Thrombocytes in first 24 hrs, mean (SD)		238.5 (97.8)	245.6 (98.1)	224.1 (95.6)	47	0.001	239.8 (102.2)	254.0 (99.8)	207.1 (100.4)	21	<0.001
Highest serum ureum in first 24 hrs, mean (SD)		7.7 (8.2)	6.7 (8.9)	9.7 (6.1)	66	<0.001	10.0 (10.3)	9.4 (7.6)	11.4 (14.7)	30	0.095
Urine output in first 24 hrs, mean (SD)		1.4 (1.0)	1.4 (1.0)	1.4 (0.8)	11	0.184	1.8 (1.1)	1.8 (1.2)	1.6 (1.0)	1	0.006
Lowest eye response in first 24 hrs, n (%)	1	29 (3.0)	10 (1.5)	19 (6.0)	7	0.002	24 (4.1)	10 (2.4)	14 (7.8)	1	0.014
	2	5 (0.5)	4 (0.6)	1 (0.3)			5 (0.8)	3 (0.7)	2 (1.1)		
	3	36 (3.7)	24 (3.7)	12 (3.8)			14 (2.4)	8 (2.0)	6 (3.3)		
	4	895 (92.7)	608 (94.1)	287 (90.0)			547 (92.7)	389 (94.9)	158 (87.8)		
Lowest motor response in first 24 hrs, n (%)	0				7	<0.001	1 (0.2)		1 (0.6)	1	0.017
	1	24 (2.5)	5 (0.8)	19 (6.0)			21 (3.6)	8 (2.0)	13 (7.2)		
	2	1 (0.1)	1 (0.2)				1 (0.2)	1 (0.2)			
	4	2 (0.2)	2 (0.3)				4 (0.7)	2 (0.5)	2 (1.1)		
	5	25 (2.6)	14 (2.2)	11 (3.4)			11 (1.9)	8 (2.0)	3 (1.7)		
	6	913 (94.6)	624 (96.6)	289 (90.6)			552 (93.6)	391 (95.4)	161 (89.4)		
Interventions											
Mechanical ventilation at ICU admission, n (%)		422 (43.4)	255 (39.2)	167 (51.9)	0	<0.001	114 (19.3)	73 (17.8)	41 (22.7)	0	0.206
Mechanical ventilation in first 24 hrs, mean (SD)		759 (78.1)	489 (75.2)	270 (83.9)	0	0.003	297 (50.3)	195 (47.6)	102 (56.4)	0	0.060

