Universiteit Leiden
The Netherlands

Predicting readmission or death after discharge from the ICU: external validation and retraining of a machine learning model

Hond, A.A.H. de; Kant, I.M.J.; Fornasa, M.; Cina, G.; Elbers, P.W.G.; Thoral, P.J.; ... ; Steyerberg, E.W.

Citation

Hond, A. A. H. de, Kant, I. M. J., Fornasa, M., Cina, G., Elbers, P. W. G., Thoral, P. J., ... Steyerberg, E. W. (2023). Predicting readmission or death after discharge from the ICU: external validation and retraining of a machine learning model. *Critical Care Medicine*, *51*(2), 291-300. doi:10.1097/CCM.0000000000005758

Version:	Publisher's Version
License:	Creative Commons CC BY-NC-ND 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3764292

Note: To cite this publication please use the final published version (if applicable).

OPEN

Predicting Readmission or Death After Discharge From the ICU: External Validation and Retraining of a Machine Learning Model

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

Anne A. H. de Hond, MSc[1,2,3]
Ilse M. J. Kant, PhD[1,3]
Mattia Fornasa, PhD[4]
Giovanni Cinà, PhD[4,5]
Paul W. G. Elbers, MD, PhD[6]
Patrick J. Thoral, MD[6]
M. Sesmu Arbous, MD, PhD[7]
Ewout W. Steyerberg, PhD[3]

OBJECTIVES: Many machine learning (ML) models have been developed for application in the ICU, but few models have been subjected to external validation. The performance of these models in new settings therefore remains unknown. The objective of this study was to assess the performance of an existing decision support tool based on a ML model predicting readmission or death within 7 days after ICU discharge before, during, and after retraining and recalibration.

DESIGN: A gradient boosted ML model was developed and validated on electronic health record data from 2004 to 2021. We performed an independent validation of this model on electronic health record data from 2011 to 2019 from a different tertiary care center.

SETTING: Two ICUs in tertiary care centers in The Netherlands.

PATIENTS: Adult patients who were admitted to the ICU and stayed for longer than 12 hours.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: We assessed discrimination by area under the receiver operating characteristic curve (AUC) and calibration (slope and intercept). We retrained and recalibrated the original model and assessed performance via a temporal validation design. The final retrained model was cross-validated on all data from the new site. Readmission or death within 7 days after ICU discharge occurred in 577 of 10,052 ICU admissions (5.7%) at the new site. External validation revealed moderate discrimination with an AUC of 0.72 (95% CI 0.67–0.76). Retrained models showed improved discrimination with AUC 0.79 (95% CI 0.75–0.82) for the final validation model. Calibration was poor initially and good after recalibration via isotonic regression.

CONCLUSIONS: In this era of expanding availability of ML models, external validation and retraining are key steps to consider before applying ML models to new settings. Clinicians and decision-makers should take this into account when considering applying new ML models to their local settings.

KEY WORDS: clinical decision support; critical care; data science; external validation; generalizability; machine learning

There has been a rapid increase in the use of machine learning (ML) techniques for prediction modeling on routinely collected hospital data (1). The ICU forms a popular application area with its high-volume data from continuously monitored patients (2, 3). ML models have been developed at the ICU to predict the onset of sepsis (4, 5), COVID-19 disease progression (6, 7), and mortality and readmission (2, 8). Clinicians increasingly encounter ML vendors that claim to revolutionize their clinical workflow, environment, and patient outcomes. Therefore, it is important that clinicians are aware of the

Copyright © 2022 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the Society of Critical Care Medicine and Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

DOI: 10.1097/CCM.0000000000005758

KEY POINTS

Question: Machine learning applications for the ICU lack rigorous external validation. We assessed the external validity and effect of retraining on the predictive performance for a certified machine learning model.

Findings: Generalizability was difficult to attain for this machine learning model despite apparent similarities between patient population, healthcare context, and model specification. Retraining with additional focus on disease severity monitoring and ICU specialty improved predictive performance.

Meaning: External validation and retraining are key steps to consider before applying machine learning models to new settings.

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

quality assessment steps that need to be taken before the local implementation of these ML models.

