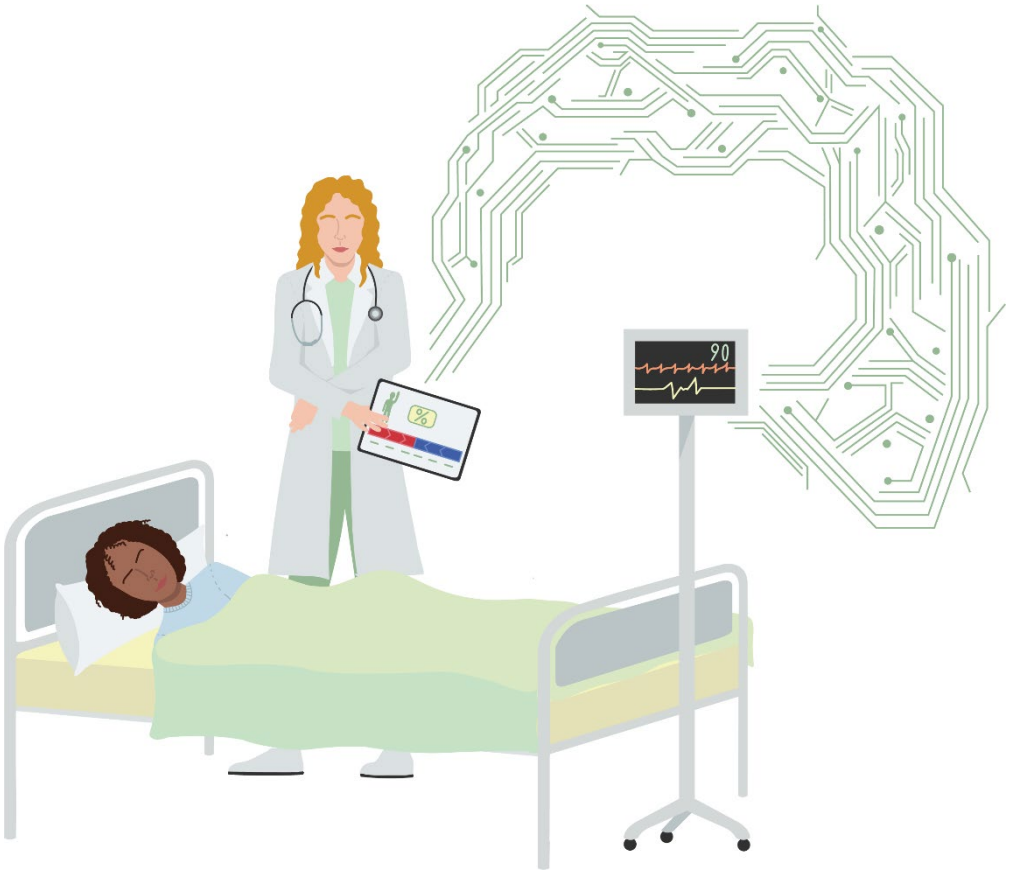
Explainable AI for Predicting ICU Readmission

S.L. van der Meijden



Master thesis Technical Medicine
Leiden University Medical Center - Department of Intensive Care Medicine
Leiden Institute of Advanced Computer Science - Explanatory Data Analysis group
March 2020 – March 2021

Universiteit Leiden
The Netherlands

TU Delft Delft University of Technology

Erasmus University Rotterdam

Cover illustration designed by Leonoor Kuiperbak

Predicting Intensive Care Unit Readmission: *Performance and Explainability of Machine Learning Algorithms*

Siri Lise van der Meijden
Student number: 4306961
March 25, 2021

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University - Delft University of Technology - Erasmus University Rotterdam

Master thesis project - TM30004 – 35 ECTS
Track: Sensing and stimulation
Faculty of Mechanical, Maritime and Materials Engineering (3mE), TU Delft
March 2020 – March 2021

Thesis committee members and supervisors

Prof. dr. ir. J. (Jaap) Harlaar	TU Delft	Chair
Dr. M.S. (Sesmu) Arbous	LUMC & TU Delft	Medical supervisor
Dr. M. (Matthijs) van Leeuwen	Leiden University	Technical supervisor
E. (Esmee) Stoop	LUMC	Daily supervisor
Dr. E.G. (Bert) Mik	Erasmus MC	External committee member

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Universiteit Leiden The Netherlands
TU Delft Delft University of Technology
Erasmus University Rotterdam

Preface

During my studies, I completed four clinical research internships at the intensive care unit (ICU), the department of transplantation surgery, the department of neurosurgery, and the neonatal intensive care unit. In my opinion, the ICU is the most exciting, inspiring, and challenging environment for a Technical Physician (to be). This is due to the complex patient population, advanced monitoring and therapeutic techniques, and the multidisciplinary team of physicians which makes every day different and educational. Therefore, I was eager to return for my master thesis to the ICU.

During my internships, I gained experience in the field of signal processing, for instance of ICU patients being mechanically ventilated. During master thesis, I wanted to take my knowledge on signal processing and data analysis one step further. Furthermore, it fascinated me that the tremendous amount of patient data monitored and stored at the ICU were not used to support the ICU physicians in difficult decision making and to reduce their high workload. Therefore, I aspired to do my master thesis on the use of clinically valuable Artificial Intelligence (AI) prediction models built on this tremendous amount of ICU patient data.

Without prior knowledge of AI, I enthusiastically started my project in March 2020 on the prediction of ICU readmission for discharge decision support. There were some challenges and difficulties last year caused by the COVID-19 pandemic. For instance, my cancelled internship on the ICU in Nagasaki, Japan, was one of the first changes in my plan with many to follow. But despite some setbacks, I had an amazing year and developed myself clinically on the ICU, and technically in data science as I learned to develop machine learning models on a large ICU database. Hopefully, my performed research the last year contributed to creating clinical value using AI-based decision support, and I am extremely thankful for the support and enthusiasm from my three main supervisors: Sesmu Arbous, Matthijs van Leeuwen, and Esmee Stoop.

In my opinion, the Technical Physician could play an important role in making the step from AI model development to implementing AI-based tools in clinical practice. I noticed the benefit of speaking the physicians' language and my clinical ICU experiences during all phases of the project. Unfortunately, my thesis ends at the start of an exciting time for the ICU of the Leiden University Medical Centre. The coming months, one of the first CE-certified AI-based decision support tools will be validated and implemented. During this prospective validation, I see an important role for the Technical Physician in training the physicians and conducting prospective (randomized) trials. For Technical Physicians working in direct patient care, adoption of AI tools will be easier due to their technical knowledge, which could also support the adoption of AI tools by their physician colleagues, speeding up the process of implementation.

I started this project without in-depth knowledge on AI, but I ended it knowing what I want to do the coming years: making real impact as a Technical Physician on the use and implementation of AI-based decision support tools.

Siri van de Meijden

Delft, 10/3/2021

Summary

Intensive Care Unit (ICU) readmission is a serious adverse event associated with high mortality rates and costs. Prediction of ICU readmission could support physicians in their decision to discharge patients from the ICU to lower care wards. Due to increasing ICU data availability, Artificial Intelligence (AI) models in the form of machine learning (ML) algorithms can be used to build high-performing decision support tools. To have impact on patient outcomes, these decision support tools should have high discriminative performance and should be explainable to the ICU physician. The goal of this thesis was to compare several types of ML models on predictive performance and explainability for the prediction of ICU readmission for discharge decision support. The scientific paper that aims to answer this question can be found in Part III of this thesis. In a broader perspective, we proposed a framework for the development and implementation of clinically valuable AI-based decision support.

First, a systematic review was conducted to examine current literature on ML prediction models for ICU readmission (Part I). We concluded that previously developed models reported inappropriate performance metrics and were not implemented in clinical practice. Furthermore, previous work did not compare explainable outcomes in terms of patient factors contributing to the risk of readmission between models. Secondly, we conducted a questionnaire among ICU physicians to investigate current discharge practices and their attitude towards the use of AI tools in their work processes (Part II). Although not all physicians agreed that the decision to discharge ICU patients is complex, most of them do believe in the clinical value of an AI-based discharge decision support tool. Thirdly, we developed several prediction models for ICU readmission and compared them on discriminative performance, calibration properties, and explainability (Part III). We concluded that advanced ML models did not outperform logistic regression in terms of discriminative performance and calibration properties. However, the explanations of XGBoost, a state-of-the-art ML algorithm, were more in line with the ICU physician's clinical reasoning compared to logistic regression and neural networks. Lastly, we designed a study protocol to prospectively evaluate the predictive performance of Pacmed Critical, a CE-certified AI-based discharge decision support tool, and that of the ICU physician (Part IV).

This thesis contributed to making the step from developing high-performing prediction models to clinical adoption of an ICU discharge decision support system. Due to small differences in discriminative power and calibration properties between models, the model best explainable to the physician and most in line with clinical reasoning should be chosen for decision support. Before final implementation, impact on patient outcomes and costs will need to be studied in prospective trials.

Table of Contents

** Part III covers the main scientific paper of this thesis.*

List of abbreviations

ABP	Arterial blood pressure
AI	Artificial Intelligence
AIOS	Physician in training (resident)
AKI	Acute Kidney Injury
ALAT	Alanine aminotransferase
ANIOS	Physician not in training (resident)
APACHE	Acute Physiology And Chronic Health Evaluation Score
APTT	Activated partial thromboplastin time
ASAT	Aspartate aminotransferase
AUC	Area under the curve of the receiver operating characteristic
AUCPR	Area under the precision recall curve
BE	Base Excess
BSE	Erythrocyte sedimentation rate
CAIRELab	Clinical Artificial Intelligence Implementation and Research Lab
CHARMS	CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies
CK	Creatinine Kinase
CNN	Convolutional Neural Network
CRF	Conditional random fields
CRP	C-Reactive Protein
CTC	Cardiothoracic surgical
CV	Cross-validation
CVP	Central venous pressure
DL	Deep Learning
ECMO	Extra Corporal Membrane Oxygenation
EHR	Electronic Health Record
FFNN	Feed-Forward Neural Network
Gamma GT	Gamma-glutamyl transferase
GB	Gradient Boosting Machines
GCS	Glasgow Coma Scale
GRU	Gated recurrent units
HR	Heart rate
IC	Intensive Care
ICD	International Statistical Classification of Diseases and Related Health Problems
ICU	Intensive Care Unit
IQR	Inter quartile range
KNN	K-Nearest Neighbour
LDH	Lactate dehydrogenase
LOS	Length of stay
LR	Logistic regression
LSTM	Long short-term memory
LUMC	Leiden University Medical Centre
MC	Medium care
MCC	Matthews correlation coefficient

MCU	Medium care unit
MCV	Mean corpuscular volume
METC	Medisch Ethische Toetsingscommissie (Medical Ethical Committee)
MEWS	Modified Early Warning Score
MIMIC-III	Medical Information Mart for Intensive Care
ML	Machine learning
MLP	Multi-layer perceptron
NIBP	Non-invasive blood pressure
NN	Neural network
PCA	Principal component analysis
PCO2	Carbon dioxide pressure
PDMS	Patient data management system
PEEP	Positive end-expiratory pressure
PPV	Positive predictive value
PT	Prothrombin time
RCT	Randomized controlled trial
RF	Random forest
RNN	Recurrent neural network
RR	Respiratory rate
SAPS	Simplified Acute Physiology Score
SHAP	Shapley Additive exPlanations
SHS	Same hospital stay
SVM	Support vector machines
SWIFT	Stability and Workload Index for Transfer
TRIPOD	Transparent Reporting of a multivariable Prediction model for Individual prognosis or Diagnosis
XGB	eXtreme Gradient Boosting
XGBoost	eXtreme Gradient Boosting

Introduction

The Intensive Care Unit (ICU) ICU has limited bed availability and expensive resources, resulting in the need to discharge patients from the ICU as soon as safely possible. In determining ICU discharge timing, a trade-off exists. In case patients are prematurely discharged to lower-care wards, deterioration of their condition may be noticed late due to limited monitoring facilities, which may result in ICU readmission. Readmission to the ICU is a serious adverse event, correlated with increased mortality rates, length of stay, and costs[1,2]. On the other hand, delaying ICU discharge impacts bed availability, affecting patients in need of intensive care, as well as costs.

The decision to discharge patients to lower care wards is one of many high stake, difficult, and often quick decisions ICU physicians need to make[3]. Currently, this decision is made by the responsible physicians and nurses based on their clinical knowledge and local hospital policies. To support the physician in determining optimal timing for ICU discharge, a decision support tool accurately predicting a patient's readmission risk could provide value. The ultimate goal of such a tool would be to reduce average length of stay, readmission rates, and costs. The prediction of ICU readmission can be reported as a risk score, similar to the Acute Physiology And Chronic Health Evaluation Score (APACHE, predicting mortality risk)[4], the Simplified Acute Physiology Score (SAPS, predicting mortality risk)[5], and the Stability and Workload Index for Transfer (SWIFT, predicting readmission risk)[6]. These scores are based on classical statistical models (e.g., logistic regression) that allow the physician to obtain insight in the contributing factors to the predicted outcome, making them highly explainable and interpretable. However, these risk scores often insufficiently generalize beyond the ICU population they were developed on[7], indicating the need for more advanced prediction models.

Due to the complexity and heterogeneity of the ICU patient population, personalized decision support tools based on Artificial Intelligence (AI) prediction models might be superior to classical risk scores. AI, in the form of Machine Learning (ML) and Deep Learning (DL) algorithms, mimics reasoning or decision-making based on real-world (patient) data[8] (Figure 1). These models can discover non-linear relations between the numerous recorded ICU parameters and relevant patient outcomes. Due to the ability of AI to handle high dimensional patient data, individualized predictions can be performed for the usage of accurate decision support tools[9]. Despite their excellent *predictive performance,* ML and DL algorithms are often referred to as 'black-box' models. In contrast to classical risk scores, the complex algorithmic structures limit *explainability,* i.e., the insight in patient factors contributing to the made prediction[10]. In recent years, much effort has been made in making AI models explainable, thereby increasing transparency and therefore promoting adoption of AI decision support tools in clinical practice[11]. Therefore, our focus is on the predictive performance and explainability of AI-based discharge decision support tools.

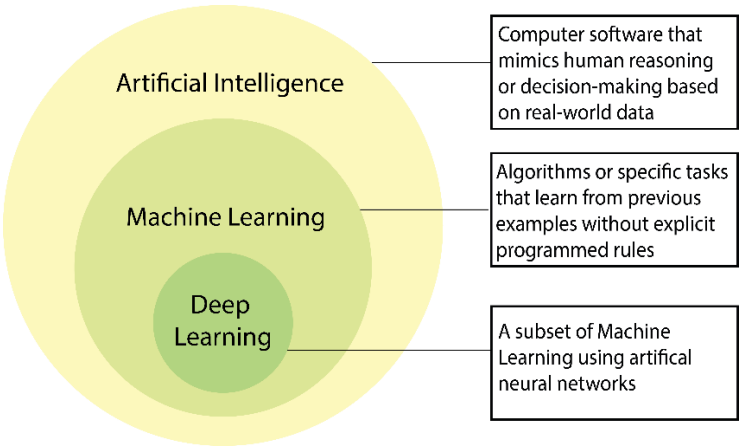


Figure 1: Artificial Intelligence, machine learning and deep learning. Adapted from [8].

1. Framework for the implementation of an AI-based clinical decision support tool

Despite the rapid increase in publications regarding the use of AI for clinical prediction modeling, only few models have been implemented in clinical practice[12,13]. Furthermore, the research so far on the clinical value of AI-based decision support is mainly limited to the field of radiology[14]. As a means to enhance the adoption of clinically meaningful decision support, we propose a framework for the implementation of an AI-based decision support tool in clinical practice as visualized in Figure 2. The framework we developed is inspired by the implementation of risk scores and other classical decision support tools[15,16]. It must be noted that it does not take the technical data integration side of the project into account, since this was out of our research scope. A summary of each phase is provided in this section.

Phase 1: Orientation

During the orientation phase, the clinical problem (in our research: ICU readmission) that might benefit from a decision support tool is defined. A (systematic) literature review performed at the start is needed to assess the availability, strengths and weaknesses of previous developed prediction tools. Next, the medical staff for whom the decision support tool is developed should be questioned regarding the clinical problem and their attitude towards AI-based decision support. If the expectation is that there is potential in using AI as decision support tool for the clinical problem, patient data are collected and explored to be used to build a prediction model.

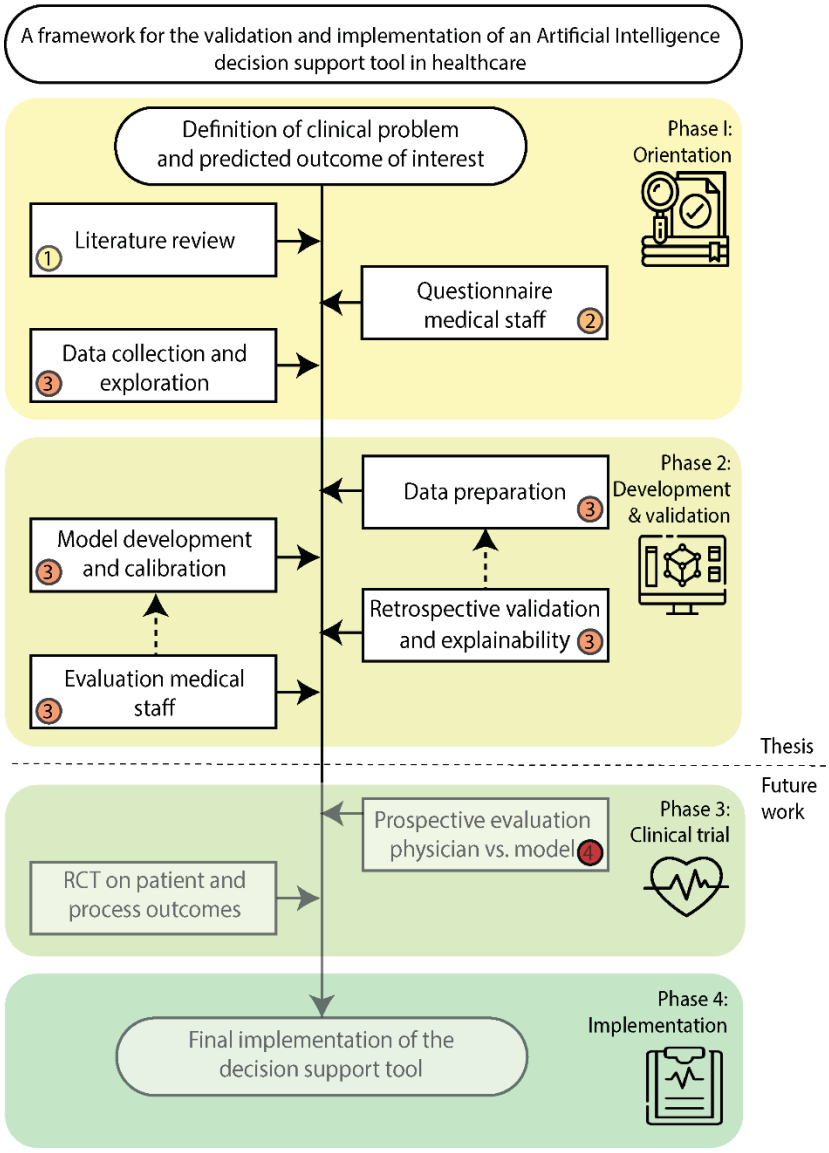


Figure 2: From clinical problem to implementation of an AI decision support tool. The numbers in the boxes correspond to the objectives and parts of this thesis. (1) = Systematic review, (2) & (4) = thesis feasibility study, (3) = Main thesis. RCT = Randomized Controlled Trial.

Phase 2: Model development and validation

Model development and validation is an iterative process, indicated by the dashed arrows in the figure pointing to a previous step. First, the collected patient data are cleaned and prepared (i.e., feature engineering and selection) to become appropriate as input data for model development. Secondly, one or more ML models are trained to predict the clinical outcome on a (retrospective) training dataset and validated on discriminative performance and calibration properties on a test dataset. Next, possibilities for making the developed models explainable to the medical staff are explored. Evaluation of the model's predictions and explainability outcomes (i.e., impactful patient factors to the predicted outcome) are evaluated with the medical staff to enhance understanding and look for potential bias[17]. In some cases, phase 2 can be skipped if an implementation-ready (commercial) decision support tool is available that meets the requirements. However, such a tool should be extensively validated and calibrated during this phase on the new patient population.

Phase 3 and 4: Clinical trial and implementation

In phase 3**,** the added clinical value of the decision support tool is evaluated during two prospective studies. The first study compares the predictive performance of the model (in 'silent-mode' i.e., not visible to the physician) to the prediction registered by the physician. For our use case, the physician is asked to predict a patient's chance of readmission at the moment of discharge. If the ML model shows superior performance over the physician, the impact of the decision support tool on patient outcomes is evaluated in a Randomized Controlled Trial (RCT)[13]. If a positive impact of the decision support tool is observed, in terms of patient outcomes and costs, the decision support tool can be implemented for clinical use. At this stage, all physicians and nurses involved should be trained in using and interpreting the predictions correctly. After final implementation, the tool's software should be managed to monitor changes in data and model performance over time.

2. Approach and research objectives

Within the described framework for the implementation of an AI discharge decision support tool, the different parts of this master thesis correspond to one or more steps in Figure 2. Our use case is the prediction of ICU readmission at the Leiden University medical Centre (LUMC). We (partly) accomplished phase 1 and 2 with the literature study (systematic review), the thesis feasibility study (questionnaire), and the modeling study (data collection until evaluation medical staff). Our developed models were made for research purposes and will most likely not be further implemented. However, phase 3 and 4 will be conducted within the LUMC during the implementation project of a CE-certified discharge decision support tool, Pacmed Critical. Pacmed Critical is a prediction tool developed in collaboration with the Amsterdam UMC, predicting a patient's risk of readmission and/or death within 7 days after ICU discharge[19,20]. Our work contributed to the first steps of validating Pacmed Critical at the ICU of the LUMC, in collaboration with the LUMC Clinical AI Implementation and Research Lab (CAIRELab).

Each thesis objective is discussed in the four parts of this thesis, comprising the literature review (TM30003), the thesis feasibility study (TM30002), and the master thesis (TM30004). The overarching aim was to contribute to the implementation of clinically meaningful AI decision support tools by combining technical skills and intensive care medicine knowledge. The four research objectives were as follows:

1. Systematically review current literature on ML models for the prediction of ICU readmission.
   Part I – ***Literature review***

2. Gain insight in current discharge practices and the ICU physician's attitude towards the integration of AI decision support tools in daily clinical practice.
   Part II – ***Thesis feasibility study 1/2***

3. Develop and compare logistic regression and advance ML models for the prediction of ICU readmission in terms of discriminative performance, calibration properties, and explainability.
   *Part III – **Main thesis***

4. Set up a protocol for the prospective evaluation of Pacmed Critical.
   *Part IV – **Thesis feasibility study 2/2***

Part III includes the main scientific paper on performance and explainability of ML models for the prediction of ICU readmission. A general discussion on the thesis objectives and future perspectives are provided at the end of this thesis.

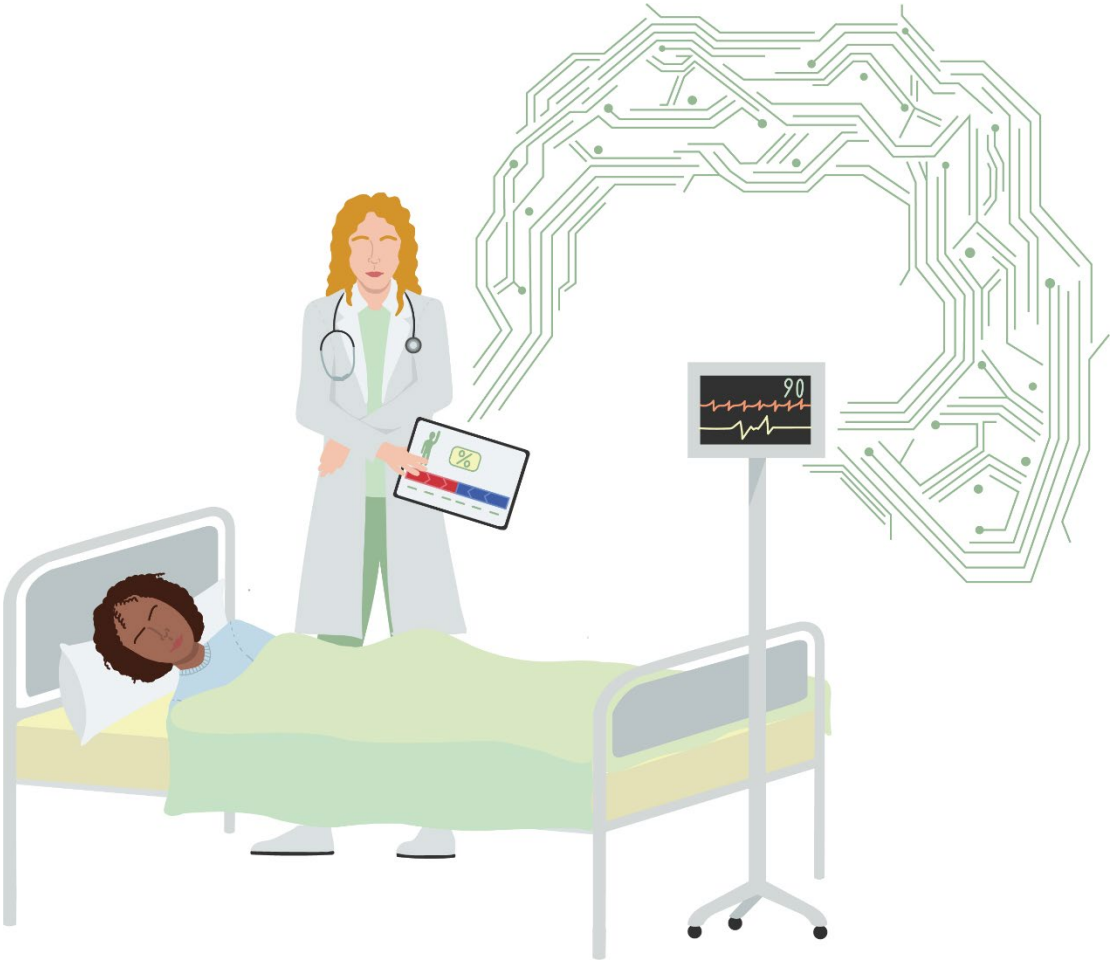
References

1. Kramer AA, Higgins TL, Zimmerman JE. The association between ICU readmission rate and patient outcomes. *Crit Care Med*. 2013;41(1):24-33. doi:10.1097/CCM.0b013e3182657b8a
2. Kramer AA, Higgins TL, Zimmerman JE. Intensive care unit readmissions in U.S. hospitals: Patient characteristics, risk factors, and outcomes. *Crit Care Med*. 2012;40(1):3-10. doi:10.1097/CCM.0b013e31822d751e
3. James FR, Power N, Laha S. Decision-making in intensive care medicine – A review. *J Intensive Care Soc*. 2018;19(3):247-258. doi:10.1177/1751143717746566
4. Akavipat P, Thinkhamrop J, Thinkhamrop B, Sriraj W. Acute physiology and chronic health evaluation (Apache) II score – the clinical predictor in neurosurgical intensive care unit. *Acta Clin Croat*. 2019;58(1):50-56. doi:10.20471/acc.2019.58.01.07
5. Le Gall JR. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA J Am Med Assoc*. 1993;270(24):2957-2963. doi:10.1001/jama.270.24.2957
6. Gajic O, Malinchoc M, Comfere TB, et al. The stability and workload index for transfer score predicts unplanned intensive care unit patient readmission: Initial development and validation. *Crit Care Med*. 2008;36(3):676-682. doi:10.1097/CCM.0B013E318164E3B0
7. Failure of the Swift Score to Predict Readmission to the ICU – SHM Abstracts. https://www.shmabstracts.com/abstract/failure-of-the-swift-score-to-predict-readmission-to-the-icu/. Accessed April 14, 2020.
8. Aunalytics – What exactly is Artificial Intelligence? - aunalytics.com/artificial-intelligence-machine-learning-and-deep-learning. Accessed March 2, 2021.
9. Caicedo-Torres W, Gutierrez J. ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU. *J Biomed Inform*. 2019;98:103269. doi:10.1016/j.jbi.2019.103269
10. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30-36. doi:10.1038/s41591-018-0307-0
11. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113. doi:10.1016/j.jbi.2020.103655
12. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. *Crit Care*. 2019;23(1):284. doi:10.1186/s13054-019-2564-9
13. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
14. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z
15. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic Progn Res 2019 31*. 2019;3(1):1-23. doi:10.1186/S41512-019-0060-Y
16. Wallace E, Smith SM, Perera-Salazar R, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC Med Inform Decis Mak*. 2011;11(1):62. doi:10.1186/1472-6947-11-62
18. Hilton CB, Milinovich A, Felix C, et al. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-0249-z
19. Intensive Care - Pacmed . https://pacmed.ai/en/projects. Accessed August 31, 2020.
20. Thoral, Patrick J , Fornasa, Mattia, de Bruin DP. Developing a Machine Learning prediction model for bedside decision support by predicting readmission or death following discharge from the Intensive Care unit,. *[PREPRINT] (version 1)*. 2020. https://www.researchsquare.com/article/rs-12522/v1.

Part I – Systematic Review

1

For the first research objective, we performed a systematic review as part of the literature study *(TM30003).* Research on the development of machine learning algorithms for the prediction of readmission after ICU discharge were evaluated on discriminative performance, calibration properties, and explainability. The findings of our systematic review contributed to the model development described in Part III.



Machine Learning for the Prediction of ICU Readmission: *A Systematic Review*

Abstract

Introduction Intensive Care Unit (ICU) readmission is a serious adverse event associated with high mortality rates and costs. The prediction of ICU readmission has shifted the last years from classical prediction modeling to using Machine Learning (ML) algorithms. The aim of this paper is to systematically review models/algorithms that predict adult ICU readmission using ML on quality, modeling strategies, and performance.

Methods We searched six databases for studies published from inception until May 17, 2020. Data extracted from the studies included source and size of datasets, predicted outcome measures, modeling strategies, variable selection methods, data pre-processing strategies, explainability methods, impactful features, and model performances. Furthermore, quality, applicability, and risk of bias were assessed using the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction modeling Studies (CHARMS).

Results Eight studies were included. Dataset sizes ranged from 2,018 to 46,252 admissions with readmission rates between 1.9 % - 24.0 %. Predicted outcomes included ICU readmission within 24 hours, 48 hours, 72 hours, 7 days, 30 days, and any time within the same hospital stay. CHARMS quality scores ranged between 65 - 88 %. Area under the receiving operator characteristic curves (AUC) ranged between 0.64-0.92. Only few papers described their method of calibration and the reporting of other performance metrics than AUC was sparse.

Conclusion Prediction of ICU readmission is performed using several types of ML algorithms. Quantitative comparison between modeling strategies could not be performed due to heterogeneity in predicted outcomes, performance metrics, and datasets used. Future studies should focus on the clinical applicability of ML models for the prediction of ICU readmission.

Keywords Intensive care unit, readmission, machine learning, decision support

1. Introduction

Due to high costs and limited bed availability, Intensive Care Unit (ICU) patients should be discharged to lower care hospital wards as soon and as safely possible. Although the decision to discharge is made with the utmost carefulness, it may result in ICU readmission due to limited monitoring and therapeutic options at lower care wards. ICU readmissions are serious adverse events related with mortality rates ranging between 26-58 % [1]. Physicians and patients could therefore benefit from identifying patients at risk of readmission before making the decision to discharge.

Hosein et al. evaluated the use of different risk stratification tools for ICU readmission in a systematic review [2]. Evaluated tools included the Modified Early Warning Score (MEWS) that predicts readmission within 72 hours of discharge, the Stability and Workload Index for Transfer (SWIFT) score, and Frost nomogram that both predict readmission within the same hospital stay. They used 5-26 variables to determine the patient's risk of readmission at discharge [2]. However, the prediction of readmission is challenging due to the heterogeneous patient population. Simplified risk scores may not be sufficient due to their inability to capture the patient's complete clinical status. The increasing amount of patient data stored in the Electronic Health Record (EHR), including a lot of (time-series) variables for large patient cohorts, may be used to develop more complex and potentially more accurate prediction algorithms.

A systematic review from Markazi-Moghaddam et al., published in 2019 (evaluating studies published until January 2017), assessed primary models (not externally validated) to predict ICU readmission [3]. Five studies were included, of which four used a logistic regression model and one a data-mining approach.

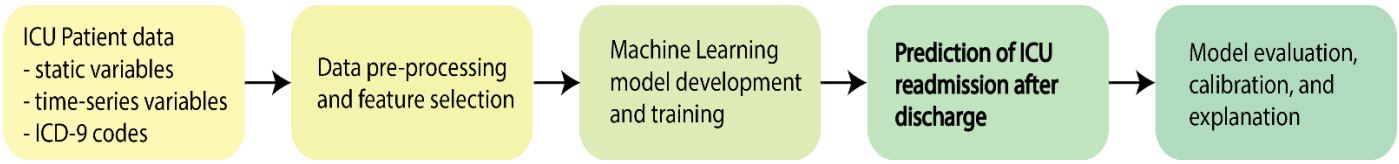


Figure 3: The process of prediction modeling using Electronic Health Record (EHR) patient data. Static variables include for example age, reason for admission, and BMI. Time-series variables include vital functions and medication that change over time and have different sample frequencies. ICD-9 codes include patient diagnoses and procedures.

Predicted outcome measures for readmission were different for all five studies; readmission within 48 hours after discharge, between 24-72 hours after discharge, any readmission within the same hospital stay, readmission related to the initial ICU admission, and readmission specific in a group of postoperative patients. Area Under the receiving operator characteristic Curve (AUC) ranged from 0.66 to 0.81 and most studies reported limited other performance metrics. Studies were evaluated to be generally of moderate to low quality according to the CHAMRS criteria (the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies).

Since the systematic review of Markazi et al., applied methods to predict ICU readmission have shifted from logistic regression to using Artifical Intelligence (AI) in the form of Machine Learning (ML) algorithms [4]. These algorithms are able to 'learn' from the thousands of different variables present in the EHR. Having learned from admission data of previously readmitted ICU patients, a ML model could accurately predict a patient's chance of readmission at the moment of discharge. Figure 3 shows the process of prediction modeling using EHR data.

To compare prediction models on their performance, several aspects need to be taken into account. The model needs to discriminate between patients at high and low risk of readmission and it should be calibrated on the population of interest to have high agreement between predictions and observations [22]. Furthermore, the model should generalize well to new populations. Prediction models are validated either internally (using k-fold cross validation or bootstrapping), or externally on a different patient dataset [3].

Compared to traditional statistical modeling (e.g., logistic regression), ML models potentially have higher predictive performance but this comes at the cost of model interpretability and explainability. Interpretability refers to the transparency of the working mechanism behind the prediction model, whereas explainability gives an explanation of the EHR features contributing to the predicted outcome for a specific patient [6]. In order to incorporate a ML prediction algorithm in clinical practice, knowing the impactful predictors (explanations) to the predictions could help physicians in the decision to trust on the algorithm or not [8]. The hypothesis is that physicians are more likely to adopt ML as decision support tools when having insight in the predictors contributing most to the reported outcome, and therefore enhancing clinical applicability [7].

Due to these concerns, advances have been made the last years in making non-interpretable and difficult to explain ('black-box') ML algorithms explainable. Therefore, we not only evaluated study quality and performance, but also explainability methods of non-interpretable ML algorithms. The aim of this study was to systematically review studies predicting readmission to adult ICU using ML models with respect to study characteristics, study quality, modeling methods, model explainability, and performance.

2. Methods

2.1. Study protocol and registration

The study protocol and study was registered in PROSPERO under CRD42021226415.

2.2. Search strategy

Six data bases (Pubmed, Embase, Web of Science, COCHRANE Library, Emcare, and Academic Search Premier) were searched from inception until May 17, 2020. Keywords used for searching included "prediction", "decision support", "intensive care unit", "Machine learning", "artificial intelligence", "readmission", and "discharge". Synonyms were used for all keywords. The full search strategy is provided in the Supplementary material Part I - Systematic Review (Page 65).

2.3. Study selection i.e., inclusion criteria

Inclusion criteria were: description of the development of a ML model for the prediction of adult (> 18 years) ICU readmission, with as primary outcome chance of readmission at any time within the same hospital stay. Studies separately predicting both readmission and mortality (and not as composite outcome) were included as well. Both retrospective and prospective studies using EHR data were included when full text was available in English.

Exclusion criteria were: prediction of another outcome than ICU readmission, validation of an existing (non-ML) risk score, descriptive studies on general risk factors for ICU readmission, review articles, and articles focused on a pediatric population. Study titles were first screened on relevance for our study objective by the author (SvdM). Afterwards, we assessed full-text on inclusion and exclusion criteria.

2.4. Data extraction

We systematically reviewed the included studies on five domains, see Supplementary material Part I - Systematic Review (Page 61) for the used data extraction tables. The first domain comprised the study characteristics, including year of publication, study type, dataset description, the number of ICU admissions (samples), definition of readmission, proportion of ICU readmission in the dataset, and the type of ML models used. Secondly, study quality, applicability and risk of bias were assessed. The modeling methods of each study were further elaborated in the third domain, including feature selection methods, number of input features, data preparation methods, handling of missing data, modeling methods, validation strategies (internal and external), and method of calibration. In the fourth domain, studies were compared on their use of methods to enhance explainability of their model and (the top 10) features contributing to the predicted outcome. Lastly, the highest performing models of each included study were compared on relevant metrics and calibration properties.

