ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Business Analytics & Quantitative Marketing

Dynamically Predicting the Mortality Hazard of Covid-19 Patients at the ICU

Name student: Matthijs Daniël Otten

Student ID number: 453401

Supervisor: Olga Kuryatnikova

Second assessor: Mikhail Zhelonkin

Supervisor from Amsterdam UMC : Harm-Jan de Grooth

Date final version: August 13, 2021

Abstract

As from the start of the Covid-19 pandemic, researchers have looked for factors that predict the mortality risk of patients. We add to this search by using Decision Trees, Random Forests and Binary Logit models to accurately predict, at any day during the stay at the ICU, the chances of dying in the coming 24-hours. We also focus on the interpretability of the models, so clinicians can compare their expert knowledge with the insights that the fitted models provide. We check whether significant differences in model performance and relations between mortality risk and explanatory variables exist among the 18 hospitals. If so, this could point to between-hospital practice variation with respect to the perceived chance of succesful recovery due to continuing the treatment of the most severely ill patients. The Decision Trees performed inferior, while the Binary Logit models and less interpretable Random Forests performed very well and similarly. Usual medical variables, such as the fraction of inspired oxygen, and dummies indicating missing data were important in model fitting. Two separate tests indicated statistically significant differences between hospitals in the relation of mortality risk with the interaction term of the Length of Stay and SOFA score, but out-of-sample prediction performance did not increase when modelling these differences.

Contents

1 Introduction

Since the Covid-19 pandemic started in 2019, medical researchers have been investigating which factors were associated with a higher mortality rate for infected patients. The resulting insights, like those found by Grasselli et al. (2020), have been crucial to warn people with increased risk of dying due to Covid-19 and help governments make vaccination plans that grant these people priority. Research on treatment of patients at the Intensive Care Unit (ICU), such as carried out by Meng et al. (2020), has been crucial as well to advise ICU doctors on providing the patients with the best possible care. Additionally, Early Warning Scores (EWS) specifically for Covid-19 have been developed (e.g. Song et al. (2020)) to flag patients after one or two days at the ICU as being at increased risk or not, based on mortality prediction over their total period of stay. Before the pandemic began, dynamic models for predicting the risk of a critical event at the ICU have been around, like those constructed for cardiac arrest by (Kennedy et al. (2015)). However, research on modelling dynamic mortality hazards of Covid-19 patients at the ICU has only led to published contributions to the literature for datasets with a few hundred patients, like by (Rieg et al. (2020)). Nevertheless, recent preliminary research has shown that it is possible to use the CovidPredict database (*CovidPredict Database* (2020)) to model the short-term mortality risks of over 2000 patients at any moment during their stay at the ICU (Smit (2021)). However, data from only 6 hospitals was used and the research did not address possible between-hospital variation in the relations between outcomes and explanatory variables. Death in the ICU often depends on the perceived futility of further treatment by the medical team. This may lead to between-hospital variation in the predictive performance of the models. Therefore, the goal of our research is to accurately model this mortality hazard for Covid-19 patients at the ICU using data from 18 hospitals in this database, while also analyzing this between-hospital variation. We use several models to predict, at any day during the ICU stay, whether a patient will pass away in the coming 24 hours or not. Due to a lack of expert knowledge in the medical field, we do not try to draw conclusions on causal relations between factors and the mortality risk, since laying these causal relations requires a thorough understanding of all the back-and-forth interactions between natural events and treatment decisions. So, we focus only on *Granger causality*, i.e. whether inputs are useful in forecasting the future

output (Granger (1969)).

We are primarily interested in accurately classifying patients that are at relative high risk of mortality. However, in the medical field, it is also especially important to be able to explain important decisions that are made in cases where wrong decisions can lead to death. Otherwise, these models will probably not be used by clinicians to increase their knowledge (Zhang et al. (2019)). This combination of the importance of accuracy and interpretability in predicting the patient's life status at the ICU leads us to posing the following research question:

How can statistical modelling and Machine Learning be used to make accurate and interpretable predictions on the dynamic mortality hazard of Covid-19 patients at the Intensive Care Unit?

We use the machine learning tool Decision Tree (DT) and statistical Binary Logit (BL) models as interpretable methods, whereas the less interpretable Random Forest (RF) is used as powerful machine learning tool for classification. It will be of interest to see if the RF outperforms the interpretable models and to what extend it can deliver insights on the relations between explanatory variables and the mortality hazard more broadly, though not allowing insight in each individual classification made.

To the best of our knowledge, time spent in the ICU at the moment of evaluation has not yet been analyzed as an independently or interactively predictive factor for short-term mortality. Therefore, we believe to bring an important extension to the literature by making the Length of Stay (LOS) at the hospital an important factor in our models, in particular as interaction term with the other inputs. These interaction terms can be used to see if relations between inputs and the output differ between days. Additionally, we investigate whether insights can be found on how relations between input data on the health status of a patient and the mortality hazard differ between hospitals. It would be of interest to see if staff at different hospitals made different decisions on when to stop treatment, if they were presented with the same patient. However, this question is too complicated to answer with our knowledge and the methods we use in this research. Therefore, we analyze whether a statistical reason exists for doing further research on possible differences in how doctors decide to stop treatment. We extend the BL models by allowing for variation among hospitals in the relations between the mortality risk and input variables. If this leads to a better model fit and higher prediction performance, we

have such a statistical reason to believe different decisions on treatment termination might have been made among hospitals. For the same end, we use a Leave-One-Hospital-Out (LOHO) approach in which we fit the RF on the data from all but one hospital. Aberrant performance in combination with high variation in the importance of a variable could also point to important differences between hospitals. This leads us to answer the following second research question:

What is the between-hospital difference in model performance and estimated relations between mortality hazard and key variables?

We used expert insights of doctors from the Amsterdam University Medical Centres (UMC) to assist in posing feasible research questions and help in selecting sets of variables of clinical interest. It will be interesting to see whether the most important explanatory variables in our models correspond to the variables that doctors consider first when assessing the health status, mortality risk and recovery chances of a patient. Therefore, we pose the third research question:

Which variables have the most predictive and explanatory power in our statistical and Machine Learning models?

In Section 2, we provide a review on relevant literature to give context on the problems we try to solve, background for the methods we use and how the research contributes to the literature. In Section 3, we describe the data that we use. In Section 4, we describe how we deal with missing data and present the variable selection procedure, the Binary Logit models, Decision Tree and Random Forest. In Section 5, the results are shown. Finally, in Section 6, we answer the research questions, show the limitations and provide suggestions for further research.

2 Literature Review

The usage of Electronic Health Records (EHR) at ICUs arose in the beginning of the 1990's and one out of many reasons to transition from paper based patient records to EHRs was the potential benefit for doing data-driven research (van der Lubbe et al. (1997); Hippisley-Cox et al. (2003)). Since then, more data has been available for researchers to

find relations between certain events at the hospital, like death or the progression of a disease, and available information on the patient and treatment.

Our research falls within the field of applying statistical and Machine Learning models to predict mortality risk at the Intensive Care Unit. Even before EHRs were broadly used in the medical field, scoring systems were developed to help doctors assess the condition of a patient not only from their expert insights. Already in 1985, when patient data was not yet commonly stored in EHRs, the Acute Physiology And Chronic Health Evaluation (APACHE) score was developed to quantify the current health status of a patient (Knaus et al. (1981)). Since then, this metric has been updated several times based on new researches to optimize its precision. We use not only the derived APACHE, but also the Sequential Organ Failure Assessment (SOFA) score as explanatory variable in our models. This SOFA is constructed from 8 measures to capture the risk of organ failure, which can lead to death (Antonelli et al. (1999)). Other methods have been developed as well to assist doctors in evaluating their patient's risks, like the Mortality Prediction Model (MPM) (Lemeshow et al. (1993)) and the Simplified Acute Physiology Score (SAPS) (Le Gall et al. (1993)).

In line with these developments of using available data to quantify the health of a patient, new methods were developed to also use available data to predict Adverse Events (AE), like cardiac arrest or death, and thereby prepare doctors or even prevent the AE from happening. These methods are known as Early Warning Scores (EWS), like the National Early Warning Score (NEWS) that uses only 7 input variables to assign a score per variable and sum these up to get the score of this metric for determining whether a patient is at severe risk of an AE (Smith et al. (2013)).

These EWSs are perfectly interpretable in the sense of allowing the doctor to understand how the risk classification is constructed from the variables. However, many more powerful risk prediction methods exist in the field of Machine Learning and Artificial Intelligence (AI). For example, not only classical linear regression is used in predicting cardiac arrest at the ICU, but also Decision Trees (DT), Support Vector Machines (SVM) and Neural Networks (NN) (Kennedy et al. (2015)). However, for many of these models, individual predictions cannot be backtracked to the input data to provide the reason why a patient is classified as she is. This not only makes it hard to explain to patients and their relatives why big decisions are made, it also gives no insights for medical staff to learn

from the relations and less opportunities to improve the model using clinical knowledge after deeply understanding it (Ahmad et al. (2018)).

The task of predicting whether a patient dies within 24 hours or survives, leads us to making binary classifications. A large range of methods and models exist to do so, from which the earlier stated DT, SVM and NN are some, but also Logistic Regression, Random Forests (RF), Adaptive Boosting, Nearest Neighbours and Bayesian methods are popular, to name a few (Kumari & Srivastava (2017)). Using Logistic Regression in the case of binary classification leads to using a Binary Logit (BL) model. This is a classic statistical model that has been used for decades and has the advantage of being interpretable. That is, for an individual observation, a change in a value of an explanatory variable leads to a multiplication of the predicted odds for the event of interest to happen. In the medical field, it is common to use the BL model as a baseline to compare the performance of another model with, like a Neural Network (Goss & Ramchandani (1998)) or Random Forest (Hsieh et al. (2018)). We will use this same strategy and use besides the BL model the RF (Breiman (2001)) as machine learning method since it has become the most frequently used machine learning classification tool in the past two decades (Kirasich et al. (2018)). The Random Forest is not interpretable at the level of an individual prediction. Therefore, we also use a Decision Tree, which is the interpretable building block of a RF since it provides binary decision rules that determine how a patient is classified based on the input data (Breiman et al. (1984)).

The Binary Logit Random Effects model exists as useful extension of the BL model for allowing relations between variables and the binary response to differ between clusters of observations (Longford (1994)). This is especially useful for answering our second research question, which we do by investigating how these relations differ between hospitals. If modelling variation in relations between mortality and variances leads to higher performance, this may suggest that different decisions are made among hospitals. On the other hand, we cannot rule out other explanations, such as differences in data collection or registration methods causing this performance variation.

3 Data

We have longitudinal data from the CovidPredict database for 2245 ICU patients from 18 Dutch hospitals, of whom 22% died in the ICU. These patients were admitted to the ICU at some day between the end of February 2020 and beginning of March 2021. The distribution of the admissions over this period can be seen in Figure 1. The number of patients present in different hospitals ranges from 42 to 242, the number of deaths from 6 to 64 and the average length of stay (LOS) at the ICU from 10 to 19. With the data, we want to predict the event that a patient dies within 24 hours from the moment of prediction. Therefore, we get a total of 29,602 daily observations, about 13 per patient. Since we need to be discreet with the private data, we cannot present a summary table with hospital specific numbers, as this could help the reader connect results to known hospitals. As explanatory variables we have demographic data, information on comorbidities (other diseases besides Covid-19), medicine records, laboratory test results, treatment choices, frequently updated health conditions (such as temperature and heart rate), the length of stay at the ICU, informative functions of several variables combined and the hospital at which a patient is located. We have to account for human errors in the data registration since many variables are quickly and manually registered. We set up rules that exclude observations containing infeasible values. With these explanatory variables, we create extra variables such as cumulative scores over time, squared values and interaction terms. Our final set of variables can be found in Appendix A.1.

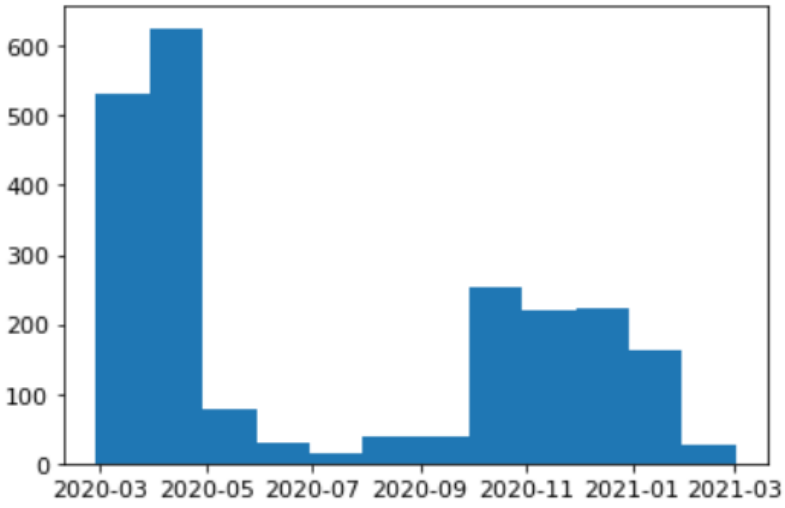


Figure 1: Histogram of ICU admissions per month

For each individual i we have different LOS (T_i) and thus unequal numbers of observations. If we do not correct for this, the model will undesirably be fitted better for those individuals with a longer LOS than those with a shorter LOS. Additionally, only about 1.7% of the observations corresponds to a a death case. This often causes bias towards predicting survival of patients (Wallace & Dahabreh (2014)). Above that, from a clinical view, correctly predicting a death instance is more important than correctly predicting survivals since an accidental death is a big problem whereas an accidental survival is not. We deal with these data problems by training the models with the final observation of the stay and one other random observation per patient, raising the number of death instances to approximately 11% of the training data, since the last observation of a patient's stay corresponds to a death case in 22% of the instances and the other observation always corresponds to a survival. Hereby we bring more balance to the training data, preventing biases towards predicting observations as survivals. When predicting out-of-sample, so for model validation or testing final performance, we use the original data. Differences between hospitals exist in the number of total patients and the distribution of the death cases. Some hospitals deliver only 2% of the patient data and others 10%. Also, the fraction of patients that die at a specific ICU ranges from 8% to 40%. When trying to explain possible differences in model performance or importance of variables among hospitals, these differences should be considered.

Since the data contains a lot of missing values, we set up a sophisticated approach to deal with this. This combines forward filling missing values based on the most recent present value, imputing missing data based on the present values of other variables and creating dummies for missing values, as we will describe in more detail in Section 4.1.

4 Methodology

We are primarily interested in making 24-hour ahead predictions on patient mortality. Each observation (i, t) for individual $i \in N$ and period $t \in T_i$, is therefore accompanied with a binary variable:

$$y_{i,t} = \begin{cases} 1, & \textit{death} \text{ occurs within 24 hours} \\ 0, & \textit{survival} \text{ for another 24 hours} . \end{cases}$$

We have J explanatory variables $X_{i,t}$, which can be split into time-invariant *static* variables (X_i^s,) that do not change over time and *dynamic* variables ($X_{i,t}^d$) that do change over time. Since we consider the possibility that relations of the explanatory variables $X_{i,t}$ with $y_{i,t}$ change over time, we include cross terms $X_{j,i,t} \cdot t$ for some explanatory variable indices j.

