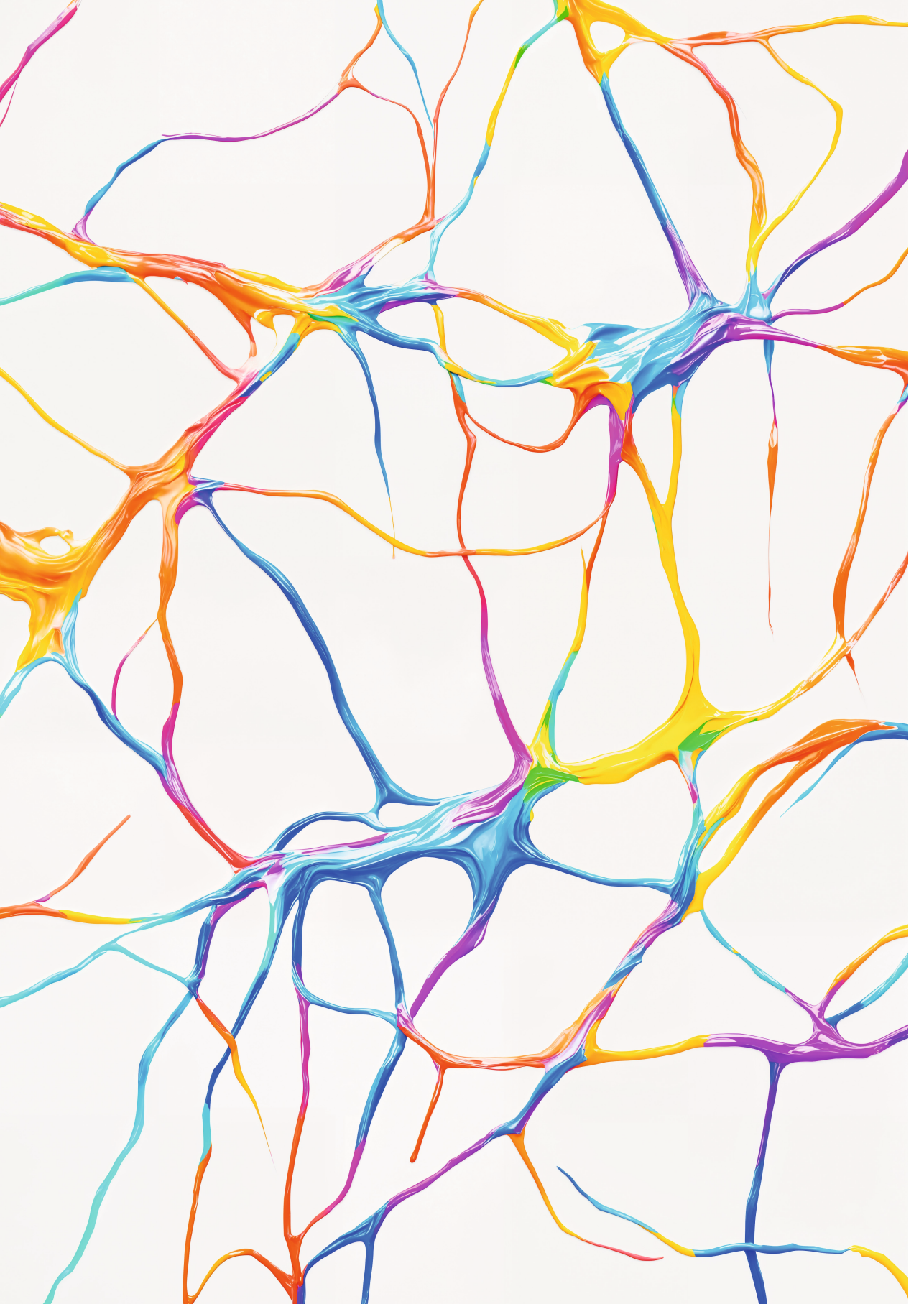
From code to clinic: theory and practice for artificial intelligence prediction algorithms

Hond, A.A.H. de



8

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

Anne A. H. de Hond, Ilse M. J. Kant, Mattia Fornasa, Giovanni Cinà, Paul W. G. Elbers, Patrick J. Thoral, M. Sesmu Arbous, and Ewout W. Steyerberg

8.1 ABSTRACT

8.1.1 Objective

Many machine learning (ML) models have been developed for application in the intensive care unit (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.

8.1.2 Design

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

8.1.3 Setting

Two ICUs in a tertiary care centers in the Netherlands.

8.1.4 Patients

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

8.1.5 Interventions

None.

8.1.6 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 (5.7%) of 10,052 ICU admissions 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.

8.1.7 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.

8.2 INTRODUCTION

There has been a rapid increase in the use of machine learning (ML) techniques for prediction modeling on routinely collected hospital data [1]. The intensive care unit (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 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 FDA approved ML models reported to have undergone multi-site assessment [12]. Moreover, 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.

8.3 MATERIALS AND METHODS

8.3.1 Patients

For the external validation, retraining, and recalibration of the Pacmed model we used EHR data from Leiden University Medical Center (Leiden UMC), a tertiary care center in the Netherlands. This data was 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, location VUmc (2017.212, date: May 2017, study title: 'Right Data, Right Now: Predicting ICU readmission rates').

8.3.2 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 that were discharged to the Medium Care Unit (MCU) as the intensity of monitoring on the ICU is comparable to that of the MCU, while on the ward the level of monitoring is much less intense. A planned surgical readmission was not considered as a readmission, but rather modelled 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, patients being transferred to the ICU of another hospital, dying at the ICU during the original admission, or receiving palliative care.