2.5. Assessment of study quality and applicability

We assessed study quality by evaluating internal validity and transparency in modeling methods. Prediction modeling studies could report high discriminative performance, but at the same time they could suffer from bias or lack of generalizability when having low study quality. The CHARMS checklist was used to assess study quality [7]. This checklist comprises a list of criteria divided over 11 domains (source of data, participants, outcome(s) to be predicted, candidate predictors, sample size, missing data, model development, model performance, results, interpretation, model evaluation, and discussion). When a criterion was fully met, two points were awarded. One point was given when a criterion was partly met, and zero when a criterion was not met at all. Besides *general study quality*, *study applicability* (study applicable for the intended use, i.e., prediction of readmission in the target population) and *risk of bias* (e.g., risk of overfitting due to design flaws) were assessed using two additional subsets of the CHARMS checklist [7]. Not all items in the CHARMS checklist were relevant for ML prediction models [3]. Thirty criteria were applicable to ML model development resulting in a maximum achievable general quality score of 60. We used 15 of these 30 criteria to assess study applicability, with a maximum of 30 points, and 16 to assess risk of bias, with a maximum of 32 points [7]. General quality, applicability, and risk of bias scores were given as percentages between 0-100%.

3. Results

3.1. Identification of eligible studies

After removal of duplicates, a total of 172 articles were identified. 139 articles could be excluded due to an irrelevant study objective. Main reasons of exclusion after reviewing abstracts and full text included the prediction of solely mortality after discharge (n = 12) and prediction of other outcome measures than ICU readmission (n = 6). See Figure 4 for other reasons of exclusion. Finally, we included eight studies [10–17].

3.2. Study characteristics

The studies included were published between 2017 and 2020. See Table A-1 (Supplementary material Part I - Systematic Review, page 61) for a summary of study characteristics. Study types were either observational or cross-sectional. Patient sample sizes differed between 2,018 inclusions [14] up to 46,252 [12]. Five studies trained and validated their model solely on the MIMIC-III database [10–13, 16], where two articles used a combination of the MIMIC database and the EHR of a local hospital [14, 15]. One study trained and validated their model solely on the data of three local hospitals [17]. The MIMIC database is a freely

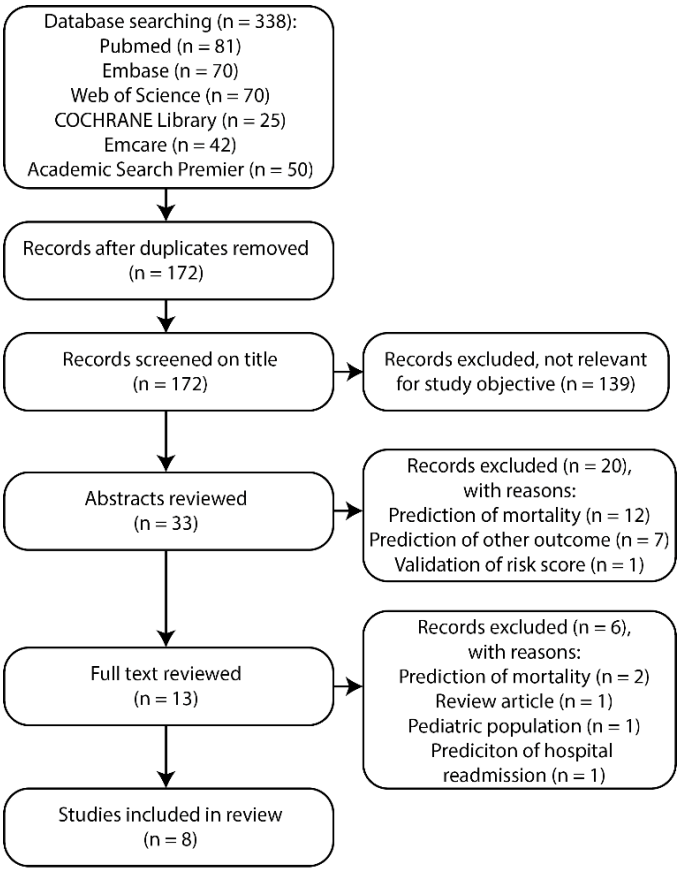


Figure 4: Flow chart of study identification and selection

accessible critical care database, including ten thousands of patient EHRs [18]. Furthermore, the included studies used several definitions for the prediction of readmission, ranging between 24 hours after discharge and 30 days after discharge. Two studies defined their predicted outcome without time-limit (i.e., readmission within the same hospital stay). The reported ICU readmission rates ranged between 1.9% (within 24 hours after discharge, [12]) to 24.0% (within 30 days after discharge [9]). See Figure 5 for an overview of readmission rates for each predicted outcome.

Different types of ML models were used among studies. In two studies, several state-of-the-art ML models were internally compared, including support vector machine (SVM), random forest (RF), and Multi-Layer Perceptron (MLP, also known as feed-forward neural networks)) [15, 16]. Three other studies focused mainly on boosting algorithms [12– 14]. Advanced Deep Learning (DL) model development was described in three papers, using recurrent neural networks (RNN), convolutional neural networks (CNN) and conditional random fields (CRF) [9–11]. Most studies compared the discriminative performance of their developed algorithms to one or more other (previous published) algorithms or risk scores, including the SWIFT score and MEWS score [8, 13] or a score obtained using logistic regression [12, 15]. All studies showed superior performance of their ML model over risk scores and LR.

3.3. Quality of studies

Studies were scored on general quality using the CHARMS criteria. Two additional subsets of criteria were used to assess applicability of models and risk of bias. See Figure 6 for an overview of study quality in percentages. Lin et al. scored highest on all aspects, with a general quality score of 88%. Venugopalan et al. had the lowest general quality score (62%), and lowest applicability score due to limited reporting of model performance, and interpretation of results. For most included studies, general study quality was high or sufficient on most aspects such as on description of participants, outcome(s) to be predicted, sample sizes, model development, reporting of results, and model evaluation. However, information on missing data and calibration measures was lacking in most studies. See Supplementary material Part I - Systematic Review (page 67) for the CHARMS scores of all included studies.

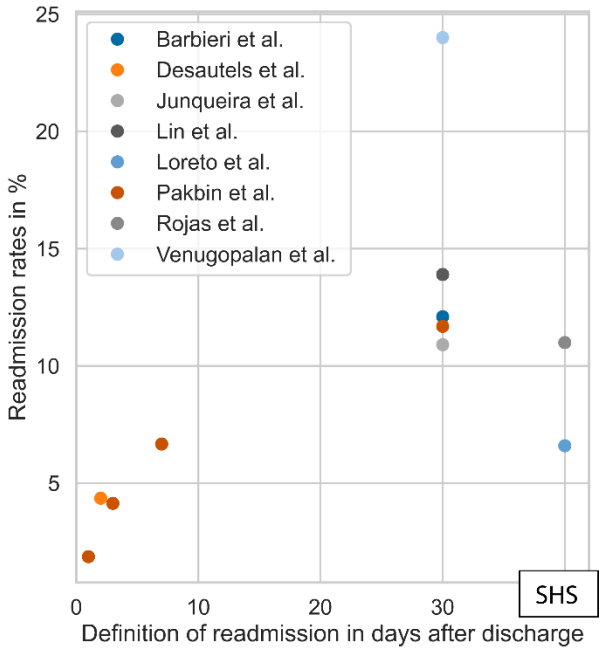


Figure 5: Readmission rates for each included study. The defined outcome for readmission in days after discharge is plotted on the x-axis against the reported readmission rate on the y-axis. Two articles, Rojas et al. and Loreto et al. defined readmission within same hospital stay (SHS) without defining with no defined limit in days after discharge. These two studies are reported separately on the right of the graph.

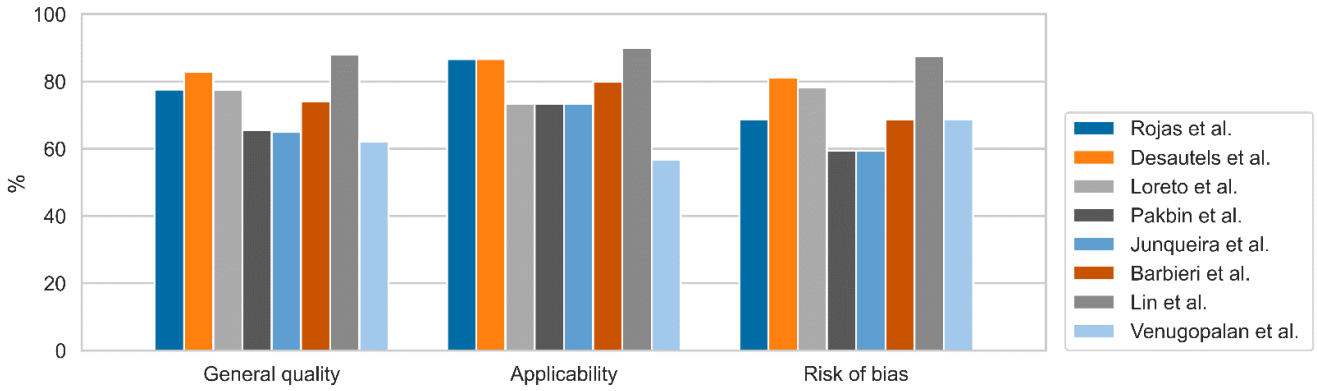


Figure 6: CHARMS scores in percentages for each included study. Two subsets of the CHARMS criteria were used to assess applicability and risk of bias.

3.4. Machine Learning modeling strategies

Using the EHR as input data source, a distinction can be made between static variables (e.g., age, BMI) and time-series variables (e.g., vital signs, laboratory results, nursing scores). Static variables are constant during the whole ICU admission, and therefore one single value is used as input to the algorithm. Time-series variables are numerical or categorical values registered with differences in sample frequencies. ICD-9 codes (Standardized Diagnosing Codes) are used to store patient diagnoses over time. In ICD-9 codes, important patient information is stored before admission (e.g., chronic diseases), but also new diagnoses assessed and interventions performed during the ICU admission, and therefore belong to the category of time-series features. The study by Junqueira et al. was the only to solely use static input variables for the prediction of readmission, where all other included studies used both static features and time-series features as input data for their model. Input data collection periods used for time-series variables differed between studies (e.g., the first hour of measurements after admission, the last 24h before discharge, or from the whole ICU stay), see Table A-2 (Supplementary material Part I - Systematic Review, page 61). Loreto et al. used multiple datasets for prediction modelling. One dataset solely included data available at ICU admission. Predicting readmission is a more difficult task when using solely data available at time of ICU admission since no information on the course of the ICU admission is used.

Different feature selection methods were used to select the features contributing mostly to the predicted outcome as input data for the model. Rojas et al. made an a priori selection of variables based on clinical experience. Other studies excluded redundant features, features with more than 50 % missing values or used L1-feature selection. L1-feature selection is used to reduce overfitting on the training dataset on redundant features, by shrinking the coefficient of the least informative features to zero. Imputation of missing data was performed using several strategies, including k-means imputation, mean, most recent values, and others. Data aggregation and vector embedding strategies were used to structure the different types of input data. Data aggregation is a method to extract features from time-series data, and vector embeddings are used to create equally sized input matrices to train the model on. The number of input features ranged between 12 [15] to 2,344 [12]. Changes in time-series variables (e.g., blood pressure) during the admission might be relevant to the prediction of readmission, DL models have the opportunity to better handle time-series data. The three studies focusing on DL used embeddings to deal with multiple sample frequencies of time-series data, to get a fixed size input data matrix for each patient. The other studies used aggregations of (e.g., mean, standard deviation) of time-series data, resulting in a few summarizing statistics for each variable.

Most studies developed several ML models and compared their performance. Four studies showed the highest performance in terms of AUC for tree-based ML models including Random Forest (RF), Gradient Boosting (GB) or XGBoost. Junqueira et al. reported highest performance for the MLP, with a marginal difference from the other models. The three papers focusing on DL predicted readmission within 30 days of discharge. Best DL models included

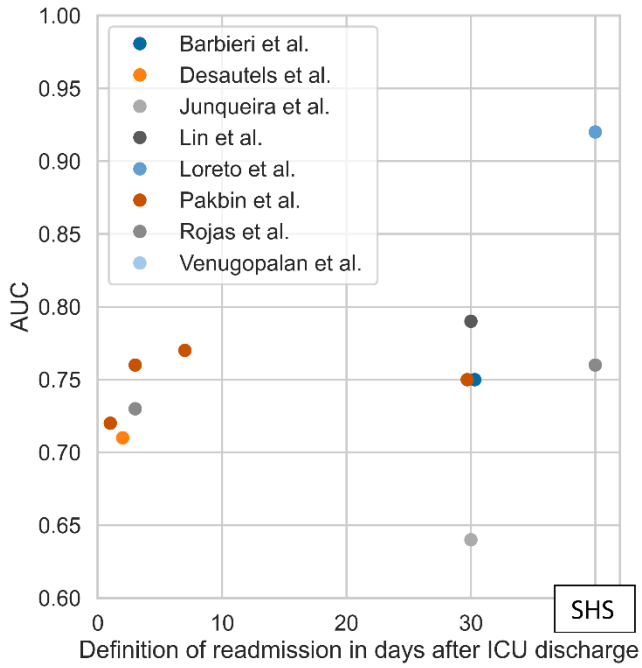


Figure 7: Highest AUC (Area under the receiving operator curve) for each model for different predictions of readmission within days after discharge. SHS = Same Hospital Stay.

Recurrent Neural Networks (RNN) with Bi-directional Gated Recurrent Units (GRU) [11], RNN with Long Short-Term Memory (LSTM) [10], and a combination of conditional random fields (CRF) as temporal model and a Feed-Forward Neural Network (FFNN) as static model [9].

Cross-validation was used to average performance over different training and test datasets. Barbieri et al. were the only using bootstrapping instead of cross-validation. Three studies performed some form of external validation by evaluating their model on the MIMIC-III dataset [13–15]. All studies compared their best model to other ML models, LR models, and/or classical scores as the Stability and Workload Index for Transfer (SWIFT) score. Desautels et al. and Pakbin et al. were the only to report methods of calibration, using LR and calibration plots.

3.5. Model explainability and predictive features

The last few years, efforts have been made in making non-interpretable ML models more explainable in terms of features contributing to the predicted outcome. Explainable models are more likely to be adapted in clinical practice because a physician is more likely to trust a prediction when knowing its impactful variables [18].

Several explainable methods were used to gain insight in contributing features to the reported outcome. The methods used and the top-10 most predictive features are shown in Table A-3 (Supplementary material Part I - Systematic Review, page 61). Glucose levels, heart rate (HR), length of stay (LOS), Glasgow Coma Scale (GCS), respiratory failure/parameters (e.g., respiratory rate, SpO2, mechanical ventilation), and multiple laboratory parameters were often named in the included studies as important predictors on the population level. Only Barbieri et al. mentioned the use of their method to be applicable for individual patient specific explanations. They concluded that by adding attention to their model, explainability was enhanced at a marginal cost in performance [11].

3.6. Performance

Reported AUC's ranged between 0.64 (Junqueira et al.) and 0.92 (Loreto et al.). However, it was not possible to compare AUC between studies due to differences in definitions of readmission. As Thoral et al. mentioned, predicting ICU readmission within 48 h is a more difficult prediction task than predicting readmission within 30 days after discharge. Due to a smaller number of patients being readmitted within 48 hours, it is more difficult to train an accurate algorithm than for patients readmitted within the same hospital stay. Furthermore, different databases were used for development of the model, making comparison challenging. For all studies that trained their model on the MIMIC-III database, Lin et al. had the highest AUC of 0.79, using a LSTM in combination with a CNN. Venugopalan et al. were the only to not report AUC as outcome measure, and used the Matthews Correlation Coefficient (MCC) with a MCC of 0.65. See Table A-4 (Supplementary material Part I - Systematic Review, page 61) for an overview of reported performance metrics. It was remarkable that Loreto et al, reporting the highest AUC, had lowest recall (sensitivity, 0.46) and precision (positive predictive value, 0.58). This could indicate a large proportion of true negatives resulting in the high AUC. In Figure 7, the AUC of each study is visualized for the different predictions of readmission in days after discharge. Pakbin et al. had the highest AUC for prediction of readmission within 72 hours (AUC = 0.76) and Lin et al. had the highest AUC for readmission within 30 days (AUC = 0.79). Multiple other performance metrics were reported for each study, but these differed among studies, see Figure 8.

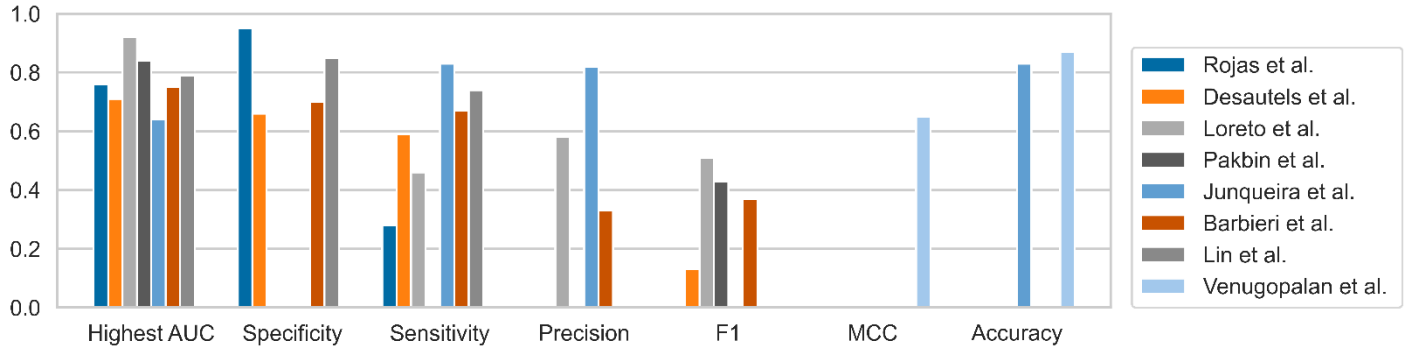


Figure 8: Reported ddiscriminative performance metrics per study (not all performance metrics of interest were given for each study). Highest performing model metrics are plotted for studies describing the development of more than one model. AUC = Area under the receiving operator curve, MCC = Matthews correlation coefficient, PPV = positive predictive value.

4. Discussion

A shift in prediction modeling of ICU readmission is observed the last few years from classical regression towards the use of machine learning. A large heterogeneity in definition of ICU readmission, modeling methods, reported performance metrics, explainability methods, input variables, and sample sizes was observed in the included studies. Direct comparison among studies could therefore not be performed, and both 'classical' ML models (mainly tree-based and boosting models) and DL models performed well. Consensus should be formed on the defined time to readmission after discharge and reporting of performance metrics. Most studies used an explainable method to get insight in the contributing features to the reported outcome, enabling 'black-box' ML models to become explainable. The next step in creating clinically valuable prediction tools would be prospective validation and implementation of these discharge decision support tools.

In the last review from Markazi et al., mainly prediction algorithms based on LR were included until January 2017. The highest AUC of the included studies was from Magruder et al. (AUC = 0.81). However, they only included four features in the final model and the sample size was small (452 patients), increasing the risk of overfitting on the dataset and reduced generalizability to other populations. The overall conclusion from this review was that the included studies had low methodological quality, including small sample sizes, lack of information on missing data handling, and remarks on model development, model evaluation and results [3]. Since 2017, there has been a shift from LR to more advanced ML models. And an improvement has been made the past years in terms of sample sizes, model evaluation, imputation of missing data, and validation methods. A similar limitation as described by Markazi et al. is that only two studies report their method of calibration.

To our knowledge, this is the first systematic review looking into the application of ML models for the prediction of adult ICU readmission. We chose to only include papers describing the development of at least one ML model, and did not compare them to papers solely based on LR or previous established risk scores. Two included studies internally compared state-of-the-art ML models to LR and reported superior predictive performance for ML compared to LR [10,11]. However, a systematic review showing no benefit of ML over LR for prediction modeling reported that this finding is not trivial [25]. Although we extensively searched six databases, it could be that not all developed models were included in this systematic review. The field of medical prediction modeling using ML is rapidly evolving and some non-peer-reviewed papers were found which aimed at making the step to clinical utility of ML models [8], but which were not included in this review. Another limitation could be the use of the CHARMS criteria, using them to assess study quality mainly focuses on whether certain aspects are reported, and not whether they actually perform well. However, transparent reporting is one of the most imported things when assessing study quality of prediction models.

Due to a relative low number of readmissions (1.9% - 24.0%), reported in the datasets, the prediction of readmission is a so-called imbalanced prediction problem. Using AUC as performance metric may

mislead the reader, due to the relatively large amounts of true negatives. Therefore, Venugopalan et al. used Matthews Correlation Coefficient (MCC) instead of AUC as performance metric, which is more suitable for imbalanced datasets. Overall, low recall (sensitivity) rates were reported where Rojas et al. had a recall of 0.28 indicating a large proportion of false negatives. They chose a cut-off specificity of 0.95, implicating a need for a large proportion of true negatives. Lin et al. mentioned that high sensitivity would be preferred because readmission is a serious event and readmitted cases should therefore not be missed by the algorithm [10]. This illustrates the debate on whether high sensitivity or specificity should be preferred in the prediction of readmission. High sensitivity at the cost of specificity could result in patients being unnecessarily kept long at the ICU, while high specificity at the cost of sensitivity could result in patients at risk of readmission being discharged too early. AUC represents the trade-off between sensitivity and specificity and is a widely used performance measure for prediction algorithms. However, for this unbalanced prediction problem, AUC is a suboptimal performance metric since it rewards a large proportion of true negatives. Therefore, a more appropriate metric would be the area under the precision-recall curve (AUCPR), which represents the trade-off between the sensitivity (recall) and positive predictive value (precision) of the model [23]. Unfortunately, none of the included studies reported the AUCPR. Because AUC was the performance metric most often reported among the included studies, some comparison could be made based on this metric. AUC ranged between 0.64 – 0.92, with the highest AUC reported by Rojas et al. However, Rojas et al. did not perform external validation and trained their model on a relatively small sample size of 9,926 patients. Therefore, the high discriminative performance could be caused by overfitting on the training dataset. Junqueira et al., having the lowest AUC, used solely static input features as input for their model. This finding implies that there is predictive information in the time-series data that are collected during a patient's ICU stay.

ICU readmission is associated with high mortality rates and hospital costs. Identifying patients at risk for readmission could assist physicians in their decision to discharge, reduce complications and might eventually contribute to lower health care costs. The use of evaluation methods looking at the clinical value of the model besides the discriminative and calibration properties of the model, such as decision curve analysis were not performed in the included studies. Decision curve analysis is used to calculate the net benefit of a model, looking at the trade-off between the harms and benefits of a model at different threshold levels at which a patient would be discharged or not discharged based on the prediction [24]. And although a relatively large number of ML prediction models have the last years been studied on the prediction of readmission, none of them have been implemented in clinical practice as far as known. Rojas et al. performed a prospective study using their ML model and asked physicians the likelihood of a patient to be readmitted on a scale of 1-10. The model outperformed the physician with an AUC of 0.78 versus 0.71 [19]. The physician might benefit from knowing the patient's risk of readmission. Discharge could be postponed for patients high at risk or they could be better monitored for a certain amount of time at the ward. Furthermore, the decision to discharge an ICU patient could be supported and even advanced for patients low at risk. However, the benefit for the patient when these algorithms were to be used, need to be established in clinical practice. The prediction of ICU readmission however is a difficult task, since the decision of readmission is made by physicians and differs between centers [20]. Therefore, the generalizability of ML models between centers is limited. Calibration should always be performed and the risk for human biases is relevant.

Another challenge in the adoption of decision support tools based on ML models in clinical practice is the explainability and interpretability of the model. ML models have the advantage over classic prediction models in that they can handle highly dimensional data. DL models can even handle different sample frequencies present in time-series data, whereas classical ML models mostly use feature aggregates of time-series variables. However, the large amount of input features used and the algorithms' complex structures result in decreased interpretability and explainability. Seven of the included studies made an effort by using feature importance methods such as partial dependence plots [13], permutation feature importance [15], and information gain [16]. The features most impactful to the population's

readmission risk were reported and differed between studies.

To conclude, with the use of EHR data collected before, and during the ICU stay of a patient, accurate prediction algorithms can be modelled using state-of-the-art ML models. Other performance metrics need to be evaluated in future studies, including the area under the precision recall curve and decision curve analysis. The next step in the field of ICU readmission prediction would be to perform clinical trials on one or multiple ML algorithms, addressing the clinical value of the prediction model in clinical practice. Furthermore, consensus should be made on what outcome(s) should be used for the prediction of readmission, to make comparison between models feasible. Both state-of-the-art ML and DL models can be used for an accurate prediction of readmission. Tree-based and boosting ML models have the advantage of being more explainable to the physician, whereas DL models are better at handling time-series data which can result in a higher performance.

References

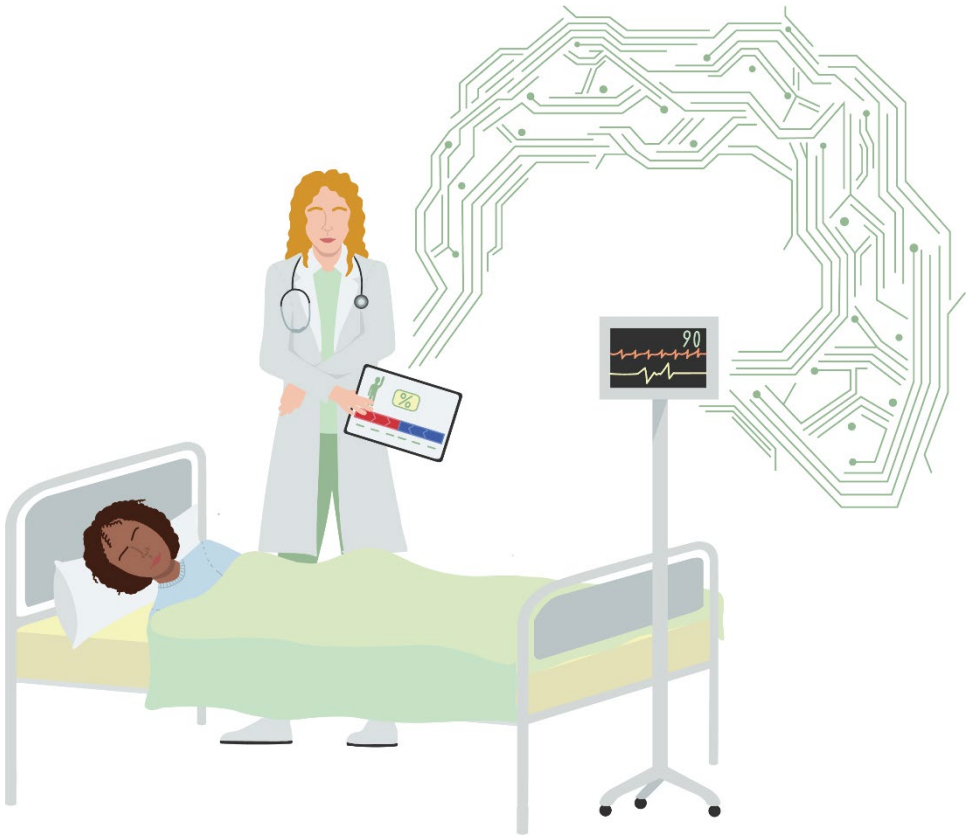
1. A. A. Kramer, T. L. Higgins, and J. E. Zimmerman. "The association between ICU readmission rate and patient outcomes". In: *Critical Care Medicine* 41.1 (Jan. 2013), pp. 24–33. issn: 00903493. doi: 10.1097/CCM.0b013e3182657b8a.
2. F. S. Hosein et al. "A systematic review of tools for predicting severe adverse events following patient discharge from intensive care units". In: *Critical Care* 17.3 (June 2013), R102. issn: 13648535. doi:10.1186/cc12747.
3. N. Markazi-Moghaddam, M. Fathi, and A. Ramezankhani. "Risk prediction models for intensive care unit readmission: A systematic review of methodology and applicability". In: *Australian Critical Care* 0.0 (2019). issn: 10367314. doi: 10.1016/j.aucc.2019.05.005.
4. D. Shillan et al. "Use of machine learning to analyse routinely collected intensive care unit data: A systematic review". In: *Critical Care* 23.1 (Aug. 2019), p. 284. issn: 1466609X. doi: 10.1186/s13054-0192564-9. url: https://ccforum.biomedcentral.com/articles/10.1186/s13054-019-2564-9.
5. H. C. Thorsen-Meyer et al. "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records". In: *The Lancet Digital Health* 2.4 (Apr. 1, 2020), e179–e191. issn: 25897500. doi: 10.1016/ S2589-7500(20)30018-2.
6. D. Doran, S. Schulz, and T. R. Besold. "What does explainable AI really mean? A new conceptualization of perspectives". In: vol. 2071. CEUR-WS, Oct. 2, 2018.
7. K. G. M. Moons et al. "Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies: The CHARMS Checklist". In: *PLoS Medicine* 11.10 (Oct. 2014), e1001744. issn: 1549-1676. doi: 10.1371/journal.pmed.1001744. url: https://dx.plos.org/10.1371/journal. pmed.1001744.
8. P. J. Thoral et al. "Developing a Machine Learning prediction model for bedside decision support by predicting readmission or death following discharge from the Intensive Care unit". In: *PREPRINT* (2020).
9. J. Venugopalan et al. "Combination of static and temporal data analysis to predict mortality and readmission in the intensive care". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*.Institute of Electrical and Electronics Engineers Inc., Sept. 2017, pp. 2570–2573. doi: 10.1109/EMBC.2017.8037382.
10. Y.-W. Lin et al. "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory." In: *PloS one* 14.7 (2019), e0218942. doi: 10.1371/journal.pone.0218942.
11. S. Barbieri et al. "Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk". In: *Scientific Reports* 10.1 (Dec. 2020). doi: 10.1038/s41598020-58053-z. arXiv: 1905.08547.
12. A. Pakbin et al. "Prediction of ICU Readmissions Using Data at Patient Discharge". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2018-July. Institute of Electrical and Electronics Engineers Inc., Oct. 2018, pp. 4932–4935. isbn: 9781538636466. doi: 10.1109/EMBC.2018. 8513181.
13. J. C. Rojas et al. "Predicting Intensive Care Unit Readmission with Machine Learning Using Electronic Health Record Data." In: *Annals of the American Thoracic Society* 15.7 (July 2018), pp. 846–853. doi: 10.1513/AnnalsATS.201710-787OC.
14. T. Desautels et al. "Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach." In: *BMJ open* 7.9 (Sept. 2017), e017199. doi: 10.1136/bmjopen-2017-017199.
15. Junqueira, A.R.B., Mirza, F. and Baig, M.M. A machine learning model for predicting ICU readmissions and key risk factors: analysis from a longitudinal health records. *Health Technol.* **9,** 297–309 (2019). https://doi.org/10.1007/s12553-019-00329-0
16. M. Loreto, T. Lisboa, and V. P. Moreira. "Early prediction of ICU readmissions using classification algorithms". In: *Computers in Biology and Medicine* 118 (Mar. 2020). doi: 10.1016/j.compbiomed. 2020.103636.
17. A. E. Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3 (2016), p. 160035.
18. C. J. Kelly et al. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC Medicine* 17.1 (Oct. 29, 2019), p. 195. issn: 17417015. doi: 10.1186/s12916-019-1426-2.
19. J. Rojas et al. "Man vs. Machine: Comparison of a Machine Learning Algorithm to Clinician Intuition for Predicting Intensive Care Unit Readmission". In:

American Thoracic Society International Conference Meetings Abstracts American Thoracic Society International Conference Meetings Abstracts. American Thoracic Society, May 2019, A2459–A2459. doi: 10.1164/ajrccm-conference.2019.199.1_meetingabstracts.a2459.

20. C. V. Cosgriff, L. A. Celi, and C. M. Sauer. “Boosting Clinical Decision-making: Machine Learning for Intensive Care Unit Discharge.” In: *Annals of the American Thoracic Society* 15.7 (July 2018), pp. 804–805. doi: 10.1513/AnnalsATS.201803205ED.
21. B. Shickel et al. “Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis”. In: *IEEE Journal of Biomedical and Health Informatics* 22.5 (Sept. 1, 2018), pp. 1589–1604. issn: 21682194. doi: 10.1109/JBHI.2017.2767063.
22. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
23. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.
24. Vickers AJ, Van calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.
25. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.

Part II - Questionnaire ICU Physicians

② We conducted a questionnaire to gain insight in current discharge practices and the ICU physician's attitude towards the integration of AI-based decision support tools (research objective 2). This part provides a summary of the questionnaire results. The full questionnaire report, as part of the thesis feasibility study (TM30002) can be found in the Supplementary material Part II – Questionnaire (page 70).



Current Discharge Practices and Physician Perspectives on the Integration of an AI-Based Discharge Decision Support Tool

- Summary thesis feasibility report -

1. Introduction

Before the start of the first phase of implementing Pacmed Critical[1] for discharge support at the ICU of the LUMC, we conducted a questionnaire amongst the ICU physicians of the LUMC. To have an impact on the physicians' decisions, their perspective on AI prediction tools is of importance[2]. Furthermore, the knowledge of current discharge practices and workflow preferences could contribute to making the tool of clinical value. The objectives of the questionnaire were to gain insight in:

1. current clinical practice to discharge ICU patients to lower care wards;
2. the physicians' attitude towards the use of decision support tools based on AI in their work processes, specifically for discharge decision support;
3. workflow preferences in terms of the appropriate place and moment of showing the prediction to the physicians; and
4. the preferred predicted outcome measure for clinical valuable decision support and influence of the predicted probability on the decision to discharge ICU patients.

2. Methods

We conducted a 21 questions survey in December 2020 amongst the intensivists (staff members), ICU fellows, physicians in training, and house doctors of the LUMC. The questionnaire included 11 statements, 6 multiple choice questions and 4 open questions. The questionnaire was conducted anonymously as we only registered the participant's function, years of ICU experience, and medical specialisation. We used a 5-point Likert scale[3] to answer the statements between strongly disagree and strongly agree. Outcomes were analysed in mean (+- standard deviation (SD)), or percentages where appropriate. See the Supplementary material Part II – Questionnaire (page 70) for the survey questions.

3. Results

A summary of the results is given in this section. See the Supplementary material Part II – Questionnaire (page 72) for the thesis feasibility report in Dutch including additional results and physician comments.

3.1. Participants

Questionnaire responses were collected between December 21, 2020 and December 29, 2020, resulting in 32 respondents. Mean ICU experience of the respondents in years was 12.5 ± 4.5 years for the intensivists and 1.1 ± 1.0 years for the other participants (Figure 9). Other participants included ICU fellows, residents (AIOS), and house doctors (ANIOS). The difference in years of ICU experience is explained by the short period (up to 2 years) the other participants are generally working at the ICU department. The level of education of the participants and their medical specialisation is visualized in Figure 10. The two largest medical specialisations of the participants were internal medicine (n = 12) and anaesthesiology (n = 9).

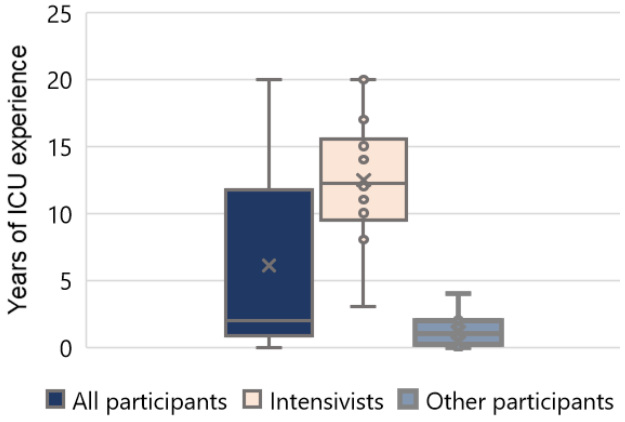


Figure 9: Years of ICU experience of participating physicians.

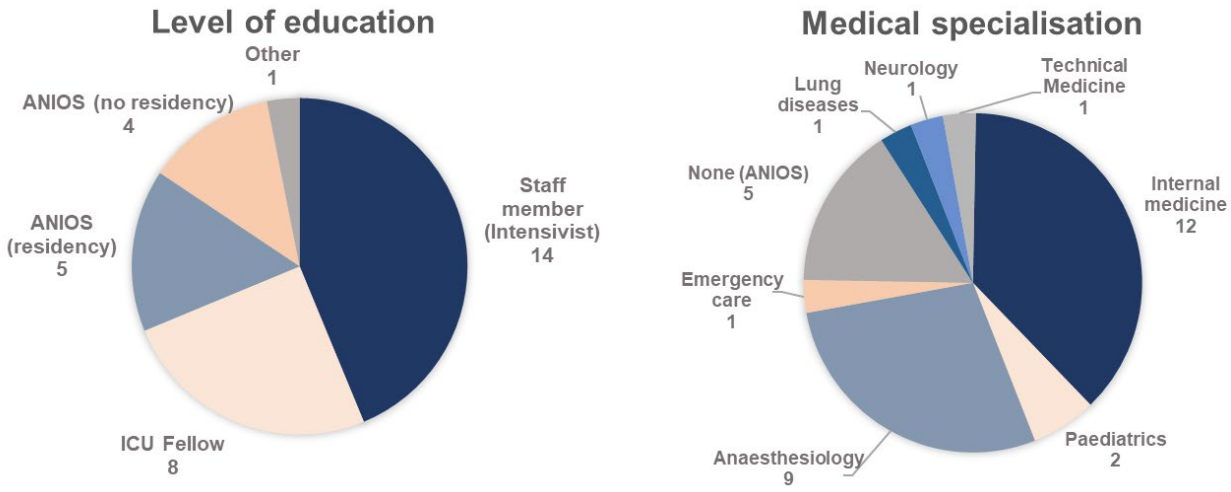


Figure 10: Level of education and medical specialisation of the questionnaire participants.