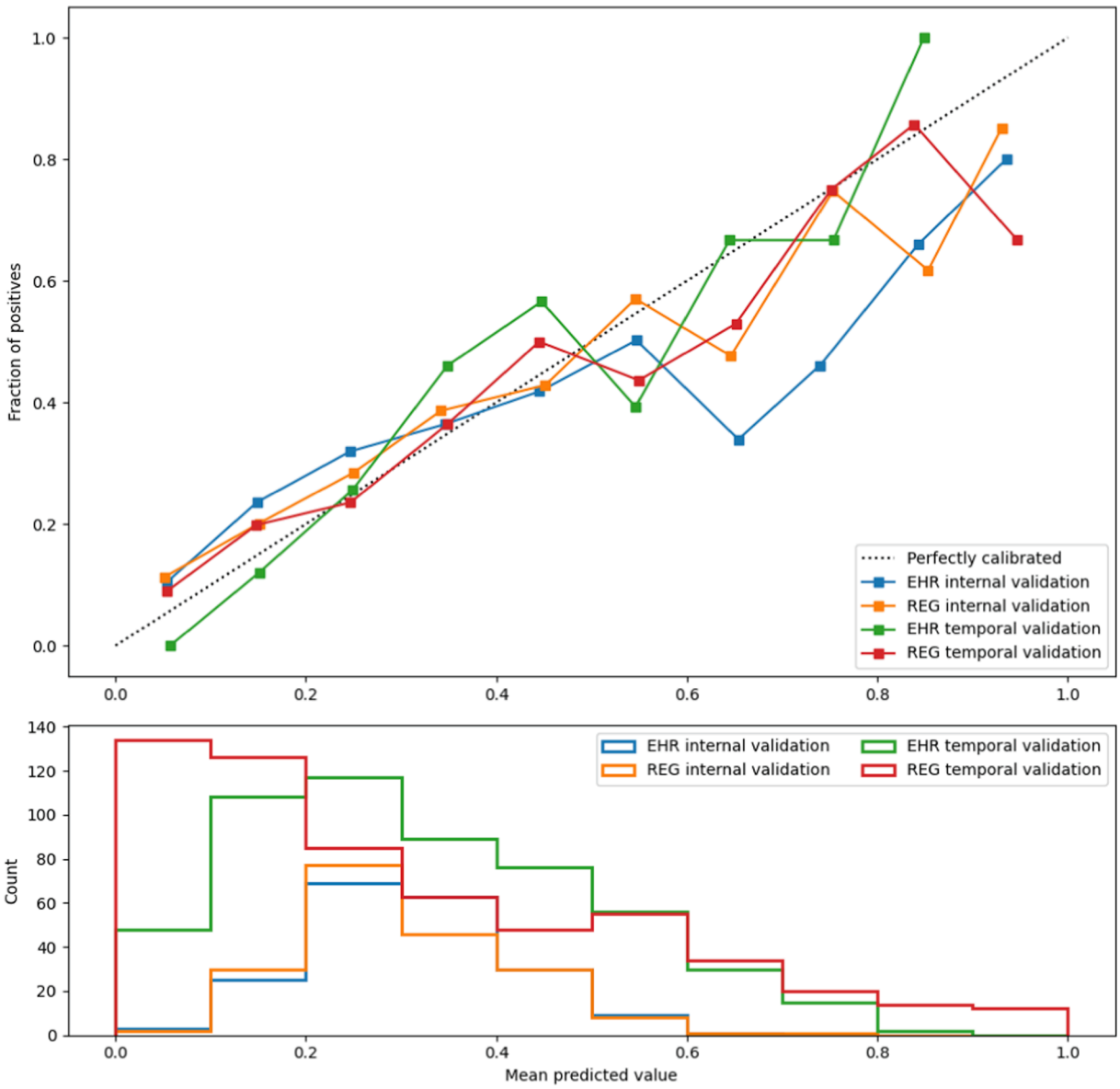


Fig. 1. Calibration curves of the REG and EHR models in the internal as well as temporal validations.

performance (AUROC = 0.71), as illustrated in Table 4. The EHR model yielded better results in the temporal validation than in the internal validation for all measures but NPV, which slightly decreased (AUROC = 0.75, see Table 3). In the temporal validation the model calibration was slightly worse than in the internal validation (Fig. 1).

Fig. 3 shows the coefficients of the REG model. Age and chronic respiratory insufficiency were found as most important risk factors. The fraction of inspired oxygen and chronic obstructive pulmonary disease (COPD) were other relevant risk factors. The REG membership model showed an acceptable performance (AUROC = 0.76), as outlined in Table 4. The predictive performance of the REG model for the temporal validation improved for all the measures compared to the internal validation, except AUPRC, which slightly decreased (AUROC = 0.75, see Table 3). The REG model showed better calibration than in the internal validation for most of the predictions (Fig. 1).

The results for the temporal validation are significantly better for the REG model than for the EHR model for all the measures ($p < 0.01$), although the EHR model had a larger improvement than the REG model from the internal to the temporal validation for AUROC (from 0.69 to 0.75, and from 0.74 to 0.75, respectively). The calibration is similarly good for both models (Fig. 1).

5. Discussion

5.1. Main findings

We assessed the predictive performance of clinical prognostic models for in-hospital mortality of ICU patients with confirmed COVID-19 using high-granular EHR data and low-granular REG data. The predictive performance in the internal validation was fair (AUROC of 0.69–0.74). In the temporal validation, the performance improves (AUROC from 0.69 to 0.75 for EHR and from 0.74 to 0.75 for REG). The membership-models' results on both datasets indicate that the case-mix was different and therefore temporal validation assess the transportability of the models. For temporal validation, transportability means that the models are stable over time. Both models are well transportable to the temporal-validation population since their performance in the temporal validation increased. Such increase may also be due to the use of 5-fold cross validation in the internal validation which resulted in reporting conservative performance: the model is trained five times with 80 % of the