8.3.3 Machine learning model

Pacmed Critical is a 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 electronic health record (EHR) data collected between 2004 and 2021 from the Amsterdam University Medical Center, 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].

8.3.4 Retraining

The Pacmed Critical model was retrained on data from the Leiden UMC with the same pipeline and modelling 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, Wisconsin, USA, 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 (supplementary Table S1). 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 cvvh blood flow) and ICU specialty.

8.3.5 Validation design

We compared the descriptive statistics on patient demographics, clinical context and type of event (readmission or death within 7 days) 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 (supplementary Table S2). 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*", supplementary Table S2). 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 2011-2015 ("*Temporal validation 1*", supplementary Table S2). In the second round of retraining, data from 2011-2017 were used for retraining ("*Temporal validation 2*", supplementary Table S2). 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*", supplementary Table S2).

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 Area Under the Receiver Operating Characteristic curve (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 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. Confidence intervals were obtained through bootstrapping (1000 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 that were readmitted or died and were identified as such) penalized for false-positive classifications (patients that 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 TRIPOD statement [29].

8.3.6 Subgroup analysis

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

8.3.7 Software

All analyses were performed in Python 3.8.0. Code for the validation analysis is available online at https://git.lumc.nl/aahdehond/pacmed-validation.

8.4 RESULTS

The Leiden UMC data consisted of a total of 10,052 ICU admissions after excluding 2198 admissions discharged to the MCU, 1980 admissions with patients dying at the ICU, 1056 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 (Figure 8.1). 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 BMI) between the original development site (Amsterdam UMC) and validation site (Leiden UMC) (Table 8.1). The average

length of ICU stay was almost a day longer at the original development site compared to 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 to development site (4.7% vs 4.3%), whereas the mortality was slightly higher at the development compared to the validation site (1.2 vs 1.0%). There were more planned surgical procedures at the validation site compared to the development site. A subset of features differed between the validation and development site in how often they were recorded (e.g., glasgow coma scale) or their median value (e.g., troponin t) (for details see supplementary Table S1). Across the different validation cohorts (Table S2-S3) there was a slight decrease in length of stay over time and a decrease in readmissions and deaths over time (supplementary Table S4).

Table 8.1 Temporal validation design throughout retraining process

	2011	2012	2013	2014	2015	2016	2017	2018	2019
External validation before retraining									
Temporal validation 1									
Temporal validation 2									
Validation after retraining									

data for retraining data for validation data for retraining and validation

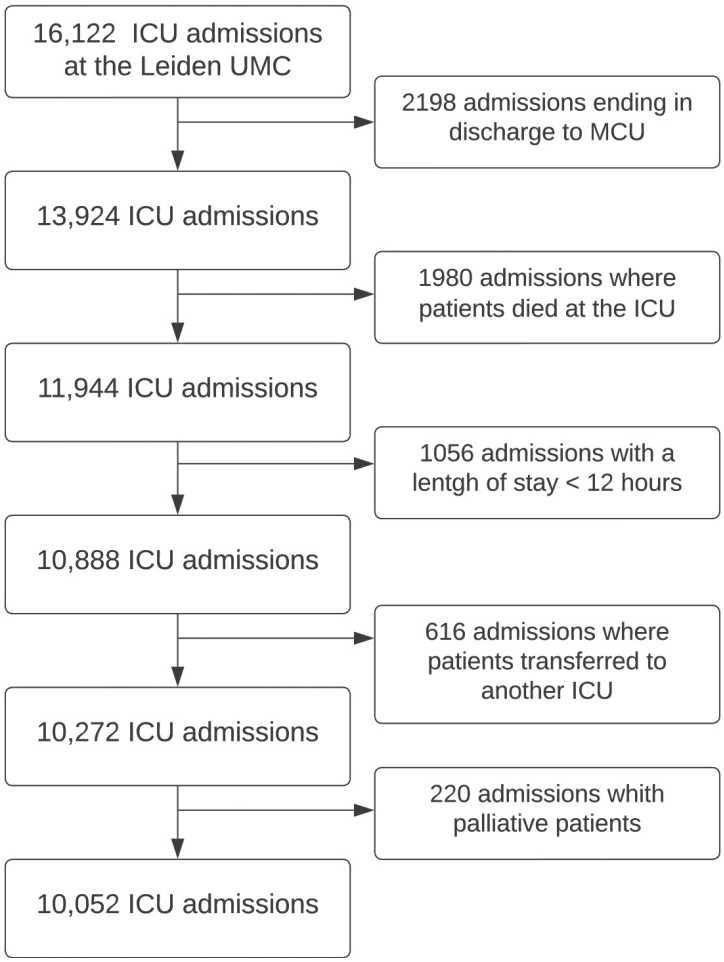


Figure 8.1 Flow chart of the ICU admissions included for external validation.

Among the 10,052 discharged patients from the ICU at the validation site, 577 (5.7%) patients experienced readmission or death within 7 days (Table 8.2). Length of ICU stay (before discharge) was notably higher for the patients who were readmitted or died compared to the patients with no such event (3.9 vs 2.2 days, Table 8.2). There were fewer surgical compared to non-surgical patients in the readmitted or dead group.