In order to assess the prediction performance of the different models, we use 80% of the patients for model training, whereas the data for 20% of the individuals will only be used in testing the model performance. For all our models, we will have to fix the values of some hyperparameters and we will have to find the optimal set of explanatory variables. Therefore, we will perform 5-fold cross-validation (CV) on the training data to tune the combination of hyperparameters and set of variables, for each model. This means we split the training data into 5 sets of 16% of the patients for validating the performance. Then, we predict the outcomes in one of these sets with a model trained on the data from the remaining 4 sets. We repeat this for each set and after averaging the performances over these 5 validation sets, we will see which combination of hyperparameters and variables yields the highest performance. More details on validation is provided in Section 4.4, after the models and performance metrics are described.

It is common to use mean prediction accuracy as measure for performance, which is the fraction of correctly classified observations. However, since we only have 1.7% of deaths over all observations, only predicting survivals leads to a very high accuracy of 98.3%. This seems high, but corresponds to never classifying a death case correctly. Additionally, falsely predicting a severe risk of death is less problematic than falsely considering a patient not to be at serious mortality risk. Therefore, we use more informative metrics, all constructed from 4 basic metrics:

- True Positives (TP): Predicted to die and does so
- False Positives (FP): Predicted to die, but survives
- True Negatives (TN): Predicted to survive and does so
- False Negatives (FN): Predicted to survive, but dies

From these 4 metrics, we create

- True Positive Rate (TPR or Recall): $\frac{TP}{TP+FN}$, what fraction of deaths is found by the model

- False Positive Rate (FPR): $\frac{FP}{FP+TN}$, what fraction of survivals is mistakenly classified as deaths

- Precision: $\frac{TP}{TP+FP}$, what fraction of death predictions corresponds to actual deaths

From these 3 metrics, we build the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision Recall Curve (AUPRC) and F2-score. The ROC curve has pairs of Recall and FPR respectively on the y-axis and x-axis, corresponding to the possible thresholds. A *threshold* is a number between 0 and 1 above which probability estimates are classified as positive cases (deaths). So, the area under this curve summarizes how well the model predicts observations based on evaluating what fraction of deaths is found by the model and what fraction of survivals is mistakenly classified as a death, irrespective of the chosen threshold. A value of 0.5 corresponds to a random guess, above 0.7 corresponds to moderate performance, above 0.8 is considered as very good and 1 means perfect prediction (Mandrekar (2010)). However, since the number of survivals is large (98.3%), an increase in false positives (predicting deaths when survival is true), will have a minor effect on the AUROC because of a large denominator in the FPR. Therefore, this metric might not give the best insight in model performance (Davis & Goadrich (2006)). Where average accuracy puts too much emphasis on correctly predicting survivals, the AUROC might put too little emphasis on it.

The Precision Recall (PR) curve has pairs of Precision and Recall respectively on the y-axis and x-axis, corresponding to all possible thresholds. The AUPRC shows the model's ability to find the actual deaths without predicting many observations as deaths that actually correspond to survivals (Boyd et al. (2013)). The difference with AUROC lies in using Precision instead of the FPR, whereby predicting death in case of survival is stronger penalized when using AUPRC. So, PRC lies between the extremes of using accuracy (almost all emphasis on survivals) and AUROC (almost no penalty for missing survivals). The baseline performance for PRC is equal to the fraction of death cases in the data (0.017), since by predicting every observation as a death, we would have a Recall of 1 and a Precision of 0.017, giving $1 * 0.017 = 0.017$ (Saito & Rehmsmeier (2015)).

As a final metric we have

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall} \tag{1}$$

which summarizes model classification and equals the harmonic mean of Precision and

Recall if $\beta = 1$. However, by using $\beta = 2$, we put twice as much weight on the Recall since we are more interested in correctly predicting deaths than not missing a survival. As the AUPRC, by using both Precision and Recall, this metric is also suitable for imbalanced classification (Weiss (2013)). As the F2-score is dependent on the threshold, we use Nelder-Mead optimization, which is described in Section 4.2, to find the optimal threshold for maximal F2. So, after fitting the model, we see which threshold gives the highest F2 in validation performance to show the maximum F2 that is attainable. However, this makes the F2 less useful to tune hyperparameters with because this maximum F2 shows maximal performance, not average performance on the validation set. For out-of-sample performance, the tuned threshold from validation phase is used. Hereby, the F2 score is a useful metric to assess the average performance on the test set.

The AUROC is not a fully representative metric for performance in the imbalanced case and the F2-score in the validation phase does not show average performance. As a result, we use the AUPRC as the leading performance metric in the validation phase and to evaluate performance on the test set.

For answering the second research question, we will modify the methods in such a way that we can measure how relations between explanatory variables and outcomes differ between hospitals. For Binary Logit models this means using Random Effects and seeing whether allowing relations to differ between hospitals improves model fit. Additionally, we train the RF by using the Leave-One-Hospital-Out (LOHO) approach, so using about 76% of the patients. We then test if the 18 models, one per missing hospital, perform differently on the test set and whether significant differences exist in the importances of variables among these models.

Unless stated otherwise, we use the Scikit-learn library for Python to fit the models (Pedregosa et al. (2011)).

4.1 Dealing with missing values

As stated, we first use three methods to deal with the missing data before we start modelling. The first is that of Last Observation Carried Forward (LOCF), in which a missing value is replaced by the most recent present value for that variable. Since we have data over time, this method is intuitive because it fills gaps when data was not observed at specific days. The method has two important assumptions. The first is that

the values which are missing do not differ that much from their last observation moment. This assumption is realistic for data on laboratory test results since these tests are carried out frequently, but are not necessarily recorded every day in each hospital. Since the LOCF approach also assumes that no important information is carried in the fact that a value is missing, this value must be Missing at Random (MAR) (Kang (2013)). Since this is a doubtful assumption for most variables, as missing values are strongly correlated to the hospital they are from or to the health status of a patient, we only use it for filling laboratory test results.

We need a second method for filling missing values that does not assume the missing values are MAR, but that the fact they are missing could be correlated to important information. Therefore, we use a method that imputes missing data based on data which is present. To do so, we use the Scikit-learn function *IterativeImputer* in Python which is equivalent to the MICE algorithm developed in R by (Van Buuren & Groothuis-Oudshoorn (2011)). Using only the observations for which the values are present, both approaches fit a linear model between the variable for which data should be imputed and the other explanatory variables:

$$x = X\beta + \epsilon, \tag{2}$$

where x is the variable for which missing values are to be imputed, X are the other explanatory variables without x and a constant, β are the coefficients expressing the linear relation between variables X and x, and ϵ is the error term depicting random variation. Based on this fitted model, through applying Ordinary Least Squares (OLS), a prediction is made for the missing value. This missing value is then replaced by an actually present value that is close to the prediction from the fitted model. By imputing missing values based on present values in other variables, we acknowledge that they are not MAR but that their absence could be correlated to important information. Specifically, the *Iterative Imputer* and MICE algorithms use Predictive Mean Matching (PMM), which has been developed by Little (1988). This algorithm has been described in steps by Vink et al. (2014) and we paraphrase it as follows:

Algorithm 1 Predictive Mean Matching

Input: N_p observations with present values for variable x. N_m observations for which values of x are missing.

1: Estimate $\hat{\beta}$ and $\hat{\epsilon}$ by applying OLS on equation (2) to the N_p present observations.
2: Get estimated variance $\sigma^{2*} = \hat{\epsilon}^T\hat{\epsilon}/A$, where A is a draw from the χ^2 distribution with $N_p - r$ degrees of freedom (with r the number of variables in X).
3: Draw $\beta^* \sim \mathcal{N}(\hat{\beta}, \sigma^{2*}(X_p^T X_p)^{-1})$, where $\mathcal{N}(\cdot)$ is the multivariate normal distribution.
4: Calculate estimations for present values as $\hat{x}_p = X_p\hat{\beta}$ and for missing values as $\hat{x}_m = X_m\beta^*$.
5: **for** $i \in N_m$ **do**
6: Calculate d distances $\Delta_d = |\hat{x}_{p,d} - \hat{x}_{m,i}|$ with $d \in N_p$
7: Randomly draw x_i from the present values in x_p that correspond to the three lowest values Δ_d.

So, we impute an existing value that corresponds to one of the closest estimated neighbours (in terms of absolute difference) of missing value x_i, by using the Bayesian approach of drawing from a posterior distribution $N(\hat{\beta}, \sigma^{2*}(X_p^T X_p)^{-1})$.

We conclude this session by describing our third method to deal with missing values, although we can fill all missing values by imputation. This is because the fact that a value is missing can bear information that could improve model fit and performance. Therefore, we also create dummy variables that equal one if a value is missing and zero if the value of the corresponding explanatory variable is present. For example, missing information on laboratory test results could imply that doctors are not that troubled about a patient's health status and therefore do not see the urge to monitor all values frequently, which is probably correlated to a lower mortality risk. After adding these dummies, we get 279 explanatory variables. The potential problem of very high dimensionality because of all these dummies is taken care of in variable deletion in the model validation phase.

4.2 Binary Logistic Regression Models

The first class of models to use will be that of Binary Logistic (BL) Regression Models. These are linear models used to predict the probability of an event while yielding estimated parameters that can be used to interpret relationships between the input variables and

the outcomes. To predict whether $y_{i,t}$ is 0 or 1, we use that

$$y_{i,t} = \begin{cases} 1, & \text{if } y^*_{i,t} > 0 \\ 0, & \text{if } y^*_{i,t} \leq 0 \end{cases}, \tag{3}$$

where we have latent response function

$$y^*_{i,t} = \beta_0 + \mu_i + X_{i,t}\beta + \epsilon_{i,t}, \tag{4}$$

where β_0 is the constant term over all observations, μ_i the patient specific constant to capture variance among observations explained by correspondence to the same individual, β denotes the relations of explanatory variables X with the mortality hazard and $\epsilon_{i,t}$ is the error term. With this latent variable, we get the following probability function

$$\begin{aligned} P_{i,t} &= \text{Prob}\{y_{i,t} = 1 \mid \beta_0, \beta, \mu_i, X_{i,t}\} \\ &= \text{Prob}\{y^*_{i,t} > 0 \mid \beta_0, \beta, \mu_i, X_{i,t}\} \\ &= \text{Prob}\{\epsilon_{i,t} > -(\beta_0 + \mu_i + X_{i,t}\beta)\} \\ &= \Lambda(-(\beta_0 + \mu_i + X_{i,t}\beta)) \\ &= [1 + \exp(-(\beta_0 + \mu_i + X_{i,t}\beta))]^{-1}, \end{aligned} \tag{5}$$

where Λ is the Logit link function because $\epsilon_{i,t} \sim \text{Logistic}(0, 1)$. The function domain is unlimited, whereas the range is between 0 and 1 (Wooldridge (2015)).

To interpret the coefficients, we should look at the odds of getting outcome *death*. We have:

$$\text{Odds}\{y_{i,t} = 1 \mid X_{i,t}\} = \exp(\beta_0 + \mu_i + X_{i,t}\beta). \tag{6}$$

Ceteris paribus, we therefore have that an increase in $X_{i,t,j}$ multiplies the odds of having outcome *death* by e^{β_j} (Harrell Jr (2015)). We thus can interpret the estimated relations between variables like age, Body Mass Index (BMI) or the presence of diabetes with the odds of dying in the next 24-hours. Since variables like age and BMI cannot take on the value 0, there is no clear interpretation of constants β_0 and μ_i.

4.2.1 Common Effect or Patient specific Random Effects

Since the observations (i, t) are not independent, we have added individual specific effects μ_i to the equation to account for the unexplained yet individual specific variance in the model. However, we first test whether the included time-invariant variables already

explain all the individual specific variance. This gives the Common Effect (CE) model as the restricted model with $\mu_i = 0$. The alternative model that includes these individual specific effects is either the Random Effects (RE) or the Fixed Effects (FE) model.

The FE model includes individual specific effects $\beta_{0,i} = \beta_0 + \mu_i$, whereby time-invariant individual specific variable parameters cannot be estimated, since including both individual specific effects and these time-invariant variables make the model non-identifiable. This is because any non-zero coefficient for a time-invariant variable can be offset by arbitrarily changing the individual-specific parameters $\beta_{0,i}$. Also, the FE model cannot be used for predictions of observations from patients that were not included in model fitting, since no $\beta_{0,i}$ is known for these new individuals. However, we do want to make predictions for patients that were not in the training data. Combining this with the fact that we are interested in interpreting relations between time-invariant variables and the outcomes as well, the FE model cannot be used for our research.

Therefore, we will use the RE model as the alternative model. We call it the Patient specific Intercept (PI) model since more Random Effect models will be used in the research. This model assumes μ_i are independent from each other and $X_{i,t}$ and that they follow a normal distribution:

$$\mu_i \sim \mathcal{N}(0, \sigma_\mu^2) \tag{7}$$

(Longford (1994)). Research has shown that distributions like the student's t or gamma distribution could be used for more robust estimations of the variance of the random effect in case of data with heavy tails or skewed data. Nevertheless, we use the normal distribution as it is most frequently used to model these random effects and no specific reason exists to doubt this normality assumption. In this manner, we allow for individual specific effects, can make predictions and get parameter estimates for the time-invariant variables. However, if the assumptions do not hold, coefficient estimates could be biased (Lee & Thompson (2008); Hsiao (2007)).

Despite the fact that the FE model is not used for prediction, in case of linear regression models, the assumptions of the PI model are usually verified by applying the Hausman test on the estimated coefficients of the PI and of the FE model (Hausman (1978)). However, the FE for a logit model can only be estimated with Conditional Logit (CL), for which only observations can be used from individuals that passed away during their stay at the ICU, reflected by the following condition on patients to use: i s.t. $0 < \sum_{t=1}^{T_i} y_{i,t} < T_i$

(Croissant & Millo (2019)). Since this implies throwing away 78% of the data, it does not allow for a valid comparison of coefficients.

To decide whether the CE or the PI model should be used in training and predicting, we fit the PI model and test with the following null-hypothesis:

$$\mathcal{H}_0 : \sigma_\mu^2 = 0 \quad \text{against} \quad \mathcal{H}_1 : \sigma_\mu^2 > 0 \ . \tag{8}$$

As shown by Stram & Lee (1994), we can test this with a Likelihood Ratio Test (LRT), using a significance level of $\alpha = 0.05$. We calculate the ratio of the likelihood from the unrestricted model (PI) over that from the restricted model (CE). The likelihood function per individual, conditional on μ_i, is

$$L_i(\mu_i) = \prod_{t=1}^{T_i} P_{i,t}^{y_{i,t}} (1 - P_{i,t})^{1-y_{i,t}}. \tag{9}$$

Since we have $\sigma_\mu = 0$ for the CE model, we can use Maximum Likelihood Estimation (MLE) to estimate β_0 and β with joint marginal likelihood function

$$L_{CE}(\theta) = \prod_{i=1}^{N} L_i(0), \tag{10}$$

where $\theta = [\beta_0, \beta, \sigma_\mu]$ and for the CE this σ_μ is thus assumed to be zero. We maximize this likelihood with Newton-CG optimization and thereby find the estimates $\hat{\theta}$. To do so, we first define the Gradient G as the vector of partial derivatives of L_{CE} with respect to the components of θ and the Hessian H as the matrix of second order partial derivatives. Then, the basic Newton direction p_k^N in which to update the k-th coefficient in θ, is the solution of

$$H_k p_k^N = -G_k. \tag{11}$$

To solve this equation and efficiently find an approximation of p_k^N, we use the Conjugate Gradient method as described in (Wright et al. (1999)). The estimates are iteratively updated in these approximated directions $\hat{p}^N$ until one of the stopping criteria is met. In our research, the optimization stops when all elements of the Gradient are smaller than 10^{-4} in absolute terms or 10,000 iterations have passed.