Before introducing these ML models in a clinical environment that is different from the development site (e.g., a different ICU, hospital, or country), we need to assess the generalizability or external validity at this site (9–11). However, few ML models have been subjected to external validation. A recent study found that less than one third of Food and Drug Administration (FDA) approved ML models reported to have undergone multisite assessment (12). Furthermore, less than 11% of prediction models developed for the ICU were externally validated (13). This is particularly problematic as correlations based on site-specific clinical practices are prone to boost local performance of ML models but may hamper generalizability to other settings (14). Similarly, shifts in the data-generating process over time at a single site can affect performance (15–18). A recent example is an ICU sepsis prediction model. This model was implemented and widely adopted before external validation showed poor discrimination and calibration, which in turn may have dangerous consequences for patients (19).

Several steps may be taken to improve model performance at a new site after external validation. First, the external validation may show poor calibration, meaning that the estimated probabilities are unreliable. Recalibration of the probability outcomes may be applied to improve the probability estimates (15, 20, 21). Second, when the external validation shows subpar discrimination, the model may be retrained on data from the external validation site. However, it remains unclear to date when and under which circumstances these steps are necessary to ensure safe and responsible introduction of ML models in local clinical settings.

We aimed to assess the external validity of a certified ML model for the ICU: Pacmed Critical (8). Pacmed Critical is a decision support tool based on a ML predictive model that estimates the probability of readmission or death within 7 days after ICU discharge. It intends to support intensivists in determining the optimal moment for discharge of a patient from the ICU to a clinical ward. Second, we aimed to assess the effect of retraining of the model on predictive performance through a temporal validation design. This study serves as a use case illustrating how the generalizability of ML models may be addressed by local retraining.

MATERIALS AND METHODS

Patients

For the external validation, retraining, and recalibration of the Pacmed model, we used electronic health record (EHR) data from Leiden University Medical Center (Leiden UMC), a tertiary care center in The Netherlands. These data were collected between 2011 and 2019. We purposefully left the year 2020 out, as COVID-19 drastically changed the composition of ICU patients and disrupted ICU care processes which might have significantly impacted model performance. This study was conducted in accordance with the Helsinki Declaration. The need for ethical approval was waived for this study by the Institutional Review Board of the Amsterdam University Medical Center (UMC), location VUmc (2017.212, date: May 2017, study title: “Right Data, Right Now: Predicting ICU readmission rates”).

Outcome

The outcome variable was defined as a readmission to the ICU or unexpected death within 7 days after discharge from the ICU to the ward. Our definition of an ICU discharge did not include patients who were discharged to the medium care unit (MCU) as the intensity of monitoring on the ICU is comparable with that of the MCU, whereas on the ward, the level of monitoring is much less intense. A planned surgical readmission was not considered as a readmission but rather modeled as one continuous ICU stay. ICU admissions with a time difference of less than 12 hours were removed from the cohort. Only the readmission was considered in the case of death after readmission. Other exclusion criteria were patients younger than 18 years old, patients being transferred to the ICU of another hospital, dying at the ICU during the original admission, or receiving palliative care.

ML Model

Pacmed Critical is a Conformité Européenne (CE)-certified decision support tool, meeting the safety, health, and environmental protection requirements of the European Union. It intends to assist intensivists in determining the optimal moment to discharge their patient from the ICU to the ward. It is a gradient boosting model that was developed and validated on EHR) data collected between 2004 and 2021 from the Amsterdam UMC, location VUmc (Amsterdam UMC), a tertiary care center in The Netherlands. The area under the receiver operating characteristic curve (AUC) at the validation cohort of Amsterdam UMC was 0.78 (95% CI 0.75–0.81). An in-depth description of the original model development and initial validation is reported elsewhere (8).

Retraining

The Pacmed Critical model was retrained on data from the Leiden UMC with the same pipeline and modeling techniques as those used for the original model developed at the Amsterdam UMC. A careful mapping was made between the feature sets of Amsterdam UMC and Leiden UMC to deal with discrepancies in recorded features between the two locations due to differences in their EHR systems (Epic, Epic Systems Corporation, Verona, WI, and HiX, Chipsoft B.V., Amsterdam, The Netherlands, respectively). Features were included for model development when good data quality could be guaranteed for the data from which the feature was computed. This led to slightly different feature lists between the two hospitals (**Table E1**, http://links.lww.com/CCM/H261). Differences in inclusion were for example caused by incomplete feature data for some of the recorded years. Leiden UMC added features related to severity monitoring (e.g., base excess mixed venous and continuous venovenous hemofiltration blood flow) and ICU specialty.