Table 8.2 Descriptive statistics by outcome event for the validation site (Leiden UMC)

	All	No event	Readmission or death
Demographics			
Total N (%)	10,052 (100.0)	9475 (94.3)	577 (5.7)
Age, mean (SD)	62.2 (14.0)	62.1 (14.0)	63.5 (13.9)
Sex (female), N (%)	3423 (34.1)	3181 (33.6)	242 (41.9)
BMI [kg/m^2], mean (SD)	26.4 (5.6)	26.5 (5.7)	25.9 (5.5)
Clinical information			
Length of stay*, mean (SD)	2.3 (4.2)	2.2 (4.0)	3.9 (5.6)
Received vasopressors/inotropes, N (%)	7119 (70.8)	6677 (70.4)	442 (76.6)
ICU specialty top 5 (%)			
Surgery	7980 (79.4)	7633 (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)

*Length of stay in days calculated before discharge.

The original model had an AUC of 0.72 (95% CI 0.67-0.76, "*External validation before retraining*", Table 8.3) 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 validation 2 on validation data (2018-2019). The final retrained model ("*Validation after retraining*", Table 8.3) obtained a discrimination of 0.79 (95% CI 0.75-0.82) on validation data (2018-2019).

Table 8.3 Predictive performance before and after retraining

Validation step	AUC	Calibration intercept	Calibration slope	Calibration loss
External validation before retraining	0.72 (0.67, 0.76)	-0.09 (-0.3, 0.12)	1.0 (0.72, 1.28)	0.01
Temporal validation 1	0.79 (0.76, 0.82)	0.07 (-0.14, 0.29)	0.95 (0.73, 1.17)	0.01
Temporal validation 2	0.79 (0.76, 0.83)	-0.0 (-0.22, 0.21)	1.02 (0.78, 1.26)	0.01
Validation after retraining	0.79 (0.75, 0.82)	-0.03 (-0.24, 0.19)	0.99 (0.77, 1.21)	0.01

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve

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 (supplementary Table S5). After recalibration via isotonic regression, the slopes and intercepts were at 1 and 0 respectively for all validation time points (Table 8.3). The calibration loss showed a minor decrease from 0.02 before recalibration to 0.01 after recalibration for all validation moments (Table 8.3 and supplementary Table S5). 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 (Figure 8.2). 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 percentage points in patients that 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 benefits of approximately 0.03 at a 5% threshold, 0.01 at a 10% threshold, and 0.002 at a 20% threshold respectively.

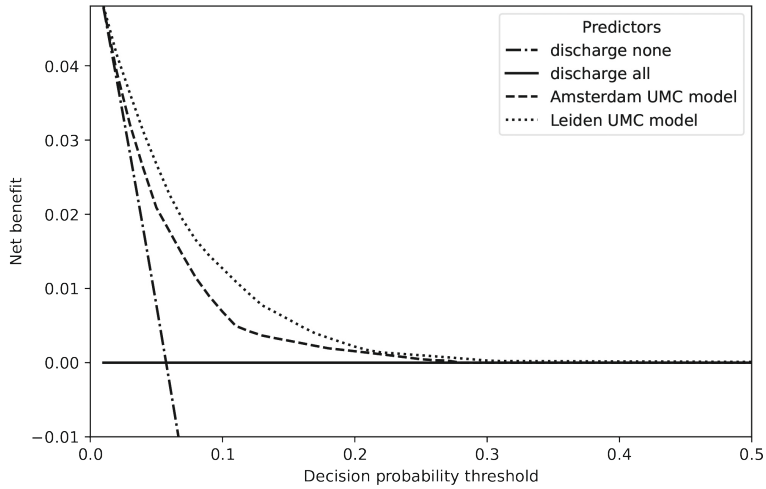


Figure 8.2 Decision curve analysis plotting Net Benefit for four discharging strategies across different threshold probabilities

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 days. 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 days and hence corresponds to the current clinical practice strategy. The 'Amsterdam 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.

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), supplementary Tables S6-S10), 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. Confidence intervals are generally large due to small sample sizes.

8.5 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 decision curve analysis 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 8.1). There were some differences in the frequency and median of the features recorded, which may indicate differences in clinical protocols at the two centers (supplementary Table S1). 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 non-transportable 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. Lastly, 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 parameters 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 (MA). 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. Moreover, 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 less than 12 hours. These exclusion criteria adhered to the strict focus on discharges from critical care to non-critical care settings, but also limits the applicability of this model for clinical practice. Moreover, 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 decisionmakers at the ICU should take this into account when considering applying new ML models to their local settings.