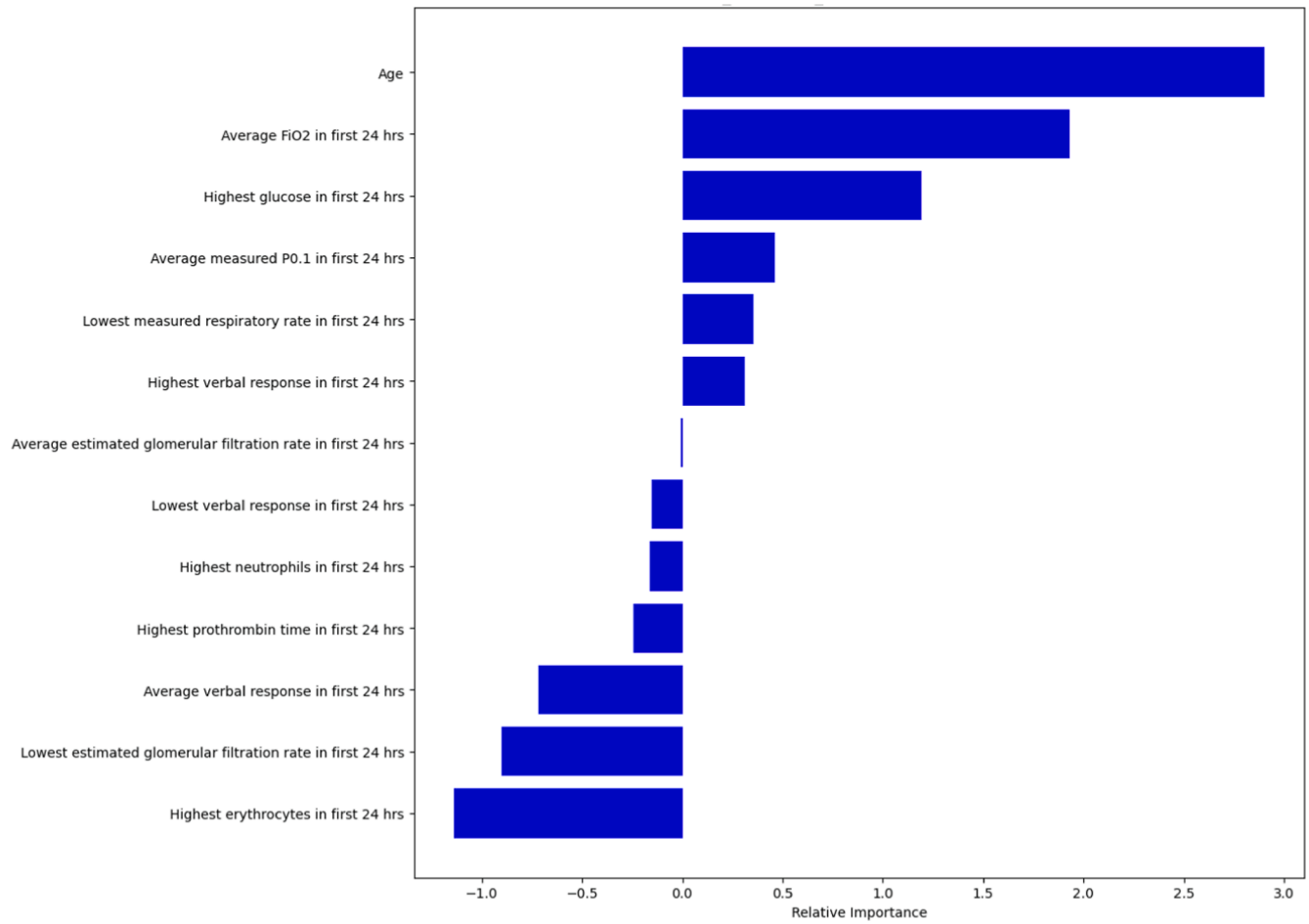


Fig. 2. Coefficients of the EHR logistic regression model. $PaCO_2$ is the partial pressure of carbon dioxide, FiO_2 is the fraction of inspired oxygen. Supplementary Table S2 includes the model description.