In order to estimate θ in the PI model, we should integrate over all possible values of μ_i in the likelihood function, for each $i \in N$. This gives the following joint marginal likelihood function:

$$L_{PI}(\theta) = \prod_{i=1}^{N} \int L_i(\mu_i) f(\mu_i) d\mu_i, \tag{12}$$

where $f(\cdot)$ is the probability density function (*pdf*) of the normal distribution with mean

0 and variance σ_μ^2. We estimate θ through maximizing this likelihood with the Nelder-Mead numerical optimization algorithm. It does not use the Gradient or Hessian of the Likelihood to find a direction towards which the coefficients should be updated, like in the Newton-CG case. When finding the K coefficient estimates in θ, it starts with $K+1$ candidate estimates of θ and iteratively replaces the candidate estimate that has the lowest Likelihood by one of three candidate estimates that were calculated using the Euclidean distance from the worst estimation to the center of all current estimates. If none of these three candidates is suitable, the space in K dimensions spanned by the $K+1$ estimates is shrunk inwards and the algorithm continues. Like with Newton-CG, we make the algorithm stop when 10,000 iterations have passed or when the absolute change in function value, relative change in function value or relative change in parameter values falls below $10^{-5}, 10^{-15}$ or 10^{-7}, respectively. (Wright et al. (1999)).

Since no closed form exists for the integrand, we rewrite equation (12) as

$$L_{PI}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{\pi}} \int L_i(\sqrt{2}\sigma_\mu \upsilon)) \exp(-\upsilon^2) d\upsilon, \tag{13}$$

to use adjusted Gauss-Hermite Quadrature (aGHQ) by numerically approximating

$$L_{PI}(\theta) \approx \prod_{i=1}^{N} \frac{1}{\sqrt{\pi}} \sum_{r=1}^{R} w_r L_i(\sqrt{2}\sigma_\mu \upsilon_r), \tag{14}$$

with degree R as the number of nodes and weights (υ_r, w_r) in the quadrature (Croissant & Millo (2019)). The weights are derived from the nodes. These nodes are in the first iteration centered around zero, but are adapted in further iterations to be centered around the conditional modes $\tilde{\mu}$, which maximize the Likelihood given estimates in the current iteration $\hat{\beta}_0^{iter}, \hat{\beta}^{iter}$ and $\hat{\sigma}_\mu^{iter}$ (Bates (2014)). By adapting the nodes and thereby the weights in the iterations, the numerical optimization is faster in finding the optimal parameters, especially when these conditional modes deviate a lot from zero (Kim et al. (2013)).

To formally test

$$-2\log(L_{CE}(\hat{\theta}_{CE})/L_{PI}(\hat{\theta}_{PI})), \tag{15}$$

which approximately follows $\sim 0.5\chi_q^2 + 0.5\chi_{q+1}^2$ under $\mathcal{H}_0$ (Stram & Lee (1994)), with q the number of random effects under the null hypothesis. So, with a significance level of $\alpha = 0.05$, we reject the null and conclude individual specific effects should be modelled if $-2\log(L_{CE}(\hat{\theta}_{CE})/L_{PI}(\hat{\theta}_{PI})) > 1.921 = 0.5(0) + 0.5(3.841)$.

4.2.2 Hospital specific Random Effects

After determining whether the CE or PI is the correct model, we also use the LRT to investigate hospital level differences between relations of X with the mortality hazard. Of special interest are the Length of Stay (LOS), the SOFA score as quantified approximation of the patient's current health status, and the interaction term between LOS and SOFA. This is because we want to see whether the data suggests that hospitals differ in how long they take and how sick patients have to be before the doctors end treatment. We want to use as many observations as possible for maximum model fit and robustness, so we must consider an FE or RE model instead of running a regression per hospital. Again, we do not use Fixed Effects since this does not yield coefficients to variables that are constant for all patients in the same hospital. For example, a certain hospital might not have data for the usage of a specific medicine, resulting in zero values for all patients. So to do the analysis, we stepwise add random effects that differ per hospital to either the CE or PI model, but only if adding the random effect significantly improves the likelihood. These added candidate variables to have their coefficients affected by random effects in the new set-up are the *Intercept*, *LOS*, *SOFA* and interaction term *SOFA∗LOS*, which are from now on related to as random effects. We call these models Hospital specific Coefficient (HC) models. So, we have L additional random effects in each candidate HC model, with $L \in \{1, 2, 3, 4\}$.

We copy the data on these L variables from X to treat as random effects into matrix Z with corresponding hospital specific coefficients γ_h for each observation corresponding to an individual that was treated in hospital h. Note that Z might contain a column of constants. Now we can rewrite equation (5) as

$$P_{i,t} = [1 + \exp(-(\beta_0 + \mu_i + X_{i,t}\beta + Z_{i,t}\gamma_h))]^{-1}, \text{ with} \tag{16}$$

$$\begin{aligned} &\gamma_h, \mu_i \sim \mathcal{N}(0, \Sigma), \\ &\Sigma = \begin{bmatrix} \sigma^2_{\gamma_1} & & & & \\ \sigma_{\gamma_1,\gamma_2} & \sigma^2_{\gamma_2} & & & \\ \dots & \dots & \dots & & \\ \sigma_{\gamma_1,\gamma_L} & \sigma_{\gamma_2,\gamma_L} & \dots & \sigma^2_{\gamma_L} & \\ \sigma_{\gamma_1,\mu} & \sigma_{\gamma_2,\mu} & \dots & \sigma_{\gamma_L,\mu} & \sigma^2_{\mu} \end{bmatrix}, \end{aligned} \tag{17}$$

where using μ_i will depend on the outcome of the earlier LRT between the CE and PI model. The new coefficients γ_h (and μ_i) are assumed to follow a multivariate normal distribution with mean 0 and covariance matrix Σ (Hsiao & Pesaran (2008)). Under the null hypothesis, all elements in Σ are zero, like in the case with one Random Effect. We use the Bonferroni correction to adjust the significance level α for multiple testing by dividing it by the number of candidate random effects in the corresponding step (Neyman & Pearson (1928)). This correction is required because testing several hypotheses at the same time with the original significance level α increases the chance of finding a significant result, though this could be due to applying more than one test. For example, when testing for 4 random effects with a significance level of 5%, the chance of wrongly rejecting the null hypothesis (error rate) is $1-(1-\alpha)^4 \approx 18.5\% > 5\%$. When using an adjusted α^* by dividing the significance level by the number of tests, the error rate does not exceed the original α (Armstrong (2014)).

We again use conditional likelihood, now also conditional on γ, as in equation (9) to get estimates $\hat{\theta} = [\hat{\beta}_0, \hat{\beta}, \hat{\Sigma}]$. The full marginal likelihood however includes L additional integrals and a product over H hospitals, whether we use the CE or PI model. This results in

$$\begin{aligned} L_{CE}^{HC}(\theta) &= \prod_{h=1}^{H} \int ... \int \prod_{i=1}^{N} L_{i,t}(0, \gamma_h) f(0, \gamma_h) d\gamma_{h,1} ... d\gamma_{h,L} \qquad \text{for CE, or} \\ L_{PI}^{HC}(\theta) &= \prod_{h=1}^{H} \int ... \int \prod_{i=1}^{N} \int L_{i,t}(\mu_i, \gamma_h) f(\mu_i, \gamma_h) d\mu_i d\gamma_{h,1} ... d\gamma_{h,L} \qquad \text{for PI,} \end{aligned} \tag{18}$$

where $f(\cdot)$ is now the *pdf* of a multivariate normal distribution with means μ_i, γ_h and covariance matrix Σ. In order to calculate the integrals, we make use of aGHQs and Nelder-Mead numerical optimization as described before, but now for L extra integrals (Longford (1994)).

Predicting with a Binary Logit model is done by inserting the new data into the estimated model, but μ_i and γ_h are not estimated in the RE models. However, we have the option of estimating the conditional modes $(\tilde{\mu}_i, \tilde{\gamma}_h)$ of the random effects based on the observed outcomes in the training data. In other words, we maximize the unnormalized posterior distribution of μ_i and γ_h, which leads to:

$$\tilde{\mu}_i, \tilde{\gamma}_h = \underset{\mu_i, \gamma_h}{\arg\max} \left(\prod_{h=1}^{H} \prod_{i=1}^{N} L(\mu_i, \gamma_h | \hat{\beta}_0, \hat{\beta}, \hat{\Sigma}) f(\mu_i, \gamma_h) \right) \tag{19}$$

We already used the conditional modes as points around which the nodes in the aGHQs

were centered, but now use them as well for out-of-sample prediction with

$$\hat{P}_{i,t} = [1 + \exp(-(\hat{\beta}_0 + X_{i,t}\hat{\beta} + Z_{i,t}\tilde{\gamma}_h))]^{-1} \tag{20}$$

Note that we have to set $\mu_i = 0$ instead of inserting $\tilde{\mu}_i$ since in the out-of-sample data, no observations are present for the individuals from the training set, but we do have observations from all 18 hospitals in the validation sets and test set (Bates (2014)).

For these RE models, we do not want to choose the specification solely based on the formal LRT, since it only uses information on in-sample performance, which can make us choose a model that performs best on the training data, but worse out-of-sample, which is called *overfitting*. We therefore also want information on out-of-sample performance to see for CE and PI what random effects to pick as extension out of the 15 possible combinations of *Intercept*, *LOS*, *SOFA* and *SOFA*∗*LOS*. Hereby, we end up with 32 sets, but for predicting on the test set we use the one that corresponds with the highest AUPRC in cross-validation, besides the RE model that resulted from the LRTs.

We fit the RE models by using the Pymer library for Python (Jolly (2018)).

4.3 Decision Tree-Based models

A Decision Tree (DT) is a model that can be used for classifying observations based on a set of binary decision rules. A huge advantage of a DT is that the corresponding logic on why a specific observation is assigned a particular class can be shown in a way which is so easily interpretable, that no expert knowlegde on Machine Learning is required to understand it. If doctors should be able to understand how a model comes to a specific classification, interpretable models are required. Combinations of individual DTs result in Random Forests (RF), which can be used to make more accurate predictions. However, this comes at the cost of losing interpretability since the way an individual classification is made by a RF cannot be explained by the combination of explanatory variables and a set of binary decision rules or coefficients. Nevertheless, methods exist to assess the importance of the different explanatory variables in the resulting RF. We refer to this as Feature Importance (FI), for which the calculation will be described for both the DT and RF in upcoming sections.

Since no parameters are involved to represent the relations between explanatory variables and outcomes, an approach like the RE model is not possible to answer the second research questions on variation between hospitals. A way to assess the variability of these

relations is to train and validate a DT or RF per hospital and then compute the performance metric and FIs. However, with on average only 124 patients per hospital and a lot of candidate variables, we think the variability to be extreme and not informative. Therefore, we use a Leave-One-Hospital-Out (LOHO) strategy where we use all but one hospital in the model building and see whether significant changes in performance and the FIs occur when a hospital is left out. To test the null hypothesis of no significant differences in performance, we calculate the median and Median Absolute Deviation with the metrics corresponding to the test set performances of the 18 hospitals sets. Assuming that these performances follow a Normal distribution, we conduct robust two-sided modified z-tests to draw conclusions on whether differences from the median are significant and so a hospital set performs notably better or worse. To do so, we first take the 18 performance scores s_h ($h \in 1, ...18$) and compute the median s_{med}. We use the sample median, as for large samples it converges to the mean of the Normal distribution, but is more robust to outliers than the sample mean. Then, we compute the Median of the Absolute Deviation (MAD):

$$\text{MAD} = median|s_h - s_{med}|. \tag{21}$$

We now get z-scores:

$$z_h = \frac{(s_h - s_{med})}{1.4826\,\text{MAD}}, \tag{22}$$

where we have the constant 1.4826 as, for large samples, $1.4826 * \text{MAD}$ converges to the standard deviation of the Normal distribution, but is more robust to outliers than the sample standard deviation (Kannan et al. (2015)). To see whether possibly significant differences in hospital performance correspond to large variation in FIs, we again conduct these modified z-tests, but now on the FI scores of the hospital sets with aberrant performance. The set of variables must include *LOS*, *SOFA* and $SOFA{*}LOS$ so that we are able to analyze their FIs and also compare conclusions with those from the RE models. We only use the RF and not the DT for LOHO, since the FIs change more smoothly in the RF than in the DT when data is being left out because of the combination of thousands of trees instead of only one. This is important since we use the assumption of normally distributed FI scores.

4.3.1 Easily interpretable: CART

Several algorithms to create a DT exist and we will use the Classification and Regression Tree (CART) technique, first introduced by Breiman et al. (1984). With CART, we will *grow* several DTs and by validation we will select the optimal one.

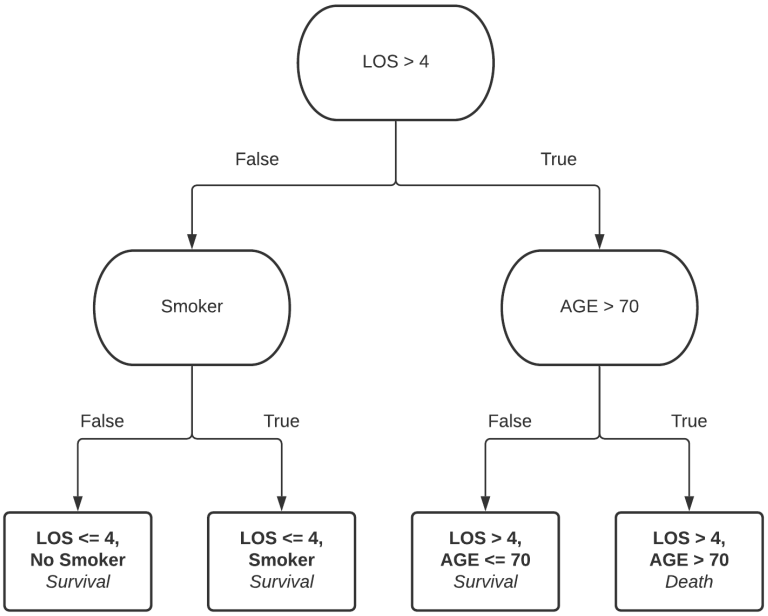


Figure 2: Example Decision Tree

As an illustration of how a DT works, we have Figure 2 where a resulting DT of depth 2 with 7 nodes is shown. For each observation, the model checks to which of the four resulting nodes (*leaves*) of the DT it belongs, and classifies the observations with the class label that belongs to that node. By default, the class label of a leaf node corresponds to the majority class of all training observations corresponding to the leaf. However, the DT can also provide a probability estimate for a class, which equals the fraction of observations of the corresponding class in the leaf node. Then, one could alter the default of classifying based on majority vote to classifying a death case if the corresponding probability estimate is larger than a certain probability *threshold*. The *splits* based on values of input variables and the final class labels corresponding to the leaf nodes make up the final DT. This final DT is found in three steps: growing DTs, pruning these DTs and selecting the optimal DT (Lewis (2000)).