Validation Design

We compared the descriptive statistics on patient demographics, clinical context, and type of event (readmission or death within 7 d) from the Amsterdam UMC with the Leiden UMC. We supplemented this analysis with information on the type of admission and mortality risk obtained from the National Intensive Care Evaluation (NICE) registry (22) for the beginning of the NICE registration (2013) up to and including 2019 for the Leiden UMC and 2021 for the Amsterdam UMC.

The predictive performance of the Pacmed model on Leiden UMC data was measured via a temporal validation design at four time points: before retraining, after the first round of retraining, after the second round of retraining, and after the third and final round of retraining (**Table E2**, http://links.lww.com/CCM/H261). The validation before retraining represents the external validation of the original gradient-boosted ML model developed on Amsterdam UMC data, validated on new, unseen data from the Leiden UMC ("External validation before retraining," Table E2, http://links.lww.com/CCM/H261). This validation was performed on the 2018–2019 Leiden UMC cohort. Temporal validation consisted of retraining the model on subsets of the Leiden UMC data and validation on the 2018–2019 Leiden UMC cohort. For the first round of retraining, the ML model was trained on data from 2011 to 2015 ("Temporal validation 1," Table E2, http://links.lww.com/CCM/H261). In the second round of retraining, data from 2011 to 2017 were used for retraining ("Temporal validation 2," Table E2, http://links.lww.com/CCM/H261). The final model was retrained on all Leiden UMC data (2011–2019). It underwent a 10-fold cross-validation after which we assessed its performance on the 2018–2019 cohort ("Validation after retraining," Table E2, http://links.lww.com/CCM/H261).

We measure the predictive performance for all validation moments along three axes: discrimination, calibration, and net benefit. Discrimination quantifies the separation between low- and high-risk subjects and was measured via the AUC (23). The AUC ranges between 0.5 and 1, with higher values indicating better discrimination. Calibration is good when the proportion of patients receiving a given risk score approximates that risk score (e.g., 40% of patients are readmitted within

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

the group of patients receiving a 40% risk of readmission) (23). Calibration was assessed through the calibration slope (1 for perfect calibration), intercept (0 for perfect calibration), and calibration loss by bins (lower loss is better) (21, 24, 25). Probability predictions were recalibrated via isotonic regression (26). Such rescaling is common for ML models for the probability estimates to better approximate the actual probability distribution. CIs were obtained through bootstrapping (1,000 samples).

A decision curve analysis (DCA) was performed to assess how the Pacmed model could impact patient care within the clinical workflow (27, 28). A DCA plots net benefit across a range of decision probability thresholds. Net benefit measures the number of true-positive classifications (patients who were readmitted or died and were identified as such) penalized for false-positive classifications (patients who were not readmitted and did not die but were identified as such). The DCA was performed with four patient discharge strategies for Leiden UMC data: discharge none, discharge all, discharge according to the original model developed at Amsterdam UMC, and discharge according to the final retrained model developed at Leiden UMC. In the reporting of our results, we followed the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement (29).

Subgroup Analysis

To assess model performance across different ICU specialties, we performed a subgroup analysis for surgery, internal medicine, cardiology, neurology, and gastroenterology patients.

Software

All analyses were performed in Python 3.8.0 (released by the Pythons Software Foundation). Code for the validation analysis is available online at https://git.lumc.nl/aahdehond/pacmed-validation.

