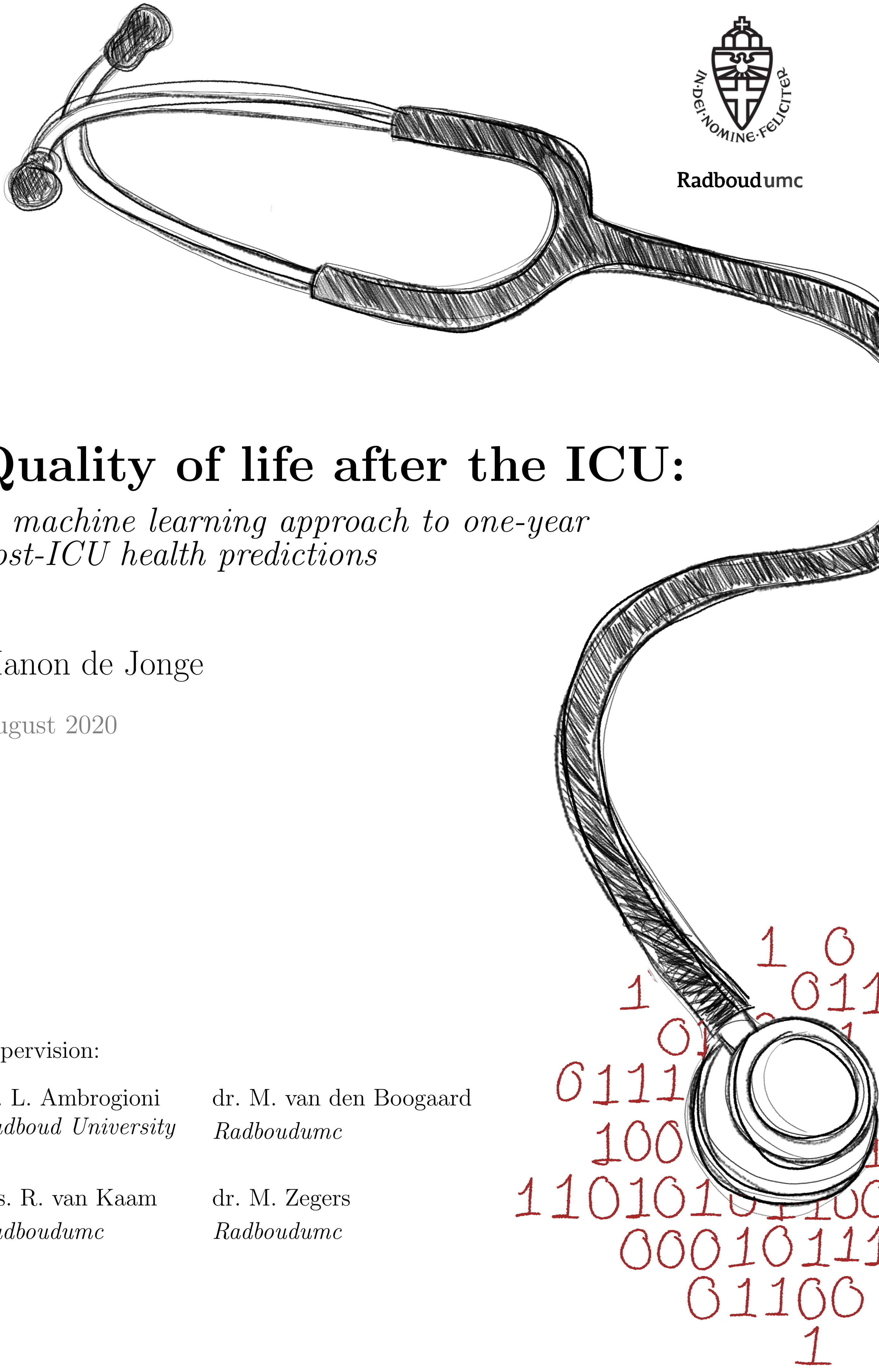
IN·DEI·NOMINE·FELICITER
Radboudumc



Quality of life after the ICU:

A machine learning approach to one-year post-ICU health predictions

Manon de Jonge

August 2020

Supervision:

dr. L. Ambrogioni
Radboud University

dr. M. van den Boogaard
Radboudumc

drs. R. van Kaam
Radboudumc

dr. M. Zegers
Radboudumc

Quality of life after the ICU

A machine learning approach to one-year post-ICU health predictions

ABSTRACT

PURPOSE - Due to the increasing number of Intensive Care Unit (ICU) survivors, the long term outcomes of the survival of a serious condition become progressively more important. With predictive modeling, the consequences of the ICU stay in terms of quality of life (QoL) can be estimated. This study aimed to improve the statistical model that is currently in use by using machine learning models and additional patient data from the Electronic Health Records (EHR).

METHODS - Data from the EHR was gathered from adult patients that were admitted to the ICU between June 2016 and January 2019. Several regression-based machine learning models, among which Random Forest and Support Vector Regression, were fitted on a combination of patient-reported data and 71 expert-selected EHR variables in order to predict the target value of change in QoL one year after admission. Results were compared to a baseline model.

RESULTS - Results from all tested regression models show that certain features that can be extracted from EHR data, such as high body temperatures and low BMIs, can have an influence on the regression-based prediction scores. However, the improvements in target prediction caused by these additions are limited. The best performing model had a decrease of 0.004 in mean absolute error (MAE) compared to the baseline results.

CONCLUSION - A machine learning-based approach to one-year post-ICU QoL prediction using EHR data was developed in this study. This approach had a small positive effect on prediction results when compared to the established statistical model, but the distinction is too small to put the model into practice without having to compromise the practicability by physicians.