Table 4
Discrimination of the EHR and REG membership models. We outline the average results for the fivefold cross validation with the standard deviation in between brackets.

Data	AUROC	AUROC interpretation	Case mix	Assessed property
REG	0.756 (0.046)	Acceptable	Different	Transportability
EHR	0.707 (0.034)	Acceptable	Different	Transportability

data as every fold correspond to one fifth (20 %) of the data. We select a final model used in the temporal validation by retraining on the whole development data. The better performance of the REG model in the internal validation, and the similar performance of both models in the temporal validation, despite more predictors in the EHR model, may be due to the greater number of missing values in the EHR dataset.

Age, fraction of inspired oxygen and glucose were the strongest predictors in the EHR model. In REG, age and chronic respiratory insufficiency were the most important predictors. Different risk factors are identified in different data sources due to the different total set of variables included. Additionally, some variables, such as the lowest verbal response in the first 24 h, although available in both datasets, were selected by the variable selection in one model but not the other.

5.2. Related work

Similar to [18], we found age and respiratory-system predictors to be predictor of mortality among COVID-19 patients. Other predictors found in other studies ranged from diverse laboratory test to comorbidities [1,19,20,21,22,23]. Izcovich et al. identified 49 valuable predictors [24], including various laboratory tests that we also identified in our EHR data, such as neutrophils, or in REG, e.g., COPD (see Figs. 2 and 3). Among these other predictors found by other studies, some were not included in our dataset (the participating hospitals did not collect or share such information). Some of earlier found comorbidities, medications (notably steroids, anticoagulants, vasopressors) and other predictors, e.g., lung compliance, ventilator volume and pressures that were included in our EHR dataset, were not selected as predictors in our models. This might be a result from dependences and correlations that are specific for our set of predictors. Various prognostic models of mortality among patients with COVID-19 have been proposed [1,18,25,26,27,28,29]. Their predictive performance varied from fair (AUROC 0.7–0.8) to excellent (AUROC $>$ 0.9). However, many studies showed high risk of bias [1] and only few temporally validate the models [27–29]. Our performance in the temporal validation is in line with another externally-validated model developed on EHR data [30], as well as other studies that take into account readily availability data [31,32]. The importance of external validation and continuous monitoring and updated of machine-learning models to ensure their long-term safety and effectiveness has also been underlined by previous studies [33,34].