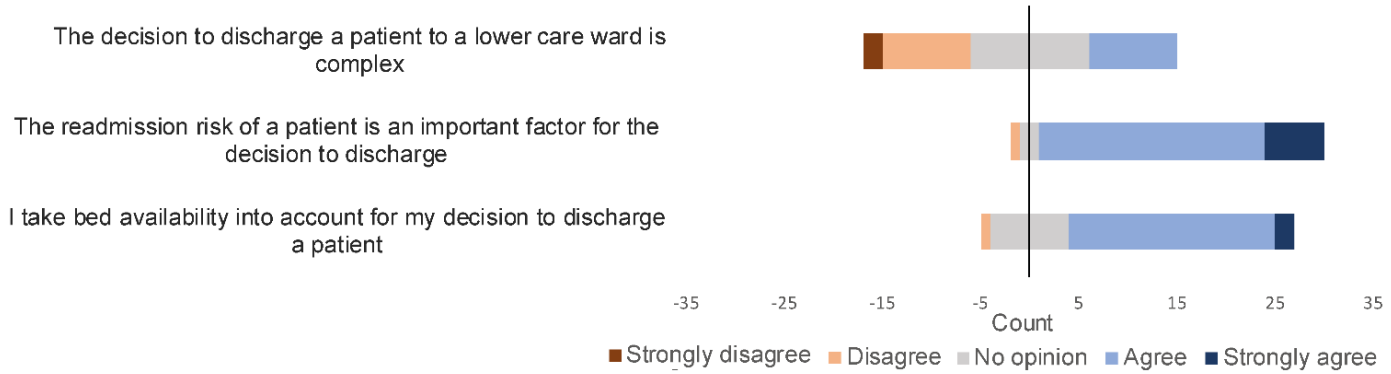


Figure 11: Statements on current discharge practices. The width of each bar indicates the number of respondents that voted for that option.

3.2. Current discharge practices

Statements regarding current ICU discharge practices of the physicians of the LUMC are visualized in Figure 11. The physicians were inconclusive on the statement regarding the complex nature of the decision to discharge. A patient's readmission risk is an important factor in their decision to discharge a patient and so is bed availability. This finding implies that a prediction tool predicting a patients readmission risk could be of clinical value when having limited bed availability.

One or more patient groups could be indicated by the participants for whom the decision and/or timing to discharge a patient to a lower care ward is perceived as most challenging, see Figure 12. Long admitted patients (75%, n = 24) and previously readmitted patients (59%, n = 19) were most often mentioned. The respondents were asked to indicate what their definition of a 'long ICU admission' was, resulting in an average of 17.6 ± 6.9 days on the ICU. Nine participants filled in one or more other patient groups, including severe weakness, reduced coughing strength, severe heart failure, afternoon or evening discharge, hematologic comorbidity, complex surgical patients, patients with no-return policy, and patients with unknown diagnosis. Furthermore, we questioned the average certainty a patient will not be readmitted after discharge (Figure 13). On a scale between 0 = not confident and 10 = fully confident, average confidence was 7.5 ± 0.9. This finding implies that the physicians discharge patients with some risk of readmission in mind.

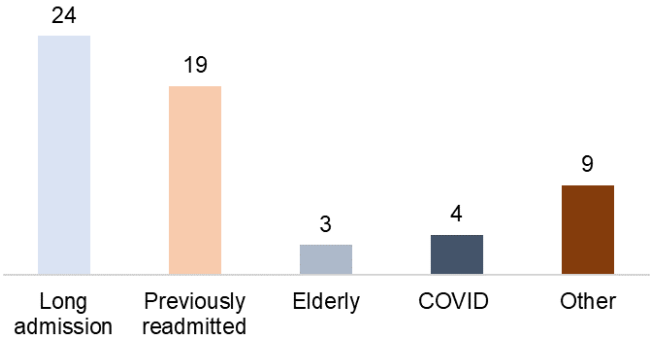


Figure 12: Patient groups for which the decision to discharge is perceived as most challenging.

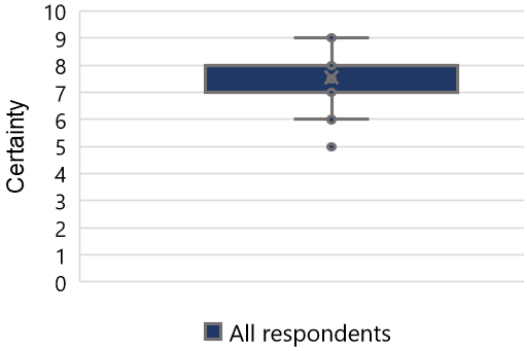


Figure 13: Average certainty a patient will not be readmitted to the ICU at moment of discharge between 0 (not confident) and 10 (fully confident).

3.3. Artificial Intelligence for discharge decision support

We asked the physicians to give their opinion on seven statements regarding the use of AI for ICU decision support, specifically for the prediction of readmission (Figure 14). Most ICU physicians of the LUMC are familiar with the concept of AI and believe that AI could support them in their work. A clear finding is that none of the respondents is afraid that AI would make their jobs unnecessary. Furthermore, 62% of the participants were neutral regarding the statement that AI understands their work sufficiently in order to support them. However, most physicians do believe in the positive value of AI-based decision support. It is important for them to have insight in the contributing factors to the patient's readmission risk. This implies the need for an explainable algorithm.

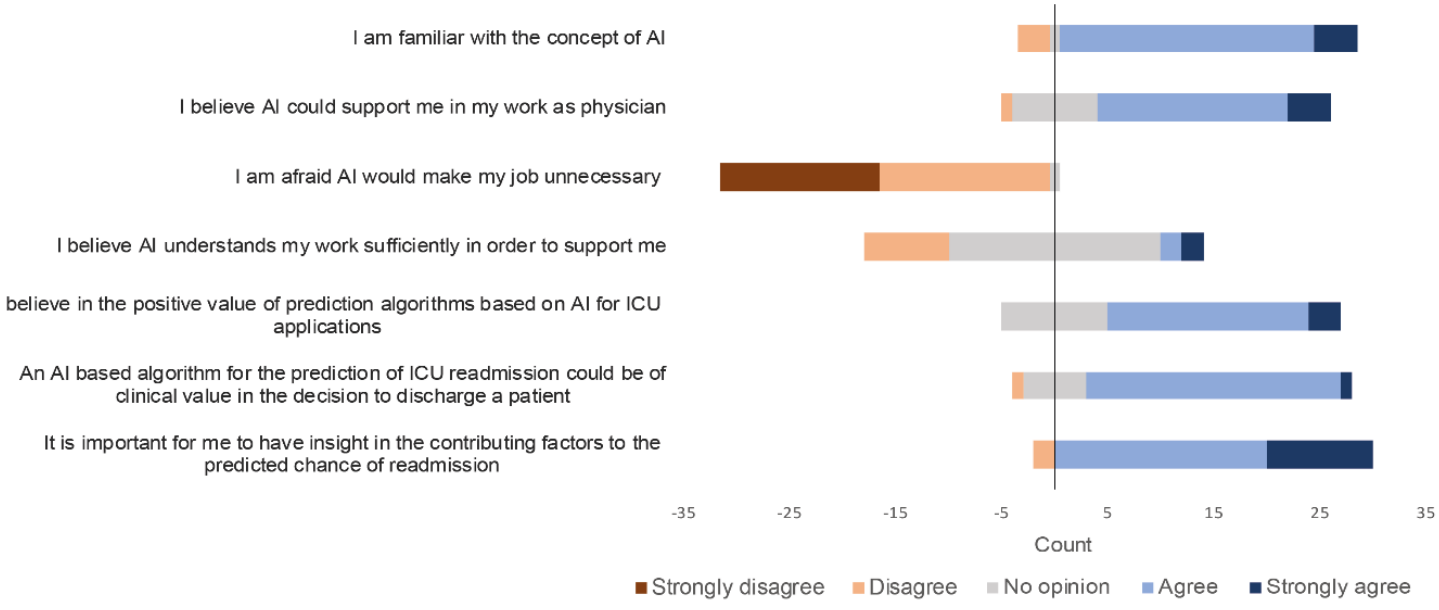


Figure 14: Statements regarding the physicians' opinion on the use of AI decision support, specifically predicting the patient's chance of readmission.

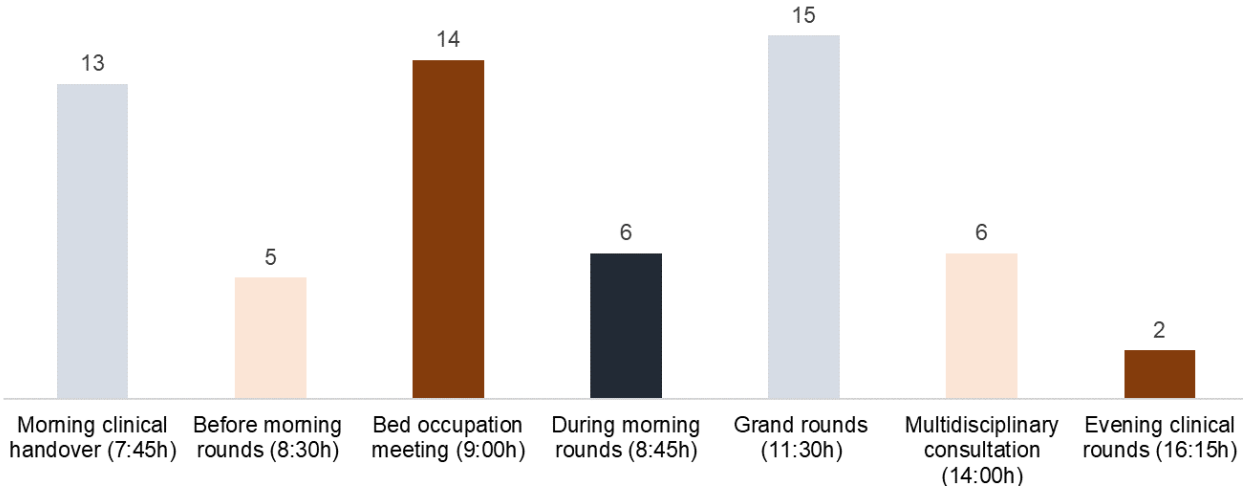


Figure 15: Preferred moments of implementing the discharge decision support tool.

3.4. Workflow preferences and outcome of interest

One or more preferred moments for displaying the discharge support tool could be indicated by the physicians, see Figure 15. The morning clinical handover, the bed occupation meeting, and the grand rounds were most often chosen. The ICU of the LUMC has a step-down ward, the medium care unit (MCU). Most physicians (84%, n = 27) would like to have the tool also to be deployed for MCU patients.

Next, the physicians' preferred predicted outcome measure was questioned (Figure 16). Half of the respondents were most interested in the patient's readmission risk alone. Pacmed Critical predicts the combined outcome of readmission and/or mortality within 7 days after discharge. These results indicate the need for Pacmed to have further evaluation with the physicians for the appropriate predicted outcome measure.

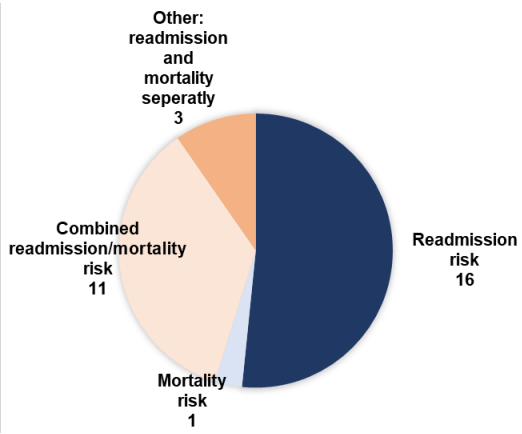


Figure 16: Outcome prediction of interest for discharge decision support.

The participants were questioned for what predicted probability of readmission or higher they would not discharge a patient, and at what predicted probability or lower they would discharge a patient. The aim of this question was to have a first indication on what the influence of a certain predicted chance of readmission would have on the physician's decision to discharge an ICU patient. The results were highly distributed, see Figure 17. On average, a predicted probability of 44.5 ± 23.4% or higher would result in postponing a patient's discharge. A predicted probability of 23.6 ± 13.8 % would result in discharge of the ICU patient. Three physicians smartly indicated that they could not fill in this question, because the readmission risk will always be taken into account with other patient factors. These results imply that the physicians need to gain experience on what is high, and what is low risk of readmission.

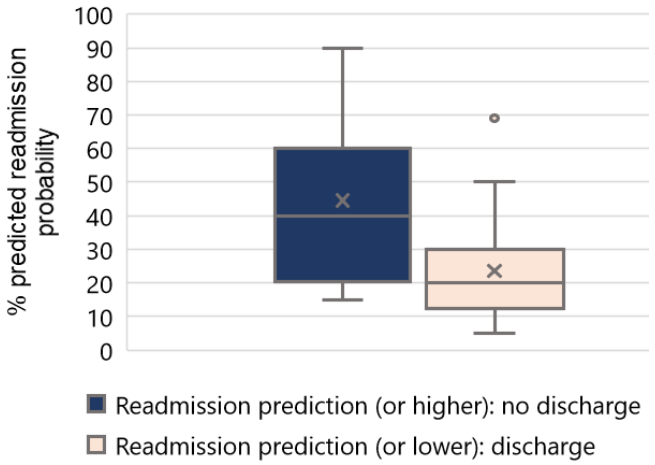


Figure 17: Hypothetical influence of predicted chance of readmission on the decision to discharge a patient.

4. Discussion

The questionnaire results gave valuable insights in the ICU physicians' discharge practices and their attitude towards the use of AI discharge decision support in their work processes. Although not all physicians consider the decision to discharge an ICU patient to a lower care ward to be complex, they do believe in the clinical value of a tool predicting a patient's chance on readmission. Pacmed Critical should be deployed on long admitted patients and previously readmitted patients to be of value for the physicians, since the decision to discharge is considered to be most complex for these groups. Furthermore, it is of importance to have insight in the patient factors underlying the prediction, highlighting the need for explainable decision support software. Further evaluation should be performed regarding the optimal outcome measure, moment, and time to implement Pacmed Critical in the daily ICU workflow.

Compared to previous research on ICU physicians' AI readiness, the physicians of the LUMC were more familiar with AI and less afraid that AI would make their jobs unnecessary[4]. Many useful suggestions on other areas that the physicians would like to have an AI prediction tool for were mentioned, indicating their awareness of the potential benefit of these tools. However, some physicians questioned the need for a discharge decision support tool and doubted whether it would be more accurate than their gut feeling. A prospective trial comparing the predictive performance of the physician to that of the algorithm should answer this question (see Part IV of this thesis).

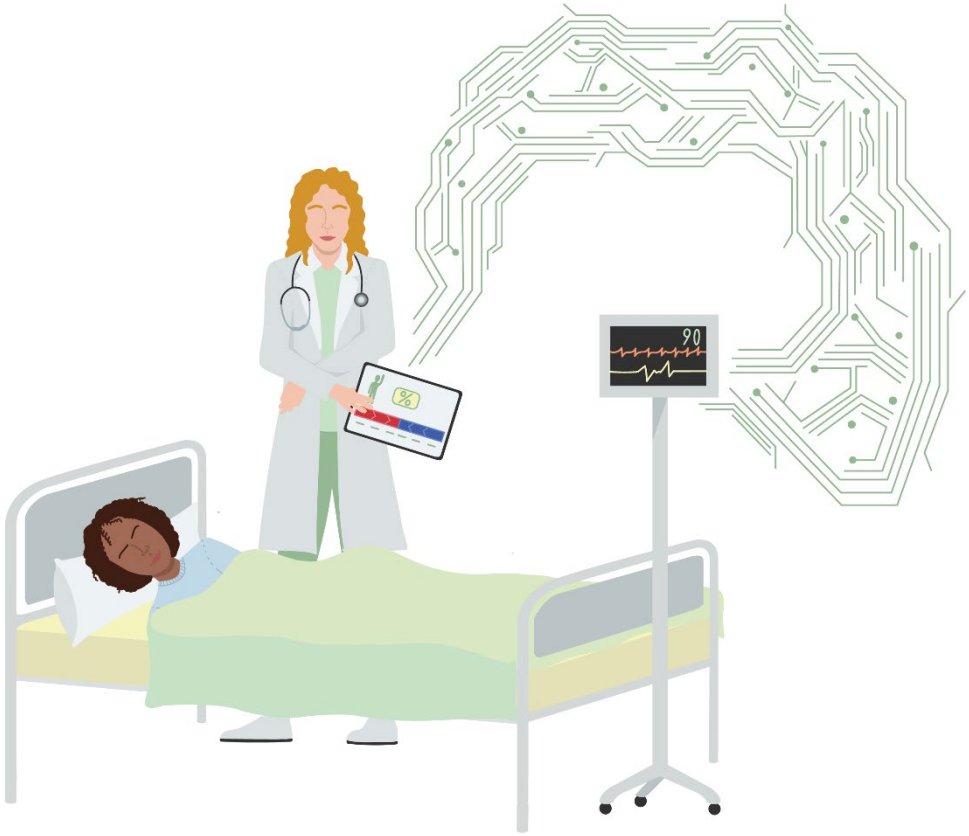
References

1. Pacmed. Intensive Care. [Internet]. Available from: https://www.pacmed.ai/nl/projects/ic. [Accessed December 29, 2020]
2. Sarwar, S., Dent, A., Faust, K. *et al.* Physician perspectives on integration of artificial intelligence into diagnostic pathology. *npj Digit. Med.* 2, 28 (2019).
3. Likert R. A technique for the measurement of attitudes. Arch psychology. 1932;22(140):55.
4. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019;21:e12422

Part III – Model Development & Validation

3

For the third and main thesis objective, we compared several machine learning on predictive performance and explainability for the prediction of ICU readmission within the LUMC. This scientific paper is the main end-product of this master thesis (TM30004). Additional modeling methods can be found in the Supplementary material Part III – Model development (page 78).



Predicting Intensive Care Unit readmission: *Performance and explainability of machine learning algorithms*

Abstract

Background Prediction of intensive care unit (ICU) readmission could support physicians in determining optimal timing for ICU discharge. Besides having high discriminative performance, prediction models need to be explainable to create trustworthy decision support. 'Black-box' machine learning (ML) models have previously outperformed logistic regression (LR), but this typically comes at the cost of explainability. To our knowledge, this is the first work comparing discriminative performance, calibration properties, and explainability between LR and state-of-the-art ML models.

Methods Boosting (XGB), neural network (NN), and LR models were trained on adult ICU patient data. Performance and calibration of ICU readmission predictions within 7 days after discharge were evaluated on a separate test dataset. The top 20 impactful features were compared between final models for explainability assessment using SHapley Additive exPlanations (SHAP) values for ML and LR coefficients. Lastly, clinical validity of models was evaluated by a panel of two ICU physicians.

Results 12,189 admissions could be included for analysis. The readmission rate within 7 days after ICU discharge was 6.7%. Final model area under the precision recall curve was 0.18 for LR and 0.17 for NN and XGB. Area under the receiver operating characteristic curve was 0.74 for all final models. Calibration properties improved after scaling the predicted probabilities, with final brier scores of 0.06. SHAP values for XGB and LR coefficients both enabled distinctive model explanations. Respiratory rate, urea levels, and C-reactive protein levels were impactful variables amongst all model types. XGB explanations were most in line with clinical reasoning according to expert opinion.

Conclusion Given the small differences in discriminative performance of XGB, NN, and LR models, explainability is of major importance in determining what model to implement for trustworthy decision support. SHAP enabled 'black-box' ML to achieve comparable explainability to LR, with impactful features more clinically relevant for XGB than LR. Future work should add additional data during model development to achieve better predictive performance, and focus on performing prospective trials to enhance meaningful decision support.

Keywords Intensive care unit, readmission, machine learning, explainability, interpretability, decision support

1. Introduction

Intensive Care Units (ICU) are dealing with limited bed availability and expensive resources, resulting in the need to discharge patients to the general hospital ward as soon and as safely possible. General wards, however, have limited monitoring and therapeutic options compared to the ICU. Due to this 'treatment gap', deterioration of a patient at the ward might be noticed late, resulting in clinical deterioration and in unplanned ICU readmission, or even death[1]. Readmission to the ICU is associated with increased mortality rates (26-58%), longer hospital stays, and higher costs, and should therefore be prevented[2]. A decision support tool that identifies patients in the ICU at high risk of readmission, could be beneficial to help the physician determine the optimal timing of discharge. In combination with clinical judgment, discharge could be postponed for patients with high readmission risk and be continued or advanced for patients with low readmission risk. Ultimately, this would result in the prevention of ICU readmissions and unnecessarily long ICU admissions.

Widespread implementation of electronic health records resulted in an increased availability of patient data, that can be used to build advanced prediction models[3]. For the prediction of ICU readmission and other healthcare-related outcomes, a shift in modeling methods was observed over the last five years from logistic regression[4] (LR) to more advanced machine learning (ML) models[5]. Compared to LR, ML can discover non-linear relations between variables, potentially resulting in higher predictive performance for the heterogeneous ICU population. Two state-of-the-art model types that have shown superior predictive performance over LR for the prediction of readmission include boosting algorithms and neural networks[6].

The potentially superior predictive power of neural networks and boosting algorithms comes at the cost of reduced explainability for the end-user due to their 'black-box' nature. Explainability of a prediction algorithm is of major importance for several reasons: 1) it provides interpretation and bias detection during model development, 2) it can discover relations between clinical variables and patient outcomes, and 3) it enhances trustworthy clinical decision support by explaining the prediction to the physician[7]. The need for explainable models resulted in the development of post-hoc algorithms that open up the 'black-box' by means of visualizing the variables' relations to the predicted outcome[8,9]. A visualization of explainable decision support for ICU discharge is provided in Figure 18.

A comparison between several types of state-of-the-art ML models and LR for the prediction of ICU readmission has been performed in previous studies[10–14]. However, in a systematic review we concluded that these studies use different, and sometimes inappropriate performance metrics to assess discriminative power and calibration measures[15]. Furthermore, the comparison between two specific types of high performing ML (boosting algorithms and neural networks) and LR has not been performed on the same patient dataset.

Apart from the limitations in performance evaluation of different model types, model explainability has not been compared for the prediction of ICU readmission. Assessment of explainability is more challenging and less objective than for discriminative performance, since the preference of the end-user (the ICU physician) needs to be taken into account[16]. To achieve fair comparison between models, we evaluated the most impactful variables among models, complemented with expert (ICU physician) opinion. This enabled us to discover risk factors associated with high readmission risk. Due to previously observed small differences in discriminative performance between models[15], we hypothesize that the most trustworthy model should be the one with high performance and explainability most in line with clinical reasoning. We aimed to predict readmission after ICU discharge using LR and state-of-the-art ML models. Secondly, we aimed to compare model performance using suitable metrics and to compare explainability outcomes (in terms of impactful patient factors to the predicted outcome) to assess clinical applicability and to identify factors associated with increased readmission risk.

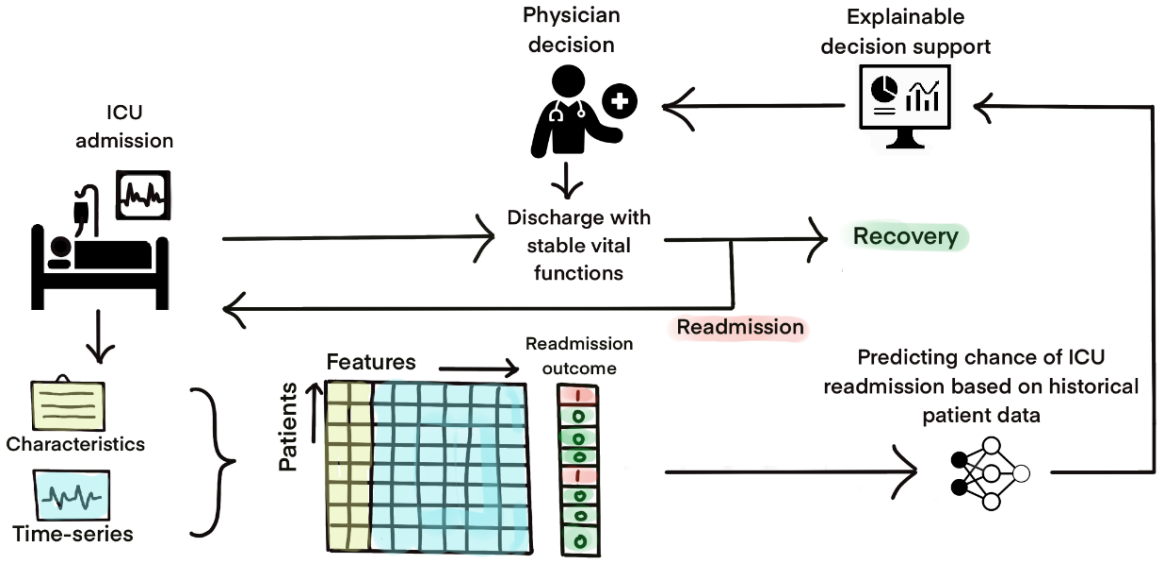


Figure 18: Schematic representation of explainable decision support by means of predicting an ICU patient's readmission risk. After ICU discharge, a patient can either recover, or deteriorate at the ward leading to ICU readmission. Patient characteristics, time-series variables (e.g., blood pressure), and readmission outcomes (1 = readmission, 0 = no readmission) collected over time are used to train prediction models. These collected variables first need to be transformed to meaningful features during pre-processing and feature engineering. For each new admission, this step is performed in order to enable the instantaneous prediction of readmission based on retrospectively trained models. The prediction should be explained to the physician by showing the contributing patient factors to the predicted outcome in order to have clinically valuable decision support.

2. Methods

We predicted ICU readmission using retrospective data from the Leiden University Medical Centre (LUMC), a 900-bed tertiary teaching hospital in the Netherlands. Modeling methods were reported according to the TRIPOD guidelines (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis)[17]. A more detailed description of modeling methods is provided in the Supplementary material Part III – Model development, page 78.

2.1. Source of data and sample size

All admissions between January 2011 and January 2020 to the ICU of the LUMC were evaluated for inclusion. We extracted and pseudo anonymized patient data from the Patient Data Management System (PDMS, MetaVision version 5 and 6, IMDsoft, Tel Aviv, Israel) and the Electronic Health Record (EHR, HiX, Chipsoft, Amsterdam, The Netherlands).

2.2. Participants

We evaluated all adult (> 18 years) ICU admissions to train and validate ML models for the prediction of ICU readmission after discharge. ICU admissions shorter than 12 hours and admissions with a direct transfer from the ICU to another hospital were excluded from analysis. We included admissions with successful discharge to the general ward or home. This resulted in the exclusion of patients that deceased at the ICU. Data regarding do-not-resuscitate and do-not-intubate orders were unavailable. Consequently, these admissions could not be excluded from analysis, which has potentially resulted in bias as these patients will not be readmitted to the ICU. We did not take into account admissions starting in 2020, due to changing circumstances (e.g., bed availability, data storage, use of other monitors) influenced by COVID-19.

2.3. Outcome

The predicted outcome was defined as readmission to the ICU or Medium Care Unit (MCU) within seven days after discharge from the ICU to a general hospital ward or home, corresponding to Thoral et al. Using seven days as readmission target, and not the ICU quality indicator of readmissions within 48 hours, enabled us to have sufficient readmission cases to train or ML models on. Furthermore, some complications resulting in readmission, e.g., sepsis or respiratory failure, typically occur later than within two days after discharge. All included admissions were labelled for the outcome of readmission (0 = no readmission (recovery), 1 = readmission), that were used to train the prediction models on.

2.4. Included predictors and feature engineering

Due to data storage and access difficulties, not all variables were available at time of model development. An overview of patient variables is provided in Table 1. Variables can be divided in static variables (patient characteristics) and time-series variables (laboratory results, vital signs, and medication data). Time-series variables are measured multiple times during an admission and have different sample frequencies. Feature engineering as described by Thoral et al.[12] was performed to capture descriptive aspects of time series variables. For this purpose, three time windows were used: the first 24 hours of the admission, the last 24 hours of the admission before discharge, and the complete admission period. For each time window, multiple aggregates were calculated. Aggregates are summary statistics that transform time-series variables into uniform features for each patient. A graphical representation of the performed feature engineering is shown in Figure 19. An overview of feature statistics, missing data, and feature engineering details is provided in the Supplementary material Part III – Model development (page 78 - 84).

2.5. Missing data and pre-processing

The occurrence of missing data is unavoidable when working with PDMS and other sources of EHR data. Three types of missing data exist: missing completely at random, missing at random, and missing not at random[18]. All types are likely to be present in EHR data. To model any information present in the missingness itself, we chose to add a feature to indicate for each time window whether the feature was missing[19]. Due to the transition between two PDMS systems and data storage difficulties, a large amount of data was missing for some variables. Therefore, we excluded variables with more than 50% missing data, that could not be explained by domain knowledge of parameters missing not at random (e.g., we did include medication that is given to only 5% of the ICU patients based on clinical judgement). Imputation of remaining missing missing values was performed using mean imputation for numerical variables and mode imputation for categorical features[20]. Other methods of imputation,

Table 1: Included variables for the prediction of ICU readmission per category. Variables were included in the final model when an acceptable level of missing data was present. PEEP = Positive End-Expiratory Pressure, NIBP = Non-invasive Blood Pressure, ABP = Arterial Blood Pressure (invasive), FiO2 = supplied oxygen content, CVP = Central venous pressure, PO2 = oxygen pressure, PCO2 = carbon dioxide pressure, PT = prothrombin time, MCV = mean corpuscular volume, LDH = lactate dehydrogenase, Gamma GT = Gamma-glutamyl transferase, CRP = C-reactive Protein, CK = Creatinine Kinase, BSE = erythrocyte sedimentation rate, BE = base Excess, ASAT = Aspartate aminotransferase, ALAT = Alanine aminotransferase, APTT = Activated partial thromboplastin time.

Category	Variable	Included in final model	Category	Variable	Included in final model
Patient characteristics	Age	✓	**Laboratory results**	Total protein	
	Gender	✓		PT	✓
	Emergency admission	✓		O2 saturation	
	Hospitalization admission source	✓		Neutrophil granulocytes	
	Treating specialty	✓		Potassium	✓
	Length of stay (ICU)	✓		Sodium	
	Length of stay prior ICU	✓		Magnesium	
Medication	Dobutamine	✓		MCV	✓
	Noradrenaline	✓		Leukocytes	✓
	Milrinon/Enoximon	✓		Lactate	✓
	Adrenaline	✓		LDH	
Vital functions	Oxygen flow	✓		Creatinine	✓
	Tidal volume			Ionized calcium	
	Temperature			Haemoglobin	✓
	SpO2			Glucose	
	Peak pressure			Gamma GT	
	PEEP			Chloride	
	NIBP			CRP	✓
	ABP	✓		CK	
	Heartrate			Bilirubin	✓
	FiO2			Bicarbonate	
	CVP			BSE	
	Respiratory rate	✓		BE	✓
Laboratory results	PO2			Amylase	
	PCO2			Alkaline phosphatase	
	pH			Albumin	
	Inorganic phosphate			ASAT	✓
	Urea	✓		ALAT	✓
	Troponin T			APTT	✓
	Thrombocytes	✓			

Including nearest neighbour and median imputation, did not result in improved predictive performance. Categorical features (e.g., hospitalization admission source) were transformed to numerical data using one-hot encoding.

Continuous numerical features (e.g., age, respiratory rate) were standard-scaled with zero mean and unit variance. A detailed description of the applied pre-processing methodology is provided in the Supplementary material Part III – Model development (page 84). Feature engineering of the included variables resulted in a total of 550 features per patient. We used logistic regression L1-feature selection as described by Thoral et al.[12] to select the most informative features. L1-feature selection is used to reduce overfitting on redundant features in the training dataset, by shrinking the coefficient of the least informative features to zero. Feature selection resulted in 416 features per patient used for model development.

2.6. Model development

For prediction model development we used three model types: logistic regression (LR), boosting algorithms (Gradient Boosting machines[21] (GB) and XGBoost[22] (XGB)), and feed-forward neural networks (NN). LR is a statistical method which is known to be highly explainable compared to boosting and NN algorithms[9,23]. However, previous research on the prediction of readmission showed that this often comes at the cost of predictive performance[10,11,14]. Model development was performed in Python, using SKlearn[24], Tensorflow[25], and Keras packages[26].

The included patients and their corresponding recorded variables in the dataset were split in 80% training and 20% test datasets. The test dataset was not used until final evaluation of the models, to prevent data-leakage and overfitting[27]. The training dataset was further split into random, stratified five-fold cross-validation (CV) sets to perform hyperparameter tuning. Stratification was performed to balance the proportion of readmitted patients in all datasets. A Bayesian optimization strategy[28] was applied to find the optimal hyperparameters for each model. Hyperparameters are the internal settings of the machine learning or deep learning algorithm, for instance the number of layers in a NN. We used the area under the precision recall curve (AUCPR) as hyperparameter tuning objective metric. To account for imbalance in the predicted outcome, we applied weighted learning for all models, except for GB in which this could not be implemented[29]. See the Supplementary material Part III – Model development (page 84-86) for an overview of our performed hyperparameter tuning. The models with optimized hyperparameters will be referred to as 'final models'.

2.7. Performance evaluation

A small proportion of patients was readmitted to the ICU, which makes the prediction of readmission an imbalanced classification problem[30]. Therefore, multiple metrics were used to evaluate predictive performance, since accuracy and area under the receiving operator curve (AUC) are not suitable metrics alone[31]. E.g., a model could predict with a 95% accuracy by predicting all patients to be not readmitted (true and false negatives) when the proportion of readmissions in the dataset is 5%. The AUCPR, recall (sensitivity), precision (positive predictive value), specificity, F1-score, brier score, and Matthews correlation coefficient (MCC) were compared between all final models. Mean (± standard deviation (SD)) performance was evaluated for all final models evaluated on the CV set, and final performance was evaluated on the test dataset.

2.8. Calibration

Calibration is the degree of agreement between the prediction and actual observation. For instance, for a perfect calibrated model, out of 100 patients with a predicted 10% chance of readmission, the actual outcome should be that 10 out of these 100 patients are actually readmitted. Predicted probabilities for the risk of readmission are known to be not correctly calibrated for ML algorithms[32]. To increase clinical utility, it is important to display the correctly calibrated probability for the outcome of a specific patient. Calibrating of the predicted outcomes was therefore performed by Platt scaling[33] the predicted probabilities on the test dataset, based on the predictions made on the training dataset.

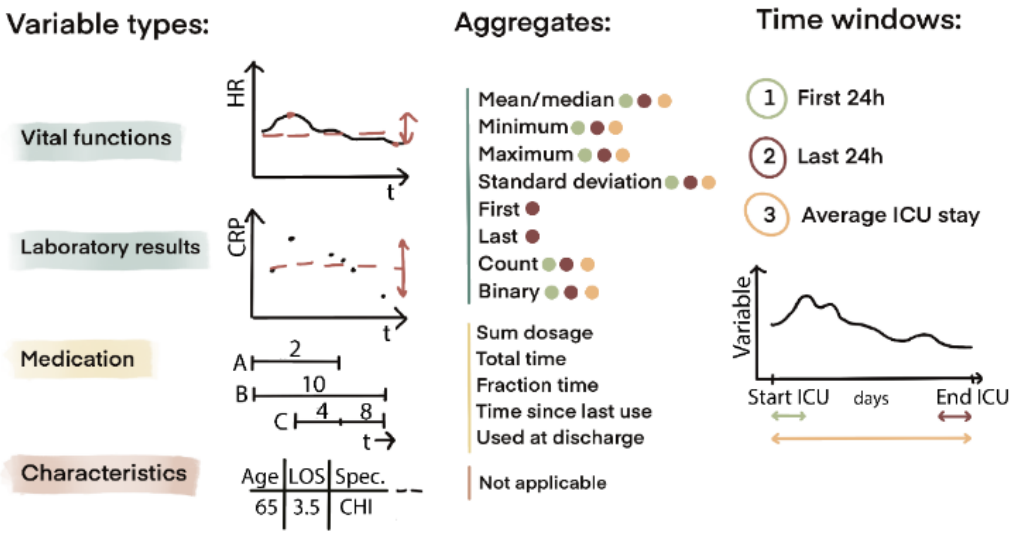


Figure 19: Feature engineering for time-series variables. Time-series variables are measured multiple times during an ICU admission with different sample frequencies. In order to get functional input features for prediction modeling, time-series features are summarized in aggregates. Three time-windows are used for the aggregates, the first 24h of the admission, the last 24h of the admission, and the complete ICU admission.

Calibration properties were evaluated in calibration plots. Brier scores were used to assess the accuracy of the probabilistic predictions i.e., the mean squared error between the observed readmission rate and the predicted chance of readmission[34].

2.9. Sensitivity analysis

We performed a sensitivity analysis on patient subgroups by means of comparing final model performances on AUC and AUCPR. This enabled the identification of patient groups for which the prediction of readmission was less accurate. Several subgroups were analysed. A division was made between medical and surgical patients (including thoracic surgical (CTC), general surgical, and neurosurgical patients). A large proportion of the admissions in our dataset were CTC patients (66% in the test dataset) which were admitted to the ICU for observation and treatment after surgery. Therefore, we also compared CTC to all other patients (including general surgical patients).

2.10. Model explainability

In order to evaluate model explainability, we compared clinical feature importance for the prediction of ICU readmission. LR is known as an explainable algorithm, since the scaled values of the coefficients can be used to assess which features are most important. Because all input features were scaled with zero mean and unit variance, absolute values could be used for evaluation of LR coefficients. We used SHAP[8] (Shapley Additive exPlanations) values to identify the most important features for boosting (SHAP Tree Explainer) and NN (deepSHAP) algorithms[35]. SHAP is a model-agnostic method for 'black-box' model explanations, which breaks down the prediction to evaluate the impact of each feature. Model-agnostic means that SHAP can be used to explain any algorithm by only evaluating the effect of changes in input features on the predicted outcome. The influence on the predicted outcome is determined by introducing each feature individually and by averaging the contribution over all patients. The total impact of a variable (e.g., all blood pressure features together) cannot be directly assessed by looking at the SHAP values, because we used feature aggregates to capture time related trends. However, SHAP values may be summed to globally evaluate total variable contributions. We used SHAP because compared to other model-agnostic methods (e.g., LIME[36]) for model explanations, it is known to have stronger agreement with human reasoning[8].

The 20 most important features according to the model's absolute coefficients (LR) and absolute SHAP values (highest performing boosting model, NN) were compared for the test dataset. Furthermore, one individual patient prediction was displayed in a SHAP force-plot to look for similarities and differences. Expert opinion (two critical care physicians) was reported to gain insight in which model was most in line with clinical reasoning. The experts were asked to indicate for each feature whether it was contradicting, irrelevant, or relevant for a patient's readmission risk (Supplementary material Part III – Model development, page 87).