Keywords *Quality of life, Critical care, Survivors, Machine learning, Prediction modeling*

INTRODUCTION

Annually, over 85,000 patients are admitted to Dutch Intensive Care Units (ICUs), frequently in life-threatening circumstances. Due to advances in critical care medicine, more patients survive their critical illness [1]. However, the survival of a serious illness does only rarely pass without consequences. It is estimated that 25% to 75% of ICU survivors experience physical problems (e.g. pain, shortness of breath, reduced muscle strength), psychological complaints (e.g. anxiety/depression), cognitive problems (e.g. memory-related) and/or problems related to daily functioning [2]. These issues often negatively influence the quality of life (QoL) and the financial and social situation of former ICU patients. The problems occurring after ICU discharge are described as part of a post-intensive care syndrome [3]. Where the emphasis of ICU healthcare professionals initially was laid on the prevention of a patient's death, the challenge now also lies in studying what the survival of a serious illness means for patients in the long term, and including these adverse consequences in the decision-making process about treatment in the ICU. Medical decisions about admission and treatment choices in the ICU (which include the most vulnerable and expensive patients in hospital care) are often made based on the experience and intuition of doctors. To a limited extent, choices are also made in consultation with the patient and their loved ones. To arrive at a more substantiated decision regarding ICU policy and treatment choices and to better inform the patients and their family about long-term consequences, the use of patient-reported outcomes of the ICU stay is of great relevance [4].

In July 2016, the MONITOR-IC study [5] (www.monitor-ic.nl) was set up with the aim of including patient-reported outcome measures in clinical decisions. The study gives insight into the long-term outcomes concerning the QoL of ICU survivors by monitoring them during a five-year follow-up period. The information gathered in MONITOR-IC can be used to identify patient factors and treatment factors that predict adverse long-term outcomes in terms of QoL. The study estimates the inclusion of 12,000 ICU patients from the Radboud University Medical Center (Radboudumc) and six other regional hospitals. As part of MONITOR-IC, a team at Radboudumc developed and validated a prediction model for the QoL of ICU survivors at one year after ICU admission. This model included patient-reported data of 1308 patients in the form of questionnaires and a set of medical variables from the NICE-database (National Intensive Care Evaluation) reported by physicians. All variables were available within the first 24 hours of a patient's stay. The explained variance (adjusted R^2) of the prediction model was 55.3% (SE = 2.6) after internal validation. This variance was achieved by executing a multivariable linear regression analysis with five predictors. This is considered a traditional statistical analysis method in medicine [6]. However, this model only incorporates data from before the ICU stay and during the first day of admission. The course of the stay of patients that were admitted for more than 24 hours was not taken into consideration. This means that, possibly, a lot of information that could be of influence on the outcome of such models are missing. This led to the belief that the use of different prediction methods in combination with additional data of the entire stay could improve the performance of the model.

With the use of machine learning models, patterns in patient data can be learned [7–9]. The objective of this study is to improve the performance of the QoL prediction model by using machine learning techniques in combination with extra physiological, pathological, drug, and treatment data from ICU patients' electronic health records (EHR) during the entire stay and to uncover the main predictive features in EHR data for changes in QoL. The study's main focus lies on the regression approach, as the original model was regression-based. However, to test classifier results, the addition of a class-based approach was made to this study.

RELATED WORK

Aside from inherently being related to the statistical predecessor of this study, there is more research that suggests the hypothesis of this study. Machine learning models have proven in many cases to be a substantial addition to modern-day healthcare. Past studies revolving around the prediction of QoL using machine learning models are typically centered towards a specific patient group, which is typically a non-ICU group. Nevertheless, these studies provide useful insights into methods and techniques that can be applied to this study's broad ICU-wide patient group. Some of this research also focuses on the inclusion of EHR data, which is a main part of this study.

EHR data for prediction purposes

One study that used EHR data for prediction studied the performance of different machine learning techniques to model heart failure [10] . To create variables from the data, some of the input data was operationalized to form an indication and duration of a certain medical condition. Secondary variables include proportions of data exceeding a certain threshold. These variables were then used to compare Support Vector Machine (SVM), Boosting, and Logistic Regression models. The study shows that the EHR data in combination with machine learning models performs reasonably well. This provides grounds for further research and uses on different data sets.

In another research [11] preoperative EHR data was used to predict postoperative delirium. Several machine learning models were compared. Their best performing Random Forest model indicated that a patient's age, alcohol and drug intake, socioeconomic status, medical issues and their severity, and their surgeon can affect the risk of delirium. These factors were all gathered from the EHR. The resulting model could be applied at the point of access to preoperative care.

These previous studies show that incorporating EHR data in the prediction process can lead to reliable results for specific patient groups. The information that machine learning extracts out of influential EHR variables can help to determine the important factors in the medical history and the hospital stay of a patient. Once these factors are known, more emphasis can be placed on handling the consequences that can be related to them.

Machine learning for prediction of QoL

Oeyen et al. [12] have developed a statistical prediction model for QoL for ICU survivors. Their model could explain 40% of the variability in QoL one year after admission by using sixteen variables extracted on the first day of ICU admission. They found that the baseline QoL (i.e. the QoL before admission) is the main predictor for long-term QoL. This approach did not make use of machine learning techniques to obtain their predictions. However, multiple medically relevant studies have concluded that incorporating machine learning into the prediction process can benefit results.

Researchers conducting a study about lung cancer have used health-related QoL in a five-year lung cancer

Variable category	Variables	EHR/NICE representation	Model representation
Demographics	Age, sex*, BMI	Date of birth, male/female, height in cm, weight in kg	Age in years at the time of admission, sex, BMI = weight / (height)^2
Clinical measurements	Body temperature, PEEP, MIP, FiO2, ICP	Order details (e.g. date) and measured values	Minimum and maximum during stay, time above or below threshold value, standard deviation
Laboratory	Hemoglobin, sodium, lactate, glucose	Order details (e.g. date) and measured values	Minimum and maximum during stay, time above or below threshold value, standard deviation
Medication	Noradrenaline, propofol, midazolam	Order details (e.g. date) and measured values	Cumulative doses compensated for LOS and number of days on medication
Stay	LOS, admission timing, admission source*	Admission date, discharge date, admission source	LOS in hours, admission within or outside of office hours, admission source
Monitor	Blood pressure, heart rate	Monitor measurement details per patient per variable	Minimum and maximum during stay, standard deviation, largest change
Other	Tracheostoma, RRT, CVA*, vulnerability*, comorbidity*	True/false variables, quantity	Boolean present/absent, number of times present

* Also part of the five-feature statistical model.

Table 1: An overview of the expert-selected variables for QoL prediction.

survival prediction model [13]. They tested several machine learning models, among which Logistic Regression and Random Forest. Their models were trained on two feature sets, the first consisting out of clinical and demographic variables and the second adds QoL factors to the first set. The findings show that models using the second feature set outperform models using the first. This indicates that QoL can improve prediction methods for lung cancer survival.