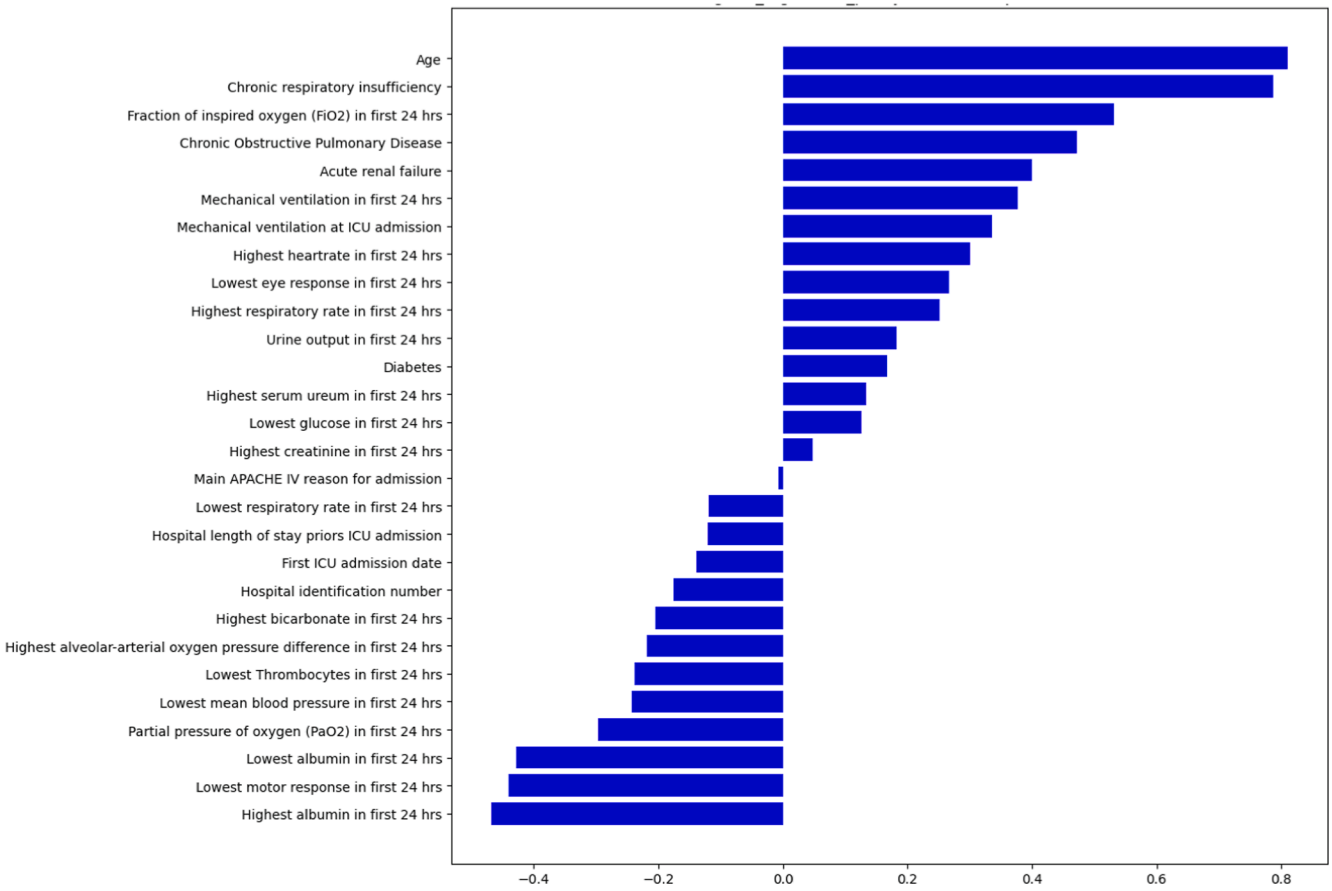


Fig. 3. Coefficients of the REG logistic regression model. Supplementary Table S3 includes the model description.

5.3. Strengths and weaknesses

We temporarily validated COVID-19 prognostic models, which is an important aspect of model evaluation [15], especially with a new disease. We used data from multiple centers and two different data sources, each with their own benefits and limitations. The EHR data source includes raw EHR data and hence more variables and more measurements per variable compared to the REG dataset. However, it was time-consuming to join all different EHR data schemas and, perhaps accordingly, missing values were frequent in the EHR data. The REG dataset is less rich in the number of variables and measures per variable but more standardized and quality-controlled.

Our study also has some limitations that need consideration in its interpretation. First, due to privacy regulations we were not able to join both datasets and link the same patients. Although we used the same ICUs there were small differences regarding included patients.

Second, the EHR data source included data only up to January 1st, 2021, so it was not possible to temporally validate its model on later waves of infection, when vaccinations became available. To compare models built on EHR and REG data we limited REG data to the same period. Six of the 19 ICUs did not provide data after June 2020 and were not included in the temporal validation. The EHR data, although a dump of several EHRs, did not include all laboratory or other individual patient variables available, and not all the hospitals provided all the variables, e.g., D-dimer was collected only by few hospitals. Although repeated measurements of the same variable (time series) were available, we aggregated them, reducing the granularity in time and potentially losing useful information. Other studies include more and different individual patient information, such as time series of laboratory values and features derived from CT images, which may explain their higher predictive performance [1].

Third, we did apply imputation and normalization before the data splitting due to using cross validation, which may introduce a bias. After 5-fold cross-validation, there would be the need of 5 different imputations (actually multiple imputations), as well as 5 different normalizations. First this would have made the difficult to identify the final imputed and normalized data: it is not straight forward which of the five imputations or normalization to use in the final model or how to properly aggregated those five imputations. Second, it would also have made the computation time explode for the EHR data, which holds over 2000 variables (before selection). After in-depth discussions and preliminary analyses, we nested the variable selection in the cross validation using a majority voting to select the variables from the five folds (a variable needs to be selected in 3 out of 5 folds to be selected in the final model), but not the rest. Given we also temporally validated the model, we believe this bias has a low impact.