3. Results

3.1. Participants and inclusion of ICU admissions

15,749 ICU admissions were identified in the dataset, of which 12,189 could be included for training and validating the model. Main reasons for exclusion were length of stay (LOS) < 12 hours (n = 1,223), patients that did not survive their ICU admission (n = 1,600), and direct transfer from the ICU to another hospital (n = 651). A flow chart of patient exclusion is presented in Figure 20. The readmission rate within 7 days after discharge was 6.7% (n = 820). Patients readmitted to the ICU were compared to non-readmitted patients more often emergency patients (49.51% vs. 36.80%, $p < 0.0001$), had longer ICU LOS (2.22 vs. 1.03 days, $p < 0.0001$), and had longer hospital LOS prior ICU admission (1.33 vs. 1.15 days, $p < 0.0001$). A summary of patient characteristics is given in Table 2. There were no significant differences in gender, age, and vasoactive drug use. The included readmitted patients were compared to the included non-readmitted patients: more often surgical patients (21.17% vs. 11.45%, $p < 0.0001$), more often internal medicine patients (8.64% vs. 6.53%, $p < 0.024$), and less often thoracic surgical patients (31.14% vs. 56.27%, $p < 0.0001$). All included ICU admissions were randomly allocated to the training dataset (n = 9,751, 80%) or the test dataset (n = 2,438, 20%). Descriptive statistics and missing data information per included variable can be found in the Supplementary material Part III – Model development (page 78-82).

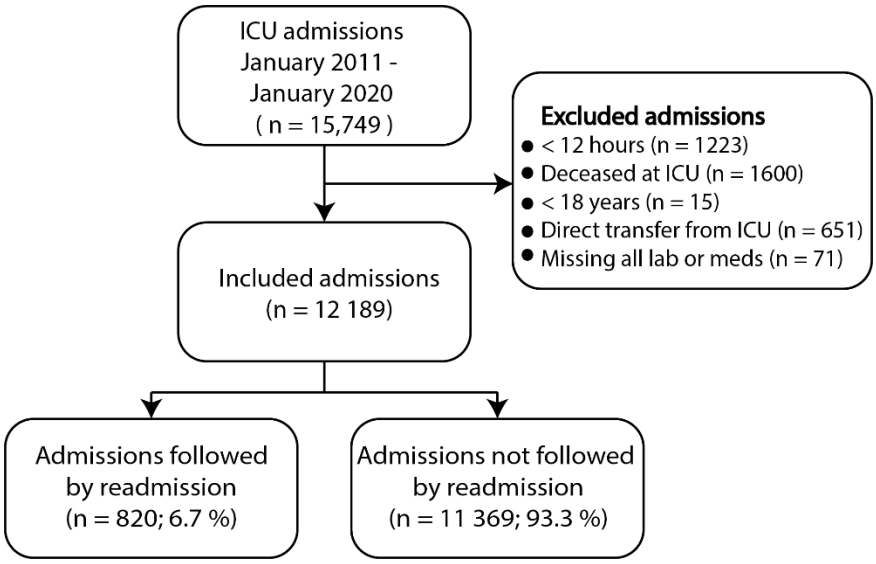


Figure 20: Flowchart of included and excluded ICU admissions.

3.2. Model specifications and performance

Hyperparameter tuning yielded the final model configurations of LR, XGB, GB, and NN models. Despite the fact that not all variables of interest were available at time of model development, acceptable performance was obtained for most models. Highest cross-validation (CV) AUCPR (0.17 ± 0.02), and MCC (0.19 ± 0.02) was obtained for both the LR and XGB model with a marginal higher performance for the LR model on the test dataset. Performance metrics are summarized in Table 3. Although GB had highest accuracy, specificity, and precision, this model was not useful due to its low recall (CV 0.03 ± 0.02, test 0.05). This indicates the model's inability to deal with class imbalance and the classification of most admissions as 'no readmission'.

Precision of all final models was low, which indicates a large proportion of false positives. Recall was highest for the NN (CV 0.70 ± 0.11, test 0.66). NN, LR, and XGB had a comparable AUC of 0.74 ± 0.02, indicating fair predictive performance. See Figure 21 for receiver operating characteristic and precision recall curves. The small differences in performance between the NN, LR, and XGB models indicate the need for explainability to assess clinical usefulness of each model. Due to the low recall of the GB model, further evaluation (calibration and explanation) was not performed for this model.

Table 2: Basic patient characteristics in percentages or median (Inter Quartile Range (IQR)). P-values were calculated using Wilcoxon Rank test and chi-squared test where appropriate. Variables marked in orange were statistically different between groups (p < 0.05).

Variable	Readmission	No readmission	p-value
Patients (%)	6.71	93.29	-
Women (%)	36.98	34.53	0.164
30-day mortality (%)	19.83	4.41	< 0.0001
Vasoactive drugs use (%)	68.73	67.82	0.616
Emergency admission (%)	49.51	36.8	< 0.0001
General surgical patients (%)	21.17	11.45	< 0.0001
Thoracic surgical patients (%)	31.14	56.27	< 0.0001
Internal medicine patients (%)	8.64	6.53	0.024
Neurosurgical patients (%)	9.85	8.06	0.081
Age in years (median (IQR))	65.12 (54.70 – 72.42)	64.88 (54.65 – 72.34)	0.711
LOS ICU in days (median (IQR))	2.22 (0.97 – 6.01)	1.03 (0.85 – 2.75)	<0.0001
LOS prior ICU in days (median (IQR))	1.33 (0.39 – 6.88)	1.15 (0.71 – 2.23)	<0.0001

Table 3: Final model performance in mean (standard deviation (SD)) after cross-validation (CV) on the training dataset. Final model performance is given on the test dataset. Best results on the test dataset are marked in orange. AUCPR = area under the curve precision recall, AUC = area under the receiver operating characteristic curve, MCC = Matthews correlation coefficient.

	Neural Network		Logistic Regression		Gradient Boosting		XGBoost	
	CV (mean (SD))	Test	CV (mean (SD))	Test	CV (mean (SD))	Test	CV (mean (SD))	Test
Train time (s)	9.94 (3.14)	-	0.21 (0.01)	-	119.3 (8.77)	-	12.14 (0.72)	-
Score time (s)	0.13 (0.11)	0.11	0.03 (0.00)	0.02	0.25 (0.05)	0.19	0.10 (0.01)	0.04
Accuracy	0.66 (0.04)	0.68	0.72 (0.01)	0.70	0.93 (0.00)	0.93	0.77 (0.02)	0.76
Precision	0.13 (0.00)	0.13	0.14 (0.01)	0.13	0.27 (0.17)	0.27	0.15 (0.01)	0.14
Recall	0.70 (0.11)	0.66	0.62 (0.04)	0.63	0.03 (0.02)	0.05	0.53 (0.02)	0.52
Specificity	0.65 (0.05)	0.68	0.72 (0.01)	0.71	0.99 (0.00)	0.99	0.78 (0.02)	0.78
AUCPR	0.16 (0.03)	0.17	0.17 (0.02)	0.18	0.15 (0.03)	0.16	0.17 (0.03)	0.17
F1-score	0.21 (0.01)	0.22	0.23 (0.01)	0.22	0.05 (0.03)	0.08	0.23 (0.01)	0.22
AUC	0.74 (0.02)	0.74	0.74 (0.01)	0.74	0.69 (0.02)	0.71	0.74 (0.01)	0.74
Brier	0.34 (0.04)	0.32	0.28 (0.01)	0.30	0.07 (0.00)	0.07	0.23 (0.02)	0.24
MCC	0.18 (0.03)	0.18	0.19 (0.02)	0.18	0.06 (0.06)	0.09	0.19 (0.02)	0.17

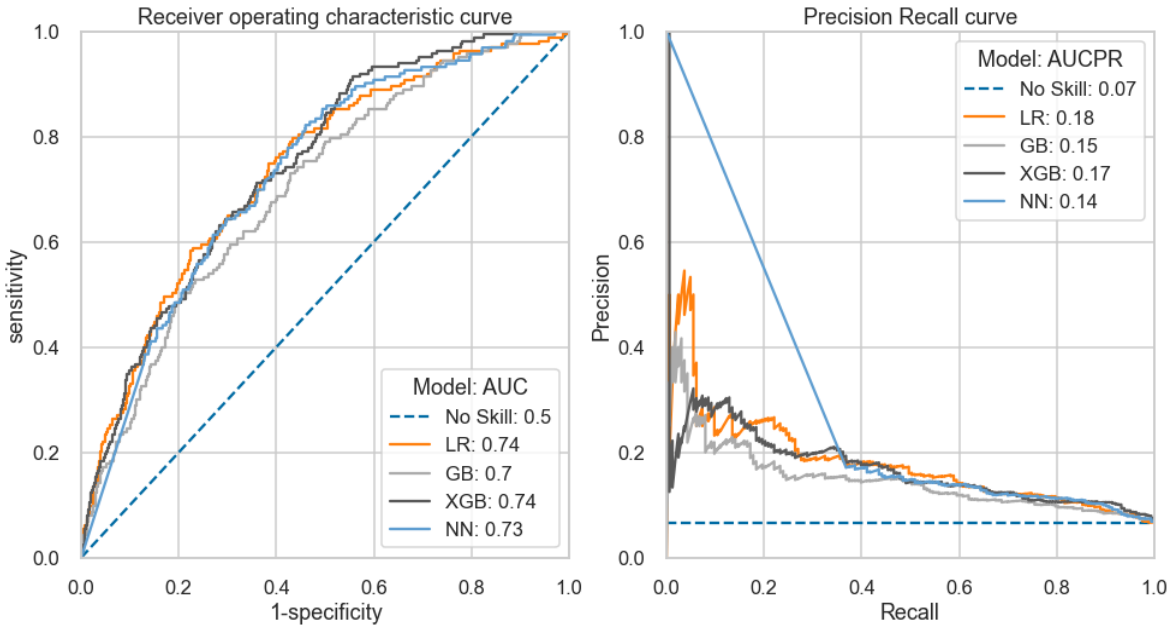


Figure 21: Receiver operating characteristics and precision recall curves. AUC = Area under the receiver operating characteristic curve, AUCPR = area under the precision recall curve, LR = Logistic Regression, GB = Gradient Boosting machines, XGB = XGBoost, NN = Neural Network.

3.3. Calibration

Uncalibrated predictions showed poor agreement between the predicted probabilities and the ratio of true positives (Figure 22, left). Similar observations were seen for the brier scores (Table 3). This finding can be explained by the low precision of the models. E.g., for the XGB model, for each 10 patients receiving a predicted chance of readmission of ~ 70%, approximately 3 patients will be actually readmitted to the ICU. Classification algorithms push a probability in the direction of 0 (no readmission) or 1 (readmission). However, the predicted probabilities will be used in clinical practice, indicating the need for correct calibration[37]. After Platt scaling the probabilities (based on the predictions in the training dataset), predicted probabilities showed better agreement (Figure 22, right). Brier scores after calibration were 0.06 for all models. The histogram in Figure 22 shows that calibration result in a larger proportion of low predicted probabilities, ranging between 0-40% instead of 10-90% pre-calibration.

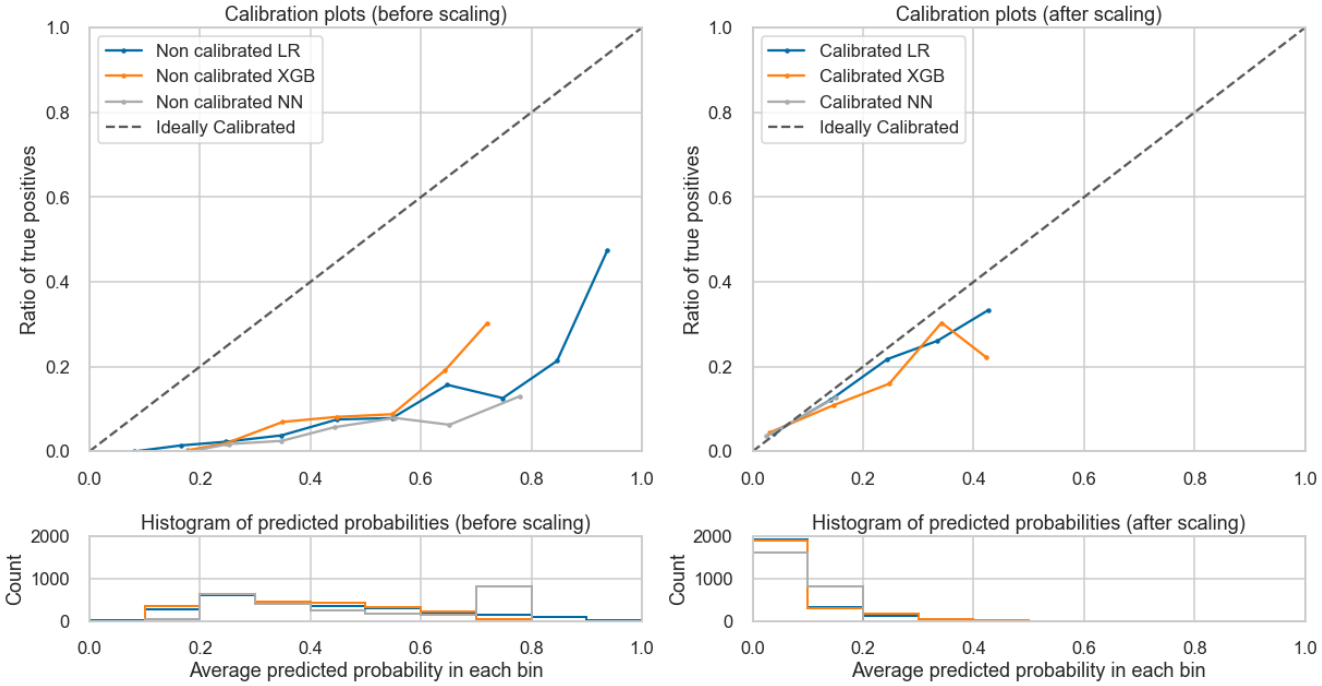


Figure 22: Calibration plots and histograms of predicted probabilities. Left: calibration plots of uncalibrated probabilities. Right: calibration plots after Platt scaling the predicted probabilities.

3.4. Sensitivity analysis

We evaluated test performances of LR, XGB and NN for the subgroups specified in Section 2.9, see Table 4. AUC and AUCPR were higher for the CTC group compared to the non-CTC group (including general surgical patients) for the NN and LR model. Another sensitivity analysis was performed for surgical and medical patients. Superior performance for all models was observed for readmission prediction for surgical patients. Differences in performance between the two groups was highest for LR and smallest for XGB.

3.4. Model explainability

3.4.1. Feature importance

Features contributing most to the prediction of readmission were evaluated for the LR, XGB, and NN models. For the LR model, the 20 largest positive and negative coefficients are displayed in Figure 23. The last measured respiratory rate before discharge had the strongest correlation with the chance of readmission. SHAP summary plots were used to evaluate the most predictive features for the XGB and NN models (Figure 24). For the XGB model, the SHAP values give clear explanations on which variables contribute to a higher and lower risk for readmission. For the NN, SHAP values are more centred around zero. It is important to note that SHAP values represent correlations and do not imply causality between features and predictions. Although differences can be observed between the most informative features across the three models, some similarities are apparent. Respiratory rate, urea (a kidney function marker), C-reactive protein (CRP, an inflammation marker), leukocytes, and Alanine aminotransferase (ALAT, a liver function marker) are in the top 20 of all three models (represented by one or more feature aggregates).

Table 4: Sensitivity analysis results. Test performance on patient subgroups. Highest performance is marked in orange. The patients in the test dataset were divided in CTC (thoracic surgical) versus non-CTC and surgical versus medical. AUCPR = Area under the curve precision recall, AUC = Area under the receiver operating curve.

	Neural Network		**Logistic Regression**		**XGBoost**	
Patient group:	*CTC*	*Other*	*CTC*	*Other*	*CTC*	*Other*
AUCPR	0.20	0.15	0.28	0.16	0.17	0.18
AUC	0.80	0.60	0.81	0.61	0.80	0.62
Patient group:	*Surgical*	*Medical*	*Surgical*	*Medical*	*Surgical*	*Medical*
AUCPR	0.19	0.13	0.25	0.14	0.18	0.16
AUC	0.74	0.60	0.75	0.64	0.77	0.66

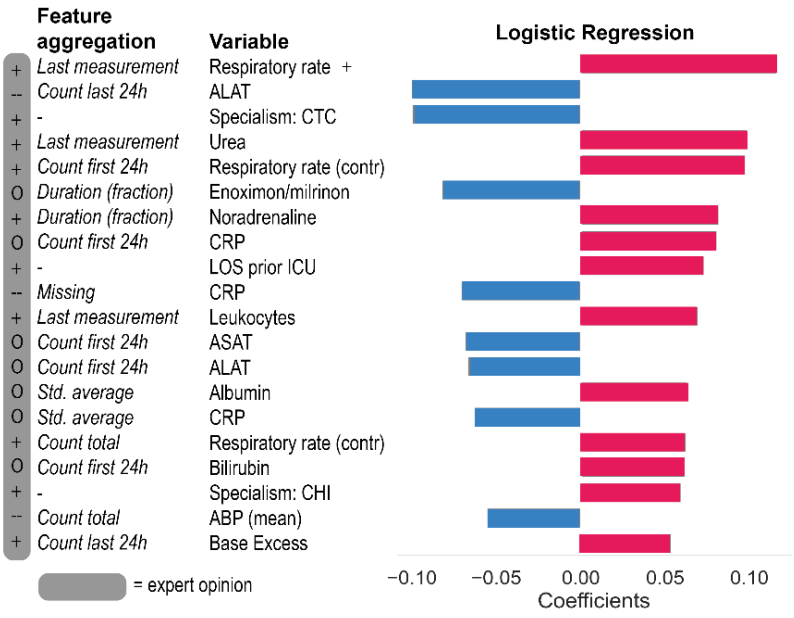


Figure 23: Top 20 logistic regression coefficients for features scaled with zero mean and unit variance. Features in red are correlated with increased risk of readmission and features in blue with decreased risk of readmission. Expert opinion is represented for each feature in the grey column: + = clinically relevant, O = clinically irrelevant, -- = contradictive with clinical practice. ALAT = Alanine aminotransferase, CTC = cardiothoracic surgery, CRP = C-reactive protein, LOS = Length of stay, ASAT = Aspartate aminotransferase, ABP = arterial blood pressure

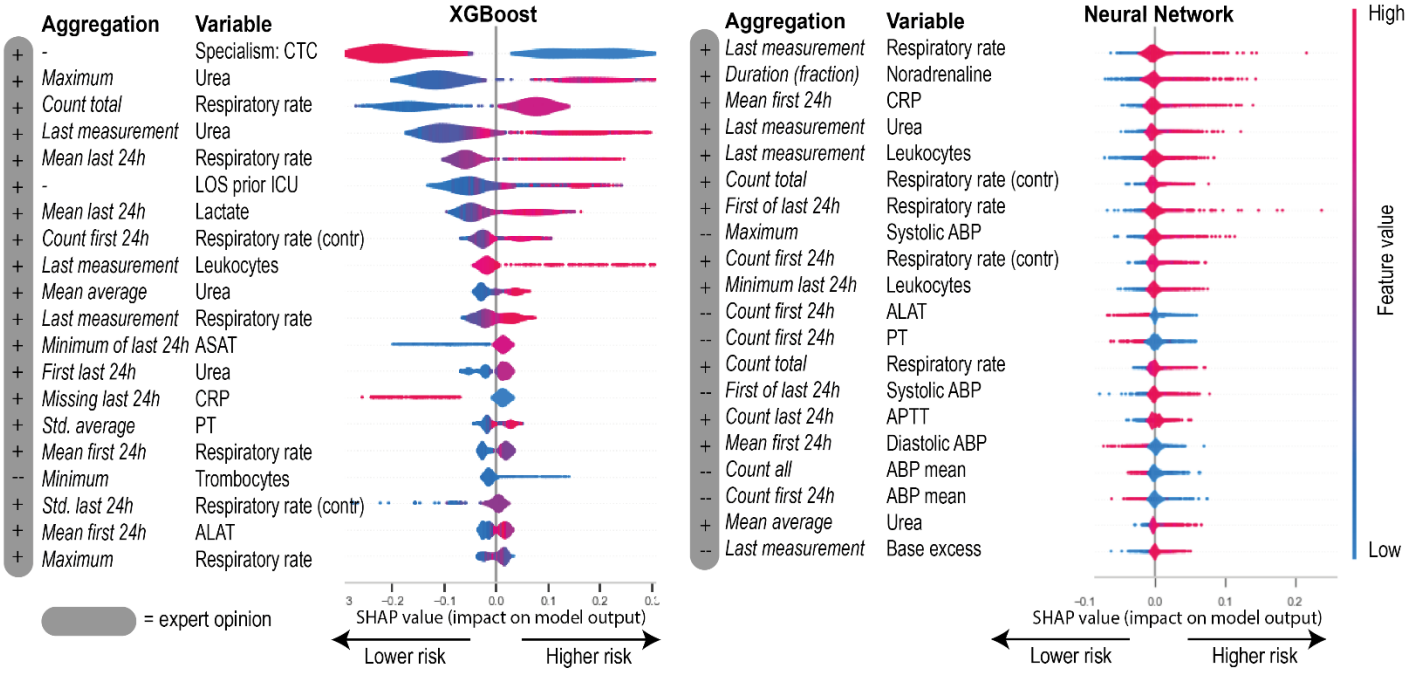


Figure 24: SHAP summary plots for the XGBoost (left) and neural network (right) model. The 20 most informative features are displayed. Each dot in the plot represents a patient with high (red) or low (blue) value for a specified variable. Dots on the right side of the y-axis indicate a correlation with high risk of readmission and vice versa. The thickness of the line is determined by the number of patients. Expert opinion is represented for each feature in the grey column: + = clinically relevant, O = clinically irrelevant, -- = contradictive with clinical practice. CTC = cardio-thoracic surgical, LOS = length of stay, ASAT = Aspartate aminotransferase, ALAT = Alanine aminotransferase, CRP = C-reactive protein, PT = prothrombin time, APTT = Activated partial thromboplastin time , ABP = arterial blood pressure.

3.4.2. Variable importance XGBoost and neural networks

As we used a total of 416 features for predictive modeling, it was infeasible to evaluate all these features for a given model. Therefore, average SHAP values of variables' aggregations were summed to gain insight in the variables most impactful to a certain model (Figure 25). Summed SHAP values are shown in absolute values (scaled between 0 and 1) and therefore only indicate the impact on the prediction and not the direction (higher or lower probability). Arterial blood pressure (ABP), urea, and respiratory rate are important features for both models. For the XGB model, patient characteristics (length of stay (LOS), specialism) are of greater importance than for the NN. Age, gender, adrenaline use, and emergency admission information have little general impact for both models. From the descriptive statistics (Table 2) it was noted that there were no significant differences for most of these variables between the two groups.

3.4.3. Patient example

One example of a patient SHAP explanation for both XGB and the NN is visualized in Figure 26. This patient was readmitted within 7 days after discharge. XGB predicted a 17.0% readmission and the NN 18.7%. The figure shows the top 10 features which pushed the prediction in positive and negative direction. For both models, the contributing features show little similarities. Although only a small proportion of features is visualized in this figure, it demonstrates the ability of SHAP to make 'black-box' ML models explainable to the physician.

3.4.4. Expert opinion

We asked two ICU physicians of the LUMC to assess what model's explanations were most clinically relevant and in line with clinical intuition (Figures 23 and 24). XGB was most in line with clinical practice, with 19 relevant features and 1 contradicting (minimum thrombocytes level). The LR model had the highest number of seemingly irrelevant features (n = 8), and the NN the highest number of contradictive features (n = 7). The physicians found the XGB model to be most in line with clinical reasoning and therefore best applicable in clinical practice. They had some difficulties with correctly interpreting the SHAP summary plots, and preferred the visualization of the LR coefficients. Both physicians mentioned that it is contra intuitive that for both LR and NN the number of laboratory measurements (count) are impactful features. This would only be relevant for haemoglobin and blood gas analysis, since the performed frequency of these measurements is correlated with a bad outcome. The experts emphasized the importance of informative and clinically relevant explanations in order to trust, and therefore use, the prediction in clinical practice. Patient examples as provided in Figure 26 were found to be clear and informative.

4. Discussion

We compared LR, boosting algorithms (GB and XGB), and NNs on their ability to provide accurate, correctly calibrated, and explainable predictions for ICU readmission. After feature engineering, boosting algorithms and NNs did not outperform LR. Using SHAP, we were able to compare state-of the-art ML model explanations (NN and XGB) to LR coefficients. We found that respiratory rate, urea levels and C-reactive protein levels were impactful predictors for all model types. According to the expert opinion of two ICU physicians, explanations of the XGB model were most clinically relevant. Due to small differences in discriminative power of state-of-the-art models, the model of which the explanations are most in line with clinical reasoning should be chosen for meaningful and safe decision support. XGB was found to be most suitable for our discharge decision support tool, as it was found to be superior in terms of predictive performance, calibration properties, and clinically relevant explanations.

4.1. Comparison with relevant literature

Previous research on the prediction of ICU readmission mainly focused on predictive performance, although some studies used SHAP[12] or other feature importance techniques to enhance model explainability[10,14,19,38–40]. To our knowledge, a comparison between model explanations has not been done before for the prediction of ICU readmission. Although a limited number of variables was available at time of model development compared to previously conducted studies, we achieved acceptable discriminative performance. AUCPR was 0.18 for LR and 0.17 for NN and XGB. AUC was 0.74 for LR, NN and XGB. Previously published papers reported AUCs ranging between 0.64[40] and 0.92[39]. AUCPR, which is a more appropriate metric to evaluate discriminative performance in imbalanced dataset[31],

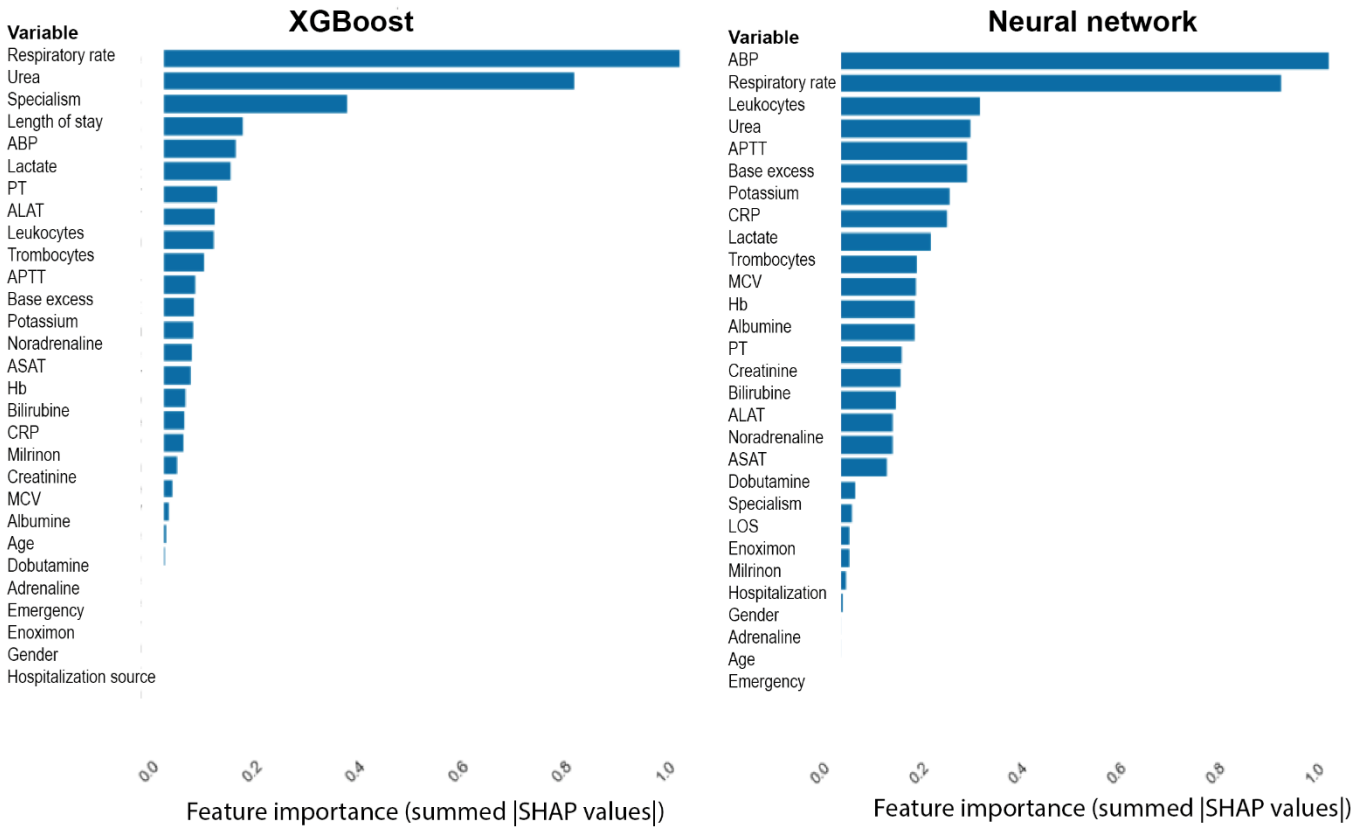


Figure 25: Summed variable importance of XGBoost (left) and neural network (right) models. Mean absolute SHAP values are summed for each variable over the feature aggregates to get insight in which variables have high correlation with the predicted outcome. Values are scaled between 0 and 1.

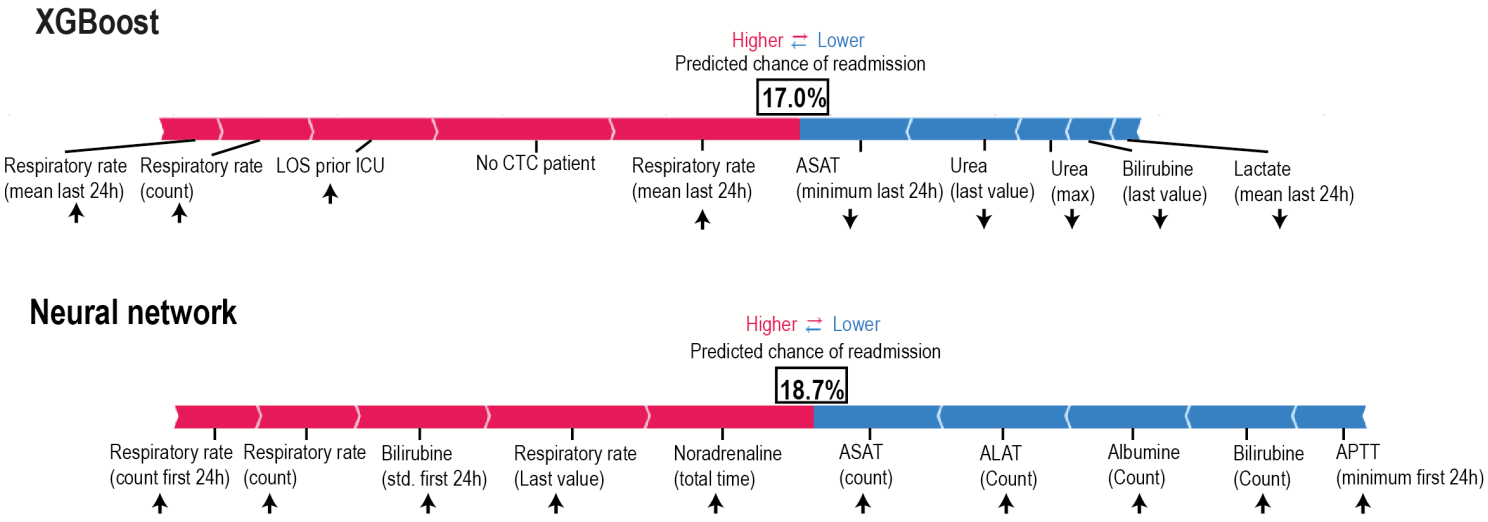


Figure 26: SHAP XGBoost (upper figure) and neural network (lower figure) force plot for the same ICU patient. The features contributing to a higher risk of readmission are visualized in red. The features contributing to a lower risk of readmission are visualized in blue. Arrows pointing upwards and downwards indicate higher or lower than average in the training dataset.

was only reported by Thoral et al., making fair comparison to other studies not feasible.

Our feature engineering and modeling methods for gradient boosting and LR were based on the methods described by Thoral et al. Their higher performance with an AUCPR of 0.20 for gradient boosting and LR models could be explained by the availability of more data at time of model development. However, Thoral et al. did not perform a final evaluation on a separate test dataset, which limits generalizability of results. Currently, the datascientists and physicians involved in the model developed by Thoral et al. are the first aiming to bring an ICU decision support tool for discharge support in clinical practice (Pacmed Critical[41]).

Our explainability results gave insight in the contributing factors to a patient predicted readmission risk. Multiple risk factors for readmission are described in the literature. For example severity of illness, high age, comorbidities, ICU admission from other hospital wards, male sex, length of ICU stay, and intensity of organ supporting therapies (the use of vasopressors, (non-invasive) mechanical ventilation, and renal replacement therapy)[42]. Although we did not use all these variables during model development, we found in contrast to literature limited impact of age and sex on our final model's predictions. More impactful variables for our predictions included respiratory rate, urea levels, C-reactive protein levels, and arterial blood pressure levels, which were also amongst the most informative features as described by Thoral et al. In these parameters, a patient's respiratory state, kidney function, inflammatory state, and cardiac function are in more or lesser extent represented. This indicates the correlation between the state of multiple physiological systems and the risk of readmission.

4.2. Limitations

Several limitations of this study have been identified. First, not all relevant variables were available at the time of model development (e.g., a representation of the neurological state and respiratory state of a patient could not be included due to data unavailability). Second, patients with a no-return-policy and palliative care patients could not be excluded from analysis. This could potentially have resulted in bias, since these patients (often with bad condition) were not eligible to be readmitted to the ICU department. Third, we observed inferior predictive performance for medical and non-thoracic surgical patients (including general surgical), compared to surgical and thoracic surgical patients. Thoracic surgical patients are less often readmitted to the ICU, and the need for discharge decision support might be lower for this group of patients. The inferior discriminative performance for medical and non-thoracic surgical patients could be explained by the heterogeneous types of patients present in this group. We observed in a sub-analysis of the SHAP explanations that the algorithms had less distinctive features for the low performing patient groups. Furthermore, we could not incorporate a variable representing oxygen saturation, which is ought to have a strong relation for medical patients with their readmission risk. Fourth, SHAP values and LR coefficients were only evaluated for the 20 most informative features for each model. A total of 416 features were used in the final models and visualizing the impact of all features would be too complex for human understanding. Therefore, the ten most predictive features for each patient could be visualized (Figure 26) to enhance patient specific explainability. It is important to note that the other 406 features contribute to the prediction to varying degrees. We observed that SHAP values for the NN model were more centered around zero and therefore less informative than those of the XGB model. A possible reason for this finding could be that the combination of features is more important for the NN's prediction. Fifth, we did not perform objective examination of explainability. More extensive evaluation of clinician preference should be performed by means of a vignette study[16]. Lastly, we only evaluated our model on retrospectively available data of patients being successfully discharged to the ward. Further prospective evaluation needs to be performed to assess the models' performance on live data. Despite these limitations, we managed to achieve acceptable predictive performance, and gained valuable insights in different model explanations.

4.3. Interpretation of findings

Our finding that LR performed slightly superior over state-of-the-art ML methods has been previously described in a systematic review by Christodoulou et al.[43]. Due to extensive feature engineering and feature selection, we captured time-related trends and used only informative features for model

development. The use of solely informative features potentially enabled LR to discover patterns in the data with comparable or even superior performance to XGB and NN. The performance of state-of-the-art ML models might outperform LR models having availability of a more complete set of variables, or by applying more advanced DL models capable of handling time-series variables. These hypotheses should be further evaluated in future studies. Although LR had highest discriminative performance, expert evaluation showed that the impactful features of XGB were more in line with clinical practice, making it more suitable for clinical applications. Generalizability of our finding that XGB's explanations were most clinically valid should be further evaluated for other datasets and use-cases. We recommend to compare explanations of several high performing ML models for each new use-case to determine what model to implement in clinical practice.

Before Platt-scaling our final predictions, we observed poor calibration properties for all models (Figure 22). This is a known issue for imbalanced classification problems[44]. Poor calibration properties could be explained by the use of weighted learning to account for class imbalance, resulting in high recall but low positive predictive value. After scaling, predicted probabilities ranged between 0-40% chance of readmission. Due to a baseline readmission rate of 6.8% in the dataset, it must be noted that a predicted probability higher than 6.8% indicates an increased risk of readmission. The effect of these predicted probabilities on clinical decision making should be further investigated.

4.4. Clinical implications

With recent advances in explainability of ML, the risk of adverse events using 'black-box' models for high-stake decisions is decreasing. However, Rudin stated that we should stop trying to explain complex models, and focus on creating sparse and interpretable models instead[45]. Even for our LR model, explanation of feature importance is complex due to the large number of features used for predictions. SHAP does allow us to visualize impactful features of ML models but is still an approximation and does not provide full transparency[16]. However, the use of simplified risk scores previously developed for the prediction of readmission showed inferior performance[12,38] and reduced generalizability[46] compared to more advanced models. This indicates the difficulty to create accurate sparse and simple models for the prediction of readmission.

Due to the complex nature of ICU readmissions, there might be a limit on maximum achievable discriminative performance[7]. Therefore, explainability of a decision support algorithm is of major importance, as it influences a clinician's trust in an algorithm[47]. A decision support tool may do more harm than good when the physician selects the wrong intervention based on misinterpreted 'black-box' predictions[7]. Besides explainability, the model should indicate out-of-domain predictions (i.e., patients with little similarities to the patients the model is trained on) to reduce the risk of harmful decisions based on uncertain predictions[48]. Taking these considerations into account, it is time to evolve from retrospective trials to prospective bed-side evaluation of discharge decision support. First, we need to compare the model's predictions to that of ICU physicians. Second, the influence of predicting ICU readmission on physician's decisions, and ultimately patient outcomes, should be evaluated during randomized controlled trials.