A degenerative cervical myelopathy study [14] oppositely uses health-related QoL. Instead of using it as a feature, class-based QoL is the outcome of their model. Their best-functioning model showed good prediction scores in their patient group after training on demographic and clinical variables. Their findings include, among other things, significantly better improvement in QoL for men that had low baseline mental health scores in comparison to women.

While no evidence could be found of studies including QoL prediction using EHR data and regression-based machine learning models, the results of the studies above suggest positive influences from both the addition of machine learning, as well as the addition of EHR data.

METHODS

The MONITOR-IC study has been approved by the research ethics committee of the Radboudumc, CMO region Arnhem-Nijmegen (number 2016–2724). All patients, or their legal representative, provided written informed consent for the use of data from their EHR. All patient data is pseudo-anonymized. This study uses EHR data and data gathered for MONITOR-IC.

Data sources

The data used in this study originate from two sources: (1) data gathered in the MONITOR-IC study, which includes questionnaire data, including patient-reported outcomes (e.g. QoL, frailty) and data from the patients' medical records (e.g. admission type, admission diagnosis, length of stay - for more details see [5]), and (2) certain variables in EHR data, such as measurements from the arterial blood gas and the administration of certain medication. Extracting the entire EHR database is expensive in terms of time and also very heavy regarding memory capacity. The EHR includes measurements that might not be of influence to QoL and can be left out. The quality of a model might improve when domain-knowledge of experts guides the process of learning [15]. This expert-augmented machine learning approach narrows down the amount of training data, which reduces time- and memory-related issues while keeping allegedly influential factors in the data based on domain-knowledge. To narrow down the number of EHR variables, ICU physicians at Radboudumc were consulted to select the variables that they know or suspect to have any influence on long-term outcomes of QoL of ICU survivors. In Table 1, the selected variables are shown. EHR data is, in its original state, not suitable for modeling. After the selection process, the variables needed to be preprocessed and structured to fit the regression models.

Preprocessing and feature selection

The process of preprocessing the data and selecting the most influential features, as well as the modeling, were carried out using the programming language Python and several of its modules that are suited for this study.

Outliers in the data were removed or corrected and missing values are replaced by group-wide averages (e.g. in temperature) or zeros (e.g. in medication dose) depending on the type of variable. ICU-related EHR data contains many parameters that change frequently over time (e.g. blood pressure). These parameter values are preprocessed to represent the changes over time rather than the actual values. Examples of resulting features are minimum and maximum (over a given time span), the standard deviation, the percentage of measurements below/above a certain threshold, and/or the largest increase/decrease (over a given time span). These derived features were extracted based on the input of experts regarding which derivatives would be informative for patient outcomes. An overview of the selection process is presented in Supplement A.

The values for QoL are gathered from MONITOR-IC questionnaire data and are calculated according to the Dutch version of EuroQol 5D (EQ-5D-5L) [16], which is a standardized measuring instrument in medical care based on five dimensions to score an individual's QoL. The EQ-5D-5L scores can range from -0.45 to 1, with higher scores indicating a better QoL.

To avoid overfitting and to increase the clarity and interpretability of the model, a number of features in the data are left out of the prediction model. By using Recursive Feature Elimination (RFE), the most influential features are selected by recursively training on an ever-shrinking set of features. The least important features are pruned. The feature set resulting in the highest adjusted R^2 is used for modeling. RFE is not used in models that use some built-in form of regularization.

Modeling

The problem that was to be solved is regression-based, as the values for changes in QoL are continuous. Within the scope of machine learning, there are several regression techniques that could be used for the prediction of changes. Linear Regression, Neural Networks, and Random Forest are examples of types of algorithms that could be used to solve the problem, each with its own assets and liabilities. Because of the availability of these techniques in Python modules, all of the candidate models could be implemented relatively quickly and a comparison could be made based on initial performance. In turn, the best performing models were tweaked to perform optimally. These models were then validated and evaluated. The models that were compared in this study are Ordinary Least Squares (OLS), Random Forest (RF) Regressor, Multilayer Perceptron (MLP) Regressor, Lasso(CV), Least Angle Regression (LARS), Elastic net regularization (ElasticNet), Huber Regressor, Ridge Regression, Automatic Relevance Determination (ARD) Regression, and Support Vector Regression (SVR) (see Supplement B). These were implemented using the scikit-learn library [17] for the programming language Python. Most of these models are linear models like the original model. The non-linear models were used to explore whether they could fit the data more accurately than linear models.

Validation and evaluation

To consistently compare this study's machine learning models to the previous model's statistical approach, the methods used in both studies were matched. The earlier model used a bootstrapping method to sample from the dataset, and used the full sample to construct the model. This approach is mimicked in this study for validation purposes. However, as this study expands on the initial methods by finding machine learning techniques that are able to more accurately catch the expected changes in QoL, a validation method that is considered as general practice in machine learning, cross-validation, was applied to judge the predictive performance of the models.

To be able to judge the machine learning models' performances and to compare the performance of the models of this study to the previous model, a method that was used to evaluate the model is the explained variance metric. Another metric commonly used to evaluate regression models and was also gathered in the previous study, the Mean Squared Error (MSE), is also taken into consideration to provide a more complete overview of the results. The Mean Absolute Error (MAE) was added as a metric in this study to give an indication of predictive performance.

RESULTS

The results of this study were based on the 1308 completed patient records of patients admitted to the ICU between July 2016 and January 2019. The patient group mainly consisted of male patients (67.9%). The median age of the patients was 65 (interquartile range (IQR) = 57-71). Most patients were admitted through planned surgery (72.7%). The median QoL of the patients before admission was 0.8 (IQR = 0.7-0.9). Further characteristics of this patient group are presented in Table 2.

	Type	Median (25% - 75%)	N (%)
Sex (male)	Binary		888 (67.9)
Age	Integer	65 (57 - 71)	
EQ-5D-5L (baseline)	Continuous	0.8 (0.7 - 0.9)	
Frailty	Integer	3 (2 - 3)	
Education level	Categorical		
- High			376 (28.7)
- Medium			574 (43.9)
- Low			358 (27.1)
Admission type	Categorical		
- Planned			951 (72.7)
- Emergency			140 (10.7)
- Medical			217 (16.6)
BMI	Continous	26.0 (23.5 - 29.0)	
Chronic conditions*	Integer		167 (12.8)
Total			1308 (100)

Table 2: Patient demographics at the time of admission. *Included chronic conditions are immunological insufficiency, malignant hematological disease, metastasized neoplasm, chronic cardiovascular insufficiency, chronic respiratory insufficiency, and chronic renal insufficiency.