Forth, we do not exclude that extensive parameter tuning of single models may provide slightly better results than automated machine learning with AutoPrognosis. However, a gap between the potential and actual use of machine learning in prognostic research exists because classical model development and tuning requires greater time and effort (and may become unfeasible in the healthcare domain). More importantly, it is hard for clinicians with no or few expertise in machine learning to do so [35]. Automated machine learning tackles these issues and our study shows how can be successfully used in the healthcare domain.

Finally, removing transferred patients from the EHR dataset may have introduced bias in the dataset because transferred patients may be healthier since they are fit for transport, or more severely ill and need treatment in a better equipped ICU. Whenever data from the referring and receiving hospital were available, data were linked to limit the exclusions.

5.4. Implications

Registries, with less-granular but readily-available and controlled data, provide better performance than high-granular EHR data in the internal validation and show similar results to EHR data in the temporal validation. Independently of the data source, model performance remains stable over time. This is an important finding because long-standing ICU registries require less effort for data-collection, integration, and processing than setting up a specific research data set from multiple EHR with different data unless such data platform is already in place, which is currently rare but upcoming (giviti.marionegri.it/portfolio/covid-19/, last access 11/02/2022) [5,36,37]. When EHRs will move to using information standards and/or FAIR data, many of the current disadvantages of EHR data may disappear, since collection, integration and processing will be eased. However, validation and quality control of data as in place in registries may still remain a challenge. High-granular EHR data could still be beneficial for other problems, such as finding the optimal combination of ventilators settings or drugs for individual patients.

5.5. Future works

We aggregated the repeated measurements of each numerical variable available. Exploiting repeated measurements should be investigated. COVID-19 patients typically have long ICU stays. Twenty-four hours might be a too-short interval to estimate patients' survival. Determining the best 'ICU trial time' requires further research. Other interesting directions of research would be exploiting different models instead of the logistic regression for the population membership discrimination, e.g. kernel based methods, as well as tracking the presence or absence of data with additional variables instead of imputing missing values, as done by a recent study [38].

6. Conclusions

In our study, temporally-validated models built on less-granular but readily-available registry data performed closely to models developed with higher-granular EHR data and showed the same transportability to a prospective COVID-19 population as model developed with higher-granular EHR data. Readily-available registry data might be a valuable resource when a rapid response is needed. Future research is needed to verify whether this finding can be confirmed for upcoming COVID-19 waves and for models focusing on other ICU patient categories.

7. Summary Table

What was already known on the topic:

- Electronic health records (EHR) typically have high granularity (multiple variables and measurements over time).
- EHR data can enable the application of advanced machine learning methods, but combining these data from multiple centers requires a considerable effort and time.
- In a sudden pandemic or crisis situation, a rapid response is needed, therefore, waiting to collect, curate and aggregate EHR data is undesirable.

What this study added to our knowledge:

- Prognostic models built on less-granular but readily-available registry data can, in particular cases as in this study, achieve performance similar to models built on high-granular EHR data
- Prognostic models built on high- and less-granular data of COVID-19 patients show equal transportability to a prospective COVID-19 population.

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.

Acknowledgments

We thank Sylvia Brinkman for her support with the extraction and understanding of the NICE data.

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.

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.

Data and code availability

All participating hospitals have access to the Dutch ICU Data Warehouse and NICE data. The NICE registry data are available under conditions as described on the NICE website at stichting-nice.nl/extractieverzoek_procedure.jsp (in Dutch). External researchers can get access to the Dutch ICU Data Warehouse in collaboration with any of the participating hospitals. The list of collaborators is available in the co-author list and in the collaborators section, through the corresponding author, and through the contact details on amsterdammedicaldatascience.nl. Research questions have to be in line with the DSA; to investigate the course of COVID-19 in the ICU and to research potential treatments. Researchers have sign a code of conduct before accessing the data.

The code used for our analyses is publicly available at bitbucket.org/aumc-kik/automl4covid.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104863.