Starting from the first node (*root*) of the DT, the CART algorithm selects for each node the split that is best in discriminating the observations in the training set. This

'best' split is the one with the highest *information gain*, corresponding to the highest decrease in the Gini Index. For binary classification, this Gini Index is computed as follows:

$$GI(\tau) = p_{\tau,0}(1 - p_{\tau,0}) + p_{\tau,1}(1 - p_{\tau,1}) = 1 - p_{\tau,0}^2 - p_{\tau,1}^2, \tag{23}$$

where $p_{\tau,r}$ corresponds to the fraction of training observations in node τ that belong to class $r \in 0, 1$ (Kiran & Serra (2017)).

So, if we have observations that fall within node τ, the split chosen for the observations in this node is the one that minimizes the average GI of the two resulting *child* nodes. In other words, we choose the split that maximizes information gain

$$\Delta GI(\tau) = GI(\tau) - \frac{n_A}{n} GI(\tau_A) - \frac{n_B}{n} GI(\tau_B), \tag{24}$$

where n is the number of observations in node τ and A, B correspond to its child nodes. To assess the relative importance of an explanatory variable j in the fitted DT compared to the importance of the other variables, we can sum all the differences in GI from splits in which j was used and divide it by the sum of $\Delta GI(\tau)$ over all τ. This gives FI_j as Feature Importance score (Menze et al. (2009)).

Candidate DTs are created by CART up till different maximum *depths*, the maximum number of splits between the root and a leaf, resulting in several DTs. We also grow a few DTs up till no further splits can be made, because in all leaf nodes either one observation is left or no binary rule for an input variable can be found to decrease the Gini Index. These identical *maximum* DTs are *overfitted* on the training data, probably yielding a lower prediction accuracy on unseen data than smaller, less complex DTs would. This is because after growing to a certain depth, the DTs have started to make splits that are so specific to the training data, that these splits result in making wrong classifications on test data. Therefore, we use a cost-complexity pruning approach on all but one of the maximum DTs with a unique complexity penalty δ per tree. So, with a DT that has been trained up till no further split could be found, we start to go back to the origin of the DT. Each node with two leaf nodes attached to it is a candidate to be *pruned*, whereby the DT gets smaller. This results in *information loss* and so an increase in the average GI, but also a loss in *overfitting* and so probably better performance when classifiying new unseen data. The candidate to prune is the one with the lowest increase in GI. The stopping

criterion for this pruning process is defined by δ, for which we will assign different values to get different unique trees. If the increase in GI caused by a potential pruning step exceeds this δ, that step is not considered. So, when all candidate nodes to prune have a corresponding increase in GI that exceeds the δ, we have found our pruned DT.

So, with DTs corresponding to different *depths* and values for δ, we can choose one DT to use for predicting on the test set. We choose the DT that performs best in the validation phase.

Unlike with Binary Logit, the classifications made by using the DT can be explained by pointing to the set of binary rules of the leaf node corresponding to the individual observation. These rules make more sense to a nurse or doctor, assuming she has no statistical knowledge of BL, than the effect of inputs on the odds ratio for a death in BL.

4.3.2 Random Forest

A single DT can classify the training data almost perfectly in an interpretable way, but tends to overfit on this training data and so has much lower performance on external data. To improve the accuracy on external data, we grow a Random Forest (RF) of DTs (Denisko & Hoffman (2018)). This RF is a collection of DTs that are trained on randomly bootstrapped subsets of the data and for which, when grown, at each node separately a random subset of input variables is available to base the split upon. The size of this random subset is equal to the square root of the total number of explanatory variables. In the end, an observation is classified according to the majority of its classifications in all separate DTs of the RF. We will tune the hyperparameter D, the total number of DTs in the random forest for optimal performance and prevention of overfitting (Breiman (2001)). Instead of majority vote classification, we will use the highest relative class frequency for classifying an observation, averaged over all trees. This relative class frequency per tree is the fraction of training samples with the same class as the observation, in the leaf node where the observation is placed. (Bostrom (2007)). Hereby, a classification made in one tree with a stronger distinction between the two classes has higher weight than in another since it corresponds to a more 'certain' classification. For example, a classification in a tree where the fraction of death cases is 0.95, has more weight in the model than where the fraction of survivals is 0.55. With the majority vote approach, the classifications of these two trees would have had equal weight in determining the final classification made

by the RF. Additionally, with this relative class frequency, one could alter the *threshold* above which an observation is classified as a death case from a 0.5 death frequency to something else.

A classification made by the RF cannot be traced back to a set of rules to show how this classification was made. Therefore, the RF is not interpretable in the sense of knowing how an individual choice was made by the model. However, we can use the same method as for the DT to quantify the importance of the explanatory variables used in the RF. For a specific explanatory variable j, its importance can be expressed by the summation over the information gain of all nodes in the RF where j is used to make a split (Menze et al. (2009)). We compute this information gain in terms of the Gini Index (see equation (24)) and therefore have for each j the Feature Importance

$$\mathrm{FI}_j = \sum_{d=1}^{D} \sum_{\tau} \Delta GI_j(\tau, d), \tag{25}$$

where d is a tree in the RF and τ a node in that tree. We use these FIs in the LOHO approach to find variance among hospitals in relations between mortality risk and the variables.

4.4 Validation phase

Before fitting the final model and predicting observations in the test set, we first choose which (scaled) variables to use, how far we grow and prune the DTs, how many trees are included in the RF, which threshold to use for getting a maximum F2-score and which variables to use in the model. To make all these decisions, we use cross-validation on the training set with AUPRC as performance metric, though conditioning on AUROC > 0.5, since setting any threshold between 0 and 1 would give an uninformative AUPRC of 0.509 ($= (1 + 0.017) * 0.5$) and AUROC of 0.5, for a model that only predicts survivals. Not setting AUROC > 0.5 would result in always picking this uninformative scenario, since an AUPRC of 0.509 is very high with $\approx 1.7\%$ of positive cases. In case of a draw between model specifications for a method in terms of AUPRC, we select out of these specifications the one with highest AUROC. If a draw still remains, we take the specification with highest F2-score out of the ones with highest AUPRC and AUROC. A potential final draw is decided by choosing the model specification that has the least

explanatory variables because lower dimensionality possibly causes less overfitting and thus higher out-of-sample performance (Plastria et al. (2008)). All three metrics have a flaw when being optimized for the DT since some specifications with only 2 leaf nodes give only 2 unique probability estimates for a death case. This also gives only 2 thresholds for in the curves and three ranges for sensible thresholds in finding the optimal F2-score. Hereby the metrics attain high levels that are often not representative for the prediction performance. Therefore, we also validate specifications for a DT conditioned on a minimum number of 8 leaf nodes.

For the numerical variables X, we calculate squared terms, cross-terms with LOS and cross-terms with LOS after squaring. After adding these variables, we come to 441 explanatory variables, including the dummies for missing values. This means we have more than one variable per 6 patients. It could be beneficial for out-of-sample performance and training speed to reduce this number and only use the most important variables. Therefore, we have 5 strategies for variable selection of which we will use the one that yields highest performance during validation for predicting on the test set. The first strategy is to use all available explanatory variables.

For the other four strategies, we first prune the set of variables by looking at the Variance-Inflation-Factors (VIF), which show the level of multicollinearity between the explanatory variables. We again consider equation (2). When fitting this equation for variable x by OLS, we get estimates $\hat{\beta}$ and thereby $\hat{x} = X\hat{\beta}$. Then, we calculate what fraction of the variance in x is explained by the other variables X:

$$R_x^2 = 1 - \frac{\sum_i^N (x_i - \hat{x}_i)^2}{\sum_i^N (x_i - \bar{x})^2}, \tag{26}$$

with N the number of observations and $\bar{x}$ the average value of variable x. The VIF of x then becomes

$$\text{VIF}_x = \frac{1}{1 - R_x^2} \tag{27}$$

from which it becomes clear that when more of the variance in an explanatory variable can be explained by the other variables, its VIF increases. A VIF of 1 implies independence with the other variables and values above 10 are considered high (Robinson & Schumacker (2009)). High multicollinearity can bias parameters in Binary Logistic Regression and for a RF it can cause that some sources of influence on the dependent variable are overly represented (for example if the condition of the lungs is depicted by several highly correlated

explanatory variables), giving this source a higher chance of being present in one of the trees than other sources, thereby making the forest less random and possibly more prone to overfitting. In three iterations we prune the set of variables, using the whole training sample together, so without cross-valiation. We first calculate the VIFs and exclude all variables for which it exceeds 100, then we do a new iteration with a threshold of 20 and a final one with a threshold of 15. Since there will be more pruning of the variable set in strategies 2 to 5, we keep more variables than is common by using 15 as a threshold instead of 10, otherwise we would throw away possibly powerful explanatory variables.

Our second strategy is *Scaling*, which means we robustly center and scale the explanatory variables that remained after pruning based on VIF. We do so by, for each variable, extracting the median and dividing by the difference between the 3rd and 1st quantile. This makes the coefficients for Binary Logit comparable, speeds up the numerical optimization in model fitting (Wright et al. (1999)) and is required before successfully applying Principal Components Analysis (PCA) (Wold et al. (1987).

Because of these advantages, we also use pruning the variable set based on VIF and scaling for the remaining three strategies. This brings us to the third strategy, which is applying *PCA*. This is in order to reduce the number of explanatory variables and prevent overfitting, while we keep the relevant information in the model by constructing the principal components (Plastria et al. (2008)). A predefined number of principal components is constructed, which can be seen as new variables that incorporate as much variation in the data while being independent from each other, thereby reducing the number of variables and summarizing the relevant information from the data in uncorrelated new variables. We lose interpretability of the relations between variables and the mortality risk, but possibly gain out-of-sample performance due to a reduction in overfitting (Wold et al. (1987)).

The fourth strategy is *MI* since we select a percentage of variables based on their level of Mutual Information (MI). This is a score that shows how strongly the mortality risk is related to an explanatory variable and can detect strong relationships even if not indicated by a high covariance (Kraskov et al. (2004)). It does so by measuring how much information two variables have about each other. The MI is calculated based on a k-nearest-neighbours approach with $k = 3$, as follows:

Algorithm 2 Mutual Information

Input: N observations with binary outcome variable y and explanatory variable x

1: **for** $i \in N$ **do**
2: N_{y_i}: set of observations for which $y_n = y_i$, $n \in N_{y_i}$
3: d: Distance in terms of the values of x between observation i and the 3rd-nearest-neighbour in N_{y_i}
4: m_i: Number of observations with distance to i smaller or equal to d, in terms of x.
5: $MI_i = \psi(N) - \psi\left(N_{y_i}\right) + \psi(3) - \psi\left(m_i\right)$
6: **end for**

7: $MI(x, y) = \text{average}(MI_i)$ =0

where $\psi(\cdot)$ is the digamma function (Ross (2014); Muqattash & Yahdi (2006)). To describe this intuitively, we consider variable x for which the average number of observations (m_i) that fall within the range of the distance (d) to the 3rd-nearest-neighbour is low. This indicates that the deaths and survivals are clustered around different values of x, which implies this variable is useful in distinguishing deaths from survivals. To get a more robust approximation of the importance of the different variables, we calculate the MI scores in a 5-fold cross-validation scheme in the training set, after which we average the MI scores. To make selections of variables with this strategy, we choose $p\%$ variables with highest MI scores, for $p \in [5, 10, 15, 25, 50, 75]$.

Unfortunately, for large amounts of explanatory variables, fitting a Random Effects model regularly runs into numerical problems. Therefore, we have final strategy *Selection* which means we get the MI scores for the explanatory variables and select the 25% of variables that have the highest average MI scores based on the cross-validation. Then we add, in collaboration with doctors from the ICU of the Amsterdam UMC, a few variables of high interest, replacing some variables that measure approximately the same condition or that have a relatively low *MI*. The variables in this *Selection* can be found in Appendix A.2.

So, we end with 5 strategies, which are all used in combination with the hyperparameters to create the DT, DTC, RF and CE model that performs best in the cross-validation phase. For RE models, only the *Selection* of variables is used, but the best combination of Random Effects is based on validation performance. Note that we never use squared terms for DTs or RFs, as these are created with binary splits and the squared terms have a one-to-one relationship with the original numerical variables, thereby adding no predictive

power to the DT or RF.

5 Numerical Results

First we will show the results from the validation phase. To find clues for important differences between hospitals on termination of treatment, we analyze for the RF the variation in performance and FIs when we use the LOHO approach, after which we investigate for the Binary Logit model the effect of treating variables as Random Effects that vary over hospitals. Then, we present the model performances on the test set to answer the primary research question on how well our models perform in mortality prediction. Finally, we report the most important coefficients and Feature Importances.

5.1 Validation performance

For getting the optimal sets of hyperparameters and variables, we carried out 5-fold cross-validation and looked at average AUPRC, while conditioning on AUROC > 0.5. We present the average AUPRC during cross-validation for the different models and variables sets in Table 1, where the values for the DT, DTC and RF correspond to their hyperparameters that yielded the highest score in combination with the corresponding variable set. The maximal AUPRC per model is depicted in bold in the table and we use the corresponding variable selection strategy for out-of-sample testing. Different sets of variables gave exactly the same DTs, which all happened with $\delta = 0.03$, as can be seen by looking at the maximal AUPRC scores of 0.347. So, they all yielded the highest AUPRC and the same AUROC and F2. Therefore, we chose from these DTs the smallest set of variables for out-of-sample testing, which was *MI* with 5% of the explanatory variables. The DT conditioned on at least 8 leaf nodes (DTC) gave 0.260 as highest AUPRC for *MI* with 75% of the explanatory variables. For the RF, we had the highest AUPRC of 0.163 for the *Selection* of variables with 10,000 trees. For the CE model, *PCA* with the highest 25% of the components yielded, with a score of 0.141, the highest AUPRC. When using *Selection* as variable selection strategy, highest AUPRC for DT (0.249) was yielded with $\delta = 0.03$ and for DTC (0.225) with $\delta = 0.0$.

Table 1: AUPRC from cross-validation

	DT	**DTC**	**RF**	**CE**
All variables	.259	.259	.152	.099
Scaling	.347	.250	.150	.092
PCA 5%	.157	.155	.050	.026
PCA 10%	.183	.183	.064	.047
PCA 15%	.176	.167	.058	.058
PCA 25%	.194	.194	.114	**.141**
PCA 50%	.204	.204	.116	.132
PCA 75%	.210	.210	.099	.130
MI 5%	**.347**	.244	.124	.102
MI 10%	.347	.238	.151	.137
MI 15%	.347	.255	.148	.121
MI 25%	.347	.245	.156	.109
MI 50%	.347	.247	.151	.102
MI 75%	.347	**.260**	.154	.097
Selection	.249	.225	**.163**	.115

For the 30 (2 times 15) candidate RE models, we calculated the AUPRC and reported the cross-validation results in Table 2, where the Random Effects are represented in the following way: PI stands for Patient specific varying Intercept and HI for Hospital specific varying Intercept. The same logic applies to *LOS* (HL), *SOFA* (HS) and *SOFA*∗*LOS* (HX). By combining these codes, you get all specifications. The model with hospital and patient specific varying intercept and hospital specific varying coefficient for the interaction between LOS and SOFA (called PIHIX) yielded 0.1236 (depicted in bold) as highest average AUPRC validation score. This is higher than for the CE and PI models, which both yielded AUPRC scores of 0.1153. So, we use this PIHIX model in predicting on the test set.