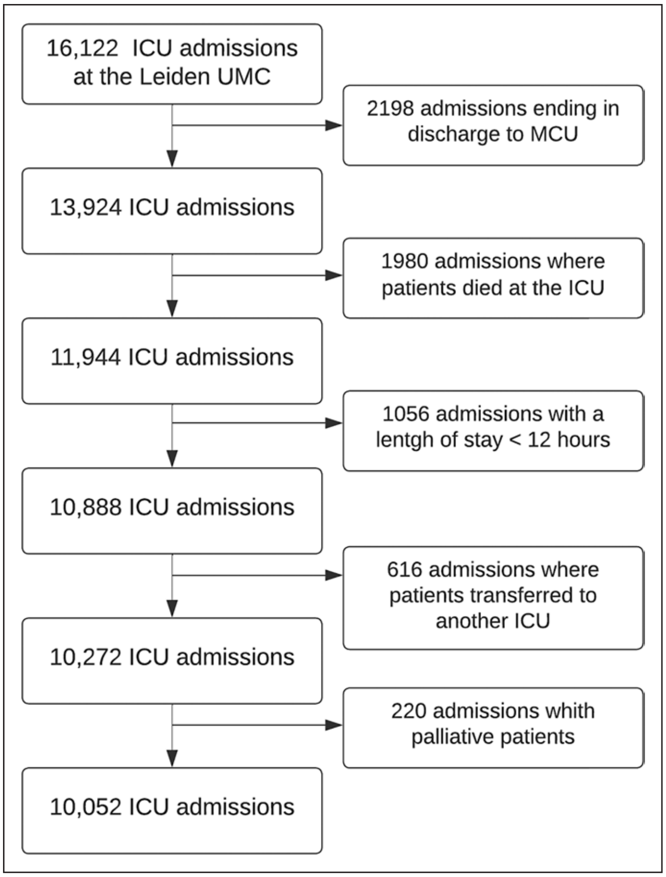
16,122 ICU admissions at the Leiden UMC
2198 admissions ending in discharge to MCU
13,924 ICU admissions
1980 admissions where patients died at the ICU
11,944 ICU admissions
1056 admissions with a lentgh of stay < 12 hours
10,888 ICU admissions
616 admissions where patients transferred to another ICU
10,272 ICU admissions
220 admissions whith palliative patients
10,052 ICU admissions

Figure 1. Flow chart of the ICU admissions included for external validation. MCU = medium care unit, UMC = University Medical Center.

RESULTS

The Leiden UMC data consisted of a total of 10,052 ICU admissions after excluding 2,198 admissions discharged to the MCU, 1,980 admissions with patients dying at the ICU, 1,056 admissions with a length of stay shorter than 12 hours, 616 admissions with patients transferred to the ICU of another hospital, and 220 admissions with patients receiving palliative care (**Fig. 1**). Approximately 0.8% of ICU admissions had a time difference of less than 12 hours and were also removed from the cohort. There were only minor differences in demographics (age, sex, and body mass index) between the original development site (Amsterdam UMC) and validation site (Leiden UMC) (**Table 1**). The average length of ICU stay was almost a day longer at the original development site compared with the validation site. The number of vasopressors or inotropes supplied were approximately the same. The percentage of readmissions within 7 days after discharge was slightly higher at the validation compared with development site (4.7% vs 4.3%), whereas the mortality was slightly higher at the development compared with the validation site (1.2% vs 1.0%). There were more planned surgical procedures at the validation site compared with the development site. A subset of features differed between the validation and development site in how

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

TABLE 1.

Descriptive Statistics for the Development Site (Amsterdam University Medical Center) and Validation Site (Leiden University Medical Center)

Electronic Health Record Data	Development Site (2004–2021)[a]	Validation Site (2011–2019)
Demographics		
Total observation, *N*	15,753	10,052
Age, mean (SD)	63.4 (14.8)	62.2 (14.0)
Sex (female), *n* (%)	4,865 (30.9)	3,423 (34.1)
Body mass index (kg/m^2), mean (SD)	26.4 (4.9)	26.4 (5.6)
Clinical information		
Length of stay (d), mean (SD)	3.1 (4.4)	2.3 (4.2)
Received vasopressors/inotropes, *n* (%)	10,807 (68.6)	7,119 (70.8)
Event		
Readmission (%)	599 (3.8)	476 (4.7)
Death (%)	205 (1.3)	103 (1.0)
Readmission or death (%)	751 (4.9)	577 (5.7)
National Intensive Care Evaluation Registry[b]	**Development Site (2013–2021)**	**Validation Site (2013–2019)**
Demographics		
Total observation *n*	11,473	16,686
Type of admission, *n* (%)		
Medical	5,612 (48.9)	6,738 (40.4)
Emergency surgery	1,684 (14.7)	2,025 (12.1)
Planned surgery	4,176 (36.4)	7,913 (47.4)
Other	1 (0.01)	10 (0.06)
Mortality risk, *n* (%)		
<30%	7,347 (64.0)	11,920 (71.4)
≥30% and <70%	1,794 (15.6)	1,453 (8.7)
≥70%	1,217 (10.6)	848 (5.1)
Missing	1,115 (9.7)	2,465 (14.8)

[a] Obtained from Thoral et al (8).

[b] Obtained from National Intensive Care Evaluation registry: https://www.stichting-nice.nl/.

often they were recorded (e.g., Glasgow Coma Scale) or their median value (e.g., troponin T) (for details see Table E1, http://links.lww.com/CCM/H261). Across the different validation cohorts (**Tables** E2 and **E3**, http://links.lww.com/CCM/H261), there was a slight decrease in length of stay over time and a decrease in readmissions and deaths over time (**Table E4**, http://links.lww.com/CCM/H261).