REFERENCES

1. Faes, L., D.A. Sim, M. van Smeden, U. Held, P.M. Bossuyt, and L.M. Bachmann, *Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins?* Frontiers in Digital Health, 2022. **4**.
2. Syed, M., S. Syed, K. Sexton, et al., *Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review.* Informatics, 2021. **8**(1): p. 16.
3. Shillan, D., J.A.C. Sterne, A. Champneys, and B. Gibbison, *Use of machine learning to analyse routinely collected intensive care unit data: a systematic review.* Crit Care, 2019. **23**(1): p. 284.
4. Moor, M., B. Rieck, M. Horn, C.R. Jutzeler, and K. Borgwardt, *Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review.* Frontiers in Medicine, 2021. **8**(348).
5. Fleuren, L.M., T.L.T. Klausch, C.L. Zwager, et al., *Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy.* Intensive Care Med, 2020. **46**(3): p. 383-400.
6. Lalmuanawma, S., J. Hussain, and L. Chhakchhuak, *Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review.* Chaos, Solitons & Fractals, 2020. **139**: p. 110059.
7. Alballa, N. and I. Al-Turaiki, *Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review.* Informatics in Medicine Unlocked, 2021. **24**: p. 100564.
8. Thoral, P.J., M. Fornasa, D.P. de Bruin, et al., *Explainable Machine Learning on AmsterdamUMCdb for ICU Discharge Decision Support: Uniting Intensivists and Data Scientists.* Critical care explorations, 2021. **3**(9): p. e0529-e0529.
9. Moons, K.G., A.P. Kengne, D.E. Grobbee, et al., *Risk prediction models: II. External validation, model updating, and impact assessment.* Heart, 2012. **98**(9): p. 691-8.
10. Steyerberg, E.W., K.G. Moons, D.A. van der Windt, et al., *Prognosis Research Strategy (PROGRESS) 3: prognostic model research.* PLoS Med, 2013. **10**(2): p. e1001381.
11. de Hond, A.A.H., A.M. Leeuwenberg, L. Hooft, et al., *Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review.* npj Digital Medicine, 2022. **5**(1): p. 2.
12. Wu, E., K. Wu, R. Daneshjou, D. Ouyang, D.E. Ho, and J. Zou, *How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals.* Nature Medicine, 2021. **27**(4): p. 582-584.
13. van de Sande, D., M.E. van Genderen, J. Huiskens, D. Gommers, and J. van Bommel, *Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit.* Intensive Care Medicine, 2021. **47**(7): p. 750-760.
14. Futoma, J., M. Simons, T. Panch, F. Doshi-Velez, and L.A. Celi, *The myth of generalisability in clinical research and machine learning in health care.* The Lancet Digital Health, 2020. **2**(9): p. e489-e492.

15. Kelly, C.J., A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, *Key challenges for delivering clinical impact with artificial intelligence.* BMC Med, 2019. **17**(1): p. 195.
16. McCradden, M.D., S. Joshi, J.A. Anderson, M. Mazwi, A. Goldenberg, and R. Zlotnik Shaul, *Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning.* J Am Med Inform Assoc, 2020. **27**(12): p. 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. The Alan Turing Institute., 2019.
18. Davis, S.E., R.A. Greevy, Jr., T.A. Lasko, C.G. Walsh, and M.E. Matheny, *Detection of calibration drift in clinical prediction models to inform model updating.* J Biomed Inform, 2020. **112**: p. 103611.
19. Wong, A., E. Otles, J.P. Donnelly, et al., *External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients.* JAMA Internal Medicine, 2021. **181**(8): p. 1065-1070.
20. Moons, K.G., D.G. Altman, Y. Vergouwe, and P. Royston, *Prognosis and prognostic research: application and impact of prognostic models in clinical practice.* Bmj, 2009. **338**: p. b606.
21. Van Calster, B., D.J. McLernon, M. van Smeden, et al., *Calibration: the Achilles heel of predictive analytics.* BMC Medicine, 2019. **17**(1): p. 230.
22. van de Klundert, N., R. Holman, D.A. Dongelmans, and N.F. de Keizer, *Data Resource Profile: the Dutch National Intensive Care Evaluation (NICE) Registry of Admissions to Adult Intensive Care Units.* International Journal of Epidemiology, 2015. **44**(6): p. 1850-1850h.
23. de Hond, A.A.H., B. van Calster, and E.W. Steyerberg, *Commentary: Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins?* Frontiers in Digital Health, 2022. **4**.
24. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* Eur Heart J, 2014. **35**(29): p. 1925-31.
25. Caruana, R. and A. Niculescu-Mizil, *Data mining in metric space: an empirical analysis of supervised learning performance criteria*, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, Association for Computing Machinery: Seattle, WA, USA. p. 69–78.
26. Zadrozny, B. and C.P. Elkan. *Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers*. in *ICML*. 2001.
27. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models.* Med Decis Making, 2006. **26**(6): p. 565-74.
28. Vickers, A.J., B. van Calster, and E.W. Steyerberg, *A simple, step-by-step guide to interpreting decision curve analysis.* Diagnostic and Prognostic Research, 2019. **3**(1): p. 18.
29. Collins, G.S., J.B. Reitsma, D.G. Altman, and K.G. Moons, *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* Bmj, 2015. **350**: p. g7594.

30. Barak-Corren, Y., P. Chaudhari, J. Perniciaro, M. Waltzman, A.M. Fine, and B.Y. Reis, *Prediction across healthcare settings: a case study in predicting emergency department disposition.* npj Digital Medicine, 2021. **4**(1): p. 169.
31. Kitamura, G. and C. Deible, *Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images.* Clin Imaging, 2020. **61**: p. 15-19.
32. Steyerberg, E.W., G.J.J.M. Borsboom, H.C. van Houwelingen, M.J.C. Eijkemans, and J.D.F. Habbema, *Validation and updating of predictive logistic regression models: a study on sample size and shrinkage.* Statistics in Medicine, 2004. **23**(16): p. 2567-2586.
33. de Vos, J., L.A. Visser, A.A. de Beer, et al., *The Potential Cost-Effectiveness of a Machine Learning Tool That Can Prevent Untimely Intensive Care Unit Discharge.* Value Health, 2022. **25**(3): p. 359-367.
34. Wald, Y., A. Feder, D. Greenfeld, and U. Shalit, *On calibration and out-of-domain generalization.* arXiv preprint arXiv:2102.10395, 2021.
35. Li, T., A.K. Sahu, A. Talwalkar, and V. Smith, *Federated Learning: Challenges, Methods, and Future Directions.* IEEE Signal Processing Magazine, 2020. **37**(3): p. 50-60.
36. Rodriguez-Barroso, N., G. Stipcich, D. Jimenez-Lopez, et al., *Federated Learning and Differential Privacy: Software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy.* Information Fusion, 2020. **64**: p. 270-292.
37. Steyerberg, E.W., A.J. Vickers, N.R. Cook, et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.