4.5. Conclusion

ICU readmission is a serious adverse event of which the occurrence might be reduced by assisting the physician in determining the optimal timing of ICU discharge. A decision support tool for the prediction of readmission could identify patients high or low at risk of readmission. To create useful bedside decision support, discriminative performance and calibration properties alone are insufficient for model assessment. We state that for creating clinical value using prediction modeling, explainability and agreement with clinical reasoning is at least as important. In contrast to previous studies, we found no superior performance of state-of-the-art ML models over LR. However, the use of more relevant variables for prediction modeling might result in higher performance for ML models. Using SHAP values, we concluded that state-of-the-art ML models are at least as explainable as LR by giving patient specific explanations. XGB explanations were more relevant for clinical practice, making it favorable over LR for clinical implementation. More extensive clinician evaluation should be performed to determine what modeling method should be implemented. The next steps in creating clinically valuable discharge

decision support are prospective evaluation and implementation of explainable models in practice. This will enable us to investigate the influence of predictions on discharge decisions, and ultimately on patient outcomes.

References

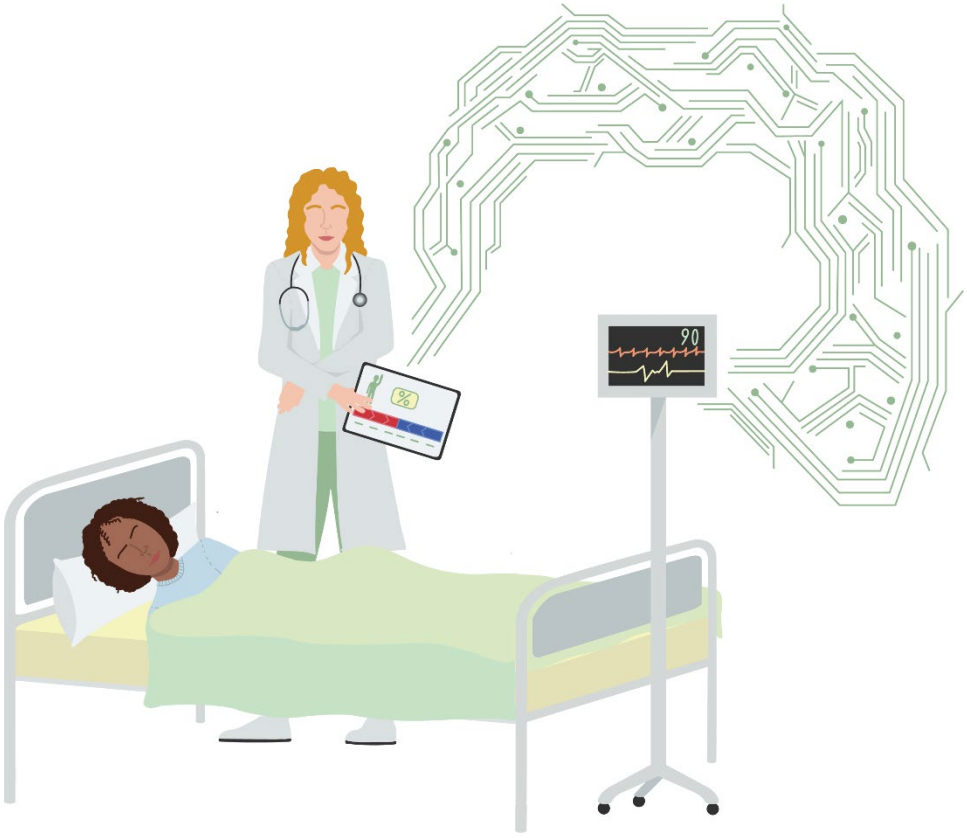
1. Rosenberg AL, Watts C. Patients readmitted to ICUs: A systematic review of risk factors and outcomes. *Chest*. 2000;118(2):492-502. doi:10.1378/chest.118.2.492
2. Kramer AA, Higgins TL, Zimmerman JE. The association between ICU readmission rate and patient outcomes. *Crit Care Med*. 2013;41(1):24-33. doi:10.1097/CCM.0b013e3182657b8a
3. Lal A, Pinevich Y, Gajic O, Herasevich V, Pickering B. Artificial intelligence and computer simulation models in critical illness. *World J Crit Care Med*. 2020;9(2):13-19. doi:10.5492/wjccm.v9.i2.13
4. Markazi-Moghaddam N, Fathi M, Ramezankhani A. Risk prediction models for intensive care unit readmission: A systematic review of methodology and applicability. *Aust Crit Care*. 2019;0(0). doi:10.1016/j.aucc.2019.05.005
5. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. *Crit Care*. 2019;23(1):284. doi:10.1186/s13054-019-2564-9
6. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. *Brief Bioinform*. 2017;19(6):1236-1246. doi:10.1093/bib/bbx044
7. Hilton CB, Milinovich A, Felix C, et al. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-0249-z
8. Lundberg SM, Allen PG, Lee S-I. *A Unified Approach to Interpreting Model Predictions*.; 2017. https://github.com/slundberg/shap. Accessed April 29, 2020.
9. Zihni E, Madai VI, Livne M, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS One*. 2020;15(4). doi:10.1371/journal.pone.0231166
10. Pakbin A, Rafi P, Hurley N, Schulz W, Harlan Krumholz M, Bobak Mortazavi J. Prediction of ICU Readmissions Using Data at Patient Discharge. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol 2018-July. Institute of Electrical and Electronics Engineers Inc.; 2018:4932-4935. doi:10.1109/EMBC.2018.8513181
11. Lin Y-W, Zhou Y, Faghri F, Shaw M, Campbell R. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS One*. 2019;14(7):e0218942. doi:10.1371/journal.pone.0218942.t006
12. Thoral, Patrick J, Fornasa, Mattia, de Bruin, Daan, Hovenkamp, Hidde, Driessen, R, Girbes, A, Hoogendoorn, Elbers P. Developing a Machine Learning prediction model for bedside decision support by predicting readmission or death following discharge from the Intensive Care unit. *Unpublished*. 2020.
13. Venugopalan J, Chanani N, Maher K, Wang MD. Combination of static and temporal data analysis to predict mortality and readmission in the intensive care. *Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf*. 2017;2017:2570-2573. doi:10.1109/EMBC.2017.8037382
14. Barbieri S, Kemp J, Perez-Concha O, et al. Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk. *Sci Rep*. 2020;10(1). doi:10.1038/s41598-020-58053-z
15. van der Meijden SL. Using Artificial Intelligence for the prediction of ICU readmission: a Systematic Review [Unpublished]. 2020.
16. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. doi:10.1016/j.jbi.2020.103655
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation*. 2015;131(2):211-219. doi:10.1161/CIRCULATIONAHA.114.014508
18. Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: Issues and methods. *J Clin Oncol*. 2012;30(26):3297-3303. doi:10.1200/JCO.2011.38.7589
19. Lin Y-W, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. Moskovitch R, ed. *PLoS One*. 2019;14(7):e0218942. doi:10.1371/journal.pone.0218942
20. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Med Res Methodol*. 2017;17(1). doi:10.1186/s12874-017-0442-1
21. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451
22. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
23. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: An overview. *J Thorac Dis*. 2019;11(Suppl 4):S574-S584. doi:10.21037/jtd.2019.01.25
24. Pedregosa F. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
25. TensorFlow | Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. https://dl.acm.org/doi/10.5555/3026877.3026899. Accessed January 29, 2021.
26. GitHub - keras-team/keras: Deep Learning for humans. https://github.com/keras-team/keras. Accessed January 29, 2021.
27. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive

models in biomedical research: A multidisciplinary view. *J Med Internet Res*. 2016;18(12). doi:10.2196/jmir.5870
28. Victoria AH, Maragatham G. Automatic tuning of hyperparameters using Bayesian optimization. *Evol Syst*. 2020;1:3. doi:10.1007/s12530-020-09345-2
29. Caicedo-Torres W, Gutierrez J. ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU. *J Biomed Inform*. 2019;98. doi:10.1016/j.jbi.2019.103269
30. Kubben P, Dumontier M, Dekker A. *Fundamentals of Clinical Data Science*.; 2019.
31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
32. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *34th International Conference on Machine Learning, ICML 2017*. Vol 3. International Machine Learning Society (IMLS); 2017:2130-2143.
33. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*. New York, New York, USA: ACM Press; 2005:625-632. doi:10.1145/1102351.1102430
34. BRIER GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Mon Weather Rev*. 1950;78(1):1-3. doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2
35. Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree Ensembles. February 2018. http://arxiv.org/abs/1802.03888. Accessed March 16, 2020.
36. Ribeiro MT, Singh S, Guestrin C. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*.
37. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *J Biomed Inform*. 2017;76:9-18. doi:10.1016/j.jbi.2017.10.008
38. Rojas J, Carey K, Edelson D. Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc*. 15:846-853.
39. Loreto M, Lisboa T, Moreira VP. Early prediction of ICU readmissions using classification algorithms. *Comput Biol Med*. 2020;118. doi:10.1016/j.compbiomed.2020.103636
40. Junqueira ARB, Mirza F, Baig MM. A machine learning model for predicting ICU readmissions and key risk factors: analysis from a longitudinal health records. *Health Technol (Berl)*. 2019;9(3):297-309. doi:http://dx.doi.org/10.1007/s12553-019-00329-0
41. Intensive Care - Pacmed . https://pacmed.ai/en/projects. Accessed August 31, 2020.
42. Ponzoni CR, Corrêa TD, Filho RR, et al. Readmission to the intensive care unit: Incidence, risk factors, resource use, and outcomes: A retrospective cohort study. *Ann Am Thorac Soc*. 2017;14(8):1312-1319. doi:10.1513/AnnalsATS.201611-851OC
43. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
44. Wallace BC, Dahabreh IJ. Class probability estimates are unreliable for imbalanced data (and how to fix them). In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. IEEE Computer Society; 2012:695-704. doi:10.1109/ICDM.2012.115
45. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
46. Failure of the Swift Score to Predict Readmission to the ICU – SHM Abstracts. https://www.shmabstracts.com/abstract/failure-of-the-swift-score-to-predict-readmission-to-the-icu/. Accessed April 14, 2020.
47. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and AI research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness. *arXiv*. December 2018. http://arxiv.org/abs/1812.10404. Accessed February 1, 2021.
48. Ruhe D, Cinà G, Tonutti M, de Bruin D, Elbers P. Bayesian Modeling in Practice: Using Uncertainty to Improve Trustworthiness in Medical Applications. June 2019. http://arxiv.org/abs/1906.08619. Accessed June 18, 2020.

Part IV – Prospective Evaluation

4

As part of the thesis feasibility study (TM30002), a study protocol was developed for the prospective validation of Pacmed Critical. This trial is planned to be performed starting around May 2021, but needs to be approved by the medical ethical committee (METC) of the LUMC.



METC protocol prospective evaluation of Pacmed Critical

Summary

Rationale: We hypothesize that a machine learning algorithm (Pacmed Critical) can more accurately predict ICU readmission than an ICU physician.

Objective: To compare the prediction of ICU readmission within 7 days after discharge of ICU physicians to the Pacmed Critical model.

Study design: Prospective observational; by means of survey questions integrated in the patient data management system (PDMS, Metavision). The prediction of readmission by the Pacmed model is performed using anonymized patient data present in the PDMS at the same moment as the physician prediction. Patient readmission and mortality data is collected up to 7 days after discharge.

Study population: ICU patients (> 18 years) in the LUMC

Main study parameters/endpoints: Discrimination, calibration and net benefit of physicians' prediction compared against the ones by Pacmed Critical.

1. Introduction and rationale

At the moment, the Intensive Care Unit (ICU) at LUMC (and other medical centres) is experiencing capacity issues (sadly illustrated by the COVID-19 outbreak). Consequently, the medical staff is pressured to discharge patients as soon as possible (when permissible) to lower care wards to free up beds. This may have detrimental effects on quality of care, patient satisfaction and physician burden and may result in ICU readmission due to limited monitoring and therapeutic options at lower care wards [1, 2].

Readmission to the ICU during the same hospital stay is correlated with increased mortality rates, longer hospital stays, and higher costs [3]. Therefore, there is a need to prevent premature ICU discharge of patients at risk of readmission. On the other hand, delayed discharge of ICU patients can result in reduced capacity affecting new patients in need of intensive care [4, 5]. Furthermore, unnecessary prolonged ICU stay may cause iatrogenic harm to the patient and may cause mental problems during rehabilitation [6]. ICU physicians might have benefit from knowing a patient's risk of readmission and/or mortality in their decision to discharge to lower care hospital wards.

The CE-certified Pacmed Critical software consists of an algorithm that could assist intensivists in determining the optimal moment for discharging their patient from the ICU [7]. The intended advantage would be optimization of discharge and thus prevention of readmission and through this better patient outcomes. The model was developed at the Amsterdam UMC and is now being validated and implemented in several other Dutch hospitals. Pacmed Critical is a so-called 'Machine Learning' model, which makes individual patient predictions based on data available of previously readmitted ICU patients.

In the first half of 2021, Pacmed Critical will be calibrated and validated on the LUMC population. First, the performance of predicting ICU readmission will be evaluated on retrospective patient data. The ultimate goal is to perform a randomized controlled trial investigating the influence of using the Pacmed Critical Software on discharge behaviour and patient outcomes.

Before implementation of the model in clinical practice and evaluating its impact, we want to compare the predictive performance of Pacmed Critical with the predictive ability of the ICU physicians of the LUMC. To enhance clinical decision making, a decision support tool should have superior performance to that of clinical judgement alone [15]. Although the performance of Machine Learning models versus physicians has been

studied for diagnosing in medical imaging [8], there has been little research prospectively comparing physician's predictive performance when it comes to patient outcomes [9, 10]. However, preliminary results of Rojas et al. showed superior performance of a machine learning model over physician's prediction of ICU readmission at moment of discharge [9].

The aim of our study is therefore to compare the predictive performance between the Pacmed Critical algorithm and the ICU physicians of the LUMC. Secondly, we want to gain insight in patient factors contributing to the physician's prediction and compare these to the patient factors given as important predictors by Pacmed Critical. Lastly, we want to investigate the level of confidence of the physician's prediction.

2. Objectives

Primary objective:
To evaluate the performance of Pacmed Critical compared to the physician's prediction of ICU readmission and/or mortality within 7 days after discharge.

Secondary objectives:

i. To compare patient factors contributing to the prediction of readmission and/or mortality of both Pacmed Critical and the physicians.
ii. To gain knowledge on the physician's level of confidence (low-medium-high) about their prediction
iii. To compare performance for multiple patient groups:
    a. Surgical and medical patients
    b. COVID-19 patients
iv. To evaluate Pacmed's predictions of readmission and/or mortality over the duration of ICU admission

3. Study Design

This study will be an observational, longitudinal study, by means of electronic survey questions implemented in the Patient Data Management System (PDMS, Metavision) of the LUMC..

The decision to discharge a patient to a lower care ward, is made by the ICU team (intensivist, fellow, resident, nurse) during daily rounds at 8.45 a.m. independently of the estimated readmission risk by Pacmed Critical, i.e. the team is blinded for the Pacmed Critical prediction. After the patient is assessed eligible for discharge, the ICU team discusses the following questions at bedside, which are filled in by one of the physicians in the discharge form in PDMS:

- What is the chance of readmission for this patient? This will be an estimate between 0 – 30%
- What are the main factors contributing to the discharge decision?
- How confident are you about the decision made?

The prediction of Pacmed Critical is based on the validated and is a percentage (0-100%) chance of readmission/mortality within 7 days. See Figure 27. The Pacmed prediction is stored at the same time as the prediction is filled in the PDMS. Seven days after actual discharge to lower care wards, for each patient the outcome (mortality/readmission to the ICU) is assessed. Because the Pacmed Critical prediction is blinded to the physician, this observational study has no influence on standard care.

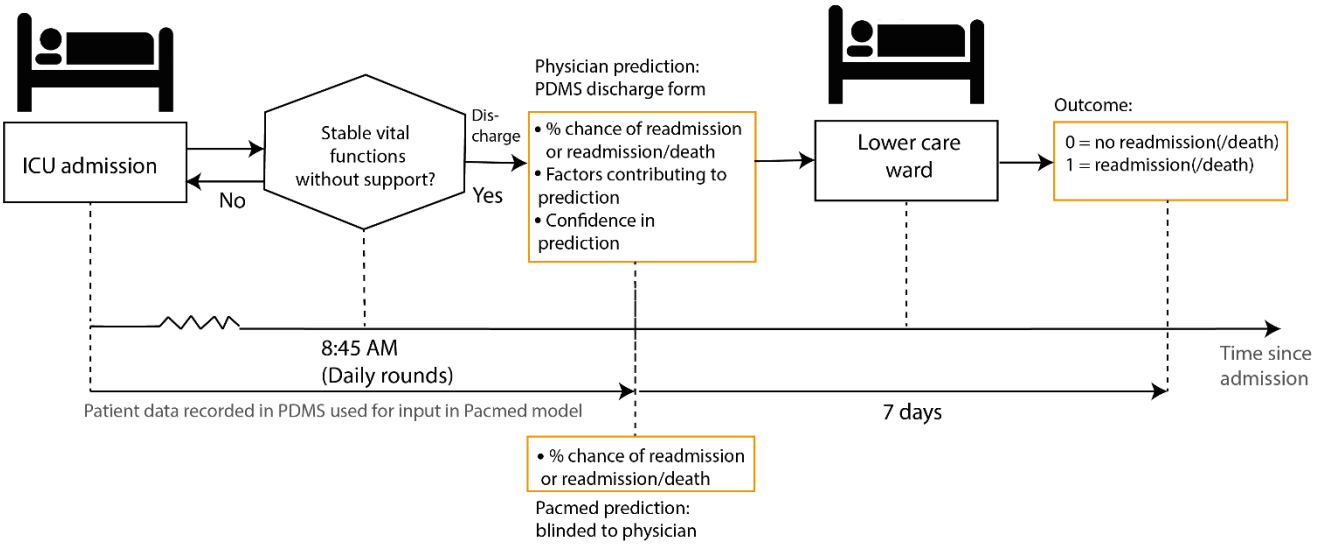


Figure 27: Patient flow and study design. Patient data is recorded and stored in the Patient Data Management System (PDMS, Metavision), including static (e.g. Age, reason for admission) and dynamic (e.g. heart rate, medication) data. During the daily rounds at 8:45 AM, the decision to (not) discharge a patient to a lower care ward is made. When assessed fit for discharge, the chance of readmission (on a scale of ....) is discussed with the intensivist and answered in the survey questions incorporated in the PDMS by the resident/fellow. If the vital functions are stable without support, and the patient is discharged to lower care wards, the physician fills in the prediction of the event, the factors contributing to the prediction, and the confidence in the prediction. The Pacmed Critical software calculates their prediction based on the recorded patient variables and is blinded for the physician. After 7 days of discharge, the outcome (readmission) is collected for each patient.

4. Study Population

4.1 Population (base)

All patients admitted to the ICU of the LUMC.

4.2 Inclusion criteria

In order to be eligible to participate in this study, a subject must meet all of the following criteria:

- Admission to the IC or Medium Care (MC) of the LUMC
- > 18 years old
- ICU admission longer than 4 hours

4.3 Exclusion criteria

A potential subject who meets any of the following criteria will be excluded from analysis in this study:

- Patients transferred to other hospitals after discharge
- Patients with a no-return to the ICU policy
- Patients receiving palliative care
- Patients that died during their ICU stay
- Patients not discharged to lower care wards due to shortage of beds

4.4 Sample size calculation

A minimum of 100 events (readmissions) and 100 non-events is suggested for validation of prognostic models [11]. Validation of the model is already performed on a large retrospective cohort of ICU patients at the LUMC, and therefore lower sample sizes for this study will be handled. A similar study comparing the prediction of AKI between physicians and a Machine Learning model, included 252 patients of which 12% (n = 30) developed AKI [10]. Other studies reported justification of a sample size of 10 events for comparison between physician prediction and a prediction model for mortality [14].

Combined readmission/mortality rate at the development site of Pacmed Critical (VUmc) within seven days after discharge was 5.3% [12]. We aim to include patients in a period of 2-4 months (300-600 patients) to have enough events for statistical analysis.

5. Treatments of subjects

Not applicable.

6. Investigational product

6.1 Name and description of investigational product(s)

The CE-certified Pacmed Critical software consists of an algorithm that could assist intensivists in determining the optimal moment for discharging their patient from the ICU. The intended advantage would be the prevention of readmission and through this better care outcomes. The model was developed at the VUmc and is now being validated and implemented in several other Dutch hospitals. Pacmed Critical is one of the E-health projects which is supported by the 'Citrienfonds' (a NFU organisation) which should be implemented in all Dutch academic hospitals. The focus of implementation of Pacmed Critical in the LUMC is to investigate the value of AI in clinical practice.

6.2 Summary of findings from non-clinical studies

Not applicable.

6.3 Summary of findings from clinical studies

Not applicable.

6.4 Summary of known and potential risks and benefits

Not applicable.

7. Non-investigational product

Not applicable.

8. Methods

8.1 Study parameters/endpoints

8.1.1 Main study parameter/endpoint

Discrimination, calibration and net benefit of physicians' prediction compared against the ones by Pacmed Critical.

8.1.2 Secondary study parameters/endpoints (if applicable)

1. Patient factors contributing to the physicians' prediction
2. Physicians' confidence in their prediction
3. Prediction performance for multiple patient groups
    a. Surgical and medical patients
    b. COVID-19 patients
4. Pacmed's predictions of readmission and/or mortality over the duration of ICU admission

8.1.3 Other study parameters (if applicable)

Patient demographics and characteristics (age, gender, source of admission, admission information, length of stay, APACHE score).

8.2 Randomisation, blinding and treatment allocation

Not applicable.

8.3 Study procedures

Survey questions are implemented in the PDMS (Metavision) of the ICU in cooperation with the ICT department.

Every morning during daily rounds at 8.45 AM, the ICU team (intensivist, fellow, residents, nurses) discusses which patients are ready for discharge to lower care wards.

The prediction if ICU readmission and/or mortality is only relevant for patients no longer in need of vital function support. At moment of discharge, the physician (usually the fellow or resident) creates a discharge letter in the PDMS. The chance of readmission/mortality, the factors contributing to this prediction, and the confidence of the prediction is filled in. The questions are discussed by the whole ICU team at bedside, and filled into the discharge form by one of the physicians:

- ❑ The chance of readmission/mortality within 7 days is [ ]%
- ❑ The chance of readmission/mortality within 7 days is *[low – average – high]*
- ❑ The factors contributing to this prediction are: [drop down menu, see the Supplementary material Part IV – study protocol (page 93)]
- ❑ I feel [*low-medium-highly]* confident about this prediction.

See Appendix A for all survey questions. Predictions are not made for patients meeting the exclusion criteria, for which the reason for exclusion is filled in the PDMS.

8.4 Withdrawal of individual subjects

Subjects can leave the study at any time for any reason if they wish to do so without any consequences. The investigator can decide to withdraw a subject from the study for urgent medical reasons.

8.4.1 Specific criteria for withdrawal (if applicable)

Not applicable.

8.5 Replacement of individual subjects after withdrawal

Not applicable.

8.6 Follow-up of subjects withdrawn from treatment

Not applicable.

8.7 Premature termination of the study

Not applicable.

9. Safety reporting

Not applicable.

10. Statistical analysis

Results will be analyzed anonymously for both participating physicians and patients. Data are presented as means and standard deviations (SD), medians and interquartile ranges (IQR), and numbers and proportions where appropriate. Statistical significance was set at $P < 0.05$. All analyses are performed using Python and SPSS.

10.1 Primary study parameters

Risk prediction of the Pacmed model will be compared to the risk prediction of the ICU physicians at moment of discharge and will be evaluated to real patient outcomes (readmission to the IC/MC). Performance outcomes include: Area under the receiving operator curve (AUC), Area under the precision recall curve (AUCPR), sensitivity, specificity, f1-score, positive predictive value.

The DeLong test is used to compare AUC for the physician's predictions and the Pacmed predictions [16]. Calibration is assessed using calibration curves. Decision curve analysis is performed to evaluate the net benefit of the model [17].

The potential added value of both the physician's prediction together with the Pacmed prediction is evaluated using multivariable logistic regression [10].

The occurrence of missing data can occur when predictions are not filled in by the physicians. We try to prevent this by implementing the predictions in the PDMS. However, if predictions are not filled in, we will exclude these cases from analysis.

10.2 Secondary study parameters

1. A sensitivity analysis will be performed on based on the following characteristics:

    a. Surgical and medical patients
    b. COVID-19 patients

    Performance measures as described under 8.1. will be evaluated for these groups.

2. Patient factors contributing to the physician's prediction:
   Descriptive statistics for factors contributing to the physicians' prediction, compared to predictors of the Pacmed algorithm.

3. Physician's confidence in their prediction performance for multiple patient groups:
   Confidence-accuracy calibration [18].

4. Pacmed predictions over time during the ICU stay for multiple time points (at time of admission, one day before discharge, two days before discharge) in descriptive statistics.

10.3 Other study parameters

Descriptive statistics will be performed on participant demographics, and on patient factors on which physicians made their prediction.

10.4 Interim analysis

Not applicable.

11. Ethical considerations

11.1 Regulation statement

The study will be conducted according to the principles of the Declaration of Helsinki and in accordance with the Medical Research Involving Human Subjects Act (WMO) and other guidelines, regulations and acts.

11.2 Recruitment and consent

This observational study has no influence on patient care, and data is handled anonymized. Therefore, no informed consent is required. Every patient and his/her family or representative receive a patient folder including the following text:

"wij werken er voortdurend aan om de kwaliteit van zorg te verbeteren en de tevredenheid van onze patiënten, familie en naasten te verbeteren. Hiervoor neemt de IC deel aan interne en externe kwaliteitsonderzoeken, verbeterprojecten en registratiesystemen. Hierbij worden in sommige gevallen ook patiëntengegevens uitgewisseld en opgeslagen. Dit gebeurt zorgvuldig en volgens de geldende privacyregels. Wij nemen bijvoorbeeld deel aan de continue Nationale Intensive Care Evaluatie (NICE) en aan periodiek onderzoek naar patiënten tevredenheid. Uw deelname is niet verplicht, laat het ons weten indien u hier bezwaar tegen heeft."

11.3 Objection by minors or incapacitated subjects (if applicable)

Not applicable.

11.4 Benefits and risks assessment, group relatedness

Not applicable.

11.5 Compensation for injury

Not applicable.

11.6 Incentives (if applicable)

Not applicable.

12. Administrative aspects, monitoring and publication

12.1 Handling and storage of data and documents

Patient data is collected from the LUMC data platform, ICT. Patient data is directly anonymized at the data platform using a pseudocode instead of the patient code. Patient data, including the variables needed for predicting with the Pacmed Critical software, physician predictions and patient outcomes (readmission, mortality). Patient is stored pseudo-anonymized at a secured Datasafe, only accessible for the investigators involved in this project within the LUMC.

12.2 Monitoring and Quality Assurance

Not applicable.

12.3 Amendments

Amendments are changes made to the research after a favourable opinion by the accredited METC has been given. All amendments will be notified to the METC that gave a favourable opinion.

12.4 Annual progress report

Not applicable.

12.5 Temporary halt and (prematurely) end of study report

The investigator/sponsor will notify the accredited METC of the end of the study within a period of 8 weeks. The end of the study is defined as the last patient's last visit.
Within one year after the end of the study, the investigator/sponsor will submit a final study report with the results of the study, including any publications/abstracts of the study, to the accredited METC.

12.6 Public disclosure and publication policy

Not applicable.

13. Structured risk analysis

Not applicable.

References

1. Brown, S.E., S.J. Ratcliffe, and S.D. Halpern, *An empirical derivation of the optimal time interval for defining ICU readmissions.* Med Care, 2013. **51**(8): p. 706-14.
2. Desautels, T., et al., *Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach.* BMJ Open, 2017. **7**(9): p. e017199.
3. Kramer, A.A., T.L. Higgins, and J.E. Zimmerman, *The association between ICU readmission rate and patient outcomes.* Crit Care Med, 2013. **41**(1): p. 24-33.
4. Cardoso, L.T., et al., *Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study.* Crit Care, 2011. **15**(1): p. R28.
5. Chalfin, D.B., et al., *Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit.* Crit Care Med, 2007. **35**(6): p. 1477-83.
6. Howell, M.D., *Managing ICU throughput and understanding ICU census.* Curr Opin Crit Care, 2011. **17**(6): p. 626-33.
7. Pacmed. Intensive Care. [Internet]. Available from: https://www.pacmed.ai/nl/projects/ic. [Accessed September 22, 2020]
8. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*2020;368:m689. doi:10.1136/bmj.m689 pmid:32213531

9. Rojas, J.C. et al. Man vs. Machine: Comparisoon of a Machine Learning Algorithm to Clinician Intuition for predicting intensive care unit readmission. American Thoracic Society 2019 International Conference, May 17-22, 2019 - Dallas, TX. Doi: 10.1164/ajrccm-conference.2019.199.1_MeetingAbstracts.A2459
10. Flechet, M., Falini, S., Bonetti, C. *et al.* Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. *Crit Care* **23,** 282 (2019). https://doi.org/10.1186/s13054-019-2563-x
11. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005 May;58(5):475-83. doi: 10.1016/j.jclinepi.2004.06.017. PMID: 15845334.
12. Patrick J. Thoral, Mattia Fornasa, Daan P. de Bruin et al. Developing a Machine Learning prediction model for bedside decision support by predicting readmission or death following discharge from the Intensive Care unit, 27 January 2020, PREPRINT (Version 1) available at Research Square [+https://doi.org/10.21203/rs.2.21940/v1+]
13. Bakker, J., Damen, J., van Zanten, A.R.H., Hubben, J.H. Criteria voor opname op en ontslag van intensive-careafdelingen. Ned Tijdschr Geneeskd. 2003 18 januari; 147(3).
14. Farinholt P, Park M, Guo Y, Bruera E, Hui D. A Comparison of the Accuracy of Clinician Prediction of Survival Versus the Palliative Prognostic Index. *J Pain Symptom Manage*. 2018;55(3):792-797. doi:10.1016/j.jpainsymman.2017.11.028
15. Cowley, L. E., Farewell, D. M., Maguire, S., & Kemp, A. M. (2019). Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic and prognostic research*, *3*, 16. https://doi.org/10.1186/s41512-019-0060-y
16. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44;837.
17. Van Calster, B., et al., *Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators.* Eur Urol, 2018. **74**(6): p. 796-804.
18. Weber N, Brewer N. Confidence-accuracy calibration in absolute and relative face recognition judgments. *J Exp Psychol Appl*. 2004;10(3):156-172.

Final discussion and future perspectives

1. Discussion

The goal of this thesis was to compare several types of AI models with respect to performance and explainability of the prediction of ICU readmission. In a broader perspective, we proposed a framework for the clinical development and implementation of clinically valuable AI-based decision support. This section provides a discussion of the model development and explainability, a reflection on the physician adoption of decision support tools, and clinical relevance of the performed research.

1.1. Model development and explainability

After systematically reviewing current literature describing model development for ICU readmission prediction (Part I), we concluded that most relevant research on the prediction of readmission was based on retrospective data. No prospective -and often no external -validation was performed, resulting in risk of overfitting and a reduced generalizability results. Furthermore, previously conducted studies lacked in reported performance metrics, calibration properties, and explainability comparison of the developed models. None of the included studies reported implementation of the developed prediction models. The research performed in this thesis contributed to creating clinical value for discharge decision support, using appropriate performance metrics and by comparing different models on explainability and calibration properties.

It would be interesting to compare our developed models (logistic regression, neural network, and boosting algorithms, see Part III) to Pacmed Critical. This was not possible within our project time frame due to implementation delays of Pacmed Critical (partly due to the COVID-19 pandemic). To our knowledge, our study is the first to perform extensive performance and explainability evaluation for the prediction of readmission for these state-of-the-art model types. As the discriminative performance and calibration properties between models were very similar, explainability becomes an important factor in deciding which model to implement. Due to the introduction of SHAP[1] values and other techniques, 'Black-box' AI models are history. To enhance adoption by physicians, it is important that the model explanations are valid and in line with clinical reasoning[2]. Furthermore, the physicians should be able to decide to not trust the prediction when explanations imply bias or erroneous assumptions[3]. Our developed XGBoost (a tree-based boosting algorithm) model showed highest agreement with clinical reasoning of ICU physicians from the LUMC. However, a limitation is that the developed ML algorithms are based on numerous features and complex calculations, making it impossible to fully interpret the working mechanism and impactful factors for each patient. Making ML models as simple as possible could enable explainability, and even increase performance when discarding clinically non-relevant parameters[4].

1.2. Adoption of decision support tools

Pacmed developed their algorithms in close collaboration with physicians from the Amsterdam UMC to create a clinically meaningful decision support tool[5]. For their model to be adopted within the LUMC, the physicians of the ICU department need to be closely involved in the implementation and evaluation process of Pacmed Critical. Some LUMC physicians mentioned in the questionnaire that they see limited benefit from such a tool, or that they strongly doubt the superior performance of such a tool over their clinical gut feeling (Part II). Therefore, the prospective study (Part IV), comparing the physicians' predictive performance to Pacmed's, will be of importance to study potential superior predictive performance. Although the physicians had some critical comments, most of them believe in the positive value of AI prediction tools. In order for the tool to have benefit, it needs to work synergistically with the physician in charge of the decision, and add to the physician's clinical perspective, knowledge, and gut feeling.

1.3. Clinical relevance

ICU readmission is a complex clinical problem that is influenced by both patient and organisational factors[6]. As one of the physicians within the LUMC stated, a patient is sometimes discharged accepting some level of readmission risk, and that it is questionable whether the patient's progression would be different by keeping them longer at the ICU. Furthermore, not all ICU readmissions can be prevented and therefore it is debatable whether

readmission prediction at discharge could always prevent readmission[7]. Because ICU readmission affects a small proportion (< 10% of the discharged patients), it might be challenging to demonstrate clinical benefit using the prediction tool. However, ICU readmission is correlated with increased hospital length of stay, and costs, and decreasing readmission rates would certainly benefit both patient and hospital. Furthermore, patients with low risk of readmission might be safely discharged sooner to the ward resulting in increasing ICU bed capacity. Moreover, one could think of other use-cases of decision support in the form of readmission predictions. These predictions could for instance be used to monitor patients more closely at the ward after discharge, resulting in adequate care and potentially prevented readmissions.

During the practical clinical part of this thesis, several readmissions were witnessed. Corresponding to what was observed in the dataset, these patients were often general surgical patients, medical patients, or patients with longer length of ICU stay. In the sensitivity analysis in Part III of this thesis, we observed inferior discriminative performance for general surgical patients compared to thoracic surgical patients. For a readmission prediction model to be of clinical value, it should be accurate for the patient groups in which readmission is most often observed. The lack of respiratory parameters (e.g., oxygen saturation) in the dataset used in Part III could explain the inferior performance in certain patient groups for which these parameters are thought to be of higher importance. In general, a limitation of our work was the high amount of missing data. Not all parameters of interest could be included for prediction modeling and predictive performance could increase for all groups when using more relevant parameters.

2. Future perspectives

The COVID-19 pandemic painfully demonstrated the limits of ICU resources, resulting in an extremely high workload for the ICU nurses and physicians[8]. AI-based decision support tools such as Pacmed Critical could be of value when having to make the decision to discharge patients to free beds for other patients in such critical circumstances. However, the circumstances during the pandemic are different than the data used to train the prediction model on. Therefore, Pacmed Critical should be validated on data from 2020 to assess generalizability to these exceptional circumstances. In the Netherlands, the pandemic has led to the first ICU wide database, icudata.nl[9]. The COVID-19 crisis demonstrated the need to collaborate in order to effectively analyse multicentre data to discover risk factors and treatment strategies. Patient data should be recorded and stored more uniformly to make decision support more easily transferable between centres[10]. During model development, we encountered the difficulty of data access, missingness, and uniformity. Having one nationwide, and in the future worldwide, used standard to record and store patient data will help in the process of model development and collaboration.

Our final models could be externally validated on other populations using an uniform multi-centre ICU database such as icudata.nl to evaluate predictive performance, calibration properties, and explainability. This would enable us to investigate how well our findings that XGBoost has both high predictive performance and its explanations are in line with clinical practice generalize to a larger population. A major part of ML model development is the data preparation and feature engineering needed to create appropriate (tabular) input data for all model types[11]. Having to execute this process for each hospital separately limits scalability of ML models as is currently observed during the implementation of Pacmed Critical at the LUMC. Another option would be to use DL models capable of handling raw patient data, but using state-of-the-art DL models often comes at the cost of explainability and need high computational power. Data uniformity, e.g., using the Fast Healthcare Interoperability Recourses (FHIR) format[12], would still be necessary to allow easy scalability of these models[13].

Besides these data quality and availability challenges, it is important to take privacy, ethical, and regulatory considerations into account. Training and validating high performing ML decision support requires large datasets ranging from 10,000 – 100,000 patient records, making it infeasible to ask each patient in retrospect for informed consent[14]. Therefore, removal of all identifiable information should take place, but a risk of re-identification remains[15]. Pacmed states that they comply to the General Data Protection Regulation (GDPR) for privacy issues, and only certified employees are allowed to access hospital data that are securely stored[16]. Furthermore, the new Medical Device Regulation (MDR) will become effective on May 26, 2021. The MDR states that decision

support software is classified as class IIa, and should therefore be CE-certified by a notified body[17]. A remark on the new MDR is that hospitals are permitted to develop and use their own algorithms, provided that the algorithm performs better than commercially available alternatives. What performance metrics to use to address this question and how the new regulations affect transfer learning between hospitals is unfortunately not stated in the MDR[18].

Taken these considerations into account, the time has come to investigate the impact in practice of AI-based decision support within the ICU. I believe that besides ethical and regulatory considerations, the largest hurdle for making impact will be whether the physicians will adopt and use the tool in their workflow. The most accurate prediction tool will not be of any value when not trusted or used by the physician it is developed for[19]. The questionnaire showed that most physicians believe in the positive value of AI-based decision support and that they had numerous ideas for what they would like to be supported on. The potential of AI enhancing the era of precision medicine should be studied in both randomized controlled trials to investigate patient impact, and by studying the complex relationship between physicians and decision support tools[2].