Among the 10,052 discharged patients from the ICU at the validation site, 577 patients (5.7%) experienced readmission or death within 7 days (**Table 2**). Length of ICU stay (before discharge) was notably higher for the patients who were readmitted or died compared with the patients with no such event (3.9 vs 2.2 d) (Table 2). There were fewer surgical compared with nonsurgical patients in the readmitted or dead group.

The original model had an AUC of 0.72 (95% CI 0.67–0.76, "External validation before retraining," **Table 3**) on validation data (2018–2019). The retrained models had improved discriminative performance with an AUC of 0.79 (95% CI 0.76–0.82) for temporal validation 1 and 0.79 (95% CI 0.76–0.83) for temporal

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

TABLE 2.
Descriptive Statistics by Outcome Event for the Validation Site (Leiden University Medical Center)

Descriptives	All	No Event	Readmission or Death
Demographics			
Total, *N* (%)	10,052 (100.0)	9,475 (94.3)	577 (5.7)
Age, mean (SD)	62.2 (14.0)	62.1 (14.0)	63.5 (13.9)
Sex (female), *n* (%)	3,423 (34.1)	3,181 (33.6)	242 (41.9)
Body mass index (kg/m^2), mean (SD)	26.4 (5.6)	26.5 (5.7)	25.9 (5.5)
Clinical information			
Length of stay[a], mean (SD)	2.3 (4.2)	2.2 (4.0)	3.9 (5.6)
Received vasopressors/inotropes, *n* (%)	7,119 (70.8)	6,677 (70.4)	442 (76.6)
ICU specialty top 5, *n* (%)			
Surgery	7,980 (79.4)	7,633 (80.6)	347 (60.1)
Internal medicine	588 (5.9)	543 (5.7)	45 (7.8)
Cardiology	327 (3.3)	295 (3.1)	32 (5.6)
Neurology	245 (2.4)	207 (2.2)	38 (6.6)
Gastroenterology	234 (2.3)	196 (2.1)	38 (6.6)

[a] Length of stay in days calculated before discharge.

TABLE 3.
Predictive Performance Before and After Retraining

Validation Step	Area Under the Receiver Operating Characteristic Curve	Calibration Intercept	Calibration Slope	Calibration Loss
External validation before retraining	0.72 (0.67–0.76)	−0.09 (−0.3 to 0.12)	1.0 (0.72–1.28)	0.01
Temporal validation 1	0.79 (0.76–0.82)	0.07 (−0.14 to 0.29)	0.95 (0.73–1.17)	0.01
Temporal validation 2	0.79 (0.76–0.83)	−0.0 (−0.22 to 0.21)	1.02 (0.78–1.26)	0.01
Validation after retraining	0.79 (0.75–0.82)	−0.03 (−0.24 to 0.19)	0.99 (0.77–1.21)	0.01

validation 2 on validation data (2018–2019). The final retrained model (“Validation after retraining,” Table 3) obtained a discrimination of 0.79 (95% CI 0.75–0.82) on validation data (2018–2019).

The models developed on data from the validation site showed calibration slopes below 1, indicating too extreme risk estimates, and intercepts above 0, indicating overall underestimation of risk (**Table E5**, http://links.lww.com/CCM/H261). After recalibration via isotonic regression, the slopes and intercepts were at 1 and 0, respectively, for all validation time points (Table 3). The calibration loss showed a minor decrease from 0.02 before recalibration to 0.01 after recalibration for all validation moments (Table 3) (Table E5, http://links.lww.com/CCM/H261). The decision curve for the model retrained at the validation site lies above the other strategies across almost the entire range of relevant probability thresholds, indicating a higher net benefit than the original model (**Fig. 2**). At a threshold of 5% for risk of readmission or death, the Leiden UMC model had a net benefit of 0.035: a net reduction of 3.5% points in patients who would have been readmitted or would have died. At a threshold of 10%, the model had a net benefit of 0.015, and at a threshold of 20%, the net benefit was 0.005. The original model had net