Following the methods used for validating the original study's five-predictor model, the linear regression model in this study was fitted on two thousand bootstrap samples using the same five predictors (baseline QoL, sex, admission type, CVA prevalence, and frailty). Adjusted R^2, MSE, and MAE were calculated over predictions from the entire data set. This resulted in a score of 0.54 for R^2, an MSE score of 0.031, and an MAE score of 0.128. These scores differ slightly from the statistical model's validation scores, which had an R^2 of 0.55 and an MSE score of 0.030. These differences may be caused by the differences in algebra behind the two methods (Python's Linear Regression model and R's linear model). The coefficients of the five predictors were similar in both models, with the baseline QoL score as the strongest predictor. Out of all predicted values, 48% differed less than 0.1 from their actual score and 82% differed less than 0.2.

Next, the predictive performances of different regression models were tested. In Table 3 the results of the ten compared models are shown. The models were first compared on their performance on the five-feature baseline data with four-fold cross-validation. Scores were calculated over predictions on unseen data. The results show that most models perform similarly with regards to adjusted R^2, with the exception of Random Forest and MLP Regressor. While it should be noted that R^2 is a questionable metric for non-linear models like these two [18], the MAEs for both models also indicate a (slight) decrease in performance when compared to the baseline. Based on MAE score, the best performing model for the five-feature data is the Huber Regressor with a MAE of 0.126.

The models were then extended with 71 EHR data features in addition to variables used in the larger models from the original study. An overview of the features that remained after feature selection by the different models is presented in Table 4. The number of features selected per model differs, as well as the features selected. Three features were selected in all of the models explored in this study: the baseline EQ-5D-5L score, the presence/absence of a cerebrovascular accident (CVA), and the highest body temperatures measured during a patient's stay. The first two are predictors from the original study, while the temperature measurement was gathered from EHR data. Other features that are considered in most of the models are sex, frailty, low BMI scores, and delirium prevalence.

The number of features selected per model after the addition of EHR data ranges from 7 to 15. In every model, the addition of EHR variables has led to a performance improvement that can be detected through the adjusted R^2, MSE, and/or MAE scores. However, these improvements are very minimal. Improved adjusted R^2 scores differ in a range of 0% to 5% from the five-feature baseline of the same model. MSE improvements reach a maximum difference of 0.003. Every model has a lower MAE using the EHR features when compared to the baseline models, with the lowest MAE being 0.125 in the Huber Regressor and SVR models. This is a difference in MAE of 0.004 compared to the linear regression model baseline. To put this score into perspective, the changes in QoL scores can theoretically range from -1.45 to 1.45 due to the way in which the Dutch EQ-5D-5L scoring method is defined. In this study, the minimum change in QoL was -1.2 and the maximum change was 1.3. The IQR of the change in QoL was -0.07 to 0.15. For the Huber Regressor, 53% of the predictions differ less than 0.1 from their actual value and 84% differ less than 0.2. These are increases of 5% and 2% respectively in comparison to the baseline linear model. In Figure 1, the predictions of the Huber Regressor with seven features are plotted against the true values for change in QoL. From this plot, it can be seen that the model does not accurately predict the lower negative and the highest positive QoL changes, while it performs passably for the other values. This pattern can be seen for all the models compared in this

	Five-feature baseline			Baseline + EHR data			
Model	adj. R^2	MSE	MAE	Nr. of features	adj. R^2	MSE	MAE
OLS	0.52	0.032	0.129	9	0.54	0.031	0.127
RF Regressor	0.51	0.033	0.130	7	0.51	0.032	0.129
MLP Regressor	0.46	0.035	0.136	11	0.51	0.032	0.127
LassoCV	0.52	0.032	0.129	8	0.53	0.031	0.128
LARS	0.52	0.032	0.129	13	0.54	0.030	0.127
ElasticNet	0.52	0.032	0.129	13	0.54	0.030	0.127
Huber Regressor	0.52	0.032	0.126	7	0.52	0.032	0.125
Ridge Regression	0.52	0.032	0.129	13	0.54	0.030	0.127
ARD Regression	0.52	0.032	0.129	15	0.53	0.031	0.127
SVR	0.52	0.032	0.127	13	0.53	0.031	0.125

Table 3: Prediction scores per model after cross-validation. Features were selected using RFE or the model's built-in regularization method.

Variable	Included in models*
Baseline EQ-5D-5L score	100%
Cerebrovascular Accident (CVA)	100%
High temperature	100%
Sex	90%
Frailty	80%
Low BMI	80%
Delirium prevalence	80%
Admission source	70%
Intracranial mass	70%
Respiratory insufficiency	70%
Low hemoglobin	60%
Malignant hematological disease	60%
Tracheostoma prevalence	40%
Nr. of registered mean inspiratory pressure (MIP) measurements	40%
Cardio Pulmonary Resuscitation (CPR)	40%
Dysrhythmia	40%
Low sodium	30%
Mechanically ventilated in the first 24 hours	10%
Length of stay (LOS)	10%

Table 4: The variables selected by RFE or a built-in selection method based on feature importance.
*Selected by RFE or built-in regularization, models that have no built-in feature importance metric were tested by hand. Inclusion in all of the ten compared models equals 100%.

study.

As the problem of QoL prediction can also be viewed from a classification point of view, a classification-based approach of this project and its results can be found in the supplementary materials (see Supplement C).

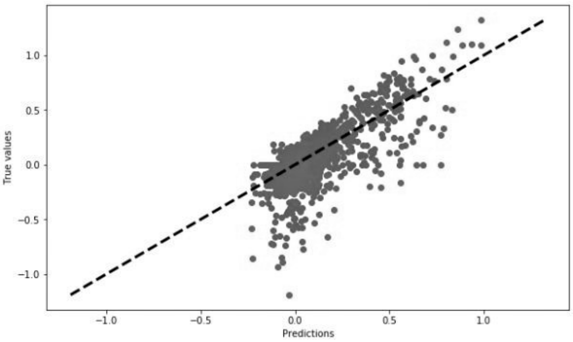


Fig. 1: The predictions of the Huber Regressor model with 7 features plotted against the true values for change in QoL.

