Optimizing an existing prediction model for quality of life one-year post-intensive care unit: An exploratory analysis

Manon de Jonge\(^1\) | Nina Wubben\(^1\) | Christiaan R. van Kaam\(^1\) | Tim Frenzel\(^1\) | Cornelia W. E. Hoedemaekers\(^1\) | Luca Ambrogioni\(^2\) | Johannes G. van der Hoeven\(^1\) | Mark van den Boogaard\(^1\) | Marieke Zegers\(^1\)

\(^1\)Department Intensive Care Medicine, Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, Netherlands
\(^2\)Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

Correspondence
Marieke Zegers, Department of Intensive Care Medicine, Radboud University Medical Center, Nijmegen, Netherlands.
Email: marieke.zegers@radboudumc.nl

Funding information
Zorginstituut Nederland

Abstract

**Background:** This study aimed to improve the PREPARE model, an existing linear regression prediction model for long-term quality of life (QoL) of intensive care unit (ICU) survivors by incorporating additional ICU data from patients’ electronic health record (EHR) and bedside monitors.

**Methods:** The 1308 adult ICU patients, aged ≥16, admitted between July 2016 and January 2019 were included. Several regression-based machine learning models were fitted on a combination of patient-reported data and expert-selected EHR variables and bedside monitor data to predict change in QoL 1 year after ICU admission. Predictive performance was compared to a five-feature linear regression prediction model using only 24-hour data (\(R^2 = 0.54\), mean square error (MSE) = 0.031, mean absolute error (MAE) = 0.128).

**Results:** The 67.9% of the included ICU survivors was male and the median age was 65.0 [IQR: 57.0–71.0]. Median length of stay (LOS) was 1 day [IQR 1.0–2.0]. The incorporation of the additional data pertaining to the entire ICU stay did not improve the predictive performance of the original linear regression model. The best performing machine learning model used seven features (\(R^2 = 0.52\), MSE = 0.032, MAE = 0.125). Pre-ICU QoL, the presence of a cerebro vascular accident (CVA) upon admission and the highest temperature measured during the ICU stay were the most important contributors to predictive performance. Pre-ICU QoL’s contribution to predictive performance far exceeded that of the other predictors.

**Conclusion:** Pre-ICU QoL was by far the most important predictor for change in QoL 1 year after ICU admission. The incorporation of the numerous additional features pertaining to the entire ICU stay did not improve predictive performance although the patients’ LOS was relatively short.

**KEYWORDS**
critical care, machine learning, prediction modeling, quality of life, survivors
1 | INTRODUCTION

Annually, over 75,000 patients are admitted to Dutch intensive care units (ICUs). Due to advances in critical care medicine, more patients survive their critical illness. However, it is estimated that 25%–75% of ICU survivors experience long-term physical, psychological, or cognitive complaints, as well as problems related to daily functioning. These issues often negatively influence ICU survivors’ quality of life (QoL) and financial and social situation. Consequently, ICU healthcare professionals’ emphasis has shifted from not only mortality prevention but also to what surviving critical illness means for patients in the long term.

Currently, medical decisions about admission and treatment in the ICU are often based on the experience and intuition of ICU physicians to fill gaps in the knowledge of ICU physicians. Though this is undoubtedly important, ICU caregivers are not always able to reliably predict their patients’ long-term QoL and prognostic uncertainty may be a barrier for physicians for discussing long-term ICU outcomes with patients and their family members. To better support ICU decision-making and ameliorate the burden of prognostic uncertainty, the use of long-term patient-reported outcomes (PROMs) after ICU stay is of great relevance.