3. Conclusion

Due to increasing patient data availability, the ICU is one of the most promising areas in the field of medical AI decision support. Our work contributed to making the step from developing high performing prediction models to clinical adoption of an ICU discharge decision support system. We executed the first steps of going from clinical problem to final implementation by means of: I - a systematic review on current literature describing the use of AI predicting ICU readmission, II - a questionnaire amongst ICU physicians regarding current discharge practices and AI readiness, III - a model development study comparing several type of (machine learning) models on performance and explainability, and IV - by designing a prospective validation study design for Pacmed Critical. Marginal differences in discriminative performance and calibration properties were observed for logistic regression, boosting algorithms, and neural networks. Therefore, we state that explainability of a model is at least as important to build trustworthy decision support and to enhance clinical adoption. Physicians should be closely involved during model development to evaluate explainability outcomes in terms of impactful features to the prediction and clinical relevance. An exciting year is coming for the ICU department of the LUMC, being one of the first to implement an AI decision support system in clinical practice. The step to clinical implementation and providing meaningful support to the physicians needs strong collaboration between (technical) physicians and data scientists, and will be studied during prospective clinical trials.

References

1. Lundberg SM, Allen PG, Lee S-I. *A Unified Approach to Interpreting Model Predictions*.; 2017. https://github.com/slundberg/shap. Accessed April 29, 2020.
2. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
3. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370
4. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
5. Citrienfonds e-health | Projecten | Pacmed (machine learning software op de IC). https://www.citrienfonds-ehealth.nl/projecten/alle-projecten/pacmed-critical-beslissingsondersteunende-machine-learning-software-op-de-i/. Accessed February 17, 2021.
6. Rosenberg AL, Hofer TP, Hayward RA, Strachan C, Watts CM. Who bounces back? Physiologic and other predictors of intensive care unit readmission. *Crit Care Med*. 2001;29(3):511-518. doi:10.1097/00003246-200103000-00008
7. Al-Jaghbeer MJ, Tekwani SS, Gunn SR, Kahn JM. Incidence and Etiology of Potentially Preventable ICU Readmissions*. *Crit Care Med*. 2016;44(9):1704-1709. doi:10.1097/CCM.0000000000001746
8. Lucchini A, Iozzo P, Bambi S. Nursing workload in the COVID-19 era. *Intensive Crit Care Nurs*. 2020;61:102929. doi:10.1016/j.iccn.2020.102929
9. icudata.nl - the Dutch ICU Data Warehouse. https://icudata.nl/. Accessed February 17, 2021.
10. Pacmed. https://www.pacmed.ai/nl/media/press/persbericht-samenwerken-en-data-delen-moet-levens-redden. Accessed February 17, 2021.
11. Press, G. Cleaning big data: most time-consuming, least enjoyable data science task, survey says. Forbes (2016). Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (Accessed 22 Oct 2017).
12. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S. & Ramoni, R. B. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J. Am. Med. Inform. Assoc. 23, 899–908 (2016).

13. Rajkomar, A., Oren, E., Chen, K. et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 1, 18 (2018). https://doi.org/10.1038/s41746-018-0029-1
14. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology*. 2020:295: 675–682.
15. Elysse Tom, Pearse A. Keane, Marian Blazes, Louis R. Pasquale, Michael F. Chiang, Aaron Y. Lee, Cecilia S. Lee, and AAO Artificial Intelligence Task Force; Protecting Data Privacy in the Age of AI-Enabled Ophthalmology. *Trans. Vis. Sci. Tech.* 2020;9(2):36
16. Pacmed - Compliance. https://pacmed.ai/nl/about/legal. Accessed March 4, 2021.
17. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Artificial Intelligence in Healthcare. Elsevier; 2020:295-336.
18. Dokter.ai - Veilige en legale AI in de zorg – deel 1: MDR of MDR-light?. https://dokter.ai/category/over-a-i/. Accessed March 4, 2021.
19. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Futur Healthc J*. 2019;6(2):94-98. doi:10.7861/futurehosp.6-2-94

Acknowledgements

I would like to express my great gratitude to the supervisors of my thesis for their support and efforts. Sesmu, thank you for your inspiring endless enthusiasm and critical clinical perspective. You taught me to think conceptually and to always keep the patient and physician in mind. Matthijs, I really appreciate your positive and constructive way of communicating, and your support when my plans changed once more. You taught me to set boundaries during my project and to keep the scientific perspective in mind. Esmee, thank you for your enthusiastic programming, (clinical) AI and daily support. It was really fun to work with you, and hopefully we will collaborate someday in the future.

Furthermore, I would like to thank Anne de Hond and Feline Spijkerboer for their supporting role in my project, my brothers Rune and Vidar for being my greatest role-models, Jesse for always being there and your great help and support during the last months of my thesis, Leonoor for the awesome cover illustration, my roommates (Marieke, Kate, Sylle, Nina, Leonoor, Saar, Lara, Sophie) for the fun and support this crazy lockdown year, my friends (Jip, Eva, Leonie, Olivier, Fleur, Willemijne, Diederik, Frederiek, Maria, Anouk, Lucas, Josine), and of course my best imaginable parents: Beppy & Louis.

Appendices

A. Supplementary material Part I - Systematic Review

Data extraction tables

Table A-1: Characteristics of included studies. MIMIC = Medical Information Mart for Intensive Care database, - = Not Reported, GB = Gradient Boosting, RF = Random Forest, NB = Naïve Bayes, SMO = sequential minimal optimization, LB = Logistic Boost, IC = Iterative Classifier, RNN = Recurrent Neural Network, LSTM = Long Short Term Memory, CNN = Convolutional Neural Network, LR = Logistic Regression, FFNN = Feed Forward Neural Network, CRF = Conditional Random Field, MLP = Multi layer perceptron

Author	Year	Study type	Dataset	Number of included patients	Propotion of ICU readmissions (%)	Predicted readmission outcome from moment of discharge	Model architecture
Rojas et al.	2018	Observational	Single-center US hospital + MIMIC III	24885	11% (same hospital stay)	ICU readmission within same hospital stay, <72h, >72h	GB
Desautels et al.	2017	Cross-sectional	Single-center UK hospital + MIMIC-III	2018	4.36%	ICU readmission within 48h	AdaBoost
Loreto et al.	2020	Observational	Multi-center Brazilian hospitals	9926	6.6%	ICU readmission within same hospital stay	RF, NB, J48, SMO, AdaBoost, LB, IC
Pakbin et al.	2018	Observational	MIMIC-III	46252	1.87% (24h), 4.14% (72h), 6.56% (7d), 11,69% (30d)	ICU readmision within 24h, 72h, 7d, 30d	XGBoost
Junqueira et al.	2019	Observational	MIMIC-III	31749	10.9%	ICU readmission within 30 days	MLP, RF, SVM
Barbieri et al.	2020	Observational	MIMIC-III	33150	12.10%	ICU readmission within 30 days	RNN
Lin et al.	2019	Observational	MIMIC-III	35334	13.9%	ICU readmission within 30 days	LSTM, CNN
Venugopalan et al.	2017	Observational	MIMIC-III	32331	24.0%	ICU readmission within 30 days	LR + FFNN, CRF

Table A-2: Summary of Machine Learning modeling strategies used.- = Not reported. GB = Gradient Boosting, RF = Random Forest, GCS = Glasgow Coma Score, PCA = Principal Component Analysis, KNN = K-nearest neighbour, RNN = Recurrent Neural Network, GRU = Gated Recurrent Unit, LSTM = Long Short Term Memory, CNN = Convolutional Neural Network, FFNN = Feed-forward Neural Network, CRF = Conditional Random Field, mRMR = Minimum-redundancy maximum-relevancy, LOS = Length of Stay, CV = cross validation, RUS = Random Majority Undersampling.

Author	Best model	Feature selection	Number of Input features	Dataprep eration	Imputati on of missing values	Period of time-series data collection	Modeling strategy	Method of internal validatio n	Metho d of calibrat ion	Performa nce compare d with	Method of external validation
Rojas et al.	GB	A priori based on clinical experience	100+	-	-	Last 24h trend of vital signs + last measured vital signs before discharge	GBM with all predictors for internal validation, and simpler GBM including age, vital signs, lab for MIMIC.	10-fold CV	-	SWIFT and MEWS	MIMIC-III database
Desaut els et al.	AdaBoo st	-	15	Binned at each hour	Most recent value	Last 5h before discharge	Transferlearni ng of AdaBoost model on local and MIMIC dataset	10-fold CV	Logistic Regress ion on training data	SWIFT	MIMIC-III database
Loreto et al.	Multipl e	Different sets (arrival, complete, PCA, wrapper).Excl uding redundant features and > 80% missing values.	134	-	Different for each algorithm	First hour after admission	Eight classification methods over different sets of attributes	10-fold CV	-	Multiple ML algorithm s, SOFA, SAPS	No, but uses data from three hospitals to train and test model on
Pakbin et al.	XGBoos t	Features with > 50% missing values dropped	2344	1-hot encoding for categorial data, aggregati on of units	Mean for numerical values, zero for booleans, categorial imputed by distributi on in the data set.	Last 24h before discharge	LR, RF, GDB and XGBoost on same feature sets	5-fold CV	Calibrat ion plot	Logistic Regressio n	No, trained and tested solely on the MIMIC-III database
Junquei ra et al.	MLP	Only static included. Features > 30% missing values, and required high medical knowledge or very sophisticated extraction technique were excluded.	12 (without feature selection) , 4 (with removal of features with <0.75% correlatio n)	Aggregati on of numerical values in categorie s.	-	No time-series used.	RF, SVM and MLP for two different datasets (data collected between 2001-2008 (first period) and 2008-2012 (second period)). Uses RUS to deal with class imbalance.	10-fold CV over first and second period.	-	RF, SVM. With and without feature selection. For first and second period.	No, trained and tested solely on the MIMIC-III database

Barbieri et al.	RNN (GRU)	All variables used	392 medication and vital signs and 23 static variables	Variables mapped to time incorperated embeddings	-	Whole stay	Multiple RNN architecture, with and without attention	Bootstrapping	-	LR	No, trained and tested solely on the MIMIC-III database
Lin et al.	RNN (LSTM) + CNN	-	59 charted events, 17 binary indicator features, 300 ICD 9 embedding, 14 demographic	1-hot encoding for categorial data, binned at each hour, ICD-9 embeddings, time series features	Last-Observation-Carried-Forward, indicator of missingness	Last 48h before discharge	Multiple RNN and CNN architectures, using different input sets	5-fold CV	-	LR, NB, RF, SVM, CNN	No, trained and tested solely on the MIMIC-III database
Venugopalan et al.	LR and FFNN + CRF	A priori selection of 87 features based on clinician input	87	Outliers removed, data binned in 6h intervals	K-means imputation	Whole stay	Combination of static models (FFNN, LR) with temporal models (CRF)	10-fold CV	-	Individual models	No, trained and tested solely on the MIMIC-III database

Table A-3: Model explainability using feature importance methods to assess the most contributing factors to the predicted outcome. Patient specific explainability is performed when a feature importance method is used to assess individual patient predictors. BUN = blood urea nitrogen, Hb = hemoglobin, GCS = Glasgow Coma Scale, HR = heart rate, RR = respiratory rate, SAPS3 = simplified acute physiology score 3, CVC = central venous catheter, MAP = mean arterial pressure, RBC = red blood cell, HF = heart failure, WBC = white blood cell.

Author	**Explainable method used**	**Patient specific explainability**	**Top 10 predictive features**
Rojas et al.	Variable importance measure using change in the Gini index, partial dependence plots.	-	BUN, braden scale, SpO2/FiO2, albumin, Hb, platelet, alk. Phosphatase, shock index, ICU length of stay, age.
Desautels et al.	-	-	-
Loreto et al.	Information Gain	-	Length of hospital stay prior to unit admission, admission source, admission type, SAPS3, chronic health status, respiratory failure (first hour), steroids use, respiratory failure, immunosupression, transplant solid organ, LOS
Pakbin et al.	Mean rank by XGBoost	-	For 72 hours: LOS, HR (last), Enteral infusion, Mech vent., arterial cath, insertion of naso. Airway, cont. mech. Vent., RR (last), bypass for heart, injection of larynx.
Junqueira et al.	Feature correlation Value Analysis using Symmetrical Uncertainty algorithm	-	LOS, number of services, service med, service SURG, service TRAUM, insurance, ethnicity, admission type, admission location, first care unit.
Barbieri et al.	Attention	Yes, using attention	Number of recent admissions, male, Ethnicity (african american), infection, infection due to CVC, desentiziation to allergens, hepatorenal syndrome, gatrostomy, plasmapheresis diabetes

Lin et al.	Feature ablation test	-	Chart events: Glucose, HR, T, GCS (eye), GCS (total), SpO2, RR, GCS(verbal), MAP
Venugopalan et al.	-	-	Static model: Hospital LOS, presence/absence of blood loss anemia, renal failure, RBC count, congestive HF. Temporal model: age, calcium, liver disease, creat, WBC count, payer group, PaCO2, SaO2, renal failure, PaO2, blood loss anemia

Table A-4: Performance metrics for best models. GB = Gradient boosting, RF = Random forest, MLP = Multi-layer perceptron, , RNN = Recurrent Neural Network, GRU = Gated Recurrent Unit, LSTM = Long Short Term Memory, CNN = Convolutional Neural Network, FFNN = Feed-forward Neural Network, CRF = Conditional Random Field, LR = logistic regression, spec. = specificity, prec. = precision.

Author	Dataset	Best model	Predicted readmission outcome	Highest AUC	Calibration	Other metrics for best model
Rojas et al.	Single-center US hospital + MIMIC	GB	< 72h, > 72h, Same hospital stay	0.73 (<72 h) 0.77 (>72h) 0.76 (same hospital stay)	-	Spec. 0.95, Recall 0.28.
Desautels et al.	Single-center UK hospital + MIMIC-III	AdaBoost	< 48h	0.71	Brier 0.04	Spec. 0.66, Recall 0.59, F1 0.13, DOR 2.86
Loreto et al.	Multi-center Brazilian hospitals	RF (with complete set)	Same hospital stay	0.92	-	Kappa 0.48, Prec. 0.58, Recall 0.46, F1 0.51
Pakbin et al.	MIMIC-III	XGBoost	<24h, <72h, <7d, <30d	0.84	Brier 0.04	F1 0.43
Junqueira et al.	MIMIC-III	MLP	< 30 days	0.64	-	Accuracy: 0.83 Prec. 0.82 Rec. 0.83
Barbieri et al.	MIMIC-III	ODE + RNN	< 30 days	0.75	-	Spec. 0.70, Recall 0.67, Prec. 0.33, F1 0.37
Lin et al.	MIMIC-III	LSTM + CNN	< 30 days	0.79	-	Spec. 0.85, Recall 0.74
Venugopalan et al.	MIMIC-III	LR + FFNN, CRF	< 30 days	-	-	Accuracy: 0.87, MCC 0.65

Systematic review search strategy

Regular references - total d.d. 17-5-2020: 172 references from:

- PubMed: 81
- Embase: 70 - 18 unique
- Web of Science: 70 - 22 unique
- COCHRANE Library: 25 - 21 unique
- Emcare: 42 - 4 unique
- Academic Search Premier: 50 - 26 unique

Embase

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=main&MODE=ovid&D=oemezd

((exp *"prediction and forecasting"/ OR Prediction.ti,ab OR Predicting.ti,ab OR Predict.ti,ab OR "Decision support".ti,ab **OR predict*.ti,ab OR exp "Decision Support System"/ OR "decision rul*".ti,ab**) AND (ICU.ti,ab OR "Intensive Care Unit".ti,ab OR "Intensive Care".ti,ab **OR exp *"Intensive Care Unit"/ OR exp *"Intensive Care"/ OR "ICUs".ti,ab**) AND ("Machine Learning".ti,ab OR "Deep Learning".ti,ab OR "Artificial Intelligence".ti,ab OR "Data analysis".ti,ab OR "Machine Learning".ti,ab **OR exp *"Machine Learning"/ OR *"Artificial Intelligence"/ OR *"Data Analysis"/ OR algorithm*.ti OR exp *"Algorithm"/**) AND ("Readmission".ti,ab OR "Discharge".ti,ab **OR *"Hospital Readmission"/ OR readmiss*.ti,ab OR readmitt*.ti,ab OR "re-admiss*".ti,ab OR "re-admitt*".ti,ab OR *"Hospital Discharge"/ OR discharg*.ti,ab**)) NOT (conference review or conference abstract).pt

Web of Science

http://isiknowledge.com/wos

70 references d.d. 14-5-2020

(TS=(Prediction OR Predicting OR Predict OR "Decision support" **OR predict* OR exp "Decision Support System" OR "decision rul*"**) AND TS=(ICU OR "Intensive Care Unit" OR "Intensive Care" **OR "Intensive Care Unit" OR "Intensive Care" OR "ICUs"**) AND (ts=("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data analysis" OR "Machine Learning" **OR "Machine Learning" OR *"Artificial Intelligence" OR *"Data Analysis") OR TI=algorithm***) AND TS=("Readmission" OR "Discharge" **OR "Hospital Readmission" OR readmiss* OR readmitt* OR "re-admiss*" OR "re-admitt*" OR "Hospital Discharge" OR discharg***)) NOT dt=(meeting abstract)

Cochrane

https://www.cochranelibrary.com/advanced-search/search-manager

25 references d.d. 14-5-2020

((Prediction OR Predicting OR Predict OR "Decision support" **OR predict* OR exp "Decision Support System" OR "decision rul*"**) AND (ICU OR "Intensive Care Unit" OR "Intensive Care" **OR "Intensive Care Unit" OR "Intensive Care" OR "ICUs"**) AND ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data analysis" OR "Machine Learning" **OR "Machine Learning" OR "Artificial Intelligence" OR "Data Analysis" OR algorithm***) AND ("Readmission" OR "Discharge" **OR "Hospital Readmission" OR readmiss* OR readmitt* OR "re-admiss*" OR "re-admitt*" OR "Hospital Discharge" OR discharg***):ti,ab,kw NOT (conference abstract):pt

Emcare

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&NEWS=n&CSC=Y&PAGE=main&D=emcr

((exp *"prediction and forecasting"/ OR Prediction.ti,ab OR Predicting.ti,ab OR Predict.ti,ab OR "Decision support".ti,ab **OR predict*.ti,ab OR exp "Decision Support System"/ OR "decision rul*".ti,ab**) AND (ICU.ti,ab OR "Intensive Care Unit".ti,ab OR "Intensive Care".ti,ab **OR exp *"Intensive Care Unit"/ OR exp *"Intensive Care"/ OR "ICUs".ti,ab**) AND ("Machine Learning".ti,ab OR "Deep Learning".ti,ab OR "Artificial Intelligence".ti,ab OR "Data analysis".ti,ab OR "Machine Learning".ti,ab **OR exp *"Machine Learning"/ OR *"Artificial Intelligence"/ OR *"Data Analysis"/ OR algorithm*.ti OR exp *"Algorithm"/**) AND ("Readmission".ti,ab OR "Discharge".ti,ab **OR *"Hospital Readmission"/ OR readmiss*.ti,ab OR readmitt*.ti,ab OR "re-admiss*".ti,ab OR "re-admitt*".ti,ab OR *"Hospital Discharge"/ OR discharg*.ti,ab**))

Academic Search Premier

http://search.ebscohost.com/login.aspx?authtype=ip,uid&profile=lumc&defaultdb=aph

50 references d.d. 14-5-2020

TI((Prediction OR Predicting OR Predict OR "Decision support" **OR predict* OR exp "Decision Support System" OR "decision rul*"**) AND (ICU OR "Intensive Care Unit" OR "Intensive Care" **OR "Intensive Care Unit" OR "Intensive Care" OR "ICUs"**) AND ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data analysis" OR "Machine Learning" **OR "Machine Learning" OR "Artificial Intelligence" OR "Data Analysis" OR algorithm***) AND ("Readmission" OR "Discharge" **OR "Hospital Readmission" OR readmiss* OR readmitt* OR "re-admiss*" OR "re-admitt*" OR "Hospital Discharge" OR discharg***)) **OR** SU((Prediction OR Predicting OR Predict OR "Decision support" **OR predict* OR exp "Decision Support System" OR "decision rul*"**) AND (ICU OR "Intensive Care Unit" OR "Intensive Care" **OR "Intensive Care Unit" OR "Intensive Care" OR "ICUs"**) AND ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data analysis" OR "Machine Learning" **OR "Machine Learning" OR "Artificial Intelligence" OR "Data Analysis" OR algorithm***) AND ("Readmission" OR "Discharge" **OR "Hospital Readmission" OR readmiss* OR readmitt* OR "re-admiss*" OR "re-admitt*" OR "Hospital Discharge" OR discharg***)) **OR** KW((Prediction OR Predicting OR Predict OR "Decision support" **OR predict* OR exp "Decision Support System" OR "decision rul*"**) AND (ICU OR "Intensive Care Unit" OR "Intensive Care" **OR "Intensive Care Unit" OR "Intensive Care" OR "ICUs"**) AND ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data analysis" OR "Machine Learning" **OR "Machine Learning" OR "Artificial Intelligence" OR "Data Analysis" OR algorithm***) AND ("Readmission" OR "Discharge" **OR "Hospital Readmission" OR readmiss* OR readmitt* OR "re-admiss*" OR "re-admitt*" OR "Hospital Discharge" OR discharg***)) **OR** AB((Prediction OR Predicting OR Predict OR "Decision support" **OR predict* OR exp "Decision Support System" OR "decision rul*"**) AND (ICU OR "Intensive Care Unit" OR "Intensive Care" **OR "Intensive Care Unit" OR "Intensive Care" OR "ICUs"**) AND ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data analysis" OR "Machine Learning" **OR "Machine Learning" OR "Artificial Intelligence" OR "Data Analysis" OR algorithm***) AND ("Readmission" OR "Discharge" **OR "Hospital Readmission" OR readmiss* OR readmitt* OR "re-admiss*" OR "re-admitt*" OR "Hospital Discharge" OR discharg***))

CHARMS scores for study quality

Table A-5: CHARMS scores for study quality. Not all CHARMS criteria were relevant for assessment of ML model development, indicated with NA (Not applicable).

Domain	Key items	Rojas et al.	Desautels et al.	Loreto et al.	Pakbin et al.	Junqueira et al.	Barbieri et al.	Lin et al.	Venugopalan et al.
Source of data	Source of data (e.g., cohort, case-control, randomized trial participants, or registry data)	2	2	2	2	2	2	2	2
Participants	Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria)	2	2	1	2	2	2	2	1
	Participant description	2	2	1	1	2	2	2	1
	Details of treatments received, if relevant	NA	NA	NA	NA	NA	NA	NA	NA
	Study dates	2	2	2	2	2	2	2	2
Outcome(s) to be predicted	Definition and method for measurement of outcome	2	2	2	2	2	2	2	2
	Was the same outcome definition (and method for measurement) used in all patients?	2	2	2	2	2	2	2	2
	Type of outcome (e.g., single or combined endpoints)	2	2	2	2	2	2	2	2
	Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?	NA	NA	NA	NA	NA	NA	NA	NA
	Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?	NA	NA	NA	NA	NA	NA	NA	NA
	Time of outcome occurrence or summary of duration of follow-up	2	2	2	2	2	2	2	1
Candidate predictors (or index tests)	Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)	2	2	2	1	1	2	2	1
	Definition and method for measurement of candidate predictors	1	2	2	2	2	1	2	1
	Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)	2	2	2	2	1	2	2	2
	Were predictors assessed blinded for outcome, and for each other (if relevant)?	NA	NA	NA	NA	NA	NA	NA	NA
	Handling of predictors in the modeling (e.g., continuous, linear, non-linear transformations or categorised)	2	1	1	2	2	2	2	1

Sample size	Number of participants and number of outcomes/events	2	1	2	2	2	2	2	2
	Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable)	1	1	1	1	1	1	2	2
Missing data	Number of participants with any missing value (include predictors and outcomes)	0	1	1	0	1	0	0	0
	Number of participants with missing data for each predictor	0	1	0	0	1	0	0	0
	Handling of missing data (e.g., complete-case analysis, imputation, or other methods)	0	2	2	1	0	0	2	2
Model development	Modeling method (e.g., logistic, survival, neural network, or machine learning techniques)	2	2	2	2	1	2	2	2
	Modeling assumptions satisfied	2	2	2	1	1	2	2	2
	Method for selection of predictors for inclusion in multivariable modeling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)	2	2	2	1	1	1	2	2
	Method for selection of predictors during multivariable modeling (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)	2	1	2	0	1	1	2	1
	Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)	2	2	0	0	0	2	2	1
Model performance	Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals	0	1	0	1	0	0	0	0
	Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used	0	2	2	1	1	2	2	0
Model	Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators)	2	2	2	2	2	2	2	2
Evaluation	In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)	NA	NA	NA	NA	NA	NA	NA	NA
Results	Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)	2	1	1	2	1	2	1	1

	Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance	1	2	1	0	1	0	2	0
	Comparison of the distribution of predictors (including missing data) for development and validation datasets	NA	NA	NA	NA	NA	NA	NA	NA
Interpretation and discussion	Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)	2	0	2	1	1	2	2	1
	Comparison with other studies, discussion of generalizability, strengths and limitations.	2	2	2	1	2	1	2	0
Total score of quality		**45**	**48**	**45**	**38**	**39**	**43**	**51**	**36**
Percentage (%)		**75**	**80**	**75**	**63.3333333**	**65**	**71.6666667**	**85**	**60**

NA: Not applicable

B. Supplementary material Part II – Questionnaire

Survey questions

Questionnaire: Artificial Intelligence voor ontslag ondersteuning IC patiënten

Het **predictiealgoritme** van Pacmed Critical voorspelt de kans dat een patiënt moet worden heropgenomen na een ontslag van de IC. Hiervoor wordt gebruik gemaakt van **Artificial Intelligence (AI),** waarbij het predictiealgoritme een voorspelling doet op basis van de gegevens in het PDMS. De komende maanden wordt geëvalueerd wat de waarde van dit instrument is, en of het zinvol is om dit instrument op onze IC te implementeren. Daarnaast wordt onderzocht hoe we het predictiealgoritme zouden kunnen gebruiken in onze workflow. Deze vragenlijst is bedoeld om inzicht te krijgen in de huidige ontslagstrategie van IC-patiënten naar de afdeling, en om inzicht te krijgen hoe jullie artsen staan tegenover het gebruik van AI ter ondersteuning van jullie werk

Bij vragen, meer informatie of suggesties kunt u contact opnamen met:

Dr. Sesmu Arbous *M.S.Arbous@lumc.nl*
Siri van der Meijden (masterstudent Technical Medicine) *S.L.van_der_Meijden@lumc.nl*

Bij voorbaat veel dank voor het invullen!

Algemeen

Functie: (Omcirkelen wat van toepassing is) ANIOS/AIOS/Fellow/Staflid
Aantal jaar werkzaam op de IC: ……..
Moederspecialisme: ……..

Ontslagstrategie en Artificial Intelligence op de IC

Stelling	Sterk mee oneens	Oneens	Neutraal	Eens	Sterk mee eens
1. Ik vind de beslissing tot ontslag van een IC patiënt naar de afdeling een complex besluit	○	○	○	○	○
2. Bij de beslissing tot ontslag van een IC patiënt is het risico op heropname een belangrijke factor in mijn overwegingen	○	○	○	○	○
3. In mijn besluit een patiënt te ontslaan is beddendruk een aanwezige factor	○	○	○	○	○
4. Ik ben bekend met het begrip Artificial Intelligence (AI)	○	○	○	○	○
5. Ik geloof dat AI mij in mijn werk zou kunnen ondersteunen	○	○	○	○	○
6. Ik ben bang dat AI mijn werk overbodig maakt	○	○	○	○	○
7. Ik geloof dat AI mijn werk goed genoeg begrijpt om mij te ondersteunen	○	○	○	○	○
8. Ik geloof in de toegevoegde waarde van predictiealgoritmen op basis van AI op de IC	○	○	○	○	○
9. Een predictiealgoritme, wat door middel van AI de kans op heropname voorspelt op basis van de patiëntgegevens in PDMS, kan van waarde zijn bij de beslissing tot ontslag	○	○	○	○	○
10. Het is voor mij belangrijk om inzicht te hebben in de patiënt factoren (bijv. bloeddruk, Hb.) waarvan het algoritme heeft vastgesteld dat ze bijdragen aan de kans op heropname bij een patiënt	○	○	○	○	○

11. Als u heeft besloten dat een patiënt ontslagen kan worden, hoe zeker bent u dan over het algemeen dat de patiënt niet heropgenomen hoeft te worden? (Patiënten met een no-return beleid uitgesloten.)

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○
(helemaal niet zeker)									(Geheel zeker)

12. **Bij de volgende patiëntgroep(en) vind ik de keuze en/of timing van ontslag het meest uitdagend (meerdere antwoorden mogelijk):**

- ❑ Lang liggende patiënten: langer dan … **(vul in)** dagen
- ❑ Eerder heropgenomen patiënten
- ❑ Patiënten boven de …. **(vul in)** jaar
- ❑ COVID patiënten
- ❑ Anders, namelijk: …….

13. **De voorspelde kans op heropname per patiënt zou weergeven moeten worden tijdens (meerdere antwoorden mogelijk):**

- ❑ De ochtend overdracht (7.45 uur)
- ❑ De verdeling van de patiënten voor de ochtend visite (8.30 uur) ❑ Het bedden overleg
- ❑ Aan bed tijdens de ochtend visite (8.45 uur)
- ❑ De grote visite (11.30 uur)
- ❑ Het MDO (14.00 uur)
- ❑ De avond overdracht (16.15 uur)

14. **De voorspelde kans op heropname zou ik willen zien bij (mogelijk) ontslag van (meerdere antwoorden mogelijk):**

- ❑ IC patiënten
- ❑ MC patiënten

15. **De voorspelde kans op heropname zou ik willen zien op de volgende plek:**

- ○ Tabblad 'overzicht' in het PDMS
- ○ Tabblad 'Status' in het PDMS
- ○ Aan te roepen als apart dashboard in het PDMS
- ○ Anders, namelijk …

16. **De volgende voorspelling(en) vind ik het meest relevant bij de beslissing tot ontslag:**

- ○ De kans op heropname binnen 7 dagen na ontslag
- ○ De kans op overlijden binnen 7 dagen na ontslag
- ○ De gecombineerde kans op overlijden/heropname binnen 7 dagen
- ○ Anders namelijk:

17. *Bij een voorspelde kans op heropname van* ☐ ***%** (of groter) zou ik mijn patiënt **niet** ontslaan naar de afdeling.*

18. *Bij een voorspelde kans op heropname van* ☐ ***%** (of kleiner) zou ik mijn patiënt **wel** ontslaan naar de afdeling.*

19. **Ik weet het niet zeker, maar ik denk dat bijna geen enkele voorspelde kans mijn gedrag zou beïnvloeden:**

Sterk mee oneens	Oneens	Neutraal	Eens	Sterk mee eens
○	○	○	○	○

20. **Voor welke andere onderwerpen die spelen bij IC zorg, zou u denken dat AI ingezet zou kunnen worden?**

……

21. **Opmerkingen/suggesties:**

……

Artificial Intelligence voor beslissingsondersteuning bij ontslag van intensive care patiënten

Resultaten enquête LUMC

Siri van der Meijden
Master thesis Technical Medicine
5 januari 2021

1. Introductie

1.1.Pacmed Critical

Het Leidsch Universitair Medisch Centrum (LUMC) is in november 2020 een traject gestart waarin op de Intensive Care (IC) beslissingsondersteunende software wordt geëvalueerd en mogelijk geïmplementeerd: Pacmed Critical [1]. Pacmed Critical voorspelt wat de kans dat een ontslagen patiënt weer moet worden heropgenomen op de IC of op de afdeling overlijdt. Momenteel wordt onderzocht of de software die door ontwikkelaar Pacmed op andere IC's is ontwikkeld, geschikt is om op de IC van het LUMC in te zetten. Er zal een grondige evaluatie van de beslissingsondersteunende software plaatsvinden met een kleine groep intensivisten (de klankbordgroep), waarmee in samenwerking met Pacmed onder andere het gebruik van de software in de dagelijkse workflow en het dashboard worden geëvalueerd.

Om bij aanvang van het project inzicht te verkrijgen in hoe artsen, werkzaam op de IC, tegenover het gebruik van Artificial Intelligence als ondersteuning bij het werk staan, is in december 2020 een korte vragenlijst van 20 vragen uitgevoerd onder de stafleden, fellows en A(N)IOS op de IC van het LUMC.

1.2.Doelen

Het doel van de enquête onder de IC artsen van het LUMC was het verkrijgen van inzicht in:

1. **Huidige overwegingen** in het maken van de beslissing tot ontslag van een IC patiënt naar de afdeling.
2. Hoe artsen tegenover het gebruik van beslissingsondersteunende software op basis van **Artificial Intelligence (AI)** staan in hun werkproces, in het bijzonder bij het ontslag proces van IC patiënten naar de afdeling
3. De gewenste plaats van het Pacmed dashboard in de **workflow** en het gewenste moment van voorspelling
4. De gewenste voorspelde **uitkomstmaat** Pacmed (heropname, mortaliteit, gecombineerd)

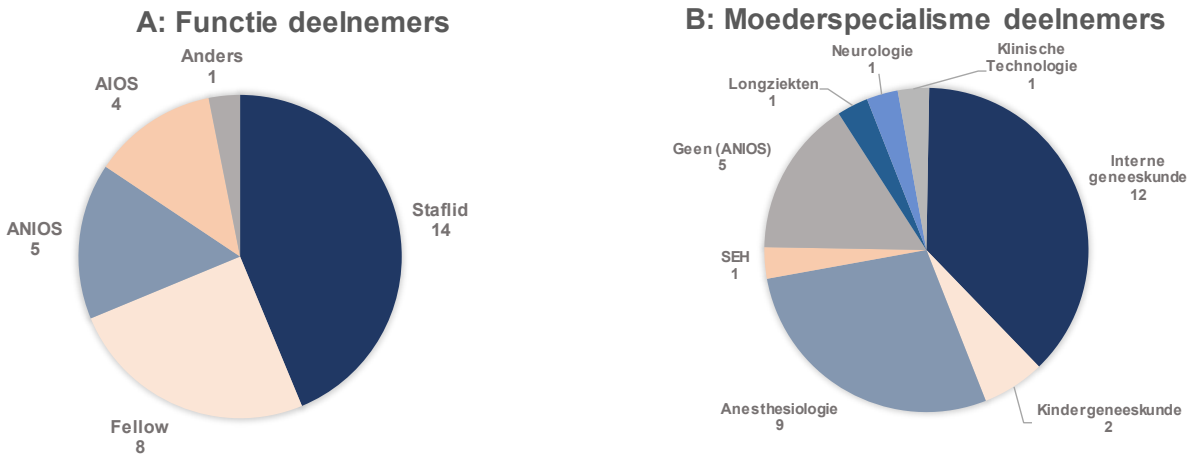
1.3.Begeleidende tekst vragenlijst ter introductie

Het **predictiealgoritme** van Pacmed Critical voorspelt de kans dat een patiënt moet worden heropgenomen na een ontslag van de IC. Hiervoor wordt gebruik gemaakt van **Artificial Intelligence (AI),** waarbij het predictiealgoritme een voorspelling doet op basis van de gegevens in het PDMS. De komende maanden wordt geëvalueerd wat de waarde van dit instrument is, en of het zinvol is om dit instrument op onze IC te implementeren. Daarnaast wordt onderzocht hoe we het predictiealgoritme zouden kunnen gebruiken in onze workflow. Deze vragenlijst is bedoeld om inzicht te krijgen in de huidige ontslagstrategie van IC-patiënten naar de afdeling, en om inzicht te krijgen hoe jullie artsen staan tegenover het gebruik van AI ter ondersteuning van jullie werk.

2. Resultaten enquête

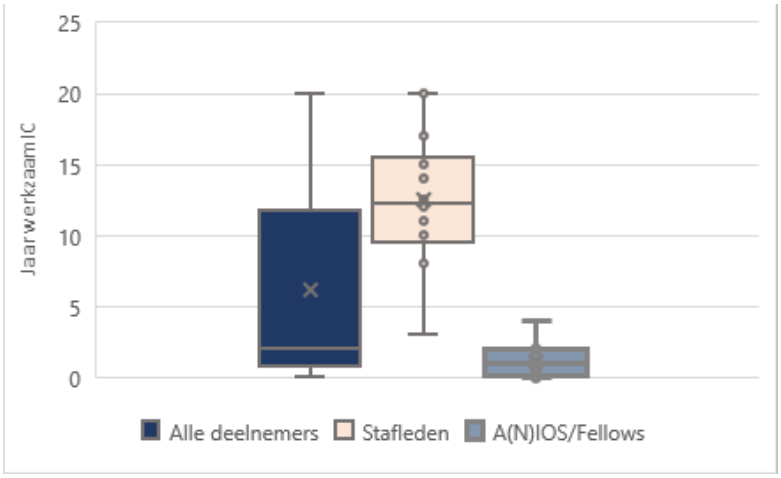
2.1.Deelnemers

In de periode 21 december 2020 tot en met 29 december 2020 is de enquête anoniem uitgevoerd op de intensive care afdeling van het LUMC. Dit resulteerde in 32 respondenten, waarvan 53.1% (n = 17) vrouw. De enquête werd uitgevoerd onder stafleden (intensivisten), fellows (intensivisten in opleiding), AIOS (artsen in opleiding tot specialist) en ANIOS (artsen niet in opleiding tot specialist). Zie Figuur 1 voor de verdeling per functie en moederspecialisme. AIOS en ANIOS werken over het algemeen voor een kortere periode van 3 maanden tot twee jaar op de IC.



Figuur B-1: A: Verdeling functie deelnemers in absolute aantallen, stafleden zijn de intensivisten van de Intensive Care. B: Verdeling moederspecialisme deelnemers in absolute aantallen.