DISCUSSION

This study has adapted an already existing model for the prediction of changes within a patient's QoL before admission vs. one year after admission. By using machine learning techniques and additional data from the EHR over the entire ICU stay, the predictive regression results could be improved slightly. While a larger effect was expected beforehand, these minimal results answer some questions that researchers and physicians had about the effects of AI methods on QoL prediction and the addition of entire stay patient data.

The original model was intentionally shrunken to use only five predictors to keep the model usable in practice without being too much administrative work for physicians but to still be able to provide sufficient predictive quality. This study expands the number of features used by the model, which requires more input effort from experts that use the model. A trade-off would have to be made to decide whether or not the increase in predictive quality is worth the extra patient data extraction. The results of this study suggest that the profit gathered from the machine learning approach with additional data is too small to put into practice. A patient's QoL prior to ICU admission stayed the strongest predictor by far. This leads to the belief that the original model is, at the time of writing, the most suitable for the problem. However, there are some limitations to this study that might be further explored in future research.

One of the obstacles encountered during this study is the number of included patient stays. Machine learning models generally need large amounts of training data to reach a reliable prediction result. Most classical machine learning data sets contain thousands up to millions of data points. In contrast, the 1308 patient stays included in this study seem limited. Initially, there was a possibility to increase the number of included patients. However, due to the situation around the COVID-19 pandemic and the pressure on the ICU, additional data extraction was impossible during this study. Related to this issue is the wide versatility within the data set of patients. Previous research that included good results for QoL prediction and/or prediction from EHR data was conducted on specific patient groups, mostly with the same or similar conditions. This, in combination with the results of this study, gives an indication that QoL is hard to predict for a patient group with highly versatile pre-ICU statuses and ICU courses. Another related issue is the fact that most of the included patients had planned surgery. These admissions usually indicate a shorter length of stay than patients with other admission sources. Their stay is often less than 24 hours long, which means that the chances of the occurrence of events during their stay that will be influential after one year are lower than for patients with non-planned admissions. A suggestion for future research would be to divide the patient data on the basis of admission source and to perform separate analyses for each source.

While the expert-based selection of variables has shown to guide the improvement of a model's predictive quality in some studies [15], it could be that the features selected in this study do not reflect the changes in QoL best. The EHR contains many more variables that could be extracted that were not extracted or even considered in this study. Also, the derived features from the extracted variables might not properly represent the consequences for the changes in QoL. Many EHR variables are best expressed as time-domain data, while all features used in this study were made to be frequency-based to keep the input features as simple as possible. To investigate the inclusion of other EHR variables, future studies could use unsupervised learning for feature selection. This also could solve the problem of ad-hoc selected variables not generalizing well [19]. Furthermore, non-EHR variables like alcohol consumption and tobacco use were not considered, even though it was expected that these could be of importance to QoL outcomes. The reason for this was the possible unreliability and inaccuracy of these variables.

CONCLUSION

Due to the increased survival rate of ICU patients, the long term outcomes of the survival of a serious condition become progressively more important. With predictive modeling, the consequences of the ICU stay in terms of QoL can be estimated. Statistical models using mostly patient-reported data have proven to provide a substantial prediction method for changes in a patient's QoL within one year after ICU admission. Previous research has shown that the use of machine learning models and/or the inclusion of EHR data can improve prediction results of healthcare-related issues. The results of this study show a slight increase in predictive performance by machine learning models using additional EHR data on top of patient-reported data. However, the increased effort by physicians needed for putting these findings into practice outweighs the slight improvement in prediction quality. It can, therefore, be concluded that the examined EHR variables in combination with the regression models that were tested do not provide sufficient added value to the already existing statistical model.

The results of this study did, however, unveil some EHR variables that influence a patient's QoL one year after admission. High temperatures, low BMI, and delirium prevalence are some of the variables of which their importance was ranked highly by the models. More research is needed to evaluate the effects of these and other influential EHR variables on patients' QoL changes after ICU admission in detail.

References

1. M. B. F. Makic, "Recovery After ICU Discharge: Post–Intensive Care Syndrome," *Critical Care Connection*, vol. 31, no. 2, pp. 172–174, 2016.

2. H. Svenningsen *et al.*, "Symptoms of Posttraumatic Stress after Intensive Care Delirium," *Research Advances in Critical Care: Targeting Patients' Physiological and Psychological Outcomes*, pp. 1–9, 2015.
3. M. Brackel-Welten, "Overlevers kritieke ziekte lopen vaak cognitieve schade op; Post-ic-syndroom wordt niet herkend," *Medisch Contact*, 2014.
4. D. M. Needham *et al.*, "Improving long-term outcomes after discharge from intensive care unit: Report from a stakeholders' conference," *Critical Care Medicine*, vol. 40, no. 2, pp. 502–509, 2012.
5. W. Geense *et al.*, "MONITOR-IC study, a mixed methods prospective multicentre controlled cohort study assessing 5-year outcomes of ICU survivors and related healthcare costs: a study protocol," *BMJ Open*, 2017.
6. K. A. Marill, "Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression," *Academic Emergency Medicine*, vol. 11, no. 1, pp. 94–102, 2008.
7. M. Wijnberge *et al.*, "Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery," *JAMA*, vol. 323, no. 11, pp. 1052–1060, 2020.
8. M. Roimi *et al.*, "Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms," *Intensive Care Medicine*, vol. 46, p. 454–462, 2019.
9. C. J. Chiew *et al.*, "Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department," *Medicine*, vol. 98, no. 6, 2018.
10. J. Wu *et al.*, "Prediction Modeling Using EHR Data: Challenges, Strategies, and a Comparison of Machine Learning Approaches," *Medical Care*, vol. 48, no. 6, pp. S106–S113, 2010.
11. A. Davoudi *et al.*, "Delirium Prediction using Machine Learning Models on Predictive Electronic Health Records Data," *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 568–573, 2017.
12. S. Oeyen *et al.*, "Development of a prediction model for long-term quality of life in critically ill patients," *Journal of Critical Care*, vol. 43, pp. 133 – 138, 2018.
13. J. Sim *et al.*, "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning," *Sci Rep*, vol. 10, pp. 133 – 138, 2020.
14. O. Khan *et al.*, "Machine learning algorithms for prediction of health-related quality-of-life after surgery for mild degenerative cervical myelopathy," *The Spine Journal*, 2020.
15. E. D. Gennatas *et al.*, "Expert-augmented machine learning," *Proceedings of the National Academy of Sciences*, vol. 117, no. 9, pp. 4571–4577, 2020.
16. EuroQol, "EQ-5D," 2020.
17. F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
18. A. Spiess *et al.*, "An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach," *BMC Pharmacol*, vol. 10, no. 6, 2010.
19. R. Miotto *et al.*, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Scientific Reports*, vol. 6, no. 26094, 2016.
20. O. Bretscher, "Linear Algebra With Applications (3rd ed.)," 1995.
21. A. Liaw and M. Wiener, "Classification and regression by randomforest," *Forest*, vol. 23, 11 2001.
22. I. Yilmaz and O. Kaynar, "Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5958 – 5966, 2011.
23. R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," pp. 18–20, 2016.
24. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
25. B. Efron *et al.*, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
26. P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
27. M. Gruber, *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. 1998.
28. D. J. C. MacKay, *Bayesian nonlinear modeling for the prediction competition*. 1994.
29. H. Drucker *et al.*, "Support Vector Regression Machines," *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1997.