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

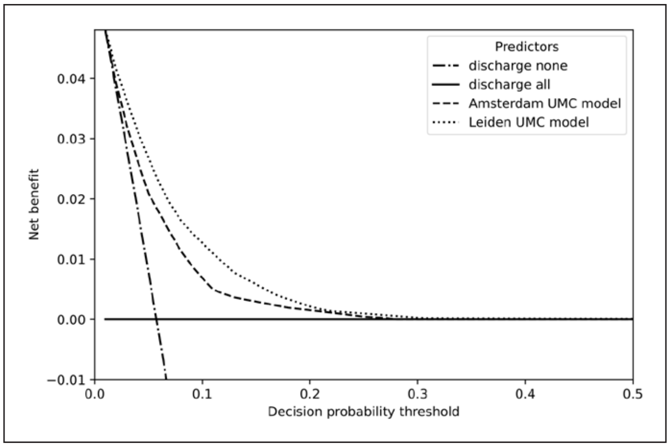

Figure 2. Decision curve analysis plotting net benefit (NB) for four discharging strategies across different threshold probabilities. Net benefit is expressed as the percentage reduction in readmission or death with respect to regular clinical practice (discharge all). The "discharge none" *line* corresponds to treating all patients as if they would be readmitted or dead within 7 d. This leads to many unnecessary prolonged ICU stays and only yields positive NB for very low threshold values (risk averseness). The "discharge all" *line* corresponds to discharging all patients as if they would not be readmitted or death within 7 d and hence corresponds to the current clinical practice strategy. The "Amsterdam University Medical Center (UMC) model" *line* corresponds to discharging according to the original model developed on Amsterdam UMC data and recalibrated for the Leiden UMC setting. The "Leiden UMC model" *line* corresponds to discharging according to the final retrained and recalibrated model developed on Leiden UMC data.

benefits of approximately 0.03 at a 5% threshold, 0.01 at a 10% threshold, and 0.002 at a 20% threshold, respectively.

Model discrimination was best for surgical and neurology patients (final model AUC of 0.79 [95% CI 0.75–0.84] and 0.84 [95% CI 0.70–0.97]) (**Tables E6–E10**, http://links.lww.com/CCM/H261) and worst for internal medicine and gastroenterology patients (final model AUC of 0.62 [95% CI 0.44–0.79] and 0.63 [95% CI 0.40–0.92]). Calibration is best for the surgical group. CIs are generally large due to small sample sizes.

DISCUSSION

This study illustrated the importance of local retraining for a specific setting to increase the applicability of a gradient boosted ML model. We confirmed the external validity of a promising ML model to predict readmission or death within 7 days after ICU discharge.

Our results indicate that retraining improved discrimination and calibration comparable to the original performance at a new site. The constant performance throughout the temporal validation indicated that there were no changes in our data (data drift) affecting performance over time. Retraining followed upon a process of extensive data preparation and harmonization (11). The need for retraining was underwritten by the DCA in which the final retrained model had a notably higher clinical usefulness than the original model. The level of heterogeneity between different sites directly relates to the generalizability of the original model to new sites. Heterogeneity between sites may for example be found in the patient populations, the healthcare context, and model specification, including the types of features included. In our case study, the model development and validation settings both treated similar patient populations and provided a similar level of care in comparable healthcare contexts (Table 1). There were some differences in the frequency and median of the features recorded, which may indicate differences in clinical protocols at the two centers (Table E1, http://links.lww.com/CCM/H261). Yet, there was considerable overlap in the feature sets used at development and validation sites. Despite these similarities, there was a clear drop in performance for the external validation in comparison to the original model results. Retraining led to markedly improved performance. We hypothesize that the drop in performance was caused by the differences in features and healthcare contexts, but this warrants further research. These results illustrate the importance of external validation and retraining, as generalizability was difficult to attain, and the exact differences between healthcare contexts driving the lack of generalizability may be hard to discern.

Retrained ML models also showed superior performance in other studies. For a ML model predicting hospital admission, the locally retrained models obtained AUCs of around 0.90 versus 0.60 for the external validations (30). For a study that aimed to identify pneumothorax patients with medical imaging, this was 0.90 versus 0.59 (31). These results underwrite that retraining and recalibration will likely be necessary when ML models are applied to a different setting. Yet, information on the external validity and necessity to recalibrate or retrain a ML model is currently not required to obtain CE-certification or FDA approval (12). Clinicians should be aware of this gap in the current regulatory requirements to prevent implementation of models with suboptimal or harmful performance.