References

[1] L.B. Wynants, G.S. Van Calster, R.D. Collins, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal, BMJ, 369: m1328.
[2] G. Grasselli, M. Greco, A. Zanella, et al., Risk Factors Associated With Mortality Among Patients With COVID-19 in Intensive Care Units in Lombardy, Italy. JAMA Intern Med. (2020), https://doi.org/10.1001/jamainternmed.2020.3539.
[3] S. Richardson, J.S. Hirsch, M. Narasimhan, et al., Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area, JAMA 323 (2020) 2052–2059, https://doi.org/10.1001/jama.2020.6775.
[4] C. Karagiannidis, C. Mostert, C. Hentschker, et al., Case characteristics, resource use, and outcomes of 10 021 patients with COVID-19 admitted to 920 German hospitals: an observational study, Lancet Respir Med 8 (2020) 853–862, https://doi.org/10.1016/S2213-2600(20)30316-7.
[5] L.M. Fleuren, D.P. de Bruin, M. Tonutti, R.C.A. Lalisang, P.W.G. Elbers, Dutch ICU Data Sharing Collaborators. Large-scale ICU data sharing for global collaboration: the first 1633 critically ill COVID-19 patients in the Dutch Data Warehouse, Intensive Care Med. (2021).

[6] N. van de Klundert, et al., Data Resource Profile: the Dutch National Intensive Care Evaluation (NICE) Registry of Admissions to Adult Intensive Care Units, Int J Epidemiol 44 (6) (2015), p. 1850-1850h.
[7] 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 9 (6) (2002) 600–611.
[8] M. Prokop, W. van Everdingen, V.T. van Rees, et al., CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19—Definition and Evaluation, Radiology 296 (2020) E97–E104, https://doi.org/10.1148/radiol.2020201473.
[9] M.C. Ottenhoff, L.A. Ramos, W. Potters, et al., Dutch COVID-PREDICT research group. Predicting mortality of individual patients with COVID-19: a multicentre Dutch cohort, BMJ Open. 11 (7) (2021) e047347, https://doi.org/10.1136/bmjopen-2020-047347.
[10] S. van Buuren, K.M.I.C.E. Groothuis-Oudshoorn, Multivariate Imputation by Chained Equations in R, Journal of Statistical Software 45 (2011).
[11] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, in Selected Papers of Hirotugu Akaike, E. Parzen, K. Tanabe, and G. Kitagawa, Editors. 1998, Springer New York: New York, NY. p. 199-213.
[12] A.M. Alaa, and M. van der Schaar, AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning. 2018.
[13] I. Vagliano, S. Brinkman, Abu-Hanna, et al. 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, International Journal of Medical Informatics, Volume 160, 2022, 104688, https://doi.org/10.1016/j.ijmedinf.2022.104688.
[14] K. Rufibach, Use of Brier score to assess binary predictions, Journal of clinical epidemiology 63 (8) (2010) 938–939.
[15] P.C. Austin, D. van Klaveren, Y. Vergouwe, D. Nieboer, D.S. Lee, E.W. Steyerberg, Geographic and temporal validity of prediction models: different approaches were useful to examine model performance, J Clin Epidemiol. 79 (2016) 76–85, https://doi.org/10.1016/j.jclinepi.2016.05.007.
[16] T.P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, K.G. Moons, A new framework to enhance the interpretation of external validation studies of clinical prediction models, J Clin Epidemiol. 68 (3) (2015 Mar) 279–289, https://doi.org/10.1016/j.jclinepi.2014.06.018.
[17] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, 3rd edition, John Wiley & Sons, 2013.
[18] F. Zhou, T. Yu, R. Du, et al., Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study, The Lancet 395 (2020) 1054–1062, https://doi.org/10.1016/S0140-6736(20)30566-3.
[19] R.K. Gupta, 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).
[20] B. Gallo Marin, et al., Predictors of COVID-19 severity: A literature review, Rev Med Virol (2020) e2146.
[21] J.A. Siordia Jr., Epidemiology and clinical features of COVID-19: A review of current literature, J Clin Virol 127 (2020), 104357.