Through the use of PROMs, prediction models that contain not only medical data but also patient-reported data can be incorporated in ICU clinical decision making. There have only been two studies that investigate the prediction of long-term QoL post-ICU. One of these is the PREPARE (PREDicting PAtients’ long-term outcome for REcovery) model, a practically usable five-variable prediction model for change in QoL 1 year after ICU admission. The PREPARE model has been developed by the authors of this study. Included predictors are pre-ICU QoL and frailty status, sex, and ICU admission following a planned surgery and a cerebro-vascular accident. Of these, pre-ICU QoL was shown to be the most important contributor to the predictive performance of the PREPARE model. All of these predictors are measurable on the first day of ICU admission. The predictive performance of the model might be improved by the incorporation of additional patient data from the electronic health record (EHR) and bedside monitor data, that pertains not only to the first day of ICU admission, but to the following days after admission as well. The further improvement may increase the validity and reliability of the model that could then be used to improve the quality of communication between physicians, patients and family members concerning long-term ICU outcomes that takes place at the end of a patient’s ICU stay.

Therefore, we aimed to improve the performance of the existing PREPARE prediction model for long-term QoL by incorporating additional physiological, pathological, drug and treatment data from the EHR, including bedside monitor data, pertaining to the entire ICU stay. To facilitate the incorporation of the new data originating from several sources, machine learning methodology was used.

2 | METHODS

2.1 | Study design, setting, and population

This is a sub-study of the MONITOR-IC study, an ongoing multicenter prospective cohort study in which long-term outcomes of ICU patients are measured up to 5 years after ICU admission. The study protocol can be referred to for more detail. The MONITOR-IC study has been approved by the research ethics committee of the Radboud University Medical Center (Radboudumc), CMO region Arnhem-Nijmegen (number 2016–2724). Informed consent was obtained from all patients or their legal representative for the use of EHR data. All patient data were pseudo-anonymized. This study uses single center data from a university hospital.

For this analysis, the follow-up period was 1 year after ICU admission. Questionnaires were submitted at admission and 1 year after admission. Patients were excluded when they had been admitted to the ICU for less than 12 h, had a life expectancy of less than 48 h, or could not read or speak the Dutch language.

2.2 | Data sources

The original prediction model (PREPARE model) was based on variables from the MONITOR-IC database including patient-reported data (e.g., QoL and frailty) and a limited number of variables from the EHR of the first 24-hour of the ICU admission (e.g., admission type, diagnosis, severity of illness, and length of stay) (Additional File 1, Table S1). The PREPARE model was developed to predict change in QoL 1 year after ICU admission and is aimed at ICU-survivors, to prepare them for life post their ICU admission. To build on the results of the first study, the same population was used in this study. During the development of the PREPARE model, multiple imputation was used to explore the consequences of independent variable missing data, but did not significantly change predictive performance and was subsequently not a part of the final methodology. As this study builds on the PREPARE model, multiple imputation was also not applied here.

In this study, the data used in the PREPARE model has been supplemented with additional EHR data (e.g., medication data, laboratory...
results and observations) and data from bedside monitors (e.g., blood pressure, body temperature) pertaining to the entire ICU stay. In order to do this, machine learning methodology was used.

Extracting the entire EHR database is expensive in terms of time and difficult regarding memory capacity. The EHR includes measurements that are likely of no influence to QoL and can be left out. Therefore, in order to narrow down the number of EHR variables, a series of expert meetings with four ICU physicians and one ICU nurse were organized. Within this process of expert-based selection, variables that these clinicians know or suspect to have any influence on long-term outcomes of QoL of ICU survivors were selected. This expert-augmented machine learning approach narrows down the amount of training data, which reduces computational issues while keeping supposed influential factors in the data based on domain-knowledge. In Additional File 1, Table S1, all selected variables are shown.

2.3 Primary outcome measure

The primary outcome measure is change in QoL: the index value on the EuroQol 5D (EQ-5D-5L) questionnaire 1 year after ICU admission minus the index value on the EQ-5D-5L questionnaire obtained at ICU admission pertaining to pre-ICU health. The EQ-5D-5L is a standardized instrument for measuring QoL in medical care based on five questions. The Dutch EQ-5D-5L calibrated index values can range from 0.45 to 1, with higher index values indicating a better QoL.