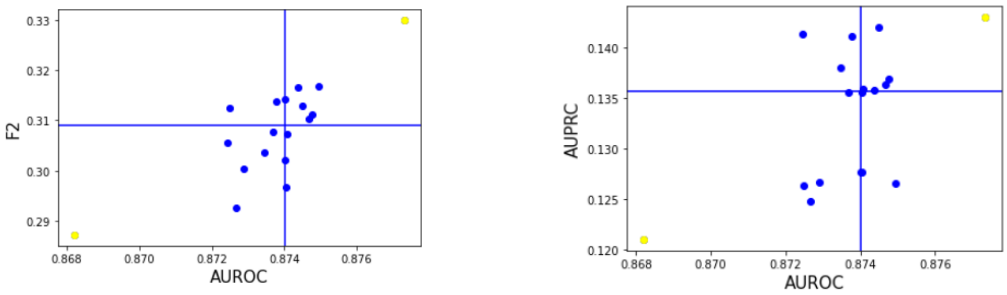
Table 2: AUPRC for various RE specifications

No Patient specific varying Intercept		**With Patient specific varying Intercept**	
Model specification	**AUPRC**	**Model specification**	**AUPRC**
CE	.1153	**PI**	.1153
HI	.1153	**PIHI**	.1153
HL	.1148	**PIHL**	.1147
HS	.1188	**PIHS**	.1188
HX	.1234	**PIHX**	.1233
HIL	.1148	**PIHIL**	.1146
HIS	.1188	**PIHIS**	.1188
HIX	.1233	**PIHIX**	**.1236**
HLS	.1172	**PIHLS**	.1173
HLX	.1229	**PIHLX**	.1227
HSX	.1230	**PIHSX**	.1232
HILS	.1173	**PIHILS**	.1170
HILX	.1233	**PIHILX**	.1228
HISX	.1234	**PIHISX**	.1235
HLSX	.1230	**PIHLSX**	.1229
HILSX	.1232	**PIHILSX**	.1227

5.2 Leave-One-Hospital-Out

We analyze the variation in hospital performance when we use the Leave-One-Hospital-Out approach with RF. Hospital names cannot be disclosed because of privacy reasons. As we are primarily special interested in the FIs of *LOS*, *SOFA* and their interaction, we must use a variable set that includes these variables. This is the case for the *Selection* of variables that was shown to yield highest RF validation performance in terms of AUPRC (with 10,000 trees), so we can carry out the LOHO approach with such a RF. The full tables with results can be found in Appendix B. We have results for 18 sets of hospitals and 3 performance metrics. We flag a set of hospitals if two of the metrics are significantly higher or lower than would be expected. With a significance level of $\alpha = 0.05$ we apply a multiple testing correction and thus divide α by 18 (hospital sets), by 3 (performance metrics) and multiply by 2 (since we need at least two significant metrics) . This results in corrected $\alpha^* \approx 0.0019$. The applied test is the modified z-tests. As can be seen in Table 3 and Figure 3a, this results in 2 outlying hospital sets, because of relative high and low AUROC and F2-scores. Their AUPRC is deviant, but not significantly. No other hospital sets corresponds to two or three significantly large or small performance metrics. Leaving hospital 'A' out leads to higher test set performance, indicating that using the hospital in

training has a negative influence on the out-of-sample performance. Leaving hospital 'G' out gives lower performance, which suggests that the corresponding hospital is important in training the optimal RF. It must be noted that for no single hospital the AUPRC was significantly aberrant, although this score was highest with 'A' and lowest with 'G' left out. The two hospitals with deviating performance scores correspond to the yellow dots in the AUPRC-AUROC plot in Figure 3b, in which we see that the relative low and high, yet insignificant, AUPRC scores are in line with the conclusions of significant low and high AUROC and F2 scores.



(a) F2-AUROC (b) AUPRC-AUROC

Figure 3: RF performance with different hospitals sets

Table 3: The only hospital sets with aberrant performances

	AUROC	**AUPRC**	**F2-score**
median	*0.874*	*0.136*	*0.309*
A	0.877^{+}	0.143	0.330^{+}
G	0.868^{-}	0.121	0.287^{-}

$^{+}$significantly larger with $p < 0.0019$
$^{-}$significantly smaller with $p < 0.0019$

To look for possible causes of the significant aberrant performance, we first turn to the distribution of deaths and survivals. From this analysis, as can be seen in Table 4, it stands out that hospitals 'A' and 'G' have a much higher death rate, while also delivering more patients than on average. This high number of dying patients could explain the lower performance by leaving 'G' out, since many death cases are thereby not used in model training, but it cannot explain why leaving 'A' out could lead to a higher performance. As long as no important information is missing in the explanatory variables, so that

Table 4: Distribution of patients

	patients	deaths	survivals	death rate
mean	*124*	*27*	*97*	*21*
A	189	64	125	34
G	156	46	110	28

the patients in this hospital are representative, there is statistical reason to believe this aberrant performance is due to a difference in relations between explanatory variables and mortality risk.

To see whether a statistical reason exists to further investigate potential differences in relations between explanatory variables and mortality risk, we take a look at the variation in FI among the three variables of special interest for the two hospital sets with aberrant performance. We apply the same z-tests as before, but now on the FIs. Since we consider two sets of hospitals and these three variables, we get corrected $\alpha^* = \frac{0.05}{2*3} \approx 0.008$. As can be seen in Table 5, all three FIs are low with hospital 'A' and high with 'G' left out. For the interaction term of SOFA with LOS, these differences are significant in both cases, for *SOFA* only with 'G' and the FIs for LOS are not significantly different. As shown in Table 6, we see that one other hospital set, 'F', yields lower p-values for differences in the FIs for the interaction term. However, 'F' has performance metrics very close to the medians because the z-tests on significant deviation from the median yield insignificant p-values of 0.56, 0.99 and 0.16 for the the tests for aberrant AUROC, AUPRC and F2, respectively. Therefore, we should not conclude that differences in FIs are necessarily related to aberrant performance. Nevertheless, the differences in performance and FI of $SOFA * LOS$ for 'A' and 'G' are significant and therefore these results give reason to further investigate the causes of these dissimilarities. If the differences in performance cannot be explained by another reason, like information in missing data, between-hospital practice variation could be a cause. We will compare these findings with those from the LRTs for RE models.

Table 5: FIs for hospital sets with aberrant performances

	LOS	**SOFA**	**SOFA x LOS**
median	*.0267*	*.0320*	*.0715*
A	.0258	.0314	.0685^{-}
G	.0275	.0340^{+}	.0750^{+}

$^{+}$significantly larger with $p < 0.008$
$^{-}$significantly smaller with $p < 0.008$

Table 6: P-values of z-tests corresponding to the 5 lowest p-values of SOFA∗LOS

	LOS	**SOFA**	**SOFA∗LOS**	**AUROC**	**AUPRC**	**F2**
F	.0551	.0000*	.0000*	.5583	.9874	.1590
G	.2762	.0012*	.0001*	.0000	.0137	.0000
A	.2591	.3265	.0008*	.0000	.2228	.0001
O	.2618	.0136	.0066*	.9406	.9746	.7493
B	.2137	.0363	.0399	.0096	.3446	.5072

* $p < 0.0019$

5.3 Binary Logit Random Effects formal tests

In order to find the correct BL model specification, we run tests on the 80% training set with the predefined *Selection* of variables. First, we check whether the intercept varies significantly between individuals when treating it as a Random Effect. The corresponding LRT yields a test statistic of 0.0 (with critical value 1.92), whereby the null hypothesis of no patient specific random intercept cannot be rejected. Hence, we focus on the CE model next.

Continuing with the CE model, so without Patient specific varying Intercepts, we test which hospital specific random effect out of 4 candidates yields the highest significant test statistic. In Table 7, we use the same codes for the different BL specifications as in Table 2. We see that adding *SOFA∗LOS* increases the log-likelihood from -693.69 to -690.48 (in bold), which corresponds to a significant test statistic of 6.41 versus the critical value of 3.16. No other HC model yielded a significant test statistic. Hence, we conclude that including a hospital specific intercept or slope for *LOS* or *SOFA* does not significantly improve the in-sample fit, although including the interaction term between *LOS* and *SOFA* does. Adding a second random effect does not increase the model fit at

all, since the log-likelihood remains -690.48. Therefore, we use this Binary Logit Hospital specific varying interaction term model (HX) for predicting on the test set.

We thus have found a statistically significant indication of a difference among hospitals in the relation of a variable with mortality risk, though only for one out of three variables and not so for a varying constant. This insight on differences in the effect of $SOFA{*}LOS$ between hospitals is in line with the results on the variance in FI for the RF. However, since only one out of the 4 HC models improved the fit, we take this result with caution. We will analyze out-of-sample performances for more insights to answer this second research question on variation in the relations between variables.

Table 7: LRT results

Model	**Log-Likelihood**	**Test statistic**	**Critical value**
CE	-693.69	-	-
PI	-693.69	0.0	1.92
HI	-693.69	0.0	3.16
HL	-692.67	2.03	3.16
HS	-692.79	1.79	3.16
HX	**-690.48**	6.41*	3.16
HIX	-690.48	0.0	7.05
HLX	-690.48	0.0	7.05
HSX	-690.48	0.0	7.05

* $p < 0.0125$

5.4 Performance on the test set

We use the models to get an indication of how well they perform in predicting mortality and which is the best. For Binary Logit, we had to make decisions on which models to use in prediction on the test set. We chose the CE (with tuned parameters) since we did not reject the null hypothesis of no variation in the intercept among patients, by using the LRT. From further LRTs, we have the HX model and from the cross-validation on AUPRC, the PIHIX model. We need to use the CE besides these HC models in order to see whether out-of-sample performance scores increase when using Random Effects.

In Table 8, we show performance when models are tuned on validation performance to use the best variable set and also the performance when the *Selection* of variables is used. The DT conditioned on a minimum of 8 leaf nodes with *Selection* of variables yields the

highest AUPRC (0.313), but this is slightly misleading since it corresponds to a tree that only yields probability estimates of 0 or 1. Hereby, only three evaluation points occur: the trivial case of predicting all observations as survivals that yields perfect precision (as no death classifications were wrong), only predicting a death for a probability estimate of 1 and predicting all observations as deaths. This causes the area under that curve to become much larger and not representative for its corresponding prediction performance. This is illustrated by comparing the AUPRC curves of the tuned CE and this DTC in Figure 4, where the area under the orange line from (0,1) to (0.57,0.05) causes the overestimation. The same problem of too few evaluation points applies to the unrestricted DT, but now because there are only two leaf nodes. The relatively very low AUROC (0.682, 0.709) and F2-scores (0.065, 0.186) however indicate the DTs perform inferior to the other models and so we ignore their too high AUPRC score.

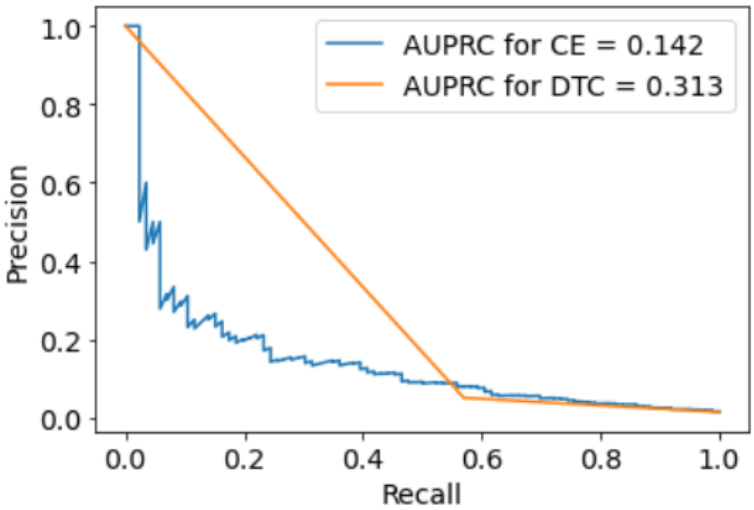


Figure 4: PR curves for CE and DTC

Then, for the remaining models, we look at the highest performance scores (in bold). We see that the CE model with 25% of its most important Principal Components performs best in our metric of highest interest with an AUPRC of 0.142. When the CE model is trained with the *Selection* of variables (in order to be comparable with the Random Effects models), the AUPRC radically drops to 0.123, though the AUROC and F2 increase.

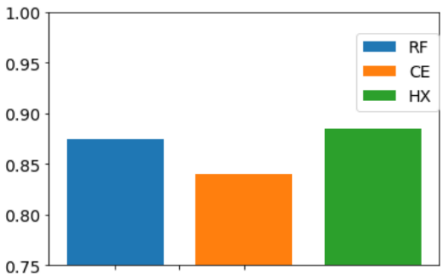
This CE with *Selection* is slightly outperformed by the HX and PIHIX only in terms of AUROC (0.885 versus 881). This AUROC score of HX (and PIHIX) is the highest of all models used, which also reflects the in-sample LRT result that the Binary Logit model is significantly improved by making the intercept vary between hospitals. Since the AUPRC

and F2 do not increase when Binary Logit is augmented to a model with Random Effects, we cannot say the potential differences in the relation between the mortality hazard and the interaction of LOS with SOFA result in aberrant performance.

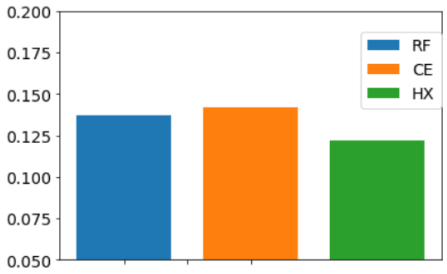
We see that the RF performs best in terms of F2 score. So, when looking at the RF, CE and HX or PIHIX, we see that each model performs best in one of the three metrics, as is displayed in Figure 5. This shows that their performance is similar, although the CE performs best for our main metric, the AUPRC.

Table 8: Performance on test set

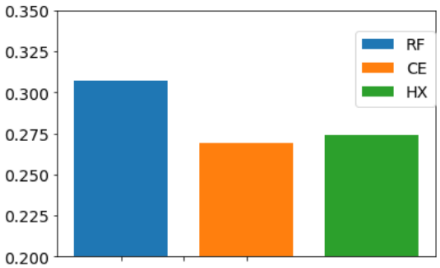
	Tuned hyperparameters			***Selection of variables***		
	AUROC	**AUPRC**	**F2-score**	**AUROC**	**AUPRC**	**F2-score**
DT	0.682	0.292	0.065	0.682	0.292	0.065
DTC	0.709	0.313	0.186	0.688	0.289	0.174
RF	0.874	0.137	**0.307**	0.874	0.137	**0.307**
CE	0.840	**0.142**	0.269	0.881	0.123	0.272
HX				**0.885**	0.122	0.274
PIHIX				**0.885**	0.122	0.272



(a) AUROC



(b) AUPRC



(c) F2

Figure 5: Performance on the test set

5.5 Coefficient estimates and Feature Importances

We want to provide opportunities for clinicians to compare their best practices in assessing the health status, mortality risk and recovery potential of patients with the variables in our models that have the highest predictive and explanatory power. Therefore, we look at all the absolute values of the coefficients for the Binary Logit models, trained on scaled data, and Feature Importance scores of the tree based models. In order to compare the importances of these variables between models, we use those models trained with the *Selection* of variables. The extensive tables with FIs, absolute coefficients and ranks for the most important variables can be found in Appendix B, whereas we only show the most important information in this section of the main paper.