SUPPLEMENTARY MATERIAL

Table S1 Features available at Amsterdam UMC and Leiden UMC

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
activated partial thromboplastin time, mean value in the last 24h	2424 (15.4)	38.0 (34.4, 43.5)	497 (4.9)	32.53 (29.6, 37)
age	17 (0.1)	66.3 (56, 74.0)	0 (0)	65 (55, 72)
alanine transaminase, mean value in the last 24h	11870 (75.4)	32 (21, 62.5)	4563 (45.4)	22 (16, 37)
albumin, mean value in the last day where it was measured	1474 (9.4)	24.5 (21, 28.0)	1330 (13.2)	31 (27, 34)
alkaline phosphatase, mean value in the last 24h	12202 (77.5)	71 (52, 109)	8311 (82.7)	71 (54, 109)
amylase, mean value in the last 24h	13194 (83.8)	54.9 (35, 91)	9548 (95)	60.25 (36.4, 120.3)
arterial bicarbonate, mean value in the last 24h	3428 (21.8)	23.6 (22.1, 25.5)	4223 (42)	23.3 (20.7, 24.3)
arterial diastolic blood pressure, mean value in the last 24h	1128 (7.2)	61.5 (55.9, 68.1)	861 (8.6)	60.9 (55.6, 67.6)
arterial systolic blood pressure, mean value in the last 24h	1126 (7.1)	126.5 (114.5, 139.1)	861 (8.6)	119.9 (109.7, 132.2)
aspartate transaminase, mean value in the last 24h	11427 (72.5)	40 (27, 65)	4419 (44)	43 (31, 65)
base excess, mean value in the last 24h	966 (6.1)	1.4 (-0.3, 2.9)	1408 (14)	-2.5 (-4.4, 0.1)
total bilirubin, mean value in the last 24h	9664 (61.3)	8 (5.6, 12)	4611 (45.9)	10 (7, 14)
body mass index, mean value in the last 24h	15711 (99.7)	26.7 (23.7, 29.4)	3351 (33.3)	25.7 (23.2, 28.7)
chloride, mean value in the last 24h	6496 (41.2)	106.9 (104, 109.8)	3632 (36.1)	104.7 (102, 107.3)
c reactive protein, mean value in the last 24h	8452 (53.7)	45 (18.8, 99)	936 (9.3)	67.1 (37.5, 116)

Table S1 Features available at Amsterdam UMC and Leiden UMC (continued)

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
ionized calcium, mean value in the last 24h	6815 (43.3)	1.2 (1.1, 1.2)	508 (5.1)	1.2 (1.1, 1.2)
cardiac output, measured in last 24 h	0 (0)	5075 (32.2)	0 (0)	351 (3.5)
cardiac output, maximum measured in the last 24h	10678 (67.8)	5.9 (5, 6.9)	9701 (96.5)	6.6 (5.4, 8)
cardiac output, minimum measured in the last 24h	10678 (67.8)	4.5 (3.7, 5.4)	9701 (96.5)	4.2 (3.5, 5)
creatinine, last value recorded over whole ICU stay	299 (1.9)	77.0 (61.0, 97.0)	127 (1.3)	75 (60, 97)
creatinine urine, overall in the last 24h	15751 (100)	5.2 (4, 6.4)	9499 (94.5)	6.7 (3.5, 10.14)
fio2, mean value in the last 24h	7534 (47.8)	40.7 (40.0, 44.7)	4397 (43.7)	37.5 (32.6, 40.4)
fluid out urine, total dose in the last 24h	0 (0)	1555 (1040, 2206.0)	0 (0)	-1250.7 (-1840.8, -832.9)
gender, male (yes/no)	0 (0)	10624 (67.4)	0 (0)	6629 (65.9)
gender, female (yes/no)	0 (0)	4865 (30.9)	0 (0)	3423 (34.1)
glasgow coma scale, total maximum in the last 24h	7431 (47.2)	15 (15, 15)	8314 (82.7)	15 (14, 15)
glasgow coma scale, total minimum in the last 24h	7431 (47.2)	15 (11, 15)	8314 (82.7)	14 (10, 15)
glucose, mean value in the last 24h	2953 (18.7)	8.0 (6.9, 9.1)	335 (3.3)	7.9 (6.9, 9.0)
heart rate, mean value in the last 24h	121 (0.8)	80.2 (71.9, 89.5)	53 (0.5)	80.5 (72.5, 90.4)
height, last value recorded over whole ICU stay	3181 (20.2)	1.75 (1.68, 1.80)	2811 (28)	1.75 (1.68, 1.8)
hemoglobin, mean value in the last 24h	401 (2.5)	6.7 (6.1, 7.4)	338 (3.4)	6.5 (5.8, 7.3)
lactate dehydrogenase, mean value in the last 24h	11684 (74.2)	264 (202, 361)	4772 (47.5)	310 (236, 400)
length of stay, hours	0 (0)	30 (25.5, 72.6)	0 (0)	23.6 (20.0, 48.2)
leukocytes, mean value in the last day where it was measured	195 (1.2)	12.7 (9.9, 15.9)	121 (1.2)	12.1 (9.5, 14.9)
magnesium, mean value in the last 24h	3066 (19.5)	0.9 (0.8, 1.1)	3512 (34.9)	0.9 (0.8, 1.0)

Table S1 Features available at Amsterdam UMC and Leiden UMC (continued)