2.4 Preprocessing and feature selection of extended data

Preprocessing the data, selection of the most influential features and modeling the data were done in Python (version 3.7) using the scikit-learn library. Outliers in the data were removed or corrected and missing values were replaced by group-wide averages or zeros depending on the type of variable. The number of missing values per variable was never more than 5% of the total number of entries per variable. Group-wide averages were used to replace missing values for bodily measurements and lab results. Missing values in medication data were replaced with zeros, as it could not be assumed that any medication was administered when no data was entered. Continuous physiological parameters (e.g., blood pressure) were preprocessed to represent the changes over time rather than the actual values.

To avoid overfitting and to increase the clarity and interpretability of the model, several features in the data were left out of the prediction model after discussions with medical professionals as well as artificial intelligence (AI) experts for reasons that included whether the medical professionals thought the variables of importance, data availability and available data quality. The selection process, as well as the variable derivatives chosen for each variable, can be viewed in Additional File 2, Table S2. Out of the remaining features, the most influential features were selected by recursively training on an ever-shrinking set of features using Recursive Feature Elimination (RFE). This is a feature ranking method that recursively considers and assigns weights to smaller sets of features. The least important features were pruned. The feature set resulting in the highest adjusted R² is used for modeling. RFE was not used in models that use regularization.

2.5 Prediction modeling

The PREPARE model was first replicated with machine learning techniques: the baseline model. Then, several machine learning techniques used for prediction were applied to the 24-hour PREPARE data as well as to the data pertaining to the entire ICU stay in order to find the best performing prediction model. The 10 machine learning models that were compared 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) (Additional File 3, Table S3). Most of these models are linear models. The non-linear models were used to explore whether they could fit the data more accurately than linear models.

The predictive performances of the different machine learning models using only 24-hour PREPARE data were compared. Then, the best performing models were tweaked to perform optimally on the extended data set based on the entire stay.

2.6 Validation and evaluation

To be able to judge the machine learning models’ performances and allow for direct comparison with the PREPARE model, the explained variance metric was used to evaluate the models after cross-validation. Four-fold cross validation was used for all models used in the current study. Another metric commonly used to evaluate regression models, the mean squared error (MSE), was also taken into consideration to provide a more complete overview of the results and to provide a comparison to the PREPARE model. The mean absolute error (MAE) was added as a metric in this study to give an indication of predictive performance. All these measures were also used to find the differences in performance between the models based on the data from the first 24 h of the ICU stay and the models based on data from the entire stay.

3 RESULTS

Between July 2016 and January 2019, a total of 4315 patients were admitted to the ICU, of whom 2291 patients were included in the
MONITOR-IC study. Of these, 780 (34.0%) were excluded because their one-year outcome measurement was missing, either due to death ($n = 184$, 8.0%), a wish to discontinue study participation ($n = 291$, 12.7%), noncompletion of the questionnaire ($n = 36$, 1.6%) or other, unknown reasons ($n = 269$, 11.7%), resulting in a dataset of 1511 patients, of whom 1308 records had no other missing non-outcome data (Figure 1). An overview of the baseline results for those who were lost to follow-up due to nonsurvival or due to reasons other than nonsurvival (Additional File 4, Table S4) shows that those who were lost to follow-up were more often female, had comparatively worse QoL index values at baseline and more chronic conditions.

### 3.1 Study population

The study group mainly consisted of male patients (67.9%). The median age of the patients was 65 years (first and third interquartile range (IQR) = 57–71). Most patients were admitted for planned surgery (72.7%). The median EQ-5D-5L index value...
### Model performance based on 24-hour data: Machine learning models

Next, the predictive performances of the different machine learning models were tested. The models were first developed using the 24-hour data and consequently compared on their performance. Scores were calculated over predictions on unseen data. A more detailed view of the performances of the different machine learning models can be found in Additional File 3, Table S3. Based on MAE score, the best performing model for the 24-hour data is the Huber Regressor. The model has an adjusted $R^2$ of 0.52, an MSE of 0.032 and an MAE of 0.126.