Het gemiddelde aantal in jaren werkervaring op de IC was 6.1 ± 6.4 jaar onder alle deelnemers. Onder de 14 stafleden was dit gemiddelde 12.5 ± 4.5 jaar en onder de overige respondenten 1.1 ± 1.0 jaar (Figuur 2).



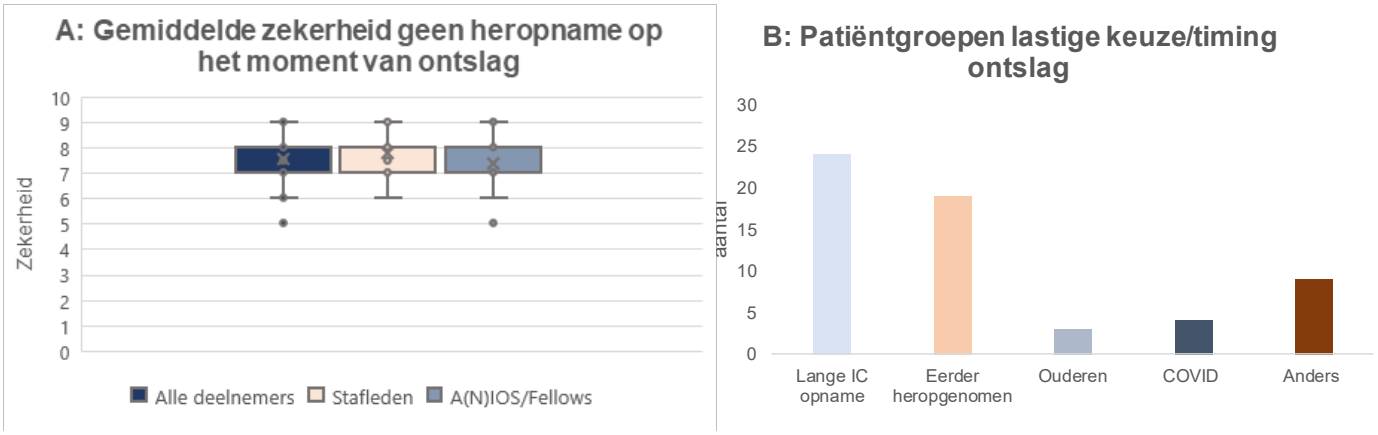
Figuur B-2: Boxplot met aantal jaren IC werkervaring voor alle deelnemers, de stafleden (intensivisten) en overige deelnemers (A(N)IOS, fellows).

2.2. Huidige ontslagstrategie

De deelnemende artsen reageerden uiteenlopend op de vraag of het ontslag van een IC patiënt naar de afdeling een complex besluit is. Op de 5-punts Likert schaal (1 = sterk mee oneens, 2 = oneens, 3 = neutraal, 4 = eens en 5 = sterk mee eens, [2]) is de gemiddelde respons 2.9 ± 0.9. Heropname wordt gezien als een belangrijke factor in de overweging tot ontslag (gemiddelde score 4.1 ± 0.6). Daarnaast speelt beddendruk mee in de beslissing tot ontslag (gemiddelde score 3.8 ± 0.6, Figuur 3). In de appendix zijn de gemiddelde scores uitgesplitst tussen stafleden en A(N)IOS/Fellows. Op de vraag hoe zeker de arts over het algemeen is dat de patiënt niet heropgenomen hoeft te worden op een schaal van 1-10 (1 = helemaal niet zeker, 10 = geheel zeker), was de gemiddelde score 7.5 ± 0.9 (Figuur 4A).

Figuur B-3: Stellingen met betrekking tot de huidige ontslagstrategie van IC patiënten. Antwoorden konden gegeven worden middels de 5-punts Likert schaal.

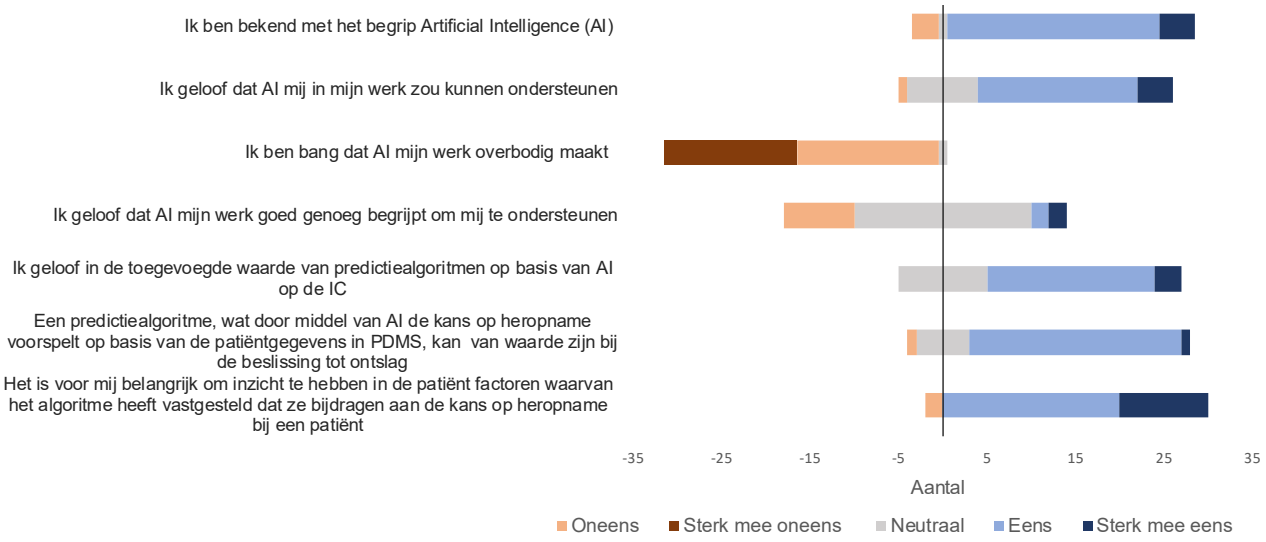
In de vragenlijst konden meerdere patiëntgroepen aangevinkt worden waarbij de beslissing tot ontslag als meest uitdagend wordt ervaren. Vijfenzeventig procent (n=24) van de respondenten benoemde dat de lang opgenomen patiënten een groep was waarbij de beslissing tot ontslag uitdagend was, waarbij zelf een definitie voor 'lang opgenomen' kon worden ingevuld (gemiddeld langer dan 17.6 ± 7.9 dagen opgenomen op de IC). Voor 59% (n = 19) was de groep eerder heropgenomen patiënten uitdagend. Oudere patiënten (9,4%, n = 3) met een zelf aangegeven leeftijdsgrens van gemiddeld 70 jaar, en COVID patiënten (13%, n = 4) werden minder vaak als optie gekozen. Negen deelnemers vulden zelf één of meerdere overige patiëntcategorieën in, waarbij ernstige spierzwakte, al dan niet in combinatie met verminderde hoeskracht, het vaakst werd genoemd (16%, n = 5). Andere groepen die werden genoemd waren patiënten < 3 maanden oud (op de kinder IC), patiënten met ernstig hartfalen, ontslag in de late namiddag/avond, forse sputumproductie, hematologische morbiditeit, 'lastige' chirurgische patiënten, patiënten met no-return/non-IC beleid, en patiënten met onbekende primaire diagnose (Figuur 4B).



Figuur B-4: A: Gemiddelde zekerheid dat een patiënt niet heropgenomen hoeft te worden bij besluit tot ontslag (patiënten met een no-return beleid uitgesloten). B: Patiëntgroepen waarbij de beslissing tot ontslag als meest uitdagend worden ervaren, meerder antwoorden mogelijk. Een gemiddelde van 17.6 ± dagen werd aangeduid als een lange IC opname.

2.3. Artificial intelligence op de IC

Een zestal stellingen zijn met de 5-point Likert Scale aan de deelnemende artsen voorgelegd om te onderzoeken hoe artsen tegenover het gebruik van AI staan in hun werkprocessen op de IC (Figuur 5). De meeste artsen zijn bekend met het begrip AI, en geloven dat AI hen in het werk zou kunnen ondersteunen. Duidelijk is dat ze niet bang zijn dat AI hun werk overbodig maakt. Tweeënzestig procent van de artsen antwoorde 'neutraal' op de stelling waarin werd bevraagd of AI goed genoeg het werk van een arts begrijpt om te kunnen ondersteunen. Zie de appendix voor de scores uitgesplitst tussen stafleden en A(N)IOS/Fellows. Het merendeel (78%) ziet de meerwaarde van AI voor het voorspellen van heropname voor IC patiënten op het moment van ontslag, waarbij het overduidelijk is dat het belangrijk is, voor 94%, om inzicht te hebben in de factoren waarop deze voorspelling is gemaakt.



Figuur B-5: Stellingen (5 punts Likert schaal) met betrekking tot het gebruik van Artificial Intelligence op de IC, met in bijzonder voor het voorspellen van heropname.

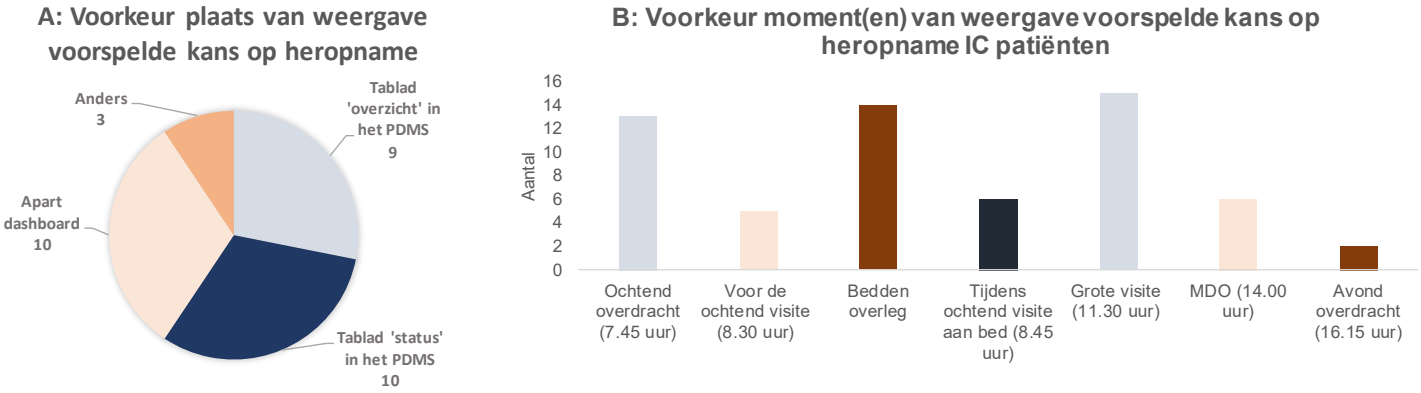
Interessant waren de responses op de vraag waarvoor AI nog meer ingezet zou kunnen worden bij IC zorg. Twintig (63%) van de respondenten gaf hierbij één of meerdere suggesties. De thema's die hierbij het vaakst terugkwamen waren het voorspellen van hemodynamische verslechteringen en/of hypotensie (n = 6), en het voorspellen van respiratoire achteruitgang, beademing en/of extubatie failure (n = 6). Daarnaast werden genoemd:

- Capaciteitsvoorspelling & voorspellen 'futility' van zorg
- Prognosebepaling hematologische patiënten
- Voorspelling dat een 80+er binnen drie dagen van de IC is (korte opnames ter overbrugging)
- Diagnoses: nierinsufficiëntie, longembolieën, Crticial Illness neuropathie/myopathie
- IC opname triage
- Medicatie review
- Diagnostische ondersteuning bij echografie
- Mortaliteitskans
- Voorspelling behoefte aan nierfunctie vervangende therapie
- Sedatieregulatie
- Medicatie met TDM
- Kans op overleven op de IC
- Protocolvorming bij aanpak circulatoire shock
- ECMO trends en inzet van ECMO
- Trends in P/F ratio of andere parameters
- Scores van beeldvorming en lab voor prognose
- Sondevoeding en metabolisme
- Beslissingsondersteuning bij wel of geen IC opname ouderen (70+ jaar).

2.4. Workflow

De voorkeur voor de plek in het PDMS (Patient Data Management Systeem, Metavision 6) waar de voorspelling weergeven moet worden verschilde tussen de deelnemende artsen, zie Figuur 6A. De opties 'Tablad 'overzicht' in het PDMS', 'Apart dashboard', en 'Tablad 'status' in het PDMS' werden ongeveer even vaak genoemd. De voorspelling zou op meerdere momenten in de workflow kunnen worden ingezet. Eén of meerder opties konden hiervoor gekozen worden, waarbij de ochtend overdracht (41%, n = 13), het beddenoverleg (44%, n = 14) en de grote visite (47%, n = 15 uur) het vaakst werden gekozen (Figuur 6B). Drie artsen gaven een andere optie, namelijk in Hix, geen voorkeur,

en een andere lay-out van PDMS waarin in één oogopslag opname diagnose, actuele problematiek, beleid en beloop weergeven worden. Deze laatste optie is vergelijkbaar met het dashboard dat Pacmed Critical levert.

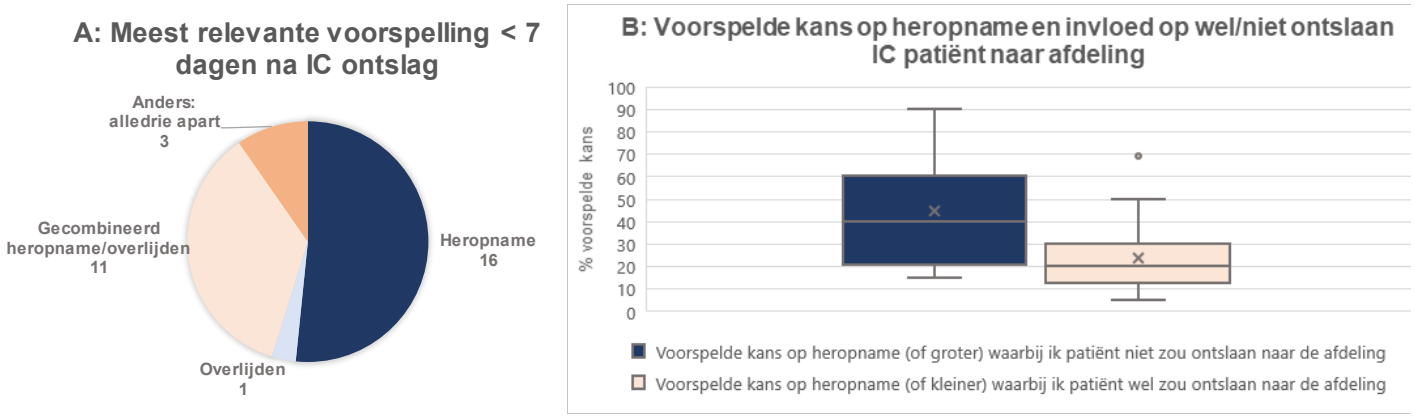


Figuur B-6: A: Voorkeur plaats van weergave voorspelde kans op heropname in het PDMS. Verdeling in absolute aantallen. B: Voorkeur moment(en) van weergave voorspelde kans op heropname IC patiënten.

2.5. Uitkomstmaat voorspelde kans op heropname

Vierentachtig procent (n = 27) van de deelnemende artsen wilt voor zowel medium care (MC) als IC patiënten de kans op heropname kunnen inzien. Pacmed Critical gebruikt de samengestelde voorspelde kans op heropname/mortaliteit binnen 7 dagen na ontslag. De respondenten werden daarom gevraagd welke voorspelde uitkomstmaat zij het meest relevant vinden bij de beslissing tot ontslag. Zestien respondenten (50%) kozen hierbij voor de kans op heropname binnen 7 dagen na ontslag, en 11 respondenten (34%) voor de gecombineerde uitkomstmaat kans op heropname/overlijden. Drie keer werd aangegeven dat alle uitkomstmaten apart weergeven moeten worden (Figuur 7A). Overige voorgestelde opties waren de kans op heropname binnen 48 uur, en de kans op heropname/overlijden binnen 30 dagen, 6 maanden en 1 jaar na ontslag.

Daarnaast werd gevraagd een schatting te geven bij welk voorspeld percentage kans op heropname dit van invloed zou zijn op het wel/niet ontslaan van een IC patiënt (Figuur 7B). De percentages die hierbij werden ingevuld waren sterk verdeeld, met een gemiddelde van 44.5 ± 23.4% of groter waarbij een patiënt **niet** ontslagen zou worden, en een gemiddelde van 23.6 ± 13.8 % of kleiner waarbij een patiënt **wel** ontslaan zou worden. Drie artsen gaven aan hier geen antwoord op te kunnen geven, gezien dit uiteraard per casus verschilt. Hierop aanvullend werd middels de stelling *"Ik weet het niet zeker, maar ik denk dat bijna geen enkele voorspelde kans mijn gedrag zou beïnvloeden"* onderzocht of de deelnemende artsen openstaan voor het gebruik van Pacmed Critical. Hierop reageerde de meeste respondenten met "Oneens" (n = 14) of "Neutraal" (n = 13).



Figuur B-7: A: Meest relevante voorspelde uitkomstmaat binnen 7 dagen na IC ontslag, verdeling in absolute aantallen. B: Verdeling voorspelde kans waarbij invloed op beslissing tot ontslag plaatsvindt.

2.7. Opmerkingen en suggesties

Een aantal nuttige suggesties en opmerkingen werden ingevuld aan het einde van de vragenlijst:

- “Zorg voor een goed vangnet op de afdeling bij een hoog voorspelde kans op heropname (bijv. regelmatige controles door de IC).”
- “IC patiënten worden bij ontslag geaccepteerd door de geschikte afdeling indien er voldoende plek is. CAVE: wellicht accepteert de afdeling een patiënt niet meer indien de voorspelde kans op heropname hoog is.”
- “Externe specialisten kunnen het als absolute maat zien en daardoor kan ontslaan lastig worden. Soms ontsla je toch met een kans op heropname, zodat je andere IC-behoeften kunt voorzien. Heropname is niet altijd vermijdbaar, het is de vraag of blijvend op de IC het beloop van de patiënt anders zou zijn. Je ontslaat soms mensen waarvan je weet dat ze terug zullen komen!”
- “Laat de kans zien bij vaste afspraken.”
- “Belangrijk om in te zien waar de score op gebaseerd is!”
- “SVP ook de kinder-IC betrekken en onderzoeken!”
- “Veel kanten aan deze materie, of dit beter wordt dan onze ervaring/gut-feeling zal ik kritisch beschouwen.”
- “Het percentage moet samen genomen worden met klinische blik. Geen absolute waarde.“

3. Discussie en conclusie

De intensive care artsen van het LUMC staan over het algemeen positief tegenover het gebruik van beslissingsondersteunende software op basis van Artificial Intelligence in hun werkproces. In vergelijking met een studie van Oh et al. naar het vertrouwen van artsen in AI, waren de artsen van de IC van het LUMC vaker bekend met AI, en waren ze minder bang dat AI het werk overbodig zou maken [3]. De beddendruk op de IC is van invloed op beslissing tot ontslag, wat tijdens de uitbraak van het SARS-CoV-2 virus extra relevant is [4]. Een tool waarbij de arts ondersteund kan worden in de beslissing tot ontslag, zou daarom juist nu extra van waarde kunnen zijn.

Vrijwel alle respondenten gaven aan het belangrijk te vinden om inzicht te hebben in de patiëntfactoren die leiden tot de voorspelling. De behoefte aan een uitlegbaar model is dus groot. De meningen zijn verdeeld over de complexheid van het besluit tot ontslag, maar duidelijk is dat kans op heropname en beddendruk een belangrijke factor zijn in de beslissing. Over het algemeen gaven artsen een score van 7.5 ± 0.9 op een schaal van 1-10 hoe zeker ze zijn dat een patiënt niet heropgenomen hoeft te worden na ontslag. De keuze en/of timing tot ontslag blijkt het lastigste bij lang liggende patiënten (langer dan gemiddeld 17.6 dagen), eerder heropgenomen patiënten en spierzwakke patiënten. Er was heterogeniteit in de responses voor de gewenste plek van implementatie van Pacmed Critical voor ontslag ondersteuning in de workflow, verdere evaluatie is hiervoor daarom gewenst in samenwerking met de klankboordgroep. Dit geld ook voor de meest relevante voorspelde uitkomstmaat (heropname, overlijden, of de combinatie van beide), waarbij duidelijk werd dat 84% van de artsen de voorspelling zowel bij IC als MC patiënten relevant vindt. De hoogte van de voorspelde kans op heropname waarbij een arts wel (gemiddeld 23.6 ± 13.8 % of kleiner) of niet (gemiddeld 44.5 ± 23.4% of groter) een patiënt zou ontslaan naar de afdeling verschilden sterk. Het wordt waardevol om de uiteindelijke verschillen in voorspelde kans van het Pacmed algoritme versus de arts bij ontslag te onderzoeken, wat gedaan zal worden middels een prospectief onderzoek. Tot slot werden veel nuttige suggesties gegeven voor de inzet van AI in de IC zorg, en werden een aantal kritische kanttekeningen geplaatst bij het voorspellen van heropname van IC patiënten.

4. Referenties

1. Pacmed. Intensive Care. [Internet]. Available from: https://www.pacmed.ai/nl/projects/ic. [Accessed December 29, 2020]
2. Likert R. A technique for the measurement of attitudes. Arch psychology. 1932;22(140):55.
3. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019;21:e12422
4. Vincent JL, Creteur J. Ethical aspects of the COVID-19 crisis: how to deal with an overwhelming shortage of acute beds. Eur Heart J Acute Cardiovasc Care. 2020;9(3):248–252.

C. Supplementary material Part III – Model development

A detailed description of available variables **(Part 1),** pre-processing of data and feature engineering **(Part 2)**, and modeling and evaluation methods **(Part 3)** of the paper *'Predicting Intensive Care Unit readmission: performance and explainability of machine learning algorithms'* are provided in this Supplementary material. The applied model development steps are simplified in Figure 28.

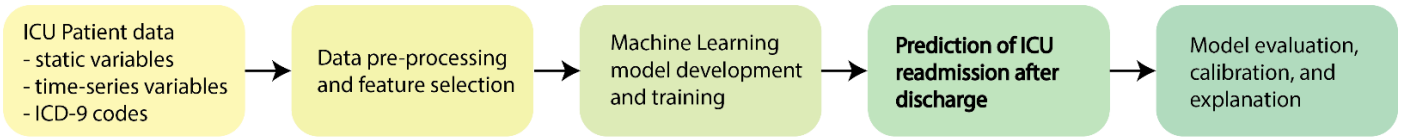


Figure 28: Steps in prediction modeling of ICU readmission. First, ICU patient data was collected and explored. Both static data (patient characteristics) and time-series data (laboratory results, vital functions, medication data) was used for prediction modeling. Second, feature engineering, feature selection, and data pre-processing was performed to get meaningful data to train the prediction models on. Third, Machine Learning algorithms (logistic regression, boosting algorithms, and neural networks) were optimized and trained. Fourth, final models used to predict ICU readmission within 7 days after discharge on an internal test dataset. Fifth, the discriminative performance and calibration properties of the final models were evaluated. Lastly, explanations were evaluated using logistic regression coefficients and SHAP values.

Part 1: Variable description and exploratory data analysis

1. Variable types

Prediction of Intensive Care Unit (ICU) readmission was performed using patient variables from the Patient Data Management System (PDMS, MetaVision version 5 and 6, IMDsoft, Tel Aviv, Israel) and the Electronic Health Record (EHR, HiX, Chipsoft, Amsterdam, The Netherlands). A division can be made between static variables and time-series variables. Static variables include patient characteristics and admission information, and do not change over the course of the ICU stay. Time-series variables include laboratory results, vital signs and device data, clinical observation and scores, diagnostics, and therapeutics[1]. Another division in variable types can be made between categorical and numerical features. Categorical features include for instance sex (male/female) and treating specialty (e.g., thoracic surgical, internal medicine). Numerical features include for instance length of stay in days and heartrate levels.

Not all variables of interest were available at time of prediction modeling, 63 out of 83 variables from all categories, except clinical observation and scores, could be used for the prediction of readmission. Due to a large amount of missing data, a subset of 29 variables was used for model development. See Table 1 in the main paper (page 34) for an overview of the included variables.

2. Descriptive statistics and statistical differences between groups

For all available variables, we explored descriptive statistics and looked for statistical differences between the group of readmitted patients and not readmitted patients. For categorical variables, percentages per subgroup were determined and Chi-Squared testing was performed for statistical testing. For numerical variables, median and interquartile ranges are provided and statistical testing was performed using Wilcoxon Rank test. It is important to note that for this part, we only tested median aggregates between the two groups. Other aggregates (e.g., minimum, maximum, standard deviation, slope) were only used during model training.

2.1. Patient characteristics

6.71% of the 12,189 included patients were readmitted to the ICU within 7 days after discharge. No significant differences in gender and vasoactive drug use between the two groups was found. Readmitted patients had significant higher 30-day mortality rates, were more often emergency admissions, and more often general surgical patients. See Figure 29 for the difference in treating specialties among readmitted and not readmitted patients. The combined outcome for readmission/mortality within 7 days after discharge was 8.6%. In Figure 30, cumulative readmissions and deaths over time after discharge are visualized. It can be seen that the curve of readmissions flattens after 7 days after discharge, indicating that this time point covers most readmissions.

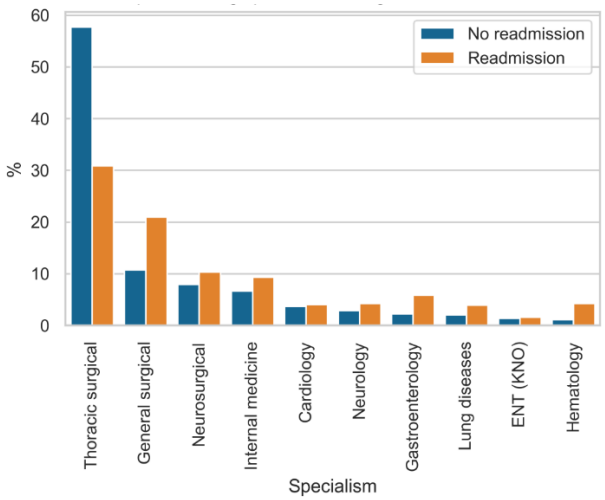


Figure 29: Differences in proportions of readmitted patients to the ICU for the main 10 treating specialties at the ICU department of the LUMC.

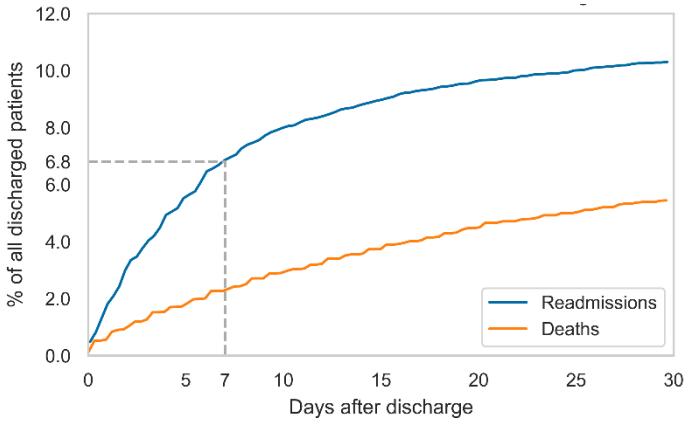


Figure 30: Cumulative readmitted and deceased patients after ICU discharge at day 0.

2.2. Laboratory results

Descriptive statistics of all available laboratory results are summarized in Table 6. Laboratory results are time-series variables and therefore measured multiple times during the ICU admission. Therefore, we looked at differences in median over the whole ICU admission. Similar results were found for last measured values before discharge. For most laboratory results, significant differences were found in median values over the whole admission although sometimes differences were small. Higher ALAT, ASAT, and alkaline phosphatase levels were found for the readmission group. This finding indicates that liver function and/or gallbladder problems are correlated with readmission risk. Also infection status is correlated with readmission risk, with higher C-reactive protein and leukocytes levels for the readmission group. Troponin levels, a marker for heart failure, were lower for the readmission group. This can be explained by the lower rate of thoracic surgical patients in the readmission group. Thoracic surgical patients often have high Troponin levels after open heart surgery.

Table 5: Laboratory measurements in median (Inter Quartile Range (IQR)) over the entire ICU admission. P-values calculated using Wilcoxon Rank test. P-values marked in orange were statistically different ($p < 0.05$).

name	Readmission median (IQR)	No readmission median (IQR)	p-value
ALAT average admission	30.75 (19.0 - 63.5)	24.0 (16.0 - 42.0)	< 0.001
APTT average admission	36.02 (31.6 - 42.41)	32.82 (29.88 - 37.31)	< 0.001
ASAT average admission	49.0 (31.04 - 93.65)	44.33 (31.0 - 72.0)	0.001
Albumin average admission	28.63 (24.52 - 33.0)	30.0 (27.0 - 34.0)	< 0.001
Alkalic Phosphatase average admission	100.67 (68.38 - 164.25)	78.0 (57.0 - 126.38)	< 0.001
Amylase average admission	66.0 (39.1 - 143.0)	66.0 (41.0 - 126.0)	0.78
BE average admission	-1.29 (-4.03 - 1.6)	-2.64 (-4.44 - -0.33)	< 0.001
BSE average admission	82.33 (30.0 - 115.5)	26.5 (9.0 - 67.0)	< 0.001
Bicarbonate average admission	23.04 (20.56 - 25.85)	22.29 (20.6 - 24.25)	< 0.001
Bilirubin Total average admission	12.0 (8.0 - 22.0)	10.0 (7.0 - 15.0)	< 0.001
CK average admission	388.0 (137.5 - 814.37)	478.42 (285.25 - 761.69)	< 0.001
CRP average admission	89.75 (48.69 - 162.57)	73.0 (41.1 - 121.43)	< 0.001
Chloride average admission	104.83 (100.1 - 108.08)	104.95 (102.29 - 107.67)	0.066
Gamma GT average admission	82.5 (38.5 - 186.0)	55.0 (25.0 - 129.25)	< 0.001
Glucose average admission	8.07 (7.28 - 9.12)	7.9 (7.01 - 8.91)	0.018

Hb average admission	6.0 (5.4 - 7.07)	6.53 (5.8 - 7.35)	< 0.001
Ionized calcium average admission	1.14 (1.1 - 1.18)	1.16 (1.12 - 1.2)	< 0.001
Potassium average admission	4.07 (3.87 - 4.35)	4.2 (3.95 - 4.49)	< 0.001
Creatinine average admission	95.17 (65.76 - 142.0)	82.0 (65.5 - 106.0)	< 0.001
LDH average admission	331.0 (233.0 - 469.23)	315.5 (236.0 - 421.0)	0.016
Lactate average admission	1.57 (1.2 - 2.02)	1.4 (1.1 - 1.82)	< 0.001
Leukocytes average admission	13.22 (10.14 - 16.94)	12.5 (9.94 - 15.47)	< 0.001
MCV average admission	90.59 (87.25 - 94.45)	90.31 (87.33 - 93.5)	0.033
Magnesium average admission	0.87 (0.74 - 1.0)	0.89 (0.76 - 1.0)	0.015
Sodium average admission	138.35 (136.0 - 141.36)	138.16 (136.2 - 140.11)	0.008
Neutrophil Granulocytes average admission	17.27 (11.42 - 80.56)	14.0 (8.32 - 72.9)	0.051
O2 Saturation average admission	94.1 (91.15 - 95.92)	92.5 (88.6 - 95.76)	< 0.001
PT average admission	17.03 (15.53 - 19.52)	16.38 (15.35 - 17.94)	< 0.001
Total protein average admission	42.0 (20.8 - 52.0)	44.0 (25.82 - 52.5)	0.104
Thrombocytes average admission	188.0 (129.86 - 263.42)	178.55 (138.82 - 231.0)	0.042
Troponin T average admission	171.05 (16.0 - 777.62)	433.0 (163.67 - 879.5)	< 0.001
Urea average admission	9.45 (6.3 - 16.29)	6.65 (5.1 - 9.5)	< 0.001
Anorganic phosphate average admission	1.22 (0.98 - 1.47)	1.11 (0.92 - 1.33)	< 0.001
pCO2 average admission	5.12 (4.64 - 5.58)	5.13 (4.76 - 5.51)	0.497
pH average admission	7.4 (7.36 - 7.45)	7.38 (7.35 - 7.42)	< 0.001
pO2 average admission	11.8 (10.55 - 13.47)	12.32 (10.93 - 14.43)	< 0.001

2.3. Medication

Medication included for analysis were vasopressor (epinephrine, noradrenaline) and inotropic drugs (dobutamine, enoximone/milrinone), see Table 7. These medication types create vasoconstriction and/or increase cardiac contractility which is needed to treat critically ill patients in shock[2]. Noradrenaline, dobutamine, and Milrinone total dosages were significantly higher for the readmission group. This finding indicates that patients readmitted to the ICU were in need of more drugs to maintain adequate blood pressure. Enoximone and Milrinone are similar agents, but surprisingly for enoximone no significant differences were found. A few years ago, the ICU department of the LUMC switched from using Enoximone to Milrinone.

Table 6: Medication total dosages in median (IQR). Enoximone and Milrinone are similar medication types but are displayed separately due to dosage differences. P-values calculated using Wilcoxon Rank test. P-values marked in orange were statistically different ($p < 0.05$).

name	Readmission median (IQR)	No readmission median (IQR)	p-value
Epinephrine sum dosage	3.77 (0.86 - 6.84)	1.73 (0.48 - 5.8)	0.184
Dobutamine sum dosage	657.45 (154.58 - 1652.8)	370.33 (142.01 - 1000.0)	0.005
Enoximone sum dosage	197.09 (93.56 - 399.93)	148.38 (74.22 - 290.22)	0.319
Milrinone sum dosage	47.87 (17.46 - 103.41)	23.31 (10.27 - 57.98)	<0.001
Noradrenaline sum dosage	12.61 (2.53 - 59.41)	2.68 (0.65 - 13.12)	<0.001

2.4. Vital functions

Vital functions representing the physiological state of a patient are closely monitored at the ICU. We found a small, but significant, difference in median arterial blood pressure (ABP) for patients readmitted to the ICU, see Table 8. Median respiratory rate and median heartrate were higher for readmitted patients, which are known to be increased for critically ill patients. Overall, difference in median vital function between the two groups were

small. It is likely that other variable statistics (minimum, maximum, slope) show larger differences between the two groups.

Table 7: Vital functions in median (IQR). P-values calculated using Wilcoxon Rank test. P-values marked in orange were statistically different (p < 0.05).

name	Readmission median (IQR)	No readmission median (IQR)	p-value
ABP mean average admission	79.96 (73.0 - 87.75)	78.5 (73.0 - 86.0)	0.004
Respiratory rate average admission	18.5 (16.0 - 21.83)	16.75 (14.5 - 19.11)	<0.001
CVP average admission	8.0 (6.0 - 11.17)	7.58 (5.5 - 10.27)	0.306
FiO2 (inst) average admission	37.5 (32.5 - 41.67)	38.04 (30.0 - 40.0)	0.216
Heartrate average admission	84.25 (75.52 - 94.22)	79.8 (70.5 - 88.5)	0.002
NIBP mean average admission	81.12 (69.62 - 94.62)	80.27 (70.38 - 93.0)	0.441
PEEP (contr) average admission	5.14 (5.0 - 6.5)	5.0 (5.0 - 6.0)	<0.001
Ppeak (contr) average admission	17.12 (15.0 - 20.31)	17.0 (15.0 - 19.5)	0.255
SpO2 average admission	97.0 (95.84 - 98.56)	97.53 (96.42 - 99.0)	0.015
Temperature 1 average admission	37.09 (36.71 - 37.4)	37.05 (36.7 - 37.4)	0.686
Tidal Volume exp (contr) average admission	542.38 (472.04 - 620.54)	549.0 (475.5 - 629.5)	0.145
Oxygen l/min (inst) average admission	3.0 (2.0 - 4.5)	2.67 (1.75 - 4.0)	<0.001

3. Missing data

Missing data is an important factor when using patient data for prediction modeling, because proportions of missing data in EHRs can be high. Variables are known to be recorded under multiple names, and in different databases. For instance, we dealt with the transition between two PDMS systems in our database, making it difficult to get all data for each variable. See Figure 31 for an overview of proportions of data missing for all included patients. There was no missing data for the included static variables. For some variables, missing data can be not at random, meaning there is a reason for missingness. E.g., adrenaline is only administered to patients with extreme low blood pressure, and tidal volume is only measured in patients receiving mechanical ventilation. These types of missing data could have a relation with the predicted outcome and can therefore not be neglected. Other high levels of missing data can be missing completely at random, due to for instance the transition in the PDMS. Data missing not at random can occur for instance in laboratory values, Troponin T is most often only measured in cardiology patients[3]. Variables included for final model development are indicated with an orange star in Figure 31.

