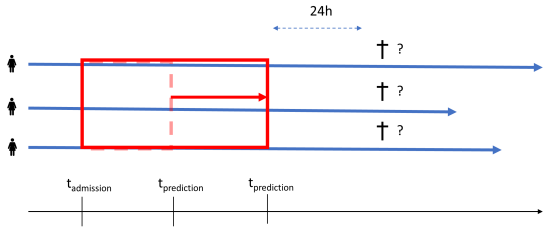
**Supplementary material: Dynamic prediction of mortality in COVID-19 patients in the intensive care unit: a retrospective multi-center cohort study**

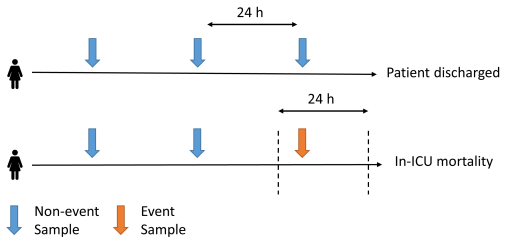
# A Supplementary figures



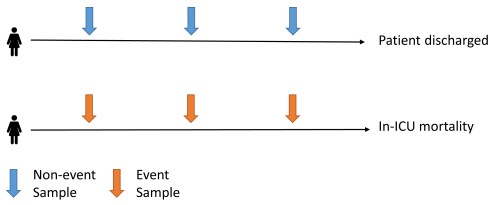

(a) Near-term mortality ($\leq$ 24 hours) modeling. Here, the model repeatedly predicts mortality to occur within 24 hours from the moment of prediction.


(b) In-ICU ('long-term') mortality modeling. Here, the model repeatedly predicts mortality to occur during the complete ICU stay.

Figure 1: Visual representation of near-term mortality modeling (a) and in-ICU mortality modeling (b).




(a) Labeling strategy for near-term mortality ($\leq$ 24 hours) modeling.




(b) Labeling strategy for in-ICU mortality modeling.

Figure 2: Visualization of the different sampling strategies.

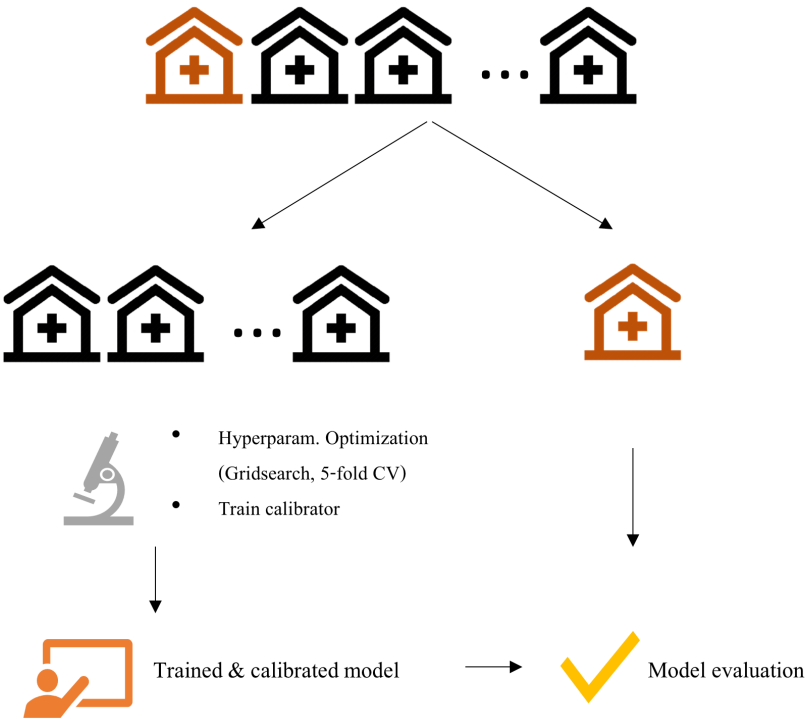



Figure 3: Leave-one-ICU-out (LOIO) cross-validation procedure.