### Model performance based on entire ICU stay data: extended machine learning models

The variable selection process based on expert opinion can be viewed in Additional File 2, Table S2. EHR data are, in their original state, not suitable for modeling. After the selection process, the variables were therefore preprocessed and structured to fit the regression models.

The models were extended with 71 EHR data features (Additional File 5, Table S5) in addition to variables measured in the first 24-hours of admission included in the PREPARE model. An overview of the features that remained after feature selection is presented in Table 2. Three features were selected in all models explored in this study: the baseline EQ-5D-5L index value, the presence/absence of a cerebrovascular accident (CVA), and the highest body temperature measured during a patient’s stay. The first two are predictors from the PREPARE model, while the temperature measurement was an additional variable gathered from the EHR. 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 more EHR data ranges from 7 to 15. In every model, the addition led to slightly improved adjusted $R^2$, MSE and/or MAE scores. Improved adjusted $R^2$ scores differed in a range of 0%-5% from the five-feature baseline model. MSE improvements reached a maximum difference of 0.003. Every model had a lower MAE using the additional EHR features when compared to the classic model, with the lowest MAE being 0.125 in the Huber Regressor and SVR models. The Huber Regressor model was (among the) best performing model(s) when using both 24-hour and extended data.

A direct comparison of model predictive performance after the addition of more EHR and bedside monitor data should only be made for models that use the same methodology (Table 3). Therefore, comparatively, the Huber Regressor model using extended data, which had an $R^2$ of 0.52, an MSE of 0.032, and a MAE of 0.125, delivered an equal $R^2$ and MSE and a very slightly improved MAE when compared to the Huber Regressor model that used only 24-hour data. Though a direct comparison with the performance of the baseline model, the classical prediction model replicating using machine learning techniques, cannot be made, its performance overall appears similar as well. Of the three perpetually selected features, baseline (pre-ICU)
**TABLE 2** The variables selected by RFE or a built-in selection method based on feature importance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Selected by models</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D index value before admission</td>
<td>100%</td>
</tr>
<tr>
<td>Cerebrovascular accident (CVA)</td>
<td>100%</td>
</tr>
<tr>
<td>Highest temperature during ICU stay</td>
<td>100%</td>
</tr>
<tr>
<td>Sex</td>
<td>90%</td>
</tr>
<tr>
<td>Frailty before admission</td>
<td>80%</td>
</tr>
<tr>
<td>Low BMI</td>
<td>80%</td>
</tr>
<tr>
<td>Delirium prevalence</td>
<td>80%</td>
</tr>
<tr>
<td>Admission source</td>
<td>70%</td>
</tr>
<tr>
<td>Intracranial mass</td>
<td>70%</td>
</tr>
<tr>
<td>Respiratory insufficiency</td>
<td>70%</td>
</tr>
<tr>
<td>Low hemoglobin</td>
<td>60%</td>
</tr>
<tr>
<td>Malignant hematological disease</td>
<td>60%</td>
</tr>
<tr>
<td>Tracheostoma prevalence</td>
<td>40%</td>
</tr>
<tr>
<td>Nr. of registered mean inspiratory pressure (MIP) measurements</td>
<td>40%</td>
</tr>
<tr>
<td>Cardio pulmonary resuscitation (CPR)</td>
<td>40%</td>
</tr>
<tr>
<td>Dysrhythm</td>
<td>40%</td>
</tr>
<tr>
<td>Low-serum sodium</td>
<td>30%</td>
</tr>
<tr>
<td>Mechanically ventilated in the first 24 h</td>
<td>10%</td>
</tr>
<tr>
<td>Length of stay (LOS)</td>
<td>10%</td>
</tr>
</tbody>
</table>

*aSelected by RFE or built-in regularization, models that have no built-in feature importance metric were tested by hand. Inclusion in all of the 10 compared models equals 100%.