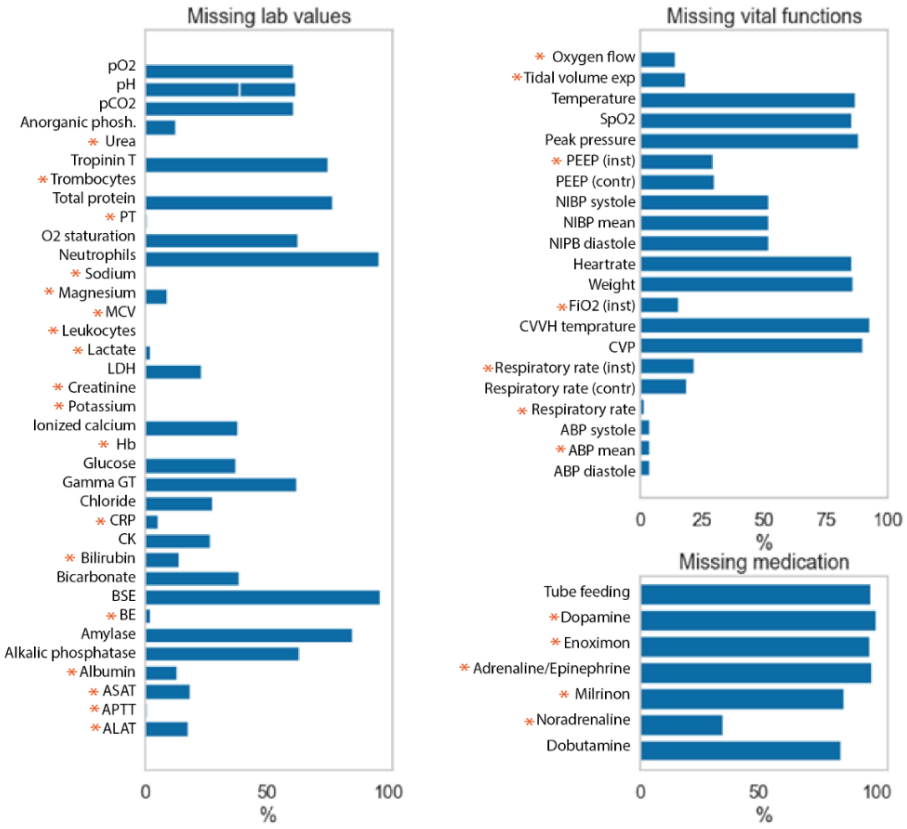


Figure 31: Overview of proportions of missing variables for all included patients.
** = included for model development.*

Part 2: Feature engineering and pre-processing

4. Data cleaning

PDMS data is prone to artefacts. Ideally, extensive data cleaning is performed to erase outliers in the dataset. For laboratory results, we could use physiological ranges as provided by Pacmed to erase outliers before feature aggregation. Outlier removal of vital functions was not possible before feature aggregation, because feature aggregates were directly created from the database. Therefore, we used median values instead of mean to correct for possible outliers. The presence of remaining outliers in our dataset could limit the performance, and for future studies, more elaborate data cleaning should be performed.

5. Feature engineering

Some algorithms (e.g., recurrent neural networks) are capable of using raw time-series data as input to train prediction models on. The use of these more advanced models was out of the scope of this research. Therefore, feature engineering was performed to capture summary statistics and time-related trends of time-series variables. Feature engineering was performed similar as described by Thoral et al.

Feature aggregates for medication included total dosage, total duration of administration, total duration of administration as fraction of LOS, and time since last use before discharge. Furthermore, a binary indicator of for use at moment of discharge was added. For laboratory results and vital signs, feature aggregates were calculated for three time windows. The first 24 hour of admission, the last 24 hour of admission, and the average of the whole admission. See Table 9 for an overview of feature aggregates used for prediction modeling on laboratory results and vital functions.

Table 8: Feature aggregates (summary statistics) of time-series variables. Three time-windows were used for calculating feature aggregates, the first 24h, the last 24h and the whole ICU admission. Adjust from Thoral et al.[1]

Feature type	Time-window	Function
Time-series: Vital signs and laboratory results	**Last 24h**	Mean or median
		Minimum
		Maximum
		Standard deviation
		First
		Last
		Count (number of measurements)
		Missing (yes/no)
		Slope
	First 24h	Mean or median
		Minimum
		Maximum
		Count (number of measurements)
		Standard deviation
		Missing (yes/no)
	Average whole admission	Mean/median
		Minimum
		Maximum
		Count (number of measurements)
		Standard deviation
		Missing (yes/no)

6. Pre-processing and feature selection

Besides feature engineering, other pre-processing steps of the data were used to get suitable input data for all prediction model types. To avoid data leakage during model development, all pre-processing steps were fitted on the training dataset before they were applied to the testing dataset[4]. Feature engineering was performed before further pre-processing. All data tables (characteristics, medication, laboratory results, vital signs) were merged into one dataset, and split in an 80% training and 20% test dataset. A pre-processing pipeline was manually built in Python through which the data was processed in the following order:

1. The column names of the dataset were categorized in categorical columns (e.g., specialization type) and numerical columns (e.g., age). Furthermore, an extra column was added for all variables with on or more missing values. For feature aggregates, a column for missingness was added for each time window.
2. A binary indicator (1 or 0) was given to indicate whether a variable was missing for each patient in the corresponding missingness column.
3. Missing data was imputed using mean imputation for numerical variables (mean value in the training dataset). Other types of imputation tested included K-nearest neighbor and median imputation, but this yielded inferior performance. More advanced types of imputation (e.g., population averages) were not evaluated.
4. Missing data was imputed using mode imputation for categorical variables.
5. All numerical variables were scaled with zero mean and unit variance (standard-scaling). This step was performed to correct for scale differences between variables.
6. Categorical features were transformed to numerical features using one-hot encoding, also known as dummy encoding. This method creates a new column for each category, with a 1 in the column for the corresponding category of a patient. One-hot encoding is often preferred over ordinal encoding (e.g., CTC = 1, CHI = 2 …) because the model can misinterpret a certain order in the categories[4].

Feature selection of all pre-processed features (1380 features all variables included 550 features for subset with variables with too much missing data dropped) was performed using L1-feature selection as described by Thoral et al. The C-parameter for regularization was set on 1.5. This resulted in respectively 703 and 416 features to train the prediction models on.

7. Labelling

Each patient in the dataset was given a label for the predicted outcome (readmission within 7 days after discharge). The labels were used to train and test the prediction algorithms on. Labelling of the dataset was not trivial because one ICU admission was often divided in multiple sub-admissions in the dataset. Therefore, all sub admissions (sub-encounters) first needed to be merged to get a start and end time of each ICU admission. We merged all sub-encounters where end and start time were less than 2 hours apart. For each patient, the label ‘readmission’ was given if a new start time of ICU admission was within 7 days of the last ICU discharge.

Part 3: Modeling and evaluation

A detailed explanation of modeling methods and model evaluation is provided in this section.

8. Experimental set-up

An experimental set-up is needed for model development, see Figure 32. After data exploration and feature engineering, the dataset (including all patients with corresponding features and labelled for the outcome readmission) was split into two subsections. Splitting of the dataset split was performed in a stratified manner; this means that the proportions of patients readmitted (positive class) and not readmitted (negative class) were equal in all sets. The training dataset was used to train pre-processing steps on (e.g., imputation values) and for model development during hyperparameter tuning. 20% of all patients in the dataset were used for finale model evaluation. By doing so, the generalizability of the models on unseen data could be evaluated. Ideally, the model performs equal on the training and test dataset, indicating that the optimum in the "Bias-variance tradeoff" is reached. An overfitted model will result in a high training/validation dataset performance, and low test performance (high variance). An underfitted model will have a large bias, because it is not possible to determine the relation between the predictors and the outcome (and a low general performance)[4].

Stratified 5-fold Cross-Validation (CV) was performed during hyperparameter tuning and final model training on the training dataset. CV is a method in which the dataset is split multiple times to get average performance on different parts of the training dataset. Each fold included a 80% training and 20% validation part of the training dataset. CV enabled us to assess generalizability of models before final model evaluation.

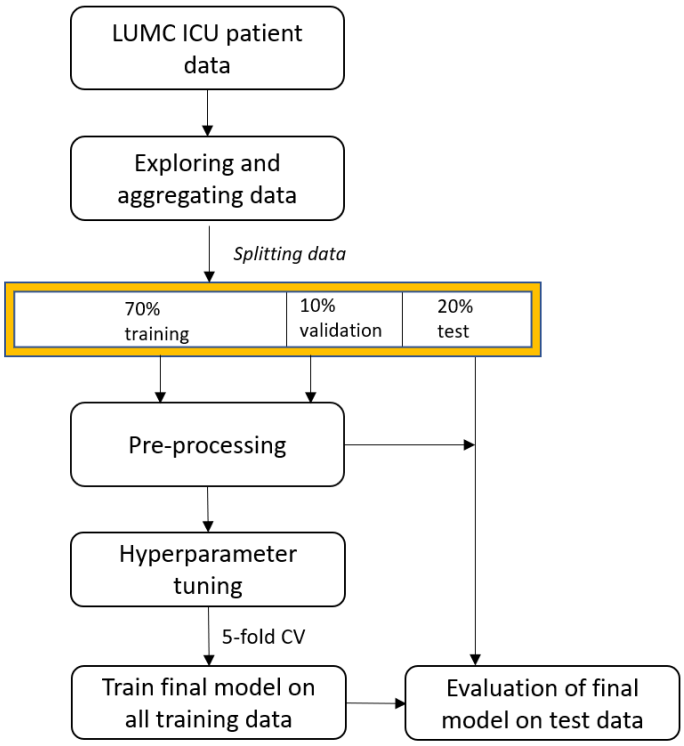


Figure 32: Experimental set-up for model development.

9. Hyperparameter tuning strategy

Hyperparameters are the internal settings of a Machine Learning model that can be adjusted to get optimal discriminative performance. For each model type, many hyperparameters can be adjusted. The process of optimizing the hyperparameters for a classification task is called hyperparameter tuning. Different types of tuning strategies can be applied. E.g., grid search tests all possible combination of hyperparameters. We used a Bayesian optimization strategy. Bayesian optimization converges to a set of hyperparameters based on previous performance, and therefore limits training time[5]. An objective metric needs to be specified to evaluate the

performance of a set of hyperparameters (e.g., 2 hidden layers in a neural network, with 16 and 32 nodes etc.). We chose area under the precision recall curve (AUCPR) as objective metric since this could be applied for all model types and is suitable for imbalanced datasets[6]. Each combination of hyperparameters was trained and tested during 5-fold CV. The combination of hyperparameters with the highest AUCPR was chosen as 'final model' configuration.

Hyperparameter tuning was performed in Keras Tuner[7], a Python library. In Keras Tuner, we manually specified for each hyperparameter a range of options, called the hyperparameter space. Furthermore, we tested multiple types of pre-processing and feature subsets. Best results were obtained using standard scaling and L1 feature selection with the subset of features with acceptable missing data.

10. Logistic Regression hyperparameter tuning

For logistic regression (LR), only one hyperparameter needed tuning: the C-parameter for L2-regularization to prevent overfitting. A range of 0.001-100 was evaluated for the C-parameter, according to Thoral et al. A small C has high regularization strength, and therefore reduces the chance of overfitting. L1-regularization, as used during feature selection, shrinks the coefficient of the least important features to zero. L2-regularization does not perform feature selection but reduces the impact of each feature by forcing the coefficients to decrease[8]. The final LR model had a C parameter of 0.002.

11. Neural network hyperparameter tuning

A feed-forward neural network (FFNN) was optimized using the hyperparameters as specified in Table 10. Settings that were unchanged during hyperparameter tuning were batch size (520), the loss function (binary cross-entropy), and the activation function (ReLU)[9]. Each configuration was trained for 100 epochs, where early stopping was performed when validation loss did not improve for more than 10 epochs.

The finale FFNN was composed out of an input layer, 4 fully connected deep layers (480, 320, 320, and 160 nodes), and a sigmoid output layer for readmission prediction between 0 and 1. Weighted training was performed to account for class imbalance. To reduce overfitting and increase generalizability, drop-out and L2-regularization was implemented[4]. L2-regularization works similar for FFNNs as for LR, by decreasing the magnitude of the weights to avoid overfitting. The drop-out hyperparameter between 0 and 1 determines the proportion of nodes that is used for training in each layer. This means that during training, a random proportion of nodes is not used, preventing the network to give high weights to certain features and therefore prevent overfitting[10].

Table 9: Hyperparameters optimized for feed forward neural networks

Hyperparameters Neural Network	Range	Intervals	Final configuration
Input layer drop-out	0.05-0.2	Log sampled	0.06
Number of deep layers	2-5	1	4
Deep layer drop-out	0.4 – 0.6	Log sampled	0.44
Number of units in each dense layer	32-512	32	480, 320, 320, 160
Batch normalization	True/false		False
L2-regularization	1e-1 - 1e-6	1e-1	0.01
Drop-out hidden layers	0.4-0.6	Log sampled	0.44
Learning rate	1e-2 – e-5	1e-1	0.01

12. Boosting hyperparameter tuning

Two types of boosting (tree-based ensemble) algorithms were used for prediction modeling. Gradient Boosting machines (GB) as described by Thoral et al., and eXtreme Gradient Boosting (XGBoost)[11]. The difference between boosting methods and other ensemble methods, e.g., random forests, is that gradient boosting builds an ensemble of trees while iteratively improving on the previous tree. Each new decision tree in the sequence focusses on improving model performance by fixing the largest prediction errors of the previous tree. XGBoost is

known for its improved regularization performance compared to GB and is one of the current state-of-the-art ML used for a wide range of applications[12].

Hyperparameters tuned for GB and XGBoost are summarized in Table 11 and Table 12. The max depth is the maximum tree depth of each decision tree, the learning rate determines how fast the algorithm learns, the number of estimators are the number of decision trees in the ensemble, the minimum samples per leaf and maximum number of features per leaf parameters control the complexity of each tree, the subsample size is the proportion of the training data used for training, and minimum child weight to limit tree depth[13]. Lastly, for XGBoost weighted learning was applied by scaling the positive class (readmissions) to account for class imbalance. This could not be performed for GB.

Table 10: Gradient Boosting Machines hyperparameters[1]

Hyperparameters	Range	Intervals	Final configuration
Max depth	3 – 9	1	0.6
Learning rate	0.01 - 0.1	Log sampled	0.1
Number of estimators	100 – 1000	50	750
Minimum samples per leaf	50 – 500	50	250
Max number of features	0.1 – 0.3	0.1	0.2
Subsample size	0.5 – 0.9	0.1	0.7

Table 11: XGBoost hyperparameters[14]

Hyperparameters	Range	Intervals	Final configuration
Number of estimators	20 – 300	10	230
Learning rate	0.01 – 0.1	Log sampled	0.01
Max depth	1 – 10	1	3
Min child weight	1-5	1	3
Scaling positive class	1 – 100	10	11

13. Performance of predicting combined outcome readmission/mortality

The primary outcome of Pacmed Critical is the combined prediction of readmission and death within 7 days after discharge. This composite outcome measure was chosen because both adverse events are competing risks and were expected to have influence on decision making for discharge. We primary trained our prediction models for the readmission outcome. A secondary analysis was performed for the combined outcome, using the same models as for the prediction of readmission. In Table 13, the performances of the final models for the prediction of the composite outcome are summarized. Overall, higher performance was obtained for the composite prediction, what can be explained by the higher number of events in the dataset for model training (822 readmissions vs. 1041 readmissions/deaths). Another reason could be that palliative care patients and patients with no-return/do-not-resuscitate orders could not be excluded from the data. Predicting the composite outcome takes into account any deaths in this patient group.

14. SHAP

SHAP (SHapley Additive exPlanations) values were used to visualize impactful predictors of 'black-box' machine learning and deep learning models. SHAP is inspired by cooperative game theory and computes the impact of each feature on the predicted outcome[15]. The Shapley value indicates how much the feature, in the context of interactions with other features, contributes to the prediction of that patient compared to the mean prediction of the population. For the prediction of readmission, the baseline prediction in our population is 6.7%. The predicted outcome for a patient is therefore the summation of the mean prediction (6.8%) with all the Shapley values combined. Each Shapley value in this case corresponds to a feature, 'pushing' the prediction lower or higher. It differs from classical explainable prediction models since it is not an isolated effect but a combined effect of the feature in combination with other features[15].

Table 12: ***Final model performance for the combined outcome (readmission/mortality) prediction. Results are given in mean (standard deviation (SD)) for cross-validation (CV) on the training dataset. Final model performance is given on the test dataset. Best results on the test dataset are marked in orange.*** *AUCPR = area under the receiver operating characteristic curve, MCC = Matthews correlation coefficient.*

	Neural Network		**Logistic Regression**		**Gradient Boosting**		**XGBoost**	
	CV (mean (SD))	Test	CV (mean (SD))	Test	CV (mean (SD))	Test	CV (mean (SD))	Test
Train time (s)	13.36 (3.41)	-	0.28 (0.01)	-	114.6 (7.27)	-	17.14 (3.53)	-
Score time (s)	0.11 (0.05)	0.11	0.04 (0.00)	0.01	0.22 (0.05)	0.37	0.14 (0.05)	0.1
Accuracy	0.66 (0.03)	0.69	0.65 (0.01)	0.65	0.91 (0.00)	0.91	0.68 (0.01)	0.67
Precision	0.15 (0.02)	0.18	0.16 (0.01)	0.17	0.30 (0.07)	0.38	0.17 (0.01)	0.18
Recall	0.66 (0.05)	0.75	0.75 (0.03)	0.79	0.06 (0.01)	0.06	0.71 (0.01)	0.78
Specificity	0.66 (0.04)	0.68	0.64 (0.01)	0.64	0.99 (0.00)	0.99	0.68 (0.01)	0.66
AUCPR	0.21 (0.02)	0.22	0.24 (0.02)	0.25	0.21 (0.02)	0.25	0.23 (0.02)	0.23
F1-score	0.25 (0.02)	0.29	0.27 (0.01)	0.28	0.10 (0.02)	0.11	0.28 (0.01)	0.29
AUC	0.73 (0.02)	0.77	0.76 (0.02)	0.78	0.73 (0.02)	0.77	0.76 (0.02)	0.77
Brier	0.34 (0.03)	0.31	0.35 (0.01)	0.35	0.09 (0.00)	0.09	0.32 (0.01)	0.33
MCC	0.19 (0.03)	0.25	0.22 (0.02)	0.25	0.10 (0.03)	0.13	0.23 (0.01)	0.25

SHAP Tree Explainer was used for XGBoost explanations and DeepSHAP was used for neural network explanations[16]. 500 patient samples from the training dataset were used as background data for the explainers. SHAP values as presented in the summary plots were then computed over the entire test dataset. For the summary plots, the top 20 features contributing most tot the predicted outcome were visualized. For patient specific visualization in so-called SHAP "force-plots", we manually extracted the top 10 features contributing most to the predicted outcome (5 most negative and 5 most positive). Because standardized values were used as input for prediction modeling, absolute feature values were not indicated in the force-plots. Therefore, we indicated with an arrow upwards or downwards whether te feature for that specific patient was higher or lower than average.

15. Expert opinion

Expert opinion was incorporated to evaluate the explanations of the different model for clinical applicability. Two ICU physicians were asked to indicate for each feature whether it was contradictive, irrelevant, or relevant for their decision to discharge a patient. Furthermore, three patient examples were incorporated showing patient specific SHAP plots for Neural Networks and XGBoost.

The following form was used for expert opinion collection:

**********Expert opinion**********

Explainability of machine learning algorithms for the prediction of ICU readmission
- Expert opinion -

1. **Background**

Prediction of intensive care unit (ICU) readmission could support physicians in determining optimal timing for ICU discharge. We found a significant higher 30-day mortality rate for patients readmitted to the ICU within 7 days after discharge (19.8% vs. 4.4%, $p < 0.001$). The hypothesis is that a decision support tool predicting a patient's risk of readmission could identify which patients are fit for discharge and which are not. First, we developed three different algorithms using a subset of patient variables available at time of model development (Table 1). Differences in discriminative model performance was small. Therefore, we focus on the explainability of the different algorithms to evaluate what model type is most in line with clinical reasoning. Expert opinion of two ICU physicians is collected by means of this form to assess what model's explanations are most clinically relevant.

Table 1: Variables included for prediction modeling.

Category	Variable included for model development	Category	Variable included for model development
Patient characteristics	Age	**Laboratory results**	Urea
	Gender		Thrombocytes
	Emergency admission		PT
	Hospitalization admission source		Potassium
	Treating specialty		MCV
	Length of stay (ICU)		Leukocytes
	Length of stay prior ICU		Lactate
Medication	Dobutamine		Creatinine
	Noradrenaline		Haemoglobin
	Milrinon/Enoximon		CRP
	Adrenaline		Bilirubin
Vital functions	Oxygen flow		BE
	ABP		Amylase
	Respiratory rate		Alkaline phosphate
			Albumin
			ASAT
			ALAT
			APTT

2. **Global predictor importance**

From these variables, a total of 416 measurements were used for model development. For the three different models, we are interested in your opinion on the top 20 most important variables. Because we used multiple types of measurements (e.g., standard deviation of CRP during first 24 hour of ICU admission), one variable can be represented by multiple measurements.

Question 1: Indicate in **Figure 1** for each measurement whether it is **contradicting, irrelevant, or relevant for a patient's readmission risk.** (E.g., a high last measured respiratory rate before discharge is often observed for patients with increased readmission risk = **relevant**). Check the box 'neutral' for no opinion.

For machine learning models, other visualization of important variables can be used to gain insight in their 'black-box' predictions. See Figure 2. Each dot in the plot represents a patient with high (red) or low (blue) value for a

specified measurement. Dots on the right side of the y-axis indicate a correlation with **high risk** of readmission and vice versa. The thickness of the line is determined by the number of patients.

Question 2: Please indicate clinical relevancy for each measurement in **Figure 2**.

Question 3: What type of visualization do you prefer? [Figure 1/Figure 2.]

Question 4: What model is overall most in line with clinical reasoning and therefore best applicable in clinical practice? [Model A/Model B/Model C]

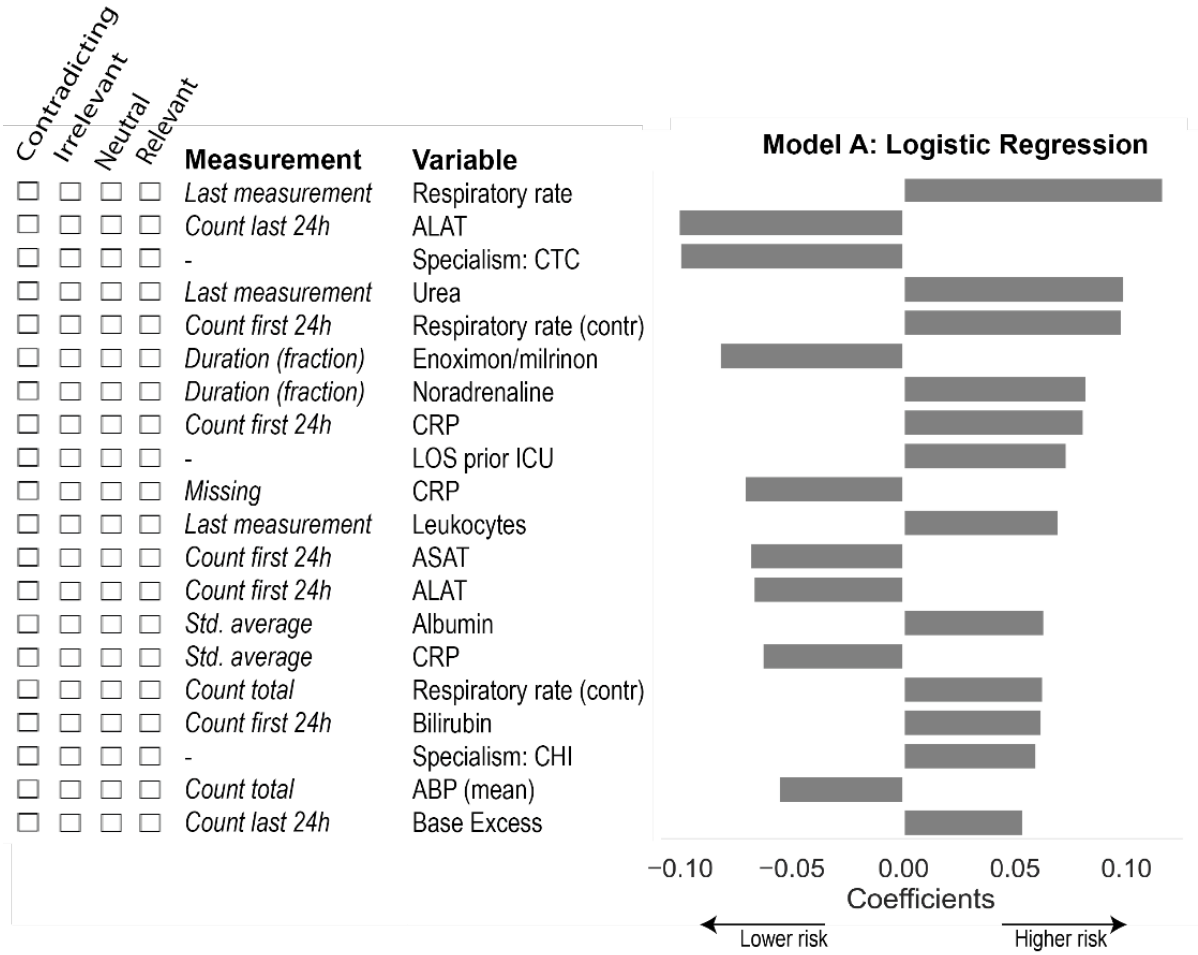


Figure C-1: ***Logistic Regression top 20 coefficients. Several type of measurements are used for each variable to capture time trends. A positive coefficient indicates correlation with high readmission risk, a negative coefficient indicates correlation with low readmission risk.*** *LOS = Length of stay. Std = standard deviation, count = number of measurements, missing = not measured.*

Figure C-2: ***XGBoost and Neural Network (two machine learning models) most important variables. The 20 most informative features are displayed. Each dot in the plot represents a patient with high (red) or low (blue) value for a specified variable. Dots on the right side of the y-axis indicate a correlation with high risk of readmission and vice versa. The thickness of the line is determined by the number of patients.*** *CTC = cardio-thoracic surgical, LOS = length of stay.*

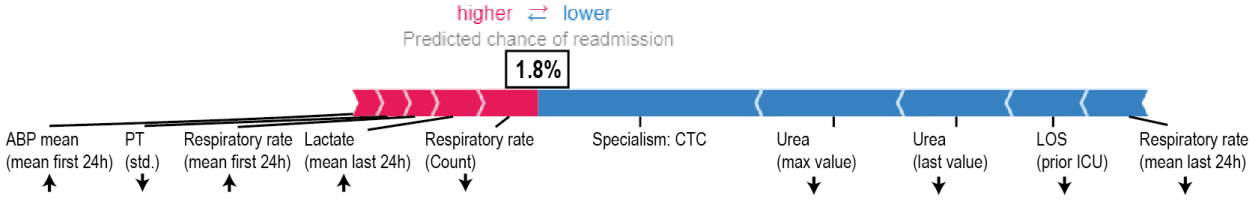
3. **Patient examples**

For model B and C, patient specific explanations can be given for each prediction of ICU readmission. In **Figure 3,** the predicted probabilities of model B and C are given in percentages for three patients. The variables in blue correspond to the variables 'pushing' the prediction lower. The variables in red correspond to the variables 'pushing' the prediction higher. A baseline readmission rate of 6.7% was observed, indicating predictions > 6.7% are high risk patients.

Question 5: Encircle for each patient whether model B or C is more clinically explainable and/or relevant.

Patient 1:

☐ **Model B: XGBoost**



☐ **Model C: Neural network**

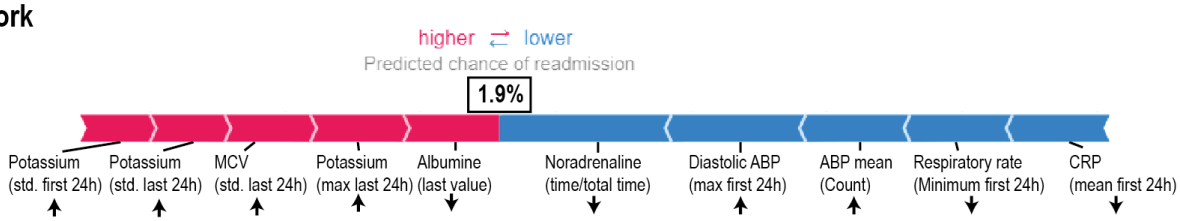
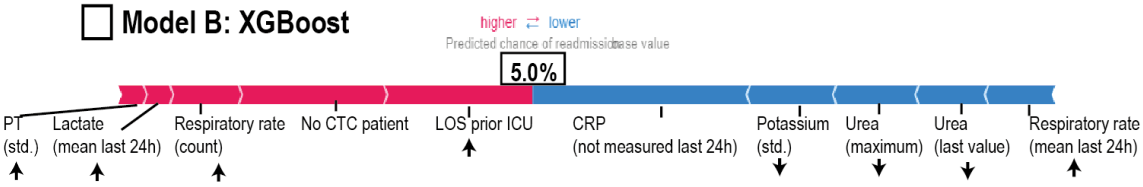


Figure C-3a: Patient specific explanations displaying the top 10 most important predictors for the readmission prediction. The variables contributing to a higher risk of readmission are visualized in red. The variables contributing to a lower risk of readmission are visualized in blue. Arrows pointing upwards and downwards indicate higher or lower than average.

Patient 2:

☐ **Model B: XGBoost**



☐ **Model C: Neural network**

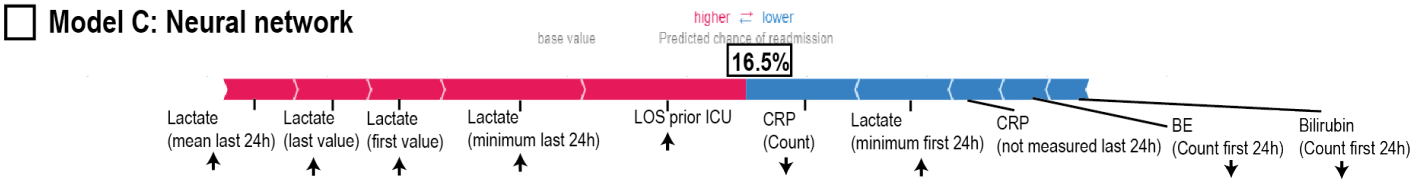


Figure C-3b: Patient specific explanations displaying the top 10 most important predictors for the readmission prediction. The variables contributing to a higher risk of readmission are visualized in red. The variables contributing to a lower risk of readmission are visualized in blue. Arrows pointing upwards and downwards indicate higher or lower than average.

Patient 3:

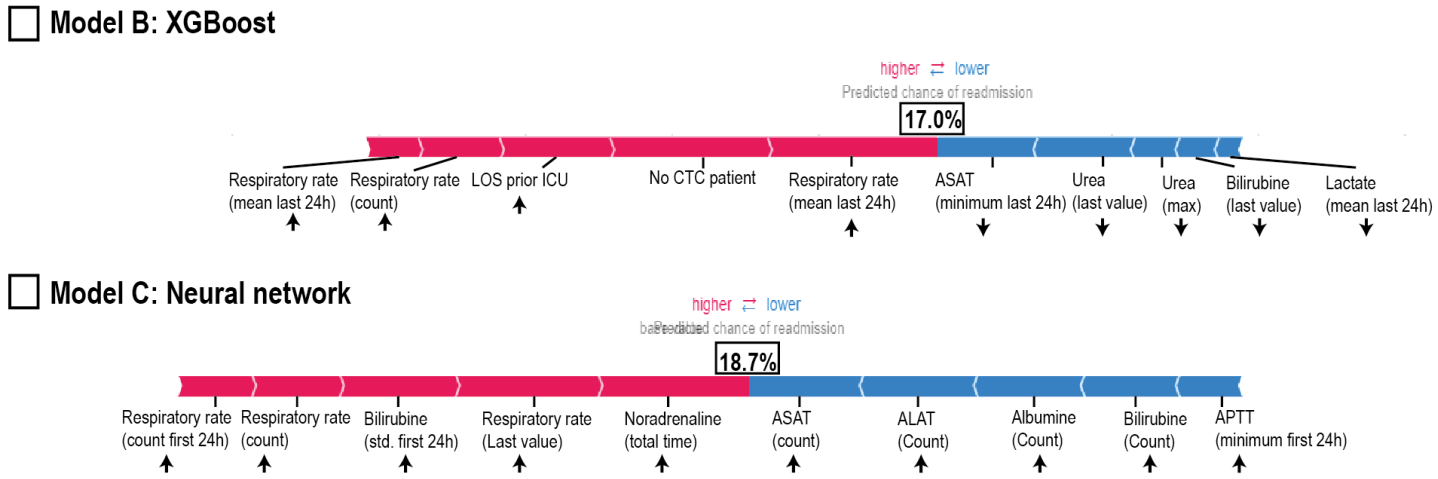


Figure C-3c: Patient specific explanations displaying the top 10 most important predictors for the readmission prediction. The variables contributing to a higher risk of readmission are visualized in red. The variables contributing to a lower risk of readmission are visualized in blue. Arrows pointing upwards and downwards indicate higher or lower than average.

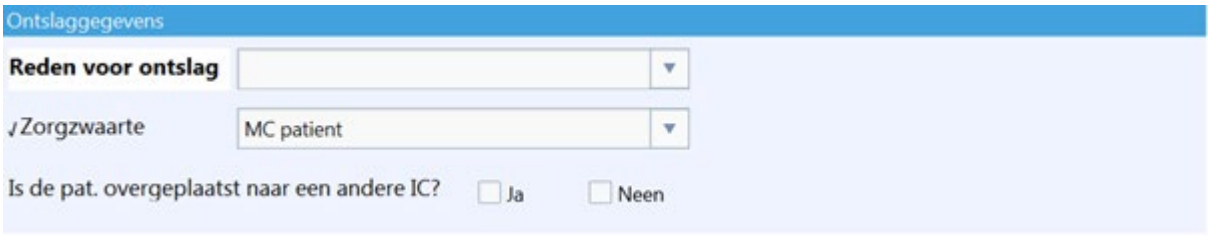
References

1. Thoral, Patrick J, Fornasa, Mattia, de Bruin, Daan, Hovenkamp, Hidde, Driessen, R, Girbes, A, Hoogendoorn, Elbers P. Developing a Machine Learning prediction model for bedside decision support by predicting readmission or death following discharge from the Intensive Care unit. *Unpublished*. 2020.
2. Saric L, Prkic I, Karanovic N. Inotropes and vasopressors. *Signa Vitae*. 2017;13:46-52. doi:10.22514/SV131.032017.6
3. Mack C, Su Z, Westreich D. Types of Missing Data. 2018.
4. Kubben P, Dumontier M, Dekker A. *Fundamentals of Clinical Data Science*.; 2019.
5. Snoek J, Larochelle H, Adams RP. *Practical Bayesian Optimization of Machine Learning Algorithms*.
6. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
7. Keras Tuner. https://keras-team.github.io/keras-tuner/. Accessed February 6, 2021.
8. (Tutorial) Regularization: Ridge, Lasso and Elastic Net - DataCamp. https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net. Accessed February 6, 2021.
9. Holmgren G, Andersson P, Jakobsson A, Frigyesi A. Artificial neural networks improve and simplify intensive care mortality prognostication: A national cohort study of 217,289 first-time intensive care unit admissions. *J Intensive Care*. 2019;7(1). doi:10.1186/s40560-019-0393-1
10. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Vol 15.; 2014. http://jmlr.org/papers/v15/srivastava14a.html. Accessed February 6, 2021.
11. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
13. Chapter 12 Gradient Boosting | Hands-On Machine Learning with R. https://bradleyboehmke.github.io/HOML/gbm.html#how-boosting-works. Accessed February 6, 2021.
14. Gao L, Ding Y. Disease prediction via Bayesian hyperparameter optimization and ensemble learning. *BMC Res Notes*. 2020;13(1):205. doi:10.1186/s13104-020-05050-0
15. Lundberg SM, Allen PG, Lee S-I. *A Unified Approach to Interpreting Model Predictions*.; 2017. https://github.com/slundberg/shap. Accessed April 29, 2020.
16. Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree Ensembles. February 2018. http://arxiv.org/abs/1802.03888. Accessed March 16, 2020.

D. Supplementary material Part IV – study protocol

Discharge survey questions

In the PDMS, under 'discharge data' (ontslag gegevens), the following questions will be asked to the fellow/intensivist at the moment of discharge:



1. Wat uw geschatte kans op heropname/mortaliteit van deze patiënt naar de IC of MC binnen 7 dagen? (Drop-down menu)

   - ❑ De patiënt heeft een no-return of abstinerend beleid en zal niet heropgenomen worden
   - ❑ De patiënt wordt overgeplaatst naar een ander ziekenhuis
   - ---
   - ❑ *Geef aan op een schaal van 1-10, waarbij 1 zeer onwaarschijnlijk en 10 zeer waarschijnlijk*[1].

2. Wat is het geschatte risico op heropname/mortaliteit van deze patient?
   - ❑ *Hoog risico*
   - ❑ *Gemiddeld risico*
   - ❑ *Laag risico*

3. *Welke vijf factoren dragen bij deze patiënt het meeste bij tot uw antwoord op vraag 1? (optioneel)*

   (Vink aan)

Categorie	Variabele[2]
Respiratoir	❑ Ademhalings frequentie ❑ PaO2 ❑ PaCO2 ❑ Beademingsduur ❑ FiO2 ❑ L/min ❑ Tijd sinds extubatie ❑ Hoestkracht
Circulatoir	❑ Hartslag ❑ Bloeddruk ❑ Vasopressie/inotropica ❑ Vochtbalans
Neurologisch	❑ EMV score ❑ Delier
Renaal	❑ Creatinine ❑ Diurese ❑ Nierfalen

Labwaarden	❑ pH ❑ hematocriet ❑ Natrium ❑ Kalium
Overig	❑ Algemene indruk ❑ Duur van verblijf op de IC ❑ Opname indicatie patiënt ❑ Voorgeschiedenis ❑ Comorbiditeiten ❑ Fast-track thorax patiënt ❑ Anders namelijk…..

Optioneel

4. Bij eventuele heropname, verwacht ik dat dit op basis van het volgende falen zal zijn[2]:
    - ❑ Circulatoire instabiliteit
    - ❑ Nieuwe of verslechtering bestaande infectie
    - ❑ Respiratoir falen
    - ❑ Bloeding
    - ❑ Encefalopathie
    - ❑ DKA/HHS
    - ❑ Zuur-base stoornis, elektrolyten stoornis, of andere lab afwijkingen
    - ❑ Verpleegkundige last
    - ❑ Anders namelijk…

Referenties

1. Rojas JC, Lyons PG, Jiang T, et al. Accuracy of Clinicians' Ability to Predict the Need for Intensive Care Unit Readmission. Ann Am Thorac Soc. 2020;17(7):847-853.
2. Heidegger CP, Treggiari MM, Romand JA; Swiss ICU Network. A nationwide survey of intensive care unit discharge practices. *Intensive Care Med*. 2005;31(12):1676-1682. doi:10.1007/s00134-005-2831-x