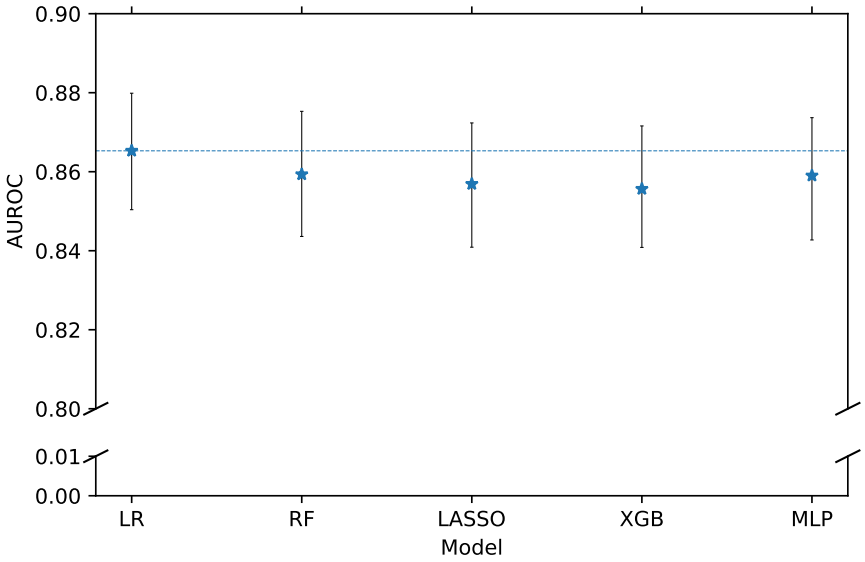



Figure 4: Overall areas under the receiver operating characteristic curves (AUROCs) yielded by the different models for near-term mortality prediction.

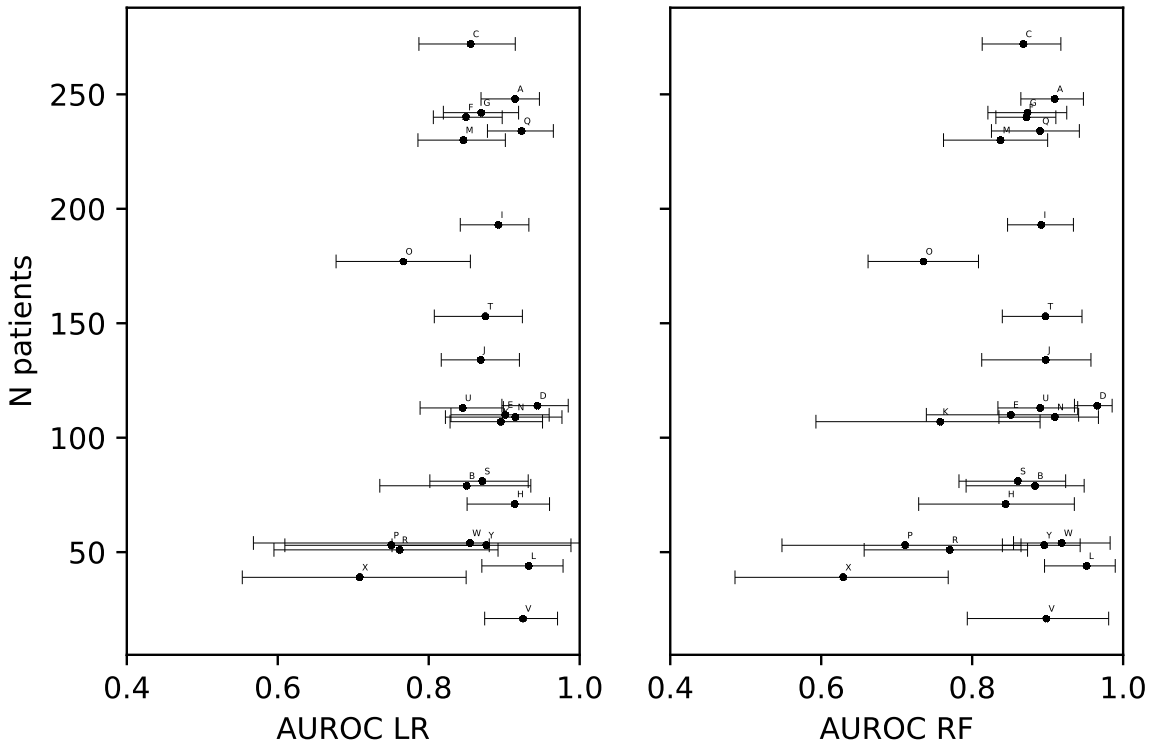
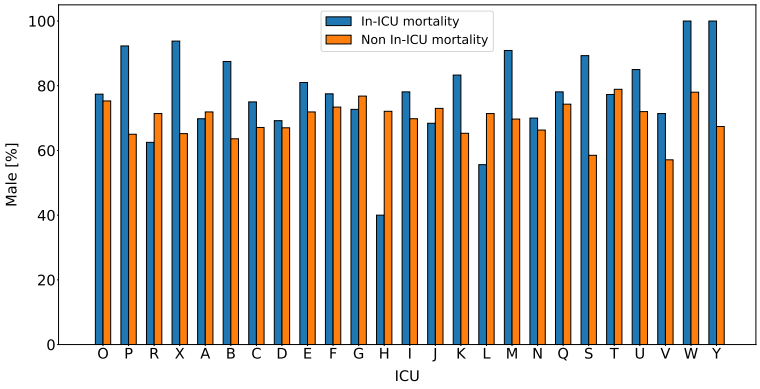
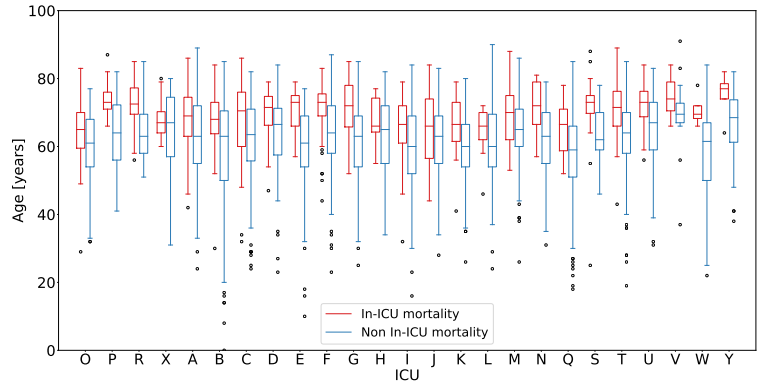