[22] J. Xu, X. Yang, L. Yang, et al., Clinical course and predictors of 60-day mortality in 239 critically ill patients with COVID-19: A multicenter retrospective study from Wuhan, China. Crit Care. 24 (2020) 394.
[23] P. Ferrando-Vivas, J. Doidge, K. Thomas, D.W. Gould, P. Mouncey, M. Shankar-Hari, J.D. Young, K.M. Rowan, D.A. Harrison, ICNARC COVID-19 Team. Prognostic Factors for 30-Day Mortality in Critically Ill Patients With Coronavirus Disease 2019: An Observational Cohort Study, Crit Care Med. 49 (1) (2021) 102–111, https://doi.org/10.1097/CCM.0000000000004740.
[24] A. Izcovich, et al., Prognostic factors for severity and mortality in patients infected with COVID-19: A systematic review, PLoS One 15 (11) (2020) e0241955.
[25] A.A. El-Solh, Y. Lawson, M. Carter, et al., Comparison of in-hospital mortality risk prediction models from COVID-19, PLOS ONE 15 (2020) e0244629.
[26] B.G. Pijls, S. Jolani, A. Atherley, et al., Demographic risk factors for COVID-19 infection, severity, ICU admission and death: a meta-analysis of 59 studies, BMJ Open 11 (2021) e044640.
[27] S.R. Knight, R.K. Gupta, A. Ho, et al., Prospective validation of the 4C prognostic models for adults hospitalised with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. Thorax. 2021 Nov 22:thoraxjnl-2021-217629. doi: 10.1136/thoraxjnl-2021-217629.
[28] S. Heber, D. Pereyra, W.C. Schrottmaier, et al., A Model Predicting Mortality of Hospitalized Covid-19 Patients Four Days After Admission: Development, Internal and Temporal-External Validation, Front Cell Infect Microbiol. 24 (11) (2022 Jan), 795026, https://doi.org/10.3389/fcimb.2021.795026.
[29] M.M. Churpek, S. Gupta, A.B. Spicer, et al., Machine Learning Prediction of Death in Critically Ill Patients With Coronavirus Disease 2019, Crit Care Explor. 3 (8) (2021 Aug 19) e0515.
[30] D. Plečko, N. Bennett, J. Mårtensson, et al., Rapid Evaluation of Coronavirus Illness Severity (RECOILS) in intensive care: Development and validation of a prognostic tool for in-hospital mortality, Acta Anaesthesiologica Scandinavica. 66 (2021), https://doi.org/10.1111/aas.13991.
[31] L. Famiglini, A. Campagner, A. Carobene, F. Cabitza, A robust and parsimonious machine learning method to predict ICU admission of COVID-19 patients Medical and Biological Engineering and Computing. (2022) doi: 10.1007/s11517-022-02543-x.
[32] Y. Gao, G.-Y. Cai, W. Fang, et al., Machine learning based early warning system enables accurate mortality risk prediction for COVID-19, Nature communications 11 (1) (2020) 1–10.
[33] F. Cabitza, A. Campagner, F. Soares, et al., The importance of being external. methodological insights for the external validation of machine learning models in medicine, Computer Methods and Programs in Biomedicine 208 (2021), 106288.
[34] J. Feng, R.V. Phillips, I. Malenica, et al., Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare, npj Digital Medicine 5 (1) (2022) 1–9.
[35] G. Luo, B.L. Stone, M.D. Johnson, et al., Automating Construction of Machine Learning Models With Clinical Big Data: Proposal Rationale and Methods, JMIR Res Protoc. 6 (8) (2017 Aug 29) e175.
[36] S. Finazzi, G. Paci, L. Antiga, et al., GiViTI-PROSAFE collaboration. PROSAFE: a European endeavor to improve quality of critical care medicine in seven countries, Minerva Anestesiol. 86 (12) (2020 Dec) 1305–1320, https://doi.org/10.23736/S0375-9393.20.14112-9.
[37] A.L. Goldberger, L.A. Amaral, L. Glass, et al., PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation. 101 (23) (2000) E215–E220, https://doi.org/10.1161/01.cir.101.23.e215.
[38] N. Tomašev, X. Glorot, J.W. Rae, et al., A clinically applicable approach to continuous prediction of future acute kidney injury, Nature. 572 (7767) (2019) 116–119.