The Binary Logit models used in predicting on the test set were all trained on robustly standardized data. Therefore, we can roughly compare the importances of different explanatory variables by looking at the absolute values of their coefficients (Menard (2011)). The regular DT only has one split, namely on the partial pressure of carbon dioxide in arterial blood (*pco2 arterial*). Since the Binary Logit models have the squared terms of age and BMI in contrast to the tree based models, the absolute values of their own and squared coefficients and FIs cannot be used for comparison. For each model, we ranked the importances of the explanatory variables. The binary variables (like dummies) yielded very low FIs for the tree based models, causing discrepancies with the three BL models. For these BL models, we have that the ranks of absolute values of coefficients are almost equal. Therefore, In Table 9 and Figure 7, we show the ten most important variables for the DTC, RF and HX, as it was the correct model after the LRTs. We highlighted the variables of RF to see the differences with the other 2 models. In the top ten most important variables for DTC, 6 correspond to the top ten from the RF. For the HX, only 4 variables from the RF appear. The HX has many dummy variables for missing values in the top 10, indicating that this information of missing values is prominently used in the model fit. This suggests vital information is missing in the data, possibly influencing the results in our research. The variable for which we found significant variation over hospitals in terms of FIs, the interaction term between SOFA and LOS, is the most important one for RF and also in the top 10 for DTC and HX. Additionally, the SOFA score itself is also in the top 10 of RF and HX, affirming the educated guess that SOFA was of high importance. On the other hand, the LOS is only found in the top ten for the DTC, but

Table 9: Variable importances

DTC	**RF**	**HX**
pCO2	**SOFA*LOS**	O2 Flow Dummy
SOFA*LOS	**pCO2**	**GCS motor**
Leukocytes	**FiO2**	Lung Compliance Dynamic Dummy
Blood pressure	**paO2/FiO2**	Peak Pressure Dummy
FiO2	**Driving Pressure**	**FiO2**
Driving Pressure	**Blood pressure**	**SOFA**
paO2/FiO2	**SOFA**	Minute Volume
LOS	**Pressure Above Peep**	Heart Rate
Phosphate	**GCS motor**	Pressure Above Peep Dummy
C reactive protein	**Leukocytes**	**SOFA*LOS**

drops to the 40th and 47th rank out of 48 variables for RF and HX. However, it must be noted that the interaction with SOFA can have a diminishing effect on the importance of LOS. So, we must say: given the interaction of SOFA with LOS, the LOS has a relative tiny importance for model fit.

To see which variables are the most important for both RF and HX, we now look at the other two variables present in both top ten lists. The first is the Glasgow Coma Scale (GCS) Motor, which measures the patient's level of motor responses to test the state of the nervous system (Teasdale & Jennett (1974). The second variable is the fraction of inspired oxygen (FiO2), which is increased by doctors when patients have an oxygen deficit in their blood. Another variable scores very high for both DTC and RF, namely the partial pressure of carbon dioxide in the arterial blood (pCO2).

After computing the conditional modes of the random interaction $SOFA{*}LOS$ for each hospital in the HX and PIHIX model, which are almost equal for the two models, we find that hospital 'I' has the highest conditional mode (-0.352) for the interaction term in absolute values. To put this in perspective, the estimated coefficient for the interaction is 0.424. We see the variation in hospital specific effects, which are the conditional modes plus the estimated coefficient, in Figure 6, with orange lines representing the estimated coefficient plus and minus the standard deviation of the Random Effect. The question remains whether this variation is caused by varying perceptions between doctors in different hospitals or by something else, like data collection differences between hospitals.

When looking at the hospitals that stood out with aberrant performance for the LOHO approach with RF, 'A' and 'G', we see that they yield the 4th (-0.218) and 5th (0.137)

highest absolute conditional mode out of 18 hospitals, implying a lot of variation for these hospitals. This consistency with the Random Effects suggests that the LOHO approach with FIs is capable of identifying hospitals with aberrant relations between variables and the outcome, but further research must show whether the approach is consistent in identifying these clusters (e.g. hospitals) that cause the significant variation in the importance of the variable of interest, such as the interaction of LOS with SOFA is in our case.

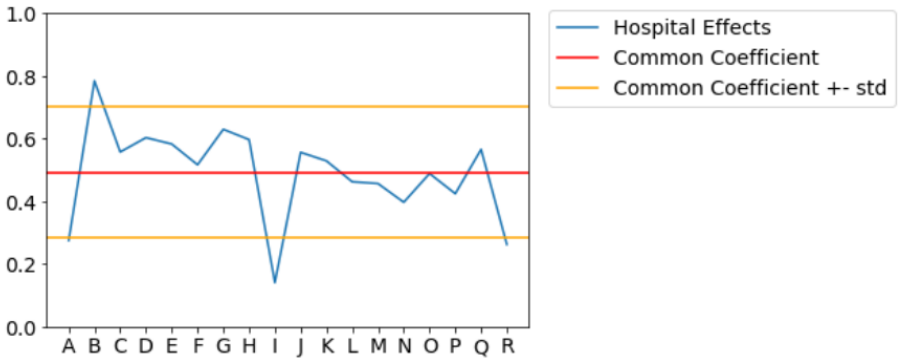
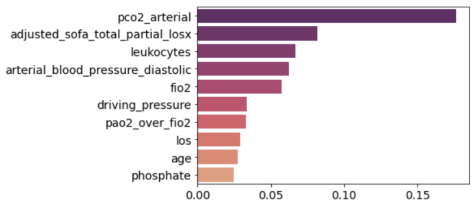
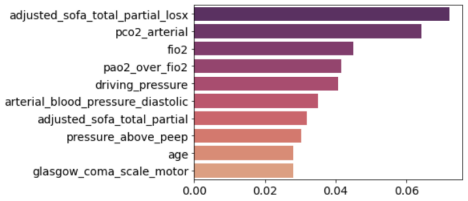


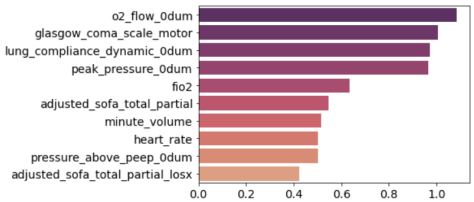
Figure 6: Hospital specific Random Effect for $SOFA * LOS$



(a) Feature Importances DTC



(b) Feature Importances RF



(c) Absolute coefficients HX

Figure 7: Importances of variables

6 Conclusion

The Binary Logit models and Random Forest performed well, we found some statistical leads for further investigation of between-hospital differences in the relations of explana-

tory variables with the 24-hour ahead mortality hazard and we indicated the most important variables in fitting the models. In this concluding section, we first extensively answer the three research questions stated in the introduction. Secondly, we discuss the limitations to this research and finally, we give suggestions for further research. Through these steps, we touch upon the question of how robust the results are.

6.1 Answers to research questions

The first research question, 'How can statistical modelling and Machine Learning be used to make accurate and interpretable predictions on the dynamic mortality hazard of Covid-19 patients at the Intensive Care Unit?', has two components: accuracy and interpretability. The model most easy to interpret because of the binary decision rules, the DT, performs inferior to the Binary Logit models and RF in terms of predicting deaths and survivals in the test set. We have seen that in terms of our primary performance metric, AUPRC, the CE model performs best (0.142) when using the first 25% of the components resulting from PCA. This 0.142 score indicates the model performs about eight times better than a random guess since the fraction of deaths among all observations (0.017) is about eight times smaller. Though the CE performs best in AUPRC, the RF does not perform much less (0.137) and it even surpasses the CE model in terms of AUROC (0.878 versus 0.840) and F2-score (0.307 versus 0.269). Although the RF cannot be used to explain individual classifications, it is interpretable in the sense that doctors can see what the most important variables are for making classifications, through the FI scores. As the CE is also interpretable at the level of explaining how, for an individual observation, input variables lead to the made classification, there seems no specific reason to use the less interpretable RF in practice. We extended the CE model by adding the interaction of LOS with SOFA as Random Effect that varies per hospital. Though this gave a significant increase in terms of in-sample fit, there was no increase in terms of AUPRC or F2-score and only a small increase in AUROC for out-of-sample performance. So, we see that the AUPRC, AUROC and F2-score were highest for three different models. Therefore, a conclusion on which model is clearly the best performing one, cannot be robustly made. The Random Effect was made interpretable by calculating the conditional modes of the Random Effect for each hospital, whereby the variance in relation between this interaction and the mortality hazard can be measured and the individual classification be understood.

So, the interpretable BL models perform very well and similar to the uninterpretable RF, but the DT, based on very easy to interpret binary decision rules, clearly underperforms.

We now focus on the second research question: 'What is the between-hospital difference in model performance and estimated relations between mortality hazard and key variables?' We looked for statistical evidence to answer the question by analyzing differences between hospitals in the coefficients and importance of the LOS, SOFA score and their interaction. We found a significant increase in in-sample fit when making this interaction term a Random Effect that differs per hospital. If hospitals differed in deciding to stop treatment, including variation on the relation between the mortality hazard, due to treatment termination, and the interaction term of the stay duration (LOS) at the ICU with the approximation of health status (SOFA) would have to lead to higher performance. This is not the case for out-of-sample performance in terms of AUPRC. Therefore, we conclude that no solid statistical argument can be made, based on solely the Random Effects models performances, for the existence of differences in how staff of different hospitals decide to stop treatment. The LOHO approach also indicated aberrant performance for two hospitals, which was found by ignoring these hospital in the training phase. These two hospitals also corresponded to significant differences in the Feature Importance of the interaction of LOS and SOFA among hospitals. This could suggest that between-hospital differences exist in the relation between mortality risk and the interaction term, but some other objections exist. First of all, a significant aberrant Feature Importance does not imply an aberrant performance, since the hospital with the lowest p-value in the test of aberrant FIs, had an average performance. Secondly, the aberrant FI could be caused by differences in the data between hospitals. For example, some hospitals have much more death cases than other hospitals. Though we found significant aberrant FIs and that in-sample performance increases due to using Random Effects, this is not sufficient to draw a solid conclusion, especially since the out-of-sample performances of the RE models do not reflect this hypothesis of varying relations among hospitals between variables and mortality hazard. Therefore, more research on these varying relations should be done to verify the suggested differences among hospitals in relation of the interaction term with mortality hazard.

Lastly, we look at the importances of the variables in model fitting so that clinicians can compare their ways to assess a patient's mortality risk with that of the models. This is

done by answering the third research question: 'Which variables have the most predictive and explanatory power in our statistical and Machine Learning models?' We saw these variables were the SOFA score, its interaction term with the LOS, the fraction of inspired oxygen, the Glasgow Coma Scale Motor score and the partial pressure of carbon dioxide, though this last one did not end in the top ten of most important variables for the BL models. The dummy variables that indicate whether a value is missing were important in fitting the BL models. This shows that important information was missing in the data, which confirms our choice of assuming they were not missing at random and also not throwing away these data points. On the other hand, it also implies all results become less reliable because of missing variables for a fraction of the observations. Perhaps the results would have been different if no data was missing. In particular, the between-hospital variations found by LOHO and RE models could (partially) be due to missing data.

6.2 Limitations

The research has several limitations that could have a negative effect on the prediction performance, ability to answer the research questions and correctly comparing the models. The first is using two different optimizers for the CE and Random Effects models, namely Newton-CG and Nelder-Mead. This was caused by the fact of having to use two different Python packages (Scikit-learn and Pymer) for model fitting. Hereby, it might be difficult to perfectly compare their performances, especially since the optimizers have different stopping criteria and one iteration in the Newton-CG is something different than an iteration in the Nelder-Mead algorithm. The CE model stopped fitting before reaching the 10,000 iterations, but the process of fitting the RE models had to be stopped at the chosen maximum of 10,000 iterations for Nelder-Mead. It was unclear how close both optimizers were to reaching the maximum likelihood. Being more precise in setting the stopping criteria could create more equality between the CE and RE models. One example is to only stop when the increase in likelihood falls below a certain threshold. Since the optimization for Newton-CG stopped when the size of the gradient fell below the threshold, an appropriate threshold for the relative change in parameter values for the Nelder-Mead could be chosen to create equality between the two optimization methods. Another option is to give the model a fixed amount of time to fit.

For these RE models, we assumed that the random effects followed a normal distribution. However, the student's t or gamma distribution could have been more appropriate, though no specific reason exists to doubt the validity of the normal distribution.

Next, we chose to deal with the problem of imbalanced data by sampling for each individual the final day at the hospital and another random day. This has some potential negative effects on the model fit, like that the health status of a patient at the last day in the ICU is not representative for the other days. Also, with about 13 observations per patient in the original data, we throw away many observations in this scenario, namely about 85% ($= 1 - (2/13)$). Other sampling scenarios could be tried, like sampling a few additional other random days or using all days and then oversampling from the observations that correspond to deaths. Using cross-validation with the training data could lead to finding the optimal sampling scenario. Also, using a different sampling scenario should be tried to check the robustness of the results and conclusions in this research.

About 22% of deaths on 2245 patients with daily observations provides abundant data for model fitting and testing, whereby the conclusions of high AUPRC performance for BL and RF are robust. However, when investigating differences per hospital, only about 27 deaths and 96 survivals are available per hospital, which is sparse. On top of that, the number of observations and distribution of deaths differs per hospital. Hereby it could be harder to find a significant difference with the LOHO in performance or FI for a hospital with a low number of patients or low fraction of deaths. Since if such a difference for a 'small' hospital exists, the test might lack the power to detect it. Therefore, a fine-tuned sampling scheme should be used that balances out the observations and deaths per hospital, such that hospital variations can be evaluated in a better way. However, this would imply oversampling, undersampling or deleting hospitals, which all have drawbacks on their own, as they imply throwing away data or could cause biases.

Another problem remains with the data since many variables come from manual registrations, which are prone to human errors. Although a team of doctors and programmers did clean the dataset and we used boundaries to truncate strange values in this research, more extensive cleaning of the dataset could make the results more accurate. Additionally, data was missing, possibly influencing the conclusions in this research. For example, missing data could have caused the significant variation in performance in the LOHO

approach.

Finally, the DT was hard to tune for a couple of reasons. First of all, the hyperparameters for maximum depth or the cost complexity pruning constant δ led to small trees of one split or huge trees. We conditioned on a minimum of leaves, but this could be done more sophisticated to get better DTs, for example by taking a set of conditions and using cross-validation to pick the best. Additionally, the AUPRC was not very appropriate for the DT, since it gave too high scores for trees with only a few unique probability estimates. The calculation of the AUPRC could be made robust to the case of only a few unique probability estimates, for example by considering the Precision to be zero instead of one when no positive (death) classification is made. The AUROC and F2 are also not optimal, but are possibly better for getting the best performing DT. Perhaps a metric can be used that was not presented in this research.

6.3 Further research

In this study, we considered making predictions per day. The data allows to increase this frequency in further research, which would have to involve modelling the variability in predictions caused by the time of day as additional factor. Based on our conclusions and limitations, we suggest a few other topics for further research. The first is to improve the way of tuning the DT, for example by validating on conditions for the DT specification or using a different performance metric in model validation. The second is to see whether a RF, CE or RE model performs best when (almost) no data is missing. On top of that, other Machine Learning methods, like SVM and Neural Networks, could be used to see whether they outperform the models used in this research.