Supplementary materials

A Table A1

Expert-based variable selection*		EHR data extraction	Included in final selection	Variable derivatives
Lab	Hemoglobine	Blood gas, POCT	Yes (blood gas)	Minimum, values below {3,4,5,6}, std dev
	Glucose	Blood gas, liquor, urine, POCT	Yes (blood gas)	Minimum, maximum, values below 4, values above 8, std dev
	pH	Blood gas, urine, POCT	No	
	Sodium	Blood gas, urine, POCT	Yes (blood gas)	Minimum, maximum, values below 130, std dev
	Creatinine	Urine	No	
	Lactate	Blood gas	Yes	Maximum, values above 2.1, std dev
Medication	Dobutamine	IV drip (several doses)	No	
	Midazolam	IV drip (several doses), bolus, nose spray, tablet	Yes (IV drip, bolus, tablet)	Cumulative (relative to LOS), nr. of days medicated
	Milrinon	IV drip (several doses)	No	
	Noradrenaline	IV drip (several doses), ampoule	Yes	Cumulative (relative to LOS), nr. of days medicated
	Propofol	IV drip (10 mg/ml), injection	Yes	Cumulative (relative to LOS), nr. of days medicated
	Rocuronium	IV drip (several doses), injection	No	
	Vasopressine	IV drip (1 unit/ml in NaCl)	No	
Measurements	Temperature	°C, measurement location	Yes (°C)	Minimum, maximum, values below 35.5, values above 38.3, std dev
	FiO2	Oxygen percentage	Yes	Minimum, maximum, std dev, average per day
	ICP	mmHG	Yes	Boolean (measured/not measured)
	ECLS	L/min	No	
	PEEP	Pressure (cm/H20)	Yes	Minimum, maximum, std dev, average per day
	GCS	Eye opening, motor response, verbal response	No	
Raw (monitor data)	Heart rate	Beats per minute (bpm)	Yes	Minimum, maximum, std dev, average per day
	Blood pressure	Systolic, diastolic, mean	Yes (mean)	Minimum, maximum, std dev, average per day
	Respiratory rate	Respirations per minute (rpm)	No	
Other	Tracheostoma	Tracheacanule prevalence	Yes	Prevalence, nr. of times
	RRT	Renal replacement therapy prevalence	Yes	Prevalence, nr. of times
	Fluid balance	Day balance	No	
	Delirium	Subtype, prevalence, comatose days	Yes (prevalence, comatose days)	Prevalence, nr. of days
	LOS	Admission date and time, discharge date and time	Yes	Nr. of hours
	Timing of admission	Admission date and time	Yes	Within office hours (yes/no)

Table A1: The variable selection process.

*Several expert-selected variables are not included in this overview as the data for these variables could not be extracted directly from the EHR database. These variables include: VAP, CRI, bacteraemia, CSZ, percutane drainage abcess, tractus digestivus bleeding, new ICU bleeding, ICU acquired weakness, decompressive craniectomy, cardiac arrest, TPV, enoximon. As experts have suggested the use of these variables, they could be included in future research to test their effect on QoL prediction.

B Regression algorithms explained

The algorithms used in this regression-based study were selected from the scikit-learn [17] library for the programming language Python on an experimental basis. This supplementary document serves as an overview and brief explanation of the used algorithms.

Ordinary Least Squares (OLS)

OLS is the method that most closely resembles the approach taken in the original study, which is why its results were used as a baseline for this study. It is a commonly used statistical method for analysis that aims to minimize the residual sum of squares between the actual data and the predictions made by the model. The OLS model used in this study, which involves multiple variables, can be represented as Equation 1 [20].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \tag{1}$$

In this equation, x represents the predictor input variable and y is the observed value, in the case of this study the y-value represents the change in QoL. β is the value of the coefficient for each predictor, which indicates the slope of the linear relationship between the predictor and the observations. β_0 is the intercept. A value of ε is assumed to be the random error. The components of this equation are used to find a linear fit for all of the data points for the n predictors.

Random Forest (RF) Regressor

The RF algorithm is an ensemble formed by multiple decision trees. Each decision tree is built using a bootstrap sample drawn from the data set with replacement. In regression, predictions are made by averaging over the predictions of all trees in the ensemble. The advantages of using multiple decision trees over a single tree are typically a lower variance and lower susceptibility to overfitting [21].

Multilayer Perceptron (MLP) Regressor

MLPs are a type of artificial neural network consisting out of an input layer, one or more hidden layers, and an output layer. They can be used for classification as well as regression. For the regression-based approach, the Softmax activation of the final layer in a classification problem is replaced by a linear activation function in order to get a real-valued output. The perceptron learns to predict by the changes in connections weights that occur after the processing of new data. The output of the MLP is essentially based on the activation function used, which can generally be presented as Equation 2.

$$x_0 = f(\sum_h x_h w_{ho}) \tag{2}$$

Here, f represents the activation function, x_h is the activation of the h^{th} hidden layer node, and w_{ho} represents the connection, or weight, for the h^{th} hidden layer node and the o^{th} output node [22]. Due to the use of multiple layers and non-linear activation functions, MLPs can distinguish non-linearly separable data.