Our study has the following implications. First, our results illustrate that generalizability cannot be taken for granted, even when the development and validation cohorts have strong similarities in terms of patient population, healthcare context, and model specification. A second implication is that when generalizability is poor, more extensive retraining may be required to improve performance at the new site, which requires substantial sample size (32). Poor generalizability of ML models from one local setting to another limits the scalability of these techniques (21). The potential of Pacmed Critical (33) may not come to fruition by nontransportable and highly tailored solutions that are labor-intensive to develop and maintain. Future research should analyze multisite datasets to explore heterogeneity in predictive relations as threats to developing generalizable models (34). Alternatively, up and coming techniques such as federated learning may prove useful in addressing the generalizability issue (35, 36). In situations where generalizable models cannot be attained, investment in data sharing infrastructure and in-hospital data science skills may help to facilitate the retraining and recalibration of these models locally. Last, the subgroup analysis showed diverging model performance across the different ICU specialties. Caution is needed when applying this model to "the ICU population" without detailed knowledge of the specific specialty case mix. Future model developments may focus on maximizing model performance across specialties by incorporating specialty specific variables and increasing the sample size of these subgroups. When applying ML models to clinical practice, clinicians should consider what case mix was considered during ML model development and whether the ML model can be safely and reliably applied to all patient groups and/or their case mix.

A strength of the current study was the use of a temporal validation design. Besides examining the effect of retraining on model performance, this design also allowed us to assess the model's sensitivity to shifts in data over time (15–18). A second strength was the complete and external EHR data for the validation after thorough data preparation in collaboration with a clinical domain expert (M.S.A). This led to a high-quality dataset. Another strength is the use of a comprehensive set of metrics to evaluate performance aspects, including calibration, discrimination, and clinical usefulness (37).

This study also had several limitations. First, the external validation was performed for one academic hospital (Leiden UMC) and one ML model (gradient boosting decision tree). Hence, our results cannot be directly extrapolated to other sites and ML techniques. Based on our findings, we anticipate that external validation and possibly retraining likely remain necessary for new implementation sites and ML techniques. Second, the models were developed with data preceding the COVID-19 pandemic to reflect "standard care." COVID-19 has drastically changed the composition of ICU patients and disrupted ICU care processes. Furthermore, COVID-19 may have changed the way critical care is practiced in non-COVID situations. Further validation is therefore needed for (post-)COVID-19 patients to use the model safely and reliably in this context. Finally, our definition of an ICU discharge excluded patients discharged to the MCU from the analysis, and those with a recorded admission of less than 12 hours. These exclusion criteria not only adhered to the strict focus on discharges from critical care to noncritical care settings but also limits the applicability of this model for clinical practice. Furthermore, the distinction between ICU and MCU may not always be clearcut. To address this limitation, future model developments should aim to incorporate ICU discharges to the MCU, and include all ICU admissions, irrespective of duration.

In conclusion, external validation can be essential to consider before clinical implementation of a ML model in a new setting. Techniques such as retraining may aid in improving model performance at a new site. Clinicians and decision-makers at the ICU should take this into account when considering applying new ML models to their local settings.

1 Department of Information Technology and Digital Innovation, Leiden University Medical Centre, Leiden, The Netherlands.

2 Department of Biomedical Informatics, Stanford Medicine, Stanford, CA.

3 Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands.

4 Pacmed, Stadhouderskade 55, Amsterdam, The Netherlands.

5 Institute of Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands.

6 Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam UMC, Amsterdam, The Netherlands.

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

7 Department of Intensive Care Medicine, Leiden University Medical Centre, Leiden, The Netherlands.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Drs. Fornasa and Cinà received funding from Pacmed; they disclosed work for hire; they disclosed the off-label product use of Pacmed Critical. Dr. Elbers disclosed that Amsterdam University Medical Center is entitled to royalties from jointly developed models. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: a.a.h.de_hond@lumc.nl