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
mcv, mean value in the last 24h	15160 (96.2)	89.0 (86.0, 93.0)	341 (3.4)	90 (87.3, 93.3)
derived minute volume, mean value in the last 24h	15607 (99.1)	7.5 (6.7, 8.3)	4964 (49.4)	7.9 (6.7, 9.5)
neutrophils, mean value in the last 24h	14986 (95.1)	9.9 (7.0, 13.0)	9941 (98.9)	9.3 (6.4, 12.6)
neutrophils percentage, mean value in the last 24h	14916 (94.7)	84.2 (77.8, 88.3)	9892 (98.4)	86.1 (78.7, 91.3)
non invasive diastolic blood pressure, mean value in the last 24h	12756 (81)	68.5 (60.5, 77.5)	6963 (69.3)	65.8 (57.1, 75.3)
non invasive mean blood pressure, mean value in the last 24h	12822 (81.4)	84.8 (75.2, 95.0)	6972 (69.4)	83.0 (73, 94.6)
non invasive systolic blood pressure, mean value in the last 24h	12755 (81)	120.4 (105.8, 135,2)	6963 (69.3)	122 (107, 139)
o2 flow, mean value in the last 24h	966 (6.1)	4.3 (2.8, 5)	1136 (11.3)	2.8 (2, 3.75)
o2 saturation, mean value in the last 24h	144 (0.9)	97.3 (96, 98.4)	43 (0.4)	96.3 (94.9, 97.5)
pao2 over fio2, mean value in the last 24h	7881 (50)	258.0 (203.2, 328.4)	4096 (40.7)	284.5 (229.3, 353.3)
arterial pco2, mean value in the last 24h	949 (6)	39.3 (36.4, 42.3)	766 (7.6)	38.4 (35.3, 41.4)
peep, mean value in the last 24h	7607 (48.3)	5 (5, 7.0)	4632 (46.1)	5 (5, 5)
pressure above peep, mean value in the last 24h	15468 (98.2)	8.5 (7.0, 10.2)	5357 (53.3)	16.3 (14.8, 19)
arterial ph, mean value in the last 24h	963 (6.1)	7.4 (7.4, 7.4)	756 (7.5)	7.4 (7.3, 7.4)
phosphate, mean value in the last 24h	3084 (19.6)	1.0 (0.8, 1.2)	4411 (43.9)	1.1 (0.9, 1.3)
arterial po2, mean value in the last 24h	1004 (6.4)	93.6 (80.2, 112.4)	765 (7.6)	89.4 (77.5, 108.0)
potassium, mean value in the last 24h	435 (2.8)	4.1 (3.9, 4.4)	486 (4.8)	4.2 (3.9, 4.5)

Table S1 Features available at Amsterdam UMC and Leiden UMC (continued)

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
rapid shallow breathing index, mean value in the last 24h	10018 (63.6)	30.7 (22.7, 42.6)	5373 (53.5)	27.2 (22.4, 34.0)
respiratory rate measured, mean value in the last 24h	1227 (7.8)	17 (14.4, 20)	296 (2.9)	16.7 (14.8, 19.3)
respiratory rate measured ventilator, mean value in the last 24h	15467 (98.2)	17.8 (16.2, 19.9)	4767 (47.4)	14.9 (13.2, 17)
richmond agitation sedation scale score, maximum in the last 24h	12156 (77.2)	0 (0, 0)	5194 (51.7)	0 (0, 0)
richmond agitation sedation scale score, minimum in the last 24h	12156 (77.2)	0 (-1, 0)	5194 (51.7)	-1 (-4, 0)
sodium, mean value in the last 24h	424 (2.7)	138.5 (136.4, 140.7)	490 (4.9)	137.8 (135.8, 139.8)
thrombocytes, mean value in the last 24h	543 (3.4)	178 (135, 239.5)	372 (3.7)	175 (134.7, 225.3)
tidal volume, mean value in the last 24h	7754 (49.2)	481.5 (425.6, 540.1)	4763 (47.4)	550 (472.7, 630)
troponin t, measured in the last 24h	0 (0)	3091 (19.6)	0 (0)	4909 (48.8)
troponin t, maximum in the last 24h	12662 (80.4)	245 (70, 467.0)	5143 (51.2)	519 (305, 914)
ventilatory ratio, mean value in the last 24h	15752 (100)	1.1 (1.1, 1.1)	6744 (67.1)	1.3 (1.1, 1.6)
weight, last value recorded over whole ICU stay	2821 (17.9)	80 (70, 90)	159 (1.6)	80 (70, 90)
cough stimulus mode, normal cough reflex most frequently recorded in the last 24h (yes/no)[a]	0 (0)	1046 (6.6)	-	-
cough stimulus mode, reasonable cough reflex most frequently recorded in the last 24h (yes/no)[a]	0 (0)	566 (3.6)	-	-

Table S1 Features available at Amsterdam UMC and Leiden UMC (continued)