EQ-SD-5L QoL index value was the most important, as it contributed most to a reduction in the MAE (results not shown).

The changes in QoL index values 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 $0.2$. This means that a shift from $5$ to $7$ features leads to increases of $5\%$ and $2\%$ in predictive performance respectively in comparison to the classic linear regression model.

In Figure 2, the predictions of the Huber Regressor with seven features are plotted against the true values for change in QoL. The plot shows 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 study.

**TABLE 3** Features included and predictive performance of the Huber Regressor model using 24-hour data compared to the Huber Regressor model using data pertaining to the whole ICU stay

<table>
<thead>
<tr>
<th>Variables selected By 24-hour model</th>
<th>By extended data model</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-SD index value before admission</td>
<td>EQ-SD index value before admission</td>
</tr>
<tr>
<td>Cerebrovascular accident (CVA)</td>
<td>Cerebrovascular accident (CVA)</td>
</tr>
<tr>
<td>ICU admission followed planned surgery</td>
<td>Highest temperature during ICU stay*</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex</td>
</tr>
<tr>
<td>Frailty score before admission</td>
<td>Frailty score before admission</td>
</tr>
<tr>
<td>Low BMI</td>
<td>Low BMI*a</td>
</tr>
<tr>
<td>Delirium prevalence</td>
<td>Delirium prevalence*</td>
</tr>
</tbody>
</table>

Predictive performance

<table>
<thead>
<tr>
<th></th>
<th>24-hour model</th>
<th>Extended data model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>MSE</td>
<td>0.031</td>
<td>0.032</td>
</tr>
<tr>
<td>MAE</td>
<td>0.128</td>
<td>0.125</td>
</tr>
</tbody>
</table>

**Abbreviations:** MSE, mean squared error; MAE, mean absolute error; $R^2$, explained variance.

*Features introduced in the extended dataset using EHR and bedside monitoring data.

This result was also seen during the development of the existing PREPARE model that used data of the first 24 h of admission. The incorporation of the additional expert-selected EHR and bedside monitoring data pertaining to the entire ICU-stay did not change this, nor did it improve model performance. Like the original PREPARE prediction model and previous literature, this study showed that the most important predictor for change in long-term QoL was the QoL before admission. For ICU survivors, QoL 1 year after ICU admission is largely determined by pre-ICU QoL. Even though the aim of this study, like the previous study about the development of the PREPARE model, was not to identify and explain the associations found, our finding that pre-ICU QoL is strongly associated with long-term QoL and other long-term health outcomes is reflected in other literature. A patient’s pre-ICU health status appears to be an important indicator for their post-ICU outcomes and should play a part in ICU decision-making and communication with patients and family members. Its importance as a predictor was such that the straightforward addition of available EHR and bedside monitoring measurements made during ICU stay do not meaningfully improve the predictive performance for a prediction model for change in QoL 1 year after ICU admission. Aside from pre-ICU QoL, high temperatures, low BMI, and delirium prevalence were some of the variables rated as important by the machine-learning models in the present study. Associations between QoL and these features have been found in previous studies. More research is needed to evaluate the effects of these features included and predictive performance of the Huber Regressor model using 24-hour data compared to the Huber Regressor model using data pertaining to the whole ICU stay.

**DISCUSSION**

This study showed that overall, pre-ICU QoL was by far the most important predictor for change in QoL 1 year after ICU-admission.
and other influential EHR variables on patients’ QoL changes after ICU admission.