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024

REFERENCES

1. Faes L, Sim DA, van Smeden M, et al: Artificial intelligence and statistics: Just the old wine in new wineskins? *Front Digital Health* 2022; 4:1–5
2. Syed M, Syed S, Sexton K, et al: Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: Systematic review. *Informatics* 2021; 8:16
3. Shillan D, Sterne JAC, Champneys A, et al: Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. *Crit Care* 2019; 23:284
4. Moor M, Rieck B, Horn M, et al: Early prediction of sepsis in the ICU using machine learning: A systematic review. *Front Med* 2021; 8:1–18
5. Fleuren LM, Klausch TLT, Zwager CL, et al: Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020; 46:383–400
6. Lalmuanawma S, Hussain J, Chhakchhuak L: Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals* 2020; 139:110059
7. Alballa N, Al-Turaiki I: Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Inf Med Unlocked* 2021; 24:100564
8. Thoral PJ, Fornasa M, de Bruin DP, et al: Explainable machine learning on AmsterdamUMCdb for ICU discharge decision support: Uniting intensivists and data scientists. *Crit Care Explor* 2021; 3:e0529–e0529
9. Moons KG, Kengne AP, Grobbee DE, et al: Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; 98:691–698
10. Steyerberg EW, Moons KG, van der Windt DA, et al; PROGRESS Group: Prognosis research strategy (PROGRESS) 3: Prognostic model research. *PLoS Med* 2013; 10:e1001381
11. de Hond AAH, Leeuwenberg AM, Hooft L, et al: Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *NPJ Digital Med* 2022; 5:2
12. Wu E, Wu K, Daneshjou R, et al: How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021; 27:582–584
13. van de Sande D, van Genderen ME, Huiskens J, et al: Moving from bytes to bedside: A systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med* 2021; 47:750–760
14. Futoma J, Simons M, Panch T, et al: The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health* 2020; 2:e489–e492
15. Kelly CJ, Karthikesalingam A, Suleyman M, et al: Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; 17:195
16. McCradden MD, Joshi S, Anderson JA, et al: Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *J Am Med Inform Assoc* 2020; 27:2024–2027
17. Leslie D: Understanding Artificial Intelligende Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. London, UK, The Alan Turing Institute, 2019
18. Davis SE, Greevy RA, Jr, Lasko TA, et al: Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform* 2020; 112:103611
19. Wong A, Otles E, Donnelly JP, et al: External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Int Med* 2021; 181:1065–1070
20. Moons KG, Altman DG, Vergouwe Y, et al: Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ* 2009; 338:b606
21. Van Calster B, McLernon DJ, van Smeden M, et al; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative: Calibration: The Achilles heel of predictive analytics. *BMC Med* 2019; 17:230
22. van de Klundert N, Holman R, Dongelmans DA, et al: Data resource profile: The Dutch National Intensive Care Evaluation (NICE) registry of admissions to adult intensive care units. *Int J Epidemiol* 2015; 44:1850–1850h
23. de Hond AAH, van Calster B, Steyerberg EW: Commentary: Artificial intelligence and statistics: Just the old wine in new wineskins? *Front Digital Health* 2022; 4:1–3
24. Steyerberg EW, Vergouwe Y: Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur Heart J* 2014; 35:1925–1931
25. Caruana R, Niculescu-Mizil A: Data mining in metric space: An empirical analysis of supervised learning performance criteria. Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, Association for Computing Machinery, August 22–25, 2004, pp. 69–78
26. Zadrozny B, Elkan CP: Obtaining Calibrated Probability Estimates From Decision Trees and Naive Bayesian Classifiers. Williamstown, MA, ICML, June 28–July 1, 2001
27. Vickers AJ, Elkin EB: Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Making* 2006; 26:565–574
28. Vickers AJ, van Calster B, Steyerberg EW: A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; 3:18
29. Collins GS, Reitsma JB, Altman DG, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ* 2015; 350:g7594

30. Barak-Corren Y, Chaudhari P, Perniciaro J, et al: Prediction across healthcare settings: A case study in predicting emergency department disposition. *NPJ Digital Med* 2021; 4:169
31. Kitamura G, Deible C: Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. *Clin Imaging* 2020; 61:15–19
32. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, et al: Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat Med* 2004; 23:2567–2586
33. de Vos J, Visser LA, de Beer AA, et al: The potential cost-effectiveness of a machine learning tool that can prevent untimely intensive care unit discharge. *Value Health* 2022; 25:359–367
34. Wald Y, Feder A, Greenfeld D, et al: On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems* 2020; 34:2215–2227
35. Li T, Sahu AK, Talwalkar A, et al: Federated learning: Challenges, methods, and future directions. *IEEE Signal Process Mag* 2020; 37:50–60
36. Rodriguez-Barroso N, Stipcich G, Jimenez-Lopez D, et al: Federated learning and differential privacy: Software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy. *Inf Fusion* 2020; 64:270–292
37. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010; 21:128–138

Downloaded from http://journals.lww.com/ccmjournal by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/21/2024