Least Absolute Shrinkage and Selection Operator (Lasso)

Lasso is a method for shrinking and variable selection in linear models. Lasso ensures that features that are irrelevant to the model are pruned, which reduces variance and also makes the model easier to interpret [23]. The algorithm achieves this goal by minimizing the prediction error for each feature. This means that the goal of Lasso can be described as solving Equation 3.

$$\underset{\beta_0,\beta}{min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to} \sum_{j=1}^{p} |\beta_j| \leq t \tag{3}$$

Here, x_i are the features, y_i indicates the target output, and β values are coefficients. Parameter t is used for tuning, it determines the amount of regularization. To find the best fitting model K-fold cross-validation, LassoCV, was performed to obtain a more accurate and reliable score.

Elastic net regularization (ElasticNet)

ElasticNet was introduced to address the limitations of Lasso. Lasso is prone to selecting only one variable from a correlated group of variables while ignoring the others. ElasticNet adds a quadratic parameter to Lasso's penalty to overcome its limitations [24].

Least Angle Regression (LARS)

Similar to Lasso and ElasticNet is LARS. This regression algorithm also performs feature selection, and in addition it works well with high-dimensional data. In its stepwise regression, it finds the most correlated feature with respect to the target [25].

Huber Regressor

The Huber Regressor is based on the Huber loss function (Equation 4) with parameter a being the difference between the observed and the predicted value [26].

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise} \end{cases} \tag{4}$$

This function ensures that the regressor is less sensitive to outliers than most other loss functions, as the function is quadratic for smaller values of a, but linear for larger values.

Ridge Regression

Ridge Regression is a linear least squares method that is similar to OLS, but introduces a bias. This bias should ensure better results on the long term in models with a large number of parameters which, in turn, should result in a lower variance [27].

Automatic Relevance Determination (ARD) Regression

ARD Regression is a type of Bayesian Ridge Regression in which coefficient weights are shifted towards zero if features are considered to be irrelevant [28]. This introduces the concept of feature relevance in the model.

Support Vector Regression (SVR)

Some years after the development of Support Vector Machines for classification, SVR was introduced [29]. The objective to be solved by SVR is presented in Equation 5.

$$\text{minimize } \frac{1}{2}\|w\|^2 \text{ subject to } |y_i - \langle w, x_i \rangle - b| \leq \varepsilon \tag{5}$$

Here, x and y are the input sample and target data. $\langle w, x_i \rangle + b$ indicates the prediction of the model. All predictions have to be within a range of ε, the tolerance range, of the actual target value.

C Quality of life after the ICU: a classification approach

INTRODUCTION

When reporting expected changes in a patient's health-related QoL, one could argue that it is not essential to know the second or third decimal of the change in a patient's EQ-5D score. What is more intuitive, especially towards patients and laymen, is an indication of whether and to what extend the QoL will improve or worsen. In the statistical regression model, the choice for a regression-based approach was deliberately made in consultation with physicians and statisticians. However, a classification-based approach was considered but not executed. In this supplementary document, the data from the regression studies are used to explore a classification approach for one-year post-ICU QoL prediction using machine learning models.

METHODS

The targets for prediction, changes in QoL, originally were continuous values. For the classification approach, the target values were converted to class labels. The bins are based on the quartiles, dividing the target values into four approximately equally sized classes based on the severity of the change in QoL. In Table C1, the ranges for the class division is presented.

Target range	Class label	N
<-0.08	0 (QoL decrease)	314
[-0.08, 0.01)	1 (Very little to no change)	339
[0.01, 0.15)	2 (Small QoL increase)	318
>= 0.15	3 (Large QoL increase)	337

Table C1: Class division for prediction of changes in QoL.

Several classification models were selected for testing. These include K-Nearest Neighbours (KNN), Stochastic Gradient Descent (SGD) Classifier, Random Forest (RF) Classifier, Logistic Regression, Multi-Layer Perceptron (MLP) Classifier, Support Vector Classifier (SVC), and Gaussian Naive Bayes. First, the data were fitted on each model separately to test its results. Next, the models were combined using a voting classifier. Voting classifiers act as a wrapper to combine multiple different models into one, which ideally results in a more robust performance. For each model, the accuracy, weighted F1-score, and weighted one-vs-rest area under the receiver operating characteristic curve (AUC-ROC) were calculated from the prediction scores.

	Five-feature baseline		
Model	Accuracy	F1	AUC-ROC
KNN	0.46	0.45	0.73
SGD Classifier	0.39	0.33	0.70
RF Classifier	0.46	0.45	0.72
Logistic Regression	0.48	0.46	0.77
MLP Classifier	0.48	0.46	0.77
SVC	0.44	0.42	0.73
Gaussian Naive Bayes	0.44	0.40	0.74
Voting classifier	0.49	0.47	0.77

Table C2: Classification scores per model after cross-validation.

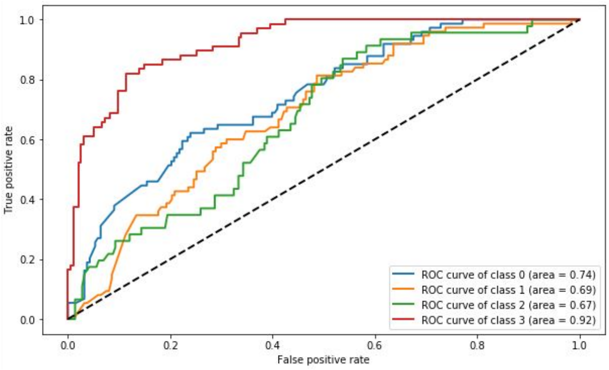


Fig. C1: The one-vs-rest ROC curves of the four classes plotted with a train/split (0.8/0.2) using the voting classifier.

RESULTS

In Table C2, the results of the classifiers are shown. The results were based on the five-feature model, as feature selection methods showed that adding additional EHR variables would not improve the results for classification.