As prognostic uncertainty fosters inaccurate expectations for future recovery and may decrease the quality of ICU communication, optimizing the use of long-term outcome data in ICU practice is of the utmost importance, especially since many former ICU patients still experience health problems a year after their admission. This is the first study in which data pertaining to the entirety of an ICU patient’s admission were used to predict changes in long-term QoL. As mentioned, the methods used to incorporate the additional data come from the field of machine learning. Previous studies that have employed such methods for both short- and long-term prediction in medicine showed an increase in predictive performance after the addition of EHR data. Others have shown that machine learning is capable of finding associations between patient characteristics (e.g., substance abuse and/or socioeconomic status) and certain medical conditions also related to QoL. The importance of pre-ICU QoL as a predictor, compared to the other predictors used in this study, might explain our finding of no improvement of predictive performance.

This study has some limitations. First, over two thirds of the study population were included after elective surgery. As these patients generally have better outcomes after their ICU-admission and are admitted for a shorter time, this sample may therefore reflect a slightly healthier cohort than the general ICU survivor population, that may have lower odds of ICU events that might have an influence on QoL 1 year after ICU admission. Also, the relatively short length of stay in our population will likely have decreased the time in which events chronicled by EHR or bedside monitor data could have an impact on long-term QoL. Second, within our methodology, compared to datasets generally used in machine learning, our population size was limited. Machine learning models generally need large amounts of training data to reach a reliable prediction result. Also, approximately a third of the patients were lost to follow-up. An overview of the baseline results for the group included the analyses compared to those who were lost to follow-up due to nonsurvival or reasons other than nonsurvival (Additional File 4, Table S4) shows that those that were lost to follow-up had comparatively worse QoL index values at baseline and more chronic conditions than the group included in the analyses. This nonresponse bias again may have caused our sample to reflect a slightly healthier cohort. However, it should be noted that the same approach with regards to missing values was taken during the development of the PREPARE model. This loss-to-follow-up has therefore not influenced the comparative performances of the PREPARE prediction model and the prediction models developed in this study.

In future research, the analysis could be replicated in a larger patient group, with a higher proportion of ICU patients with a medical admission and with a larger variety in length of stay, to verify the results. Furthermore, several expert-selected variables were not included in this study as the data for these variables could not be extracted from the EHR database. Also, the derived features from the extracted variables might not properly represent the consequences for the changes in QoL. However, as pre-ICU QoL topped all other bedside monitor and EHR variables in terms of predictive importance, the influence of these omissions is not estimated to be large.

All in all, pre-ICU QoL has high-predictive value in predicting long-term QoL. Our findings here suggest that the incorporation of pre-ICU QoL information in prediction models for long-term QoL renders the inclusion of more comprehensive physiological data unnecessary. Therefore, future implementation of a prediction model for long-term QoL in critical care practice to improve the quality of communication between physicians, patients and family members about long-term ICU outcomes might be able to be achieved using the comparatively less complicated 5-feature linear regression PREPARE model.

5 | CONCLUSION

This study showed that pre-ICU QoL was by far the most important predictor for change in QoL 1 year after ICU-admission in a prediction model that used data pertaining to the entire ICU stay. The
incorporation of readily available additional straightforward expert-selected EHR and bedside monitor measurements pertaining to the entire ICU stay did not improve the predictive performance of an existing prediction model using only data from the first 24 h of admission, although it is important to note that the length of stay of this population was relatively short. Future research might improve the predictive performance by employing a larger cohort and more expert-selected predictors. Until then, the implementation of QoL outcomes into critical care practice might be able to be achieved using a comparatively less complicated model that is understandable for clinicians.

FUNDING INFORMATION
This work was supported by the National Health Care Institute (2018026879). The National Health Care Institute was not involved in the design of the study, nor with the data collection, analysis, interpretation or writing of the manuscript.

CONFICT OF INTEREST
The author declares there is no potential conflicts of interest.

DATA AVAILABILITY STATEMENT
The datasets used and analyzed in this study are available from the corresponding author on reasonable request.

ORCID
Nina Wubben https://orcid.org/0000-0001-6461-0068

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.