Modelling the Random Effects to follow a student's t or gamma distribution could be tried to improve the RE model performance and more accurately identify between-hospital variation in coefficient estimates. Also, using different variables, like the FiO2, as candidate Random Effects could give new insights in the variation of relations between mortality hazard and explanatory variables.

As another method to find these variations, the LOHO approach should be further developed and refined on new data. Additionally, repeating the research after balancing out the number of patients and death cases among hospitals could alter the conclusions from the LOHO approach, leading to new insights. To draw fairer conclusions on the dif-

ferences in performance between CE and RE models, a new research with better stopping criteria must be conducted or the same optimization algorithm should be used for all BL models. Connected to that, investigating the influence of differences in stopping criteria on the performances of CE and RE models might be useful to show which criteria are the most appropriate and how robust our results are with the current stopping criteria.

In any case, building dynamic mortality hazard models for ICUs from several different countries would be interesting, as it could shed light on possible differences between countries on how they deal with patients at the ICU. Such a research could also include the LOHO approach to test its value and be used for trying out other methods to investigate heterogeneity in relations between the outcome and explanatory variables.

References

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 acm international conference on bioinformatics, computational biology, and health informatics* (pp. 559–560).

Antonelli, M., Moreno, R., Vincent, J. L., Sprung, C., Mendoca, A., Passariello, M., ... Osborn, J. (1999). Application of sofa score to trauma patients. *Intensive care medicine*, *25*(4), 389–394.

Armstrong, R. A. (2014). When to use the b onferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508.

Bates, D. (2014). Computational methods for mixed models. *LME4: Mixed-effects modeling with R*, 99–118.

Bostrom, H. (2007). Estimating class probabilities in random forests. In *Sixth international conference on machine learning and applications (icmla 2007)* (pp. 211–216).

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: point estimates and confidence intervals. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 451–466).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Taylor and Francis.

CovidPredict Database. (2020). `https://covidpredict.org/`. (Accessed: 2021-04-24)

Croissant, Y., & Millo, G. (2019). *Panel data econometrics with r*. Wiley Online Library.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on machine learning* (pp. 233–240).

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690–1692.

Goss, E. P., & Ramchandani, H. (1998). Survival prediction in the intensive care unit: a comparison of neural networks and binary logit regression. *Socio-Economic Planning Sciences*, *32*(3), 189–198.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Grasselli, G., Greco, M., Zanella, A., Albano, G., Antonelli, M., Bellani, G., ... others (2020). Risk factors associated with mortality among patients with covid-19 in intensive care units in lombardy, italy. *JAMA internal medicine*, *180*(10), 1345–1355.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251–1271.

Hippisley-Cox, J., Pringle, M., Cater, R., Wynn, A., Hammersley, V., Coupland, C., ... Johnson, C. (2003). The electronic patient record in primary care—regression or progression? a cross sectional study. *Bmj*, *326*(7404), 1439–1443.

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *Test*, *16*(1), 1–22.

Hsiao, C., & Pesaran, M. H. (2008). Random coefficient models. In *The econometrics of panel data* (pp. 185–213). Springer.

Hsieh, M. H., Hsieh, M. J., Chen, C.-M., Hsieh, C.-C., Chao, C.-M., & Lai, C.-C. (2018). Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. *Scientific reports*, *8*(1), 1–7.

Jolly, E. (2018). Pymer4: connecting r and python for linear mixed modeling. *Journal of Open Source Software*, *3*(31), 862.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402.

Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231–238.

Kennedy, C. E., Aoki, N., Mariscalco, M., & Turley, J. P. (2015). Using time series analysis to predict cardiac arrest in a pediatric intensive care unit. *Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, *16*(9), e332.

Kim, Y., Choi, Y.-K., & Emery, S. (2013). Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. *The American Statistician*, *67*(3), 171–182.

Kiran, B. R., & Serra, J. (2017). Cost-complexity pruning of random forests. In *International symposium on mathematical morphology and its applications to signal and image processing* (pp. 222–232).

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.

Knaus, W. A., Zimmerman, J. E., Wagner, D. P., Draper, E. A., & Lawrence, D. E. (1981). Apache-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, *9*(8), 591–597.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.

Kumari, R., & Srivastava, S. K. (2017). Machine learning: A review on binary classification. *International Journal of Computer Applications*, *160*(7).

Lee, K. J., & Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in medicine*, *27*(3), 418–434.

Le Gall, J.-R., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, *270*(24), 2957–2963.

Lemeshow, S., Teres, D., Klar, J., Avrunin, J. S., Gehlbach, S. H., & Rapoport, J. (1993). Mortality probability models (mpm ii) based on an international cohort of intensive care unit patients. *Jama*, *270*(20), 2478–2486.

Lewis, R. J. (2000). An introduction to classification and regression tree (cart) analysis. In *Annual meeting of the society for academic emergency medicine in san francisco, california* (Vol. 14).

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics & Data Analysis*, *17*(1), 1–15.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Menard, S. (2011). Standards for standardized logistic regression coefficients. *Social Forces*, *89*(4), 1409–1428.

Meng, L., Qiu, H., Wan, L., Ai, Y., Xue, Z., Guo, Q., ... others (2020). Intubation and ventilation amid the covid-19 outbreak: Wuhan's experience. *Anesthesiology*, *132*(6), 1317–1332.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*(1), 1–16.

Muqattash, I., & Yahdi, M. (2006). Infinite family of approximations of the digamma function. *Mathematical and computer modelling*, *43*(11-12), 1329–1336.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20A*, 175-240.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Plastria, F., De Bruyne, S., & Carrizosa, E. (2008). Dimensionality reduction for clas-

sification. In *International conference on advanced data mining and applications* (pp. 411–418).

Rieg, S., von Cube, M., Kalbhenn, J., Utzolino, S., Pernice, K., Bechet, L., ... others (2020). Covid-19 in-hospital mortality and mode of death in a dynamic and non-restricted tertiary care model in germany. *PloS one*, *15*(11), e0242127.

Robinson, C., & Schumacker, R. E. (2009). Interaction effects: centering, variance inflation factor, and interpretation issues. *Multiple linear regression viewpoints*, *35*(1), 6–11.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PloS one*, *9*(2), e87357.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Smit, J. (2021). Prediction models for adverse outcomes in covid-19 patients.

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, *84*(4), 465–470.

Song, C.-Y., Xu, J., He, J.-Q., & Lu, Y.-Q. (2020). Covid-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. *MedRxiv*.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 1171–1177.

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, *304*(7872), 81–84.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, *45*(1), 1–67.

van der Lubbe, I., Van der Loo, R., & Burgerhout, J. (1997). Flexible electronic patient record: first results from a dutch hospital. In *Medical informatics europe'97* (pp. 246–251). IOS Press.

Vink, G., Frank, L. E., Pannekoek, J., & Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61–90.

Wallace, B. C., & Dahabreh, I. J. (2014). Improving class probability estimates for imbalanced data. *Knowledge and information systems*, *41*(1), 33–52.

Weiss, G. M. (2013). Foundations of imbalanced learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 13–41.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Wright, S., Nocedal, J., et al. (1999). Numerical optimization. *Springer Science*, *35*(67-68), 26-27.

Zhang, X., Qian, B., Li, Y., Yin, C., Wang, X., & Zheng, Q. (2019). Knowrisk: an interpretable knowledge-guided model for disease risk prediction. In *2019 ieee international conference on data mining (icdm)* (pp. 1492–1497).

Appendices

A Variables

A.1 All explanatory variables

['los', 'female', 'age', 'bmi', 'weight', 'height', 'acute_kidney_injury', 'cardiovascular_insufficiency',
'chronic_dialysis', 'chronic_renal_insufficiency', 'cirrhosis',
'copd', 'diabetes', 'hematologic_malignancy', 'immunodeficiency',
'neoplasm', 'respiratory_insufficiency', 'acute_kidney_injury_nan',
'cardiovascular_insufficiency_nan', 'chronic_dialysis_nan',
'chronic_renal_insufficiency_nan', 'cirrhosis_nan', 'copd_nan',
'diabetes_nan', 'hematologic_malignancy_nan',
'immunodeficiency_nan', 'neoplasm_nan',
'respiratory_insufficiency_nan', 'apache_age_score',
'apache_blood_pressure_score', 'apache_creatinine_score',
'apache_heart_rate_score', 'apache_hematocrit_score',
'apache_leukocytes_score', 'apache_operative_score',
'apache_oxygenation_score', 'apache_ph_score',
'apache_potassium_score', 'apache_respiration_rate_score',
'apache_sodium_score', 'apache_temperature_score',
'apache_partial', 'arterial_blood_pressure_diastolic',
'arterial_blood_pressure_mean', 'arterial_blood_pressure_systolic',
'heart_rate', 'activated_partial_thromboplastin_time',
'alanine_transaminase', 'albumin', 'alkaline_phosphatase',
'aspartate_transaminase', 'base_excess', 'bicarbonate_unspecified',
'bilirubin_total', 'c_reactive_protein', 'calcium',
'calcium_ionised', 'chloride', 'creatine_kinase', 'creatinine',
'd_dimer', 'eosinophils', 'eosinophils_percentage',
'estimated_glomerular_filtration_rate',
'gamma_glutamyl_transferase', 'glucose', 'hematocrit',
'hemoglobin', 'lactate_dehydrogenase', 'leukocytes', 'lymphocytes',
'lymphocytes_percentage', 'magnesium', 'monocytes',

'monocytes_percentage', 'pco2_arterial', 'ph_arterial',
'phosphate', 'po2_arterial', 'potassium', 'so2_arterial', 'sodium',
'thrombocytes', 'ureum', 'ureum_over_creatinine', 'intubated',
'intubated_cum', 'driving_pressure', 'end_tidal_co2', 'fio2',
'lung_compliance_dynamic', 'lung_compliance_static',
'mean_pressure', 'mechanical_power', 'mechanical_power_per_kg',
'minute_volume', 'minute_volume_derived', 'o2_flow',
'o2_saturation', 'pao2_over_fio2', 'peak_pressure', 'peep',
'pressure_above_peep', 'rapid_shallow_breathing_index',
'respiratory_rate_measured',
'respiratory_rate_measured_ventilator', 'respiratory_rate_set',
'tidal_volume', 'tidal_volume_per_kg', 'ventilatory_ratio',
'glasgow_coma_scale_eye', 'glasgow_coma_scale_motor',
'glasgow_coma_scale_total', 'glasgow_coma_scale_verbal',
'temperature', 'adjusted_sofa_cardiovascular',
'adjusted_sofa_coagulation', 'adjusted_sofa_liver',
'adjusted_sofa_nervous', 'adjusted_sofa_renal',
'adjusted_sofa_respiratory', 'adjusted_sofa_total',
'adjusted_sofa_total_partial',
'arterial_blood_pressure_diastolic_0dum',
'arterial_blood_pressure_mean_0dum',
'arterial_blood_pressure_systolic_0dum', 'heart_rate_0dum',
'activated_partial_thromboplastin_time_0dum',
'alanine_transaminase_0dum', 'albumin_0dum',
'alkaline_phosphatase_0dum', 'aspartate_transaminase_0dum',
'base_excess_0dum', 'bicarbonate_unspecified_0dum',
'bilirubin_total_0dum', 'c_reactive_protein_0dum', 'calcium_0dum',
'calcium_ionised_0dum', 'chloride_0dum', 'creatine_kinase_0dum',
'creatinine_0dum', 'd_dimer_0dum', 'eosinophils_0dum',
'eosinophils_percentage_0dum',
'estimated_glomerular_filtration_rate_0dum',
'gamma_glutamyl_transferase_0dum', 'glucose_0dum',

'hematocrit_0dum', 'hemoglobin_0dum', 'lactate_dehydrogenase_0dum',
'leukocytes_0dum', 'lymphocytes_0dum',
'lymphocytes_percentage_0dum', 'magnesium_0dum', 'monocytes_0dum',
'monocytes_percentage_0dum', 'pco2_arterial_0dum',
'ph_arterial_0dum', 'phosphate_0dum', 'po2_arterial_0dum',
'potassium_0dum', 'so2_arterial_0dum', 'sodium_0dum',
'thrombocytes_0dum', 'ureum_0dum', 'ureum_over_creatinine_0dum',
'driving_pressure_0dum', 'end_tidal_co2_0dum', 'fio2_0dum',
'lung_compliance_dynamic_0dum', 'lung_compliance_static_0dum',
'mean_pressure_0dum', 'mechanical_power_0dum',
'mechanical_power_per_kg_0dum', 'minute_volume_0dum',
'minute_volume_derived_0dum', 'o2_flow_0dum', 'o2_saturation_0dum',
'pao2_over_fio2_0dum', 'peak_pressure_0dum', 'peep_0dum',
'pressure_above_peep_0dum', 'rapid_shallow_breathing_index_0dum',
'respiratory_rate_measured_0dum',
'respiratory_rate_measured_ventilator_0dum',
'respiratory_rate_set_0dum', 'tidal_volume_0dum',
'tidal_volume_per_kg_0dum', 'ventilatory_ratio_0dum',
'glasgow_coma_scale_eye_0dum', 'glasgow_coma_scale_motor_0dum',
'glasgow_coma_scale_total_0dum', 'glasgow_coma_scale_verbal_0dum',
'temperature_0dum', 'adjusted_sofa_cardiovascular_0dum',
'adjusted_sofa_coagulation_0dum', 'adjusted_sofa_liver_0dum',
'adjusted_sofa_nervous_0dum', 'adjusted_sofa_renal_0dum',
'adjusted_sofa_respiratory_0dum', 'adjusted_sofa_total_0dum',
'adjusted_sofa_total_partial_0dum', 'med_antibiotics',
'med_antiinfectives_and_antiseptics_for_local_oral_treatment',
'med_carbohydrates', 'med_ceftriaxone', 'med_ciprofloxacin',
'med_clonidine', 'med_dalteparin', 'med_dexamethasone',
'med_electrolytes', 'med_furosemide', 'med_general_nutrients',
'med_haloperidol', 'med_insulin_aspart', 'med_macrogol',
'med_macrogol_combinations', 'med_magnesium_sulfate',
'med_metoclopramide', 'med_metoprolol', 'med_midazolam',