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
cough stimulus mode, weak cough reflex most frequently recorded in the last 24h (yes/no)[a]	0 (0)	212 (1.3)	-	-
blood temperature, mean value in the last 24h[b]	10246 (65)	36.7 (36.5, 37)	-	-
internal temperature, maximum in the last 24h[b]	12420 (78.8)	37.1 (36.5, 37.6)	-	-
internal temperature, minimum in the last 24h[b]	12420 (78.8)	36.5 (36, 37.0)	-	-
ureum, mean value in the last 24h[c]	7256 (46.1)	6.6 (4.9, 9.4)	-	-
ureum over creatinine, mean value in the last 24h[d]	7286 (46.3)	0.1 (0.1, 0.1)	-	-
base excess mixed venous, mean value in the last 24h[e]	-	-	9773 (97.2)	-4.6 (-6, -2.85)
creatinine kinase, last value recorded over whole ICU stay[f]	-	-	2419 (24.1)	532 (318, 857)
cvvh blood flow, last value recorded over whole ICU stay[e]	-	-	9796 (97.5)	180 (160, 200)
gamma glutamyl transferase, mean value in the last 24h[g]	-	-	8267 (82.2)	47 (22, 117)
Icu specialty, cardio surgery (yes/no)[h]	-	-	0 (0)	6147 (61.2)
Icu specialty, cardiology (yes/no)[h]	-	-	0 (0)	327 (3.3)
Icu specialty, endocrinology (yes/no)[h]	-	-	0 (0)	13 (0.1)
Icu specialty, ent (yes/no)[h]	-	-	0 (0)	139 (1.4)
Icu specialty, gastroenterology (yes/no)[h]	-	-	0 (0)	234 (2.3)
Icu specialty, gynecology (yes/no)[h]	-	-	0 (0)	46 (0.5)

Table S1 Features available at Amsterdam UMC and Leiden UMC (continued)

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
Icu specialty, hematology (yes/no)[h]	-	-	0 (0)	103 (1)
Icu specialty, infectious_diseases (yes/no)[h]	-	-	0 (0)	32 (0.3)
Icu specialty, intensive_care (yes/no)[h]	-	-	0 (0)	25 (0.2)
Icu specialty, internal_medicine (yes/no)[h]	-	-	0 (0)	588 (5.8)
Icu specialty, nephrology (yes/no)[h]	-	-	0 (0)	82 (0.8)
Icu specialty, neurology (yes/no)[h]	-	-	0 (0)	245 (2.4)
Icu specialty, neurosurgery (yes/no)[h]	-	-	0 (0)	748 (7.4)
Icu specialty, obstetrics (yes/no)[h]	-	-	0 (0)	24 (0.2)
Icu specialty, oncology (yes/no)[h]	-	-	0 (0)	13 (0.1)
Icu specialty, other (yes/no)[h]	-	-	0 (0)	12 (0.1)
Icu specialty, pediatrics (yes/no)[h]	-	-	0 (0)	2 (0)
Icu specialty, pulmonary (yes/no)[h]	-	-	0 (0)	160 (1.6)
Icu specialty, surgery (yes/no)[h]	-	-	0 (0)	1053 (10.5)
Icu specialty, urology (yes/no)[h]	-	-	0 (0)	27 (0.3)
arterial lactate, last value recorded over whole ICU stay[e]	-	-	356 (3.5)	1.3 (1, 1.7)
venous lactate, last value recorded over whole ICU stay[e]	-	-	4172 (41.5)	1.3 (1, 1.8)
venous pco2, mean value in the last 24h[i]	-	-	5541 (55.1)	43.1 (39.8, 46.6)
venous ph, mean value in the last 24h[i]	-	-	5540 (55.1)	7.3 (7.3, 7.4)
venous po2, mean value in the last 24h[i]	-	-	5543 (55.1)	37.3 (33.8, 41.3)

Table S1 Features available at Amsterdam UMC and Leiden UMC (continued)

Feature	Included in the Amsterdam UMC model		Included in the Leiden UMC model	
	Missing, N (%)	Median (IQR) / N (%)	Missing, N (%)	Median (IQR) / N (%)
respiratory rate difference, mean value in the last 24h[j]	-	-	5302 (52.7)	0 (-0.5, 1.2)
respiratory rate set, mean value in the last 24h[j]	-	-	5294 (52.7)	14 (12.0, 15.8)
unspecified temperature, mean value in the last 24h[b]	-	-	1025 (10.2)	36.9 (36.5, 37.3)

Features were included for model development at each individual location 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.

[a]Cough stimulus mode was not recorded for all years of interest at the LUMC and therefore excluded.

[b]Blood temperature was measured similarly at the two centers but recorded under different names and was split up in minimum and maximum for the AUMC (internal temperature and blood temperature at the AUMC and unspecified temperature at the LUMC).

[c]Ureum was missing for the LUMC, but overall kidney function was measured at both centers by creatinine urine.

[d]Ureum over creatinine was not systematically collected at the LUMC and therefore excluded.

[e]Base excess mixed venous, cvvh blood flow, and arterial lactate, and venous lactate were all added for the LUMC to capture additional information regarding severity monitoring. These variables were not available for all years of interest at the time of model development at the AUMC and therefore excluded.

[f]Creatinine kinase was not available for all years of interest at the time of model development at the AUMC and therefore excluded.

[g]Gamma glutamyl transferase was not included at the AUMC, but both centers included alkaline phosphatase as another indicator for the effect of drug toxicity on liver function and/or intrahepatic bile duct function.

[h]ICU specialty was not included at the AUMC and included at the LUMC to be reflective of the differences in clinical course across different specialties.

[i]LUMC chose to model the difference between venous and arterial blood saturation in addition to arterial blood saturation which was captured at both centers.

[j]LUMC chose to model respiratory rate difference and set in addition to respiratory rate which was captured at both centers.