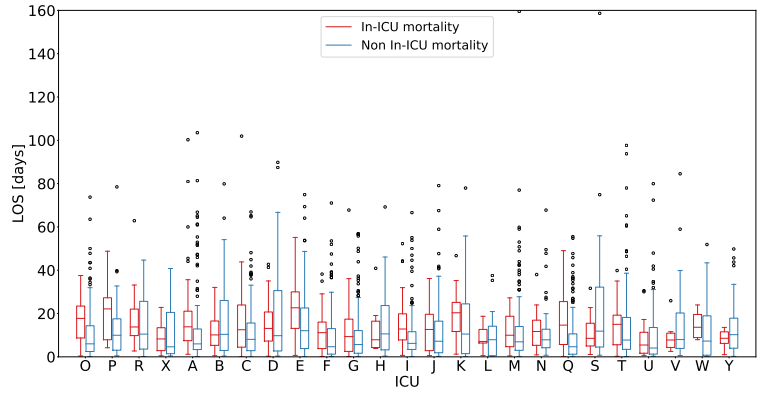

Figure 5: Near-term mortality: areas under the receiver-operating-curve (AUROCs) with 95% CIs for the logistic regression (LR) and random forest (RF) model, validated on the different ICUs, sorted by sample size.




(a)




(b)




(c)

Figure 6: Comparison of sex at birth (a), age (b) and length-of-stay (LOS) (c) of ICUs O, P, R and X compared to the remaining ICUs.