'med_morphine', 'med_nadroparin', 'med_norepinephrine',
'med_oxazepam', 'med_pantoprazole', 'med_paracetamol',
'med_potassium_chloride', 'med_propofol', 'med_remifentanil',
'med_rocuronium_bromide', 'med_salbutamol_and_ipratropium_bromide',
'med_sodium_chloride', 'med_sodium_phosphate', 'med_sufentanil',
'med_antibiotics_0dum',
'med_antiinfectives_and_antiseptics_for_local_oral_treatment_0dum',
'med_carbohydrates_0dum', 'med_ceftriaxone_0dum',
'med_ciprofloxacin_0dum', 'med_clonidine_0dum',
'med_dalteparin_0dum', 'med_dexamethasone_0dum',
'med_electrolytes_0dum', 'med_furosemide_0dum',
'med_general_nutrients_0dum', 'med_haloperidol_0dum',
'med_insulin_aspart_0dum', 'med_macrogol_0dum',
'med_macrogol_combinations_0dum', 'med_magnesium_sulfate_0dum',
'med_metoclopramide_0dum', 'med_metoprolol_0dum',
'med_midazolam_0dum', 'med_morphine_0dum', 'med_nadroparin_0dum',
'med_norepinephrine_0dum', 'med_oxazepam_0dum',
'med_pantoprazole_0dum', 'med_paracetamol_0dum',
'med_potassium_chloride_0dum', 'med_propofol_0dum',
'med_remifentanil_0dum', 'med_rocuronium_bromide_0dum',
'med_salbutamol_and_ipratropium_bromide_0dum',
'med_sodium_chloride_0dum', 'med_sodium_phosphate_0dum',
'med_sufentanil_0dum', 'prone_position_only',
'supine_position_only', 'both_positions_today', 'no_info_position',
'no_hymo_info', 'no_labo_info', 'no_resp_info',
'no_vent_info_though_intubated', 'no_neur_info', 'no_temp_info',
'no_med_info', 'no_sofa_info', 'los_p2', 'age_p2', 'bmi_p2',
'weight_p2', 'height_p2', 'apache_partial_p2',
'arterial_blood_pressure_diastolic_p2',
'arterial_blood_pressure_mean_p2',
'arterial_blood_pressure_systolic_p2', 'heart_rate_p2',
'activated_partial_thromboplastin_time_p2',

'alanine_transaminase_p2', 'albumin_p2', 'alkaline_phosphatase_p2',
'aspartate_transaminase_p2', 'base_excess_p2',
'bicarbonate_unspecified_p2', 'bilirubin_total_p2',
'c_reactive_protein_p2', 'calcium_p2', 'calcium_ionised_p2',
'chloride_p2', 'creatine_kinase_p2', 'creatinine_p2', 'd_dimer_p2',
'eosinophils_p2', 'eosinophils_percentage_p2',
'estimated_glomerular_filtration_rate_p2',
'gamma_glutamyl_transferase_p2', 'glucose_p2', 'hematocrit_p2',
'hemoglobin_p2', 'lactate_dehydrogenase_p2', 'leukocytes_p2',
'lymphocytes_p2', 'lymphocytes_percentage_p2', 'magnesium_p2',
'monocytes_p2', 'monocytes_percentage_p2', 'pco2_arterial_p2',
'ph_arterial_p2', 'phosphate_p2', 'po2_arterial_p2',
'potassium_p2', 'so2_arterial_p2', 'sodium_p2', 'thrombocytes_p2',
'ureum_p2', 'ureum_over_creatinine_p2', 'intubated_cum_p2',
'driving_pressure_p2', 'end_tidal_co2_p2', 'fio2_p2',
'lung_compliance_dynamic_p2', 'lung_compliance_static_p2',
'mean_pressure_p2', 'mechanical_power_p2',
'mechanical_power_per_kg_p2', 'minute_volume_p2',
'minute_volume_derived_p2', 'o2_flow_p2', 'o2_saturation_p2',
'pao2_over_fio2_p2', 'peak_pressure_p2', 'peep_p2',
'pressure_above_peep_p2', 'rapid_shallow_breathing_index_p2',
'respiratory_rate_measured_p2',
'respiratory_rate_measured_ventilator_p2',
'respiratory_rate_set_p2', 'tidal_volume_p2',
'tidal_volume_per_kg_p2', 'ventilatory_ratio_p2',
'glasgow_coma_scale_eye_p2', 'glasgow_coma_scale_motor_p2',
'glasgow_coma_scale_total_p2', 'glasgow_coma_scale_verbal_p2',
'temperature_p2', 'adjusted_sofa_total_partial_p2',
'med_antibiotics_p2',
'med_antiinfectives_and_antiseptics_for_local_oral_treatment_p2',
'med_carbohydrates_p2', 'med_ceftriaxone_p2',
'med_ciprofloxacin_p2', 'med_clonidine_p2', 'med_dalteparin_p2',

'med_dexamethasone_p2', 'med_electrolytes_p2', 'med_furosemide_p2',
'med_general_nutrients_p2', 'med_haloperidol_p2',
'med_insulin_aspart_p2', 'med_macrogol_p2',
'med_macrogol_combinations_p2', 'med_magnesium_sulfate_p2',
'med_metoclopramide_p2', 'med_metoprolol_p2', 'med_midazolam_p2',
'med_morphine_p2', 'med_nadroparin_p2', 'med_norepinephrine_p2',
'med_oxazepam_p2', 'med_pantoprazole_p2', 'med_paracetamol_p2',
'med_potassium_chloride_p2', 'med_propofol_p2',
'med_remifentanil_p2', 'med_rocuronium_bromide_p2',
'med_salbutamol_and_ipratropium_bromide_p2',
'med_sodium_chloride_p2', 'med_sodium_phosphate_p2',
'med_sufentanil_p2', 'age_losx', 'age_p2_losx', 'female_losx',
'apache_partial_losx', 'apache_partial_p2_losx',
'arterial_blood_pressure_diastolic_losx',
'arterial_blood_pressure_diastolic_p2_losx',
'arterial_blood_pressure_mean_losx',
'arterial_blood_pressure_mean_p2_losx',
'arterial_blood_pressure_systolic_losx',
'arterial_blood_pressure_systolic_p2_losx', 'heart_rate_losx',
'heart_rate_p2_losx', 'ph_arterial_losx', 'ph_arterial_p2_losx',
'o2_saturation_losx', 'o2_saturation_p2_losx', 'so2_arterial_losx',
'so2_arterial_p2_losx', 'fio2_losx', 'fio2_p2_losx',
'pao2_over_fio2_losx', 'pao2_over_fio2_p2_losx', 'ureum_losx',
'ureum_p2_losx', 'creatinine_losx', 'creatinine_p2_losx',
'temperature_losx', 'temperature_p2_losx',
'adjusted_sofa_total_partial_losx',
'adjusted_sofa_total_partial_p2_losx', 'potassium_losx',
'potassium_p2_losx', 'po2_arterial_losx', 'po2_arterial_p2_losx',
'glasgow_coma_scale_motor_losx',
'glasgow_coma_scale_motor_p2_losx', 'glasgow_coma_scale_eye_losx',
'glasgow_coma_scale_eye_p2_losx', 'thrombocytes_losx',
'thrombocytes_p2_losx', 'lactate_dehydrogenase_losx',

'lactate_dehydrogenase_p2_losx', 'pco2_arterial_losx',
'pco2_arterial_p2_losx', 'leukocytes_losx', 'leukocytes_p2_losx',
'mechanical_power_losx', 'mechanical_power_p2_losx']

A.2 Selection of variables

['pco2_arterial', 'driving_pressure', 'lung_compliance_static', 'fio2', 'pressure_above_peep',
'glasgow_coma_scale_motor', 'arterial_blood_pressure_diastolic', 'mechanical_power_per_kg',
'minute_volume', 'estimated_glomerular_filtration_rate', 'pao2_over_fio2', 'glasgow_coma_scale_verbal',
'peak_pressure_0dum', 'lung_compliance_dynamic', 'tidal_volume_per_kg', 'leukocytes', 'potassium', 'peep', 'pressure_above_peep_0dum', 'base_excess', 'albumin', 'intubated_cum', 'o2_flow',
'ureum', 'los', 'med_norepinephrine', 'alkaline_phosphatase', 'phosphate', 'magnesium',
'intubated', 'heart_rate', 'med_sodium_chloride', 'med_norepinephrine_0dum',
'lung_compliance_dynamic_0dum', 'lymphocytes_percentage', 'activated_partial_thromboplastin_time',
'c_reactive_protein', 'respiratory_rate_measured_ventilator_0dum', 'rapid_shallow_breathing_index',
'estimated_glomerular_filtration_rate_0dum', 'age', 'respiratory_rate_measured', 'med_propofol',
'creatinine', 'so2_arterial_0dum', 'o2_flow_0dum', 'adjusted_sofa_total_partial', 'female', 'bmi',
'adjusted_sofa_total_partial_losx', 'age_p2', 'bmi_p2']

B Tables

Table A1: Hospital performance variation with RF

	AUROC	**AUPRC**	**F2-score**
A	0.877321	0.142951	0.329912
B	0.872445	0.141321	0.305556
C	0.873680	0.135605	0.307798
D	0.874680	0.136314	0.310345
E	0.874938	0.126531	0.316804
F	0.874380	0.135792	0.316537
G	0.868200	0.121034	0.287206
H	0.874773	0.136904	0.311203
I	0.873771	0.141063	0.313808
J	0.873462	0.137996	0.303644
K	0.874051	0.127706	0.296610
L	0.872477	0.126359	0.312500
M	0.874487	0.141974	0.312935
N	0.872893	0.126659	0.300401
O	0.874069	0.135888	0.307377
P	0.874018	0.127638	0.302198
Q	0.872671	0.124822	0.292588
R	0.874029	0.135527	0.314246

Table A2: Variable importance in terms of FI and absolute coefficient

	DT	DTC	RF	CE	HX	PIHIX
o2_flow_0dum	0.0	0.000	0.001	1.086	1.086	1.125
glasgow_coma_scale_motor	0.0	0.018	0.028	-1.008	-1.008	-0.985
lung_compliance_dynamic_0dum	0.0	0.000	0.002	-0.971	-0.971	-0.970
peak_pressure_0dum	0.0	0.010	0.006	-0.966	-0.966	-0.871
fio2	0.0	0.057	0.045	0.635	0.635	0.645
adjusted_sofa_total_partial	0.0	0.019	0.032	-0.548	-0.548	-0.606
heart_rate	0.0	0.023	0.026	0.504	0.504	0.496
minute_volume	0.0	0.017	0.024	0.516	0.516	0.493
pressure_above_peep_0dum	0.0	0.000	0.002	-0.503	-0.503	-0.463
adjusted_sofa_total_partial_losx	0.0	0.082	0.072	0.424	0.424	0.444
albumin	0.0	0.008	0.020	-0.371	-0.371	-0.415
pao2_over_fio2	0.0	0.033	0.042	-0.348	-0.348	-0.369
respiratory_rate_measured_ventilator_0dum	0.0	0.000	0.002	0.373	0.373	0.357
peep	0.0	0.011	0.017	-0.393	-0.393	-0.342
lung_compliance_dynamic	0.0	0.010	0.010	-0.320	-0.320	-0.335
pco2_arterial	1.0	0.176	0.064	0.305	0.305	0.309
estimated_glomerular_filtration_rate_0dum	0.0	0.004	0.004	-0.301	-0.301	-0.307
med_norepinephrine_0dum	0.0	0.007	0.003	-0.243	-0.243	-0.299
estimated_glomerular_filtration_rate	0.0	0.016	0.019	-0.314	-0.314	-0.290
so2_arterial_0dum	0.0	0.000	0.002	0.367	0.367	0.262
arterial_blood_pressure_diastolic	0.0	0.062	0.035	-0.264	-0.264	-0.260
glasgow_coma_scale_verbal	0.0	0.018	0.015	0.311	0.311	0.234
tidal_volume_per_kg	0.0	0.000	0.015	-0.252	-0.252	-0.231
leukocytes	0.0	0.067	0.027	0.223	0.223	0.225
intubated	0.0	0.000	0.002	-0.247	-0.247	-0.212
ureum	0.0	0.022	0.027	0.133	0.133	0.153
intubated_cum	0.0	0.020	0.026	0.138	0.138	0.130
magnesium	0.0	0.018	0.020	-0.101	-0.101	-0.104
activated_partial_thromboplastin_time	0.0	0.015	0.019	0.107	0.107	0.098
driving_pressure	0.0	0.034	0.041	-0.105	-0.105	-0.094
lung_compliance_static	0.0	0.004	0.012	-0.085	-0.085	-0.084
phosphate	0.0	0.025	0.024	0.074	0.074	0.068
respiratory_rate_measured	0.0	0.021	0.023	0.063	0.063	0.064
creatinine	0.0	0.003	0.019	-0.065	-0.065	-0.056
c_reactive_protein	0.0	0.024	0.020	0.074	0.074	0.055
pressure_above_peep	0.0	0.011	0.030	0.051	0.051	0.054
base_excess	0.0	0.017	0.022	-0.034	-0.034	-0.041
female	0.0	0.000	0.002	0.026	0.026	0.029
med_propofol	0.0	0.002	0.009	-0.032	-0.032	-0.026
los	0.0	0.029	0.027	0.001	0.001	0.022
o2_flow	0.0	0.002	0.004	0.018	0.018	0.019
mechanical_power_per_kg	0.0	0.005	0.021	-0.002	-0.002	-0.019
med_sodium_chloride	0.0	0.008	0.006	-0.021	-0.021	-0.016
potassium	0.0	0.011	0.021	0.008	0.008	0.014
lymphocytes_percentage	0.0	0.016	0.017	-0.019	-0.019	-0.011
rapid_shallow_breathing_index	0.0	0.020	0.010	-0.003	-0.003	-0.003
med_norepinephrine	0.0	0.002	0.022	0.001	0.001	0.002
alkaline_phosphatase	0.0	0.014	0.018	0.007	0.007	-0.002

Table A3: Ranks of variable importance

	DTC	RF	CE	HX	PIHIX
o2_flow_0dum	46	48	1	1	1
glasgow_coma_scale_motor	17	9	2	2	2
lung_compliance_dynamic_0dum	41	44	3	3	3
peak_pressure_0dum	29	38	4	4	4
fio2	5	3	5	5	5
adjusted_sofa_total_partial	16	7	6	6	6
heart_rate	11	14	8	8	7
minute_volume	21	15	7	7	8
pressure_above_peep_0dum	44	42	9	9	9
adjusted_sofa_total_partial_losx	2	1	10	10	10
albumin	31	23	13	13	11
pao2_over_fio2	7	4	15	15	12
respiratory_rate_measured_ventilator_0dum	43	45	12	12	13
peep	28	30	11	11	14
lung_compliance_dynamic	30	34	16	16	15
pco2_arterial	1	2	19	19	16
estimated_glomerular_filtration_rate_0dum	36	40	20	20	17
med_norepinephrine_0dum	33	41	24	24	18
estimated_glomerular_filtration_rate	22	26	17	17	19
so2_arterial_0dum	45	47	14	14	20
arterial_blood_pressure_diastolic	4	6	21	21	21
glasgow_coma_scale_verbal	18	32	18	18	22
tidal_volume_per_kg	47	31	22	22	23
leukocytes	3	10	25	25	24
intubated	42	43	23	23	25
ureum	12	12	27	27	26
intubated_cum	14	13	26	26	27
magnesium	19	22	30	30	28
activated_partial_thromboplastin_time	24	25	28	28	29
driving_pressure	6	5	29	29	30
lung_compliance_static	35	33	31	31	31
phosphate	9	16	32	32	32
respiratory_rate_measured	13	17	35	35	33
creatinine	37	27	34	34	34
c_reactive_protein	10	24	33	33	35
pressure_above_peep	26	8	36	36	36
base_excess	20	19	37	37	37
female	48	46	39	39	38
med_propofol	40	36	38	38	39
los	8	11	47	47	40
o2_flow	38	39	42	42	41
mechanical_power_per_kg	34	21	46	46	42
med_sodium_chloride	32	37	40	40	43
potassium	27	20	43	43	44
lymphocytes_percentage	23	29	41	41	45
rapid_shallow_breathing_index	15	35	45	45	46
med_norepinephrine	39	18	48	48	47
alkaline_phosphatase	25	28	44	44	48