Table S2 Temporal validation design throughout retraining process

	2011	2012	2013	2014	2015	2016	2017	2018	2019
External validation before retraining									
Temporal validation 1									
Temporal validation 2									
Validation after retraining									

data for retraining | data for validation | data for retraining and validation

Table S3 Sample size across years at the validation site (Leiden UMC)

	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Sample size	845	1119	1238	1302	1298	1201	1211	996	842	10052

Table S4 Descriptive statistics for the temporal validation cohorts

	All (2011-2019)	Cohort 2018-2019
Demographics		
Total observation N (%)	10052 (100.0)	1838 (18.3)
Age, mean (SD)	62.2 (14.0)	62.1 (13.6)
Sex (female), N (%)	3423 (34.1)	617 (33.6)
BMI [kg/m^2], mean (SD)	26.4 (5.6)	26.7 (5.7)
Clinical information		
Lenth of stay (days), mean (SD)	2.3 (4.2)	2.2 (4.6)
Received vasopressors/inotropes, N (%)	7119 (70.8)	1443 (78.5)
ICU specialty top 5 (%)		
Surgery	7980 (79.4)	1505 (81.9)
Internal medicine	588 (5.9)	85 (4.6)
Cardiology	327 (3.3)	69 (3.8)
Neurology	245 (2.4)	53 (2.9)
Gastroenterology	234 (2.3)	41 (2.2)
Event		
Readmission (%)	467 (4.7)	78 (4.2)
Death (%)	103 (1.0)	14 (0.8)
Readmission or death (%)	577 (5.7)	92 (5.0)

Table S5 Predictive performance for different validation cohorts before recalibration by isotonic regression

	AUC	Calibration intercept	Calibration slope	Calibration loss
External validation before retraining	0.71 (0.66, 0.76)	0.61 (0.4, 0.82)	1.05 (0.75, 1.35)	0.02
Temporal validation 1	0.78 (0.75, 0.82)	0.27 (0.05, 0.49)	0.76 (0.59, 0.94)	0.02
Temporal validation 2	0.78 (0.74, 0.82)	0.38 (0.15, 0.6)	0.76 (0.59, 0.93)	0.02
Validation after retraining	0.78 (0.75, 0.82)	0.41 (0.18, 0.63)	0.77 (0.6, 0.94)	0.02

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve

Table S6 Predictive performance for different validation cohorts within the surgical ICU specialty

	AUC	Calibration intercept	Calibration slope	Calibration loss
External validation before retraining	0.72 (0.66, 0.78)	-0.3 (-0.57, -0.03)	1.01 (0.66, 1.36)	0.02
Temporal validation 1	0.8 (0.76, 0.84)	0.08 (-0.19, 0.35)	0.95 (0.69, 1.21)	0.01
Temporal validation 2	0.8 (0.76, 0.84)	0.02 (-0.25, 0.29)	0.99 (0.71, 1.27)	0.01
Validation after retraining	0.79 (0.75, 0.84)	0.01 (-0.26, 0.28)	0.98 (0.72, 1.23)	0.01

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve

Table S7 Predictive performance for different validation cohorts within the internal medicine ICU specialty

	AUC	Calibration intercept	Calibration slope	Calibration loss
External validation before retraining	0.7 (0.53, 0.87)	0.12 (-0.68, 0.91)	0.93 (-0.08, 1.95)	0
Temporal validation 1	0.55 (0.45, 0.66)	-0.17 (-0.96, 0.61)	0.83 (-0.58, 2.23)	0
Temporal validation 2	0.6 (0.45, 0.76)	-0.18 (-0.96, 0.6)	0.5 (-0.92, 1.92)	0
Validation after retraining	0.62 (0.44, 0.79)	-0.27 (-1.06, 0.51)	0.69 (-0.66, 2.04)	0

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve

Table S8 Predictive performance for different validation cohorts within the cardiology ICU specialty

	AUC	Calibration intercept	Calibration slope	Calibration loss
External validation before retraining	0.62 (0.49, 0.75)	1.09 (0.42, 1.76)	0.68 (-0.22, 1.59)	0
Temporal validation 1	0.68 (0.54, 0.82)	0.64 (-0.04, 1.32)	0.89 (0.08, 1.7)	0
Temporal validation 2	0.7 (0.53, 0.84)	0.55 (-0.12, 1.22)	1.14 (0.13, 2.15)	0
Validation after retraining	0.69 (0.52, 0.83)	0.5 (-0.17, 1.18)	0.97 (0.06, 1.89)	0

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve

Table S9 Predictive performance for different validation cohorts within the neurology ICU specialty

	AUC	Calibration intercept	Calibration slope	Cali. loss
External validation before retraining	0.84 (0.74, 0.93)	1.04 (0.24, 1.83)	1.45 (0.4, 2.51)	0
Temporal validation 1	0.81 (0.69, 0.92)	0.64 (-0.14, 1.42)	3.14 (0.1, 6.18)	0
Temporal validation 2	0.87 (0.76, 0.97)	0.27 (-0.51, 1.06)	2.77 (0.71, 4.84)	0
Validation after retraining	0.84 (0.7, 0.97)	0.25 (-0.53, 1.02)	2.36 (0.65, 4.07)	0

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve

Table S10 Predictive performance for different validation cohorts within the gastroenterology ICU specialty

	AUC	Calibration intercept	Calibration slope	Cali. loss
External validation before retraining	0.41 (0.19, 0.71)	-0.11 (-1.3, 1.09)	-0.14 (-1.77, 1.48)	0
Temporal validation 1	0.59 (0.39, 0.78)	-0.49 (-1.69, 0.71)	0.51 (-1.21, 2.23)	0
Temporal validation 2	0.72 (0.54, 0.85)	-0.49 (-1.67, 0.7)	2.08 (-1.46, 5.61)	0
Validation after retraining	0.63 (0.4, 0.92)	-0.57 (-1.76, 0.62)	0.95 (-1.09, 2.99)	0

Abbreviations: *AUC* Area Under the Receiver Operating Characteristics Curve