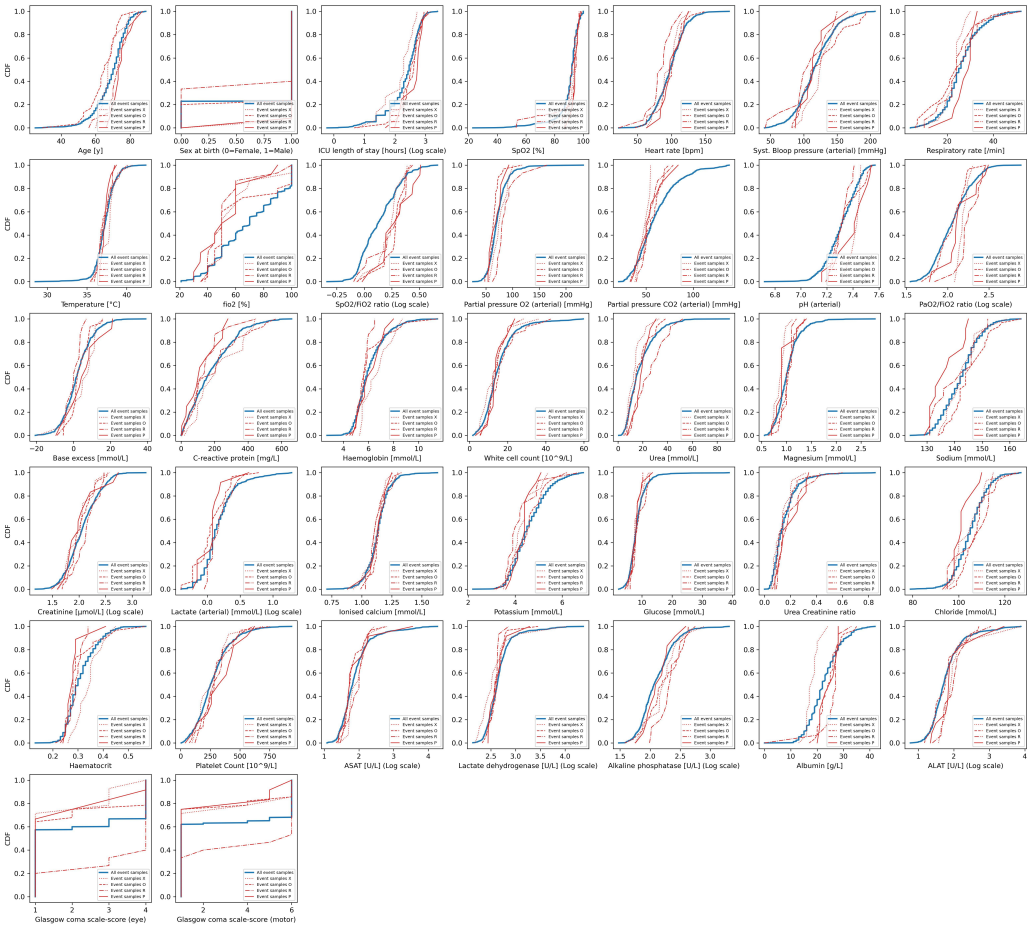
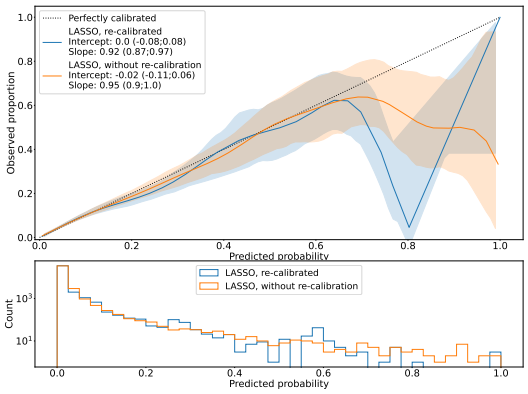
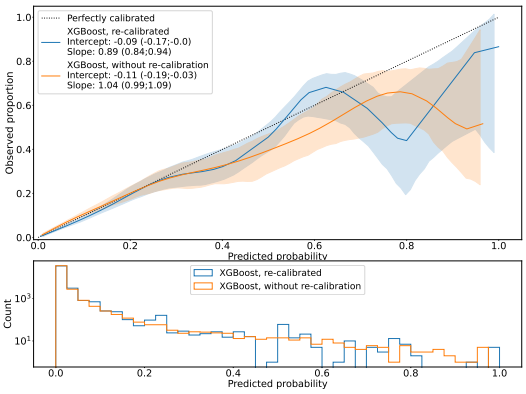


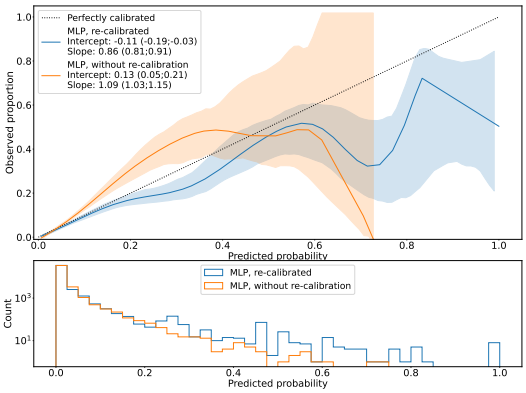

Figure 7: Cumulative distributions for all predictors based on the samples taken within 24 hours of ICU death ('event samples') of patients from ICU O(N=31), P(N=13), R(N=16) and X(N=16) in red. The cumulative distribution based on event samples of patients from all ICUs (N=667) is plotted as a reference.




(a) LASSO




(b) XGBoost




(c) MLP

Figure 8: Results near-term (24 hour) mortality modeling: smoothed flexible calibration curves for the (a) logistic regression with L1 regularization (LASSO), (b) Gradient Boosting (XGBoost) and (c) multilayer perceptron (MLP) models, with and without re-calibration using isotonic regression. Shaded areas around the curves represent the 95%CIs. In the bottom plots, histograms of the predictions.

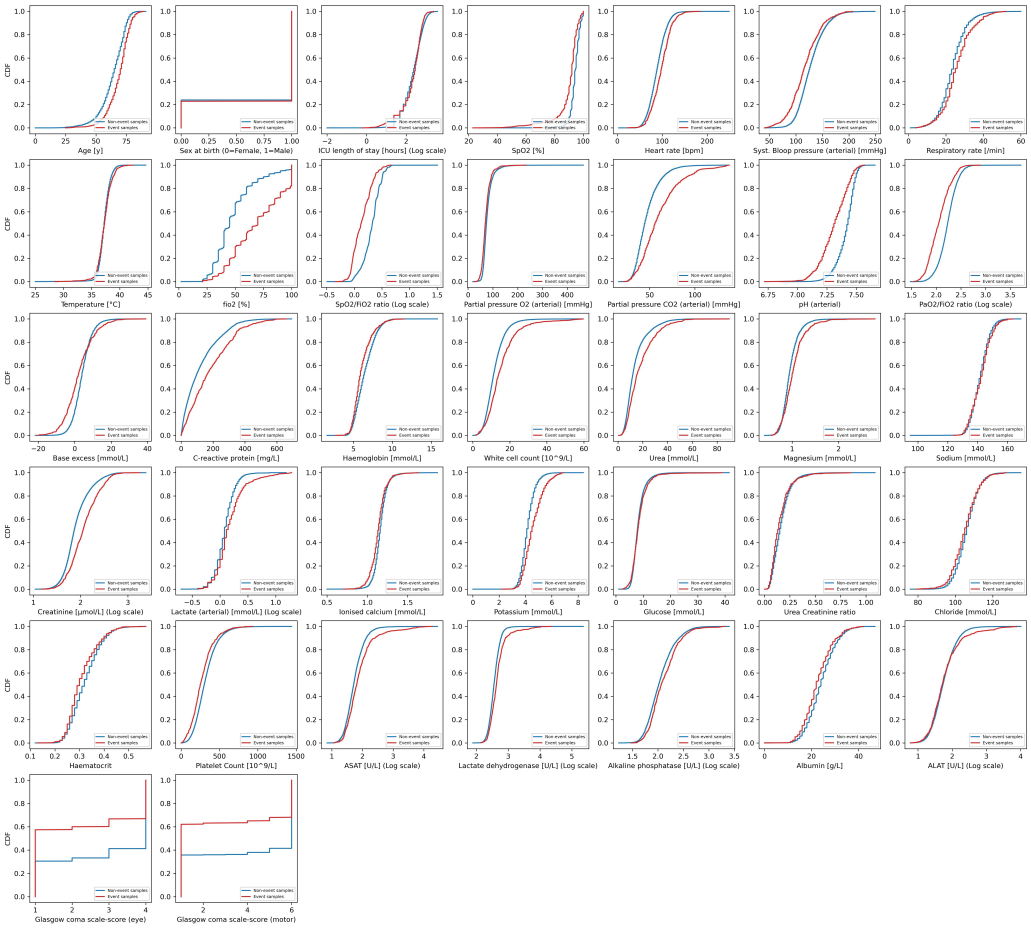



Figure 9: Cumulative distributions of different predictors from all included patients, based on samples taken within 24 hours of ICU death ('event samples') in blue and all other ('non-event') samples in red.

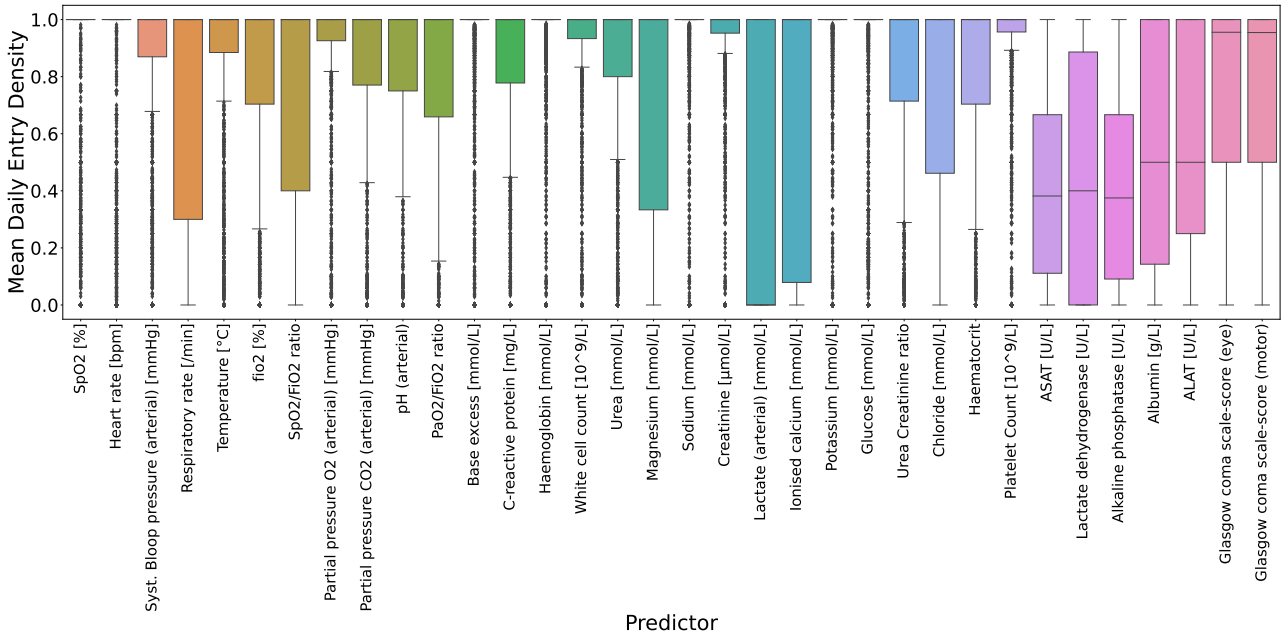



Figure 10: Boxplots of the daily entry density (i.e., fractions of non-empty daily measurements) distributions for each candidate predictor.

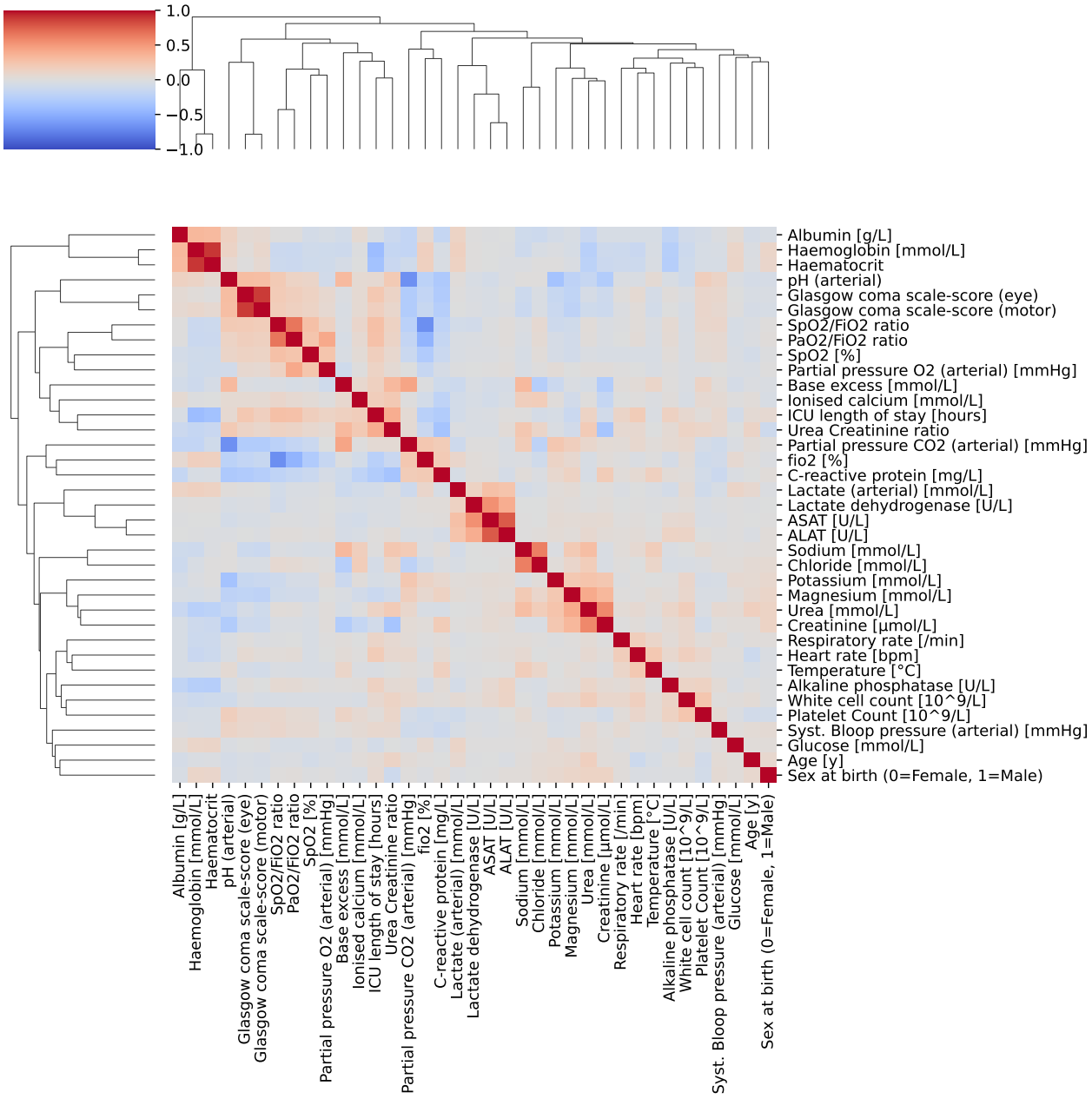
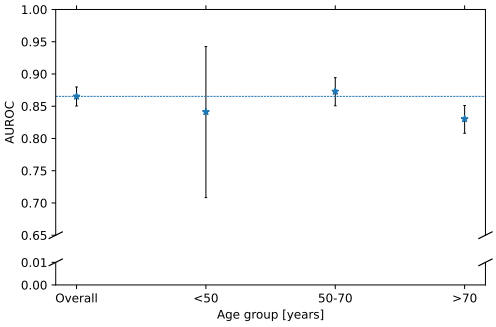