For this problem, a classifier would perform at chance level with accuracy scores of 0.25. All classifier models tested performed above chance level. The best-performing model, the voting classifier, returned an accuracy of 0.49, an F1-score of 0.47, and a weighted one-vs-rest AUC-ROC of 0.77. The plot in Figure C1 shows that the predictions for the class with large QoL increases contribute most to the relatively decent AUC-ROC score overall. The scores for the other classes are lower and are considered acceptable at best. When looking into the predictions, 49% of the predictions exactly match their true label. The other predictions are largely represented (36.5%) by misclassification of just one class difference (e.g. a true label of 0 while the predicted class is 1). The rest (14.5%) is off by more than one class.

DISCUSSION

The prediction of patients' QoL at one year after ICU admission was initially treated as a regression problem. In this supplementary document, the problem was treated as a classification problem in order to see whether the performance of classifiers on this data set could be of sufficient quality to be used in practice at the ICU. The results show that an accuracy of 49% could be reached for the four-class problem. While this indicates that the model performed better than chance for this classification problem, the score also shows that more work is needed before the model can be put into practice.

The bins for the classes were created based upon the distribution of IQRs within the target QoL values. This decision was made to equalize the size of all classes, so that class imbalance and skewness in the results are prevented. Another bin division could have led to different classifier results. However, an underlying issue for the classification approach in this problem is that the class predictions made by the classifier do not take into account the continuous values on which the classes are based. In other words, if a data point falls under class 1 because of a change in QoL value of 0.009 and another data point is class 2 because of the change in QoL of 0.01, the difference between the two patients corresponding to these data points could be minimal. However, the classifier is trained to learn that these are two different classes. This could be the reason that the AUC-ROC scores for classes 0 and 3 are higher than the scores for classes 1 and 2. Likely because of the relatively small range for change in QoL for classes 1 and 2, it was more difficult for the classifier to differentiate between the two classes.

EHR data did not have a positive effect on the classification scores. In comparison to the regression approach, in which some EHR variables had a positive effect on the prediction results, this was surprising. The EHR variables that have shown to be of positive influence to the model in the regression study, did not contribute to higher classification scores in this supplementary study. It is suspected that these variables do decrease the continuous absolute distance between the predicted and true labels, but often in such a way that the prediction falls just over the edge of the class range into another class for the classification approach. The finding concerning the additional EHR data further supports the conclusion made in the regression approach in which it was stated that the five-feature model that is currently in use is the most effective and efficient until further research proves otherwise.

D An AI student's guide to ICU projects

by Lisette Boeijenk and Manon de Jonge

The Radboudumc ICU was unknown territory for AI students until recently. No AI student had started an internship or project at the ICU before the beginning of 2020. The ICU at Radboudumc was mostly used to guiding students from other disciplines, such as (Technical) Medicine. As the manners, rules, and regulations differ between almost every university program, the guidelines coming from the AI department were mostly unknown to the ICU staff and the student's external supervisors. During the first projects, several obstacles regarding the flow of the projects were encountered by both the students and the ICU staff. This document serves as a guide for students that are planning to start their project at the Radboudumc ICU, in order to prevent the same issues from arising and to ease the start of new projects. The guide can be extended by others who have finished their projects at this department, which makes it a living document.

Before the project

Once the student has decided on an internship or project within the ICU, there are several steps to take before the starting date to make for a smooth start. Make sure to:

- Contact all the involved people (e.g. supervisors, physicians, PhDs) early, before the actual starting date. Also include your AI track leader or the person responsible for the approval of your project.
- Ask the supervisor(s) at the ICU to help with requesting and setting up a Radboudumc-account and get an employee pass. They can help put some pressure on the right people, as the process of requesting an account can be a bit slow.
- Let the supervisor(s) at the ICU know what is needed for this specific project. Examples are a specific programming environment, a digital research environment (DRE), or extra processing power.
- Keep in mind the compensation for projects within AI for Health; a compensation of 500 euros is offered for 6 months. Usually, the interns of other disciplines at the Radboudumc do not receive compensation. This can cause some confusion within the ICU and HR, so make sure to communicate well towards the people registering this compensation.
- Inform about the data needed for the project. If the data set needed for the project is not yet available or has not yet been gathered from the database, start by already discussing with the supervisors about what kind of data would be needed. This prevents getting stuck without data for a while because the ICT staff of the ICU is usually very busy and this project might not have their priority at the moment.

All of the above will also be helpful for writing the internship or thesis proposal.

During the project

While every project has its own specific scenarios and potential obstacles, the recommendations mentioned below are of use to most ICU projects. In no specific order, keep in mind the following:

- Physicians are usually busy people. It might be hard to arrange a meeting with them, as they also often have night shifts. When a meeting is arranged, be prepared to process a lot of information at once. It is wise to take notes or to assign a secretary for the meeting.
- Almost everybody within the ICU is willing to help out if the project gets stuck for whatever reason. Most of them are very approachable, so do not hesitate to walk by their office or to send them an email whenever their help is needed.
- Once more specific details are known about the data required for the project, make sure to pass them in the most detailed way to the person that is extracting the data from the database. An example: for EHR data, there are multiple ways of measuring glucose values (blood gas, urine, etc.) and only one of these values might be needed for the project. The supervisors or involved physicians will probably know which value is needed. This prevents an unnecessarily high workload for the ICT staff.
- Chances are high that, once the required data is extracted, it does not hold the format that can be used in the project. A large part of the project time will likely be spent on data preparation. It can be easy to underestimate this part of the project.

Useful materials

During the first AI projects within the ICU, the need arose for a place to easily store and share documents between all the people involved in the project. This resulted in the creation of a Google Drive shared folder that is accessible by students, supervisors, and the involved physicians. This Google Drive folder can be used for storing repeatedly used documents, updating the project team with presentations, or for keeping minutes accessible for all. Parts of the research of previous AI ICU projects are stored on there as well. Access to this drive has to be granted, as it is not publicly accessible. ICU Research staff will be able to grant access. There is also the option of storing files in the local ICU drive accessible via a Radboudumc account. This can be used to store files that include patient data or otherwise sensitive information that should not be put online.

Further information

For specific information regarding an AI internship or thesis, please refer to the Rules and Regulations that can be found on the intranet page for AI BSc and MSc students.

Good luck with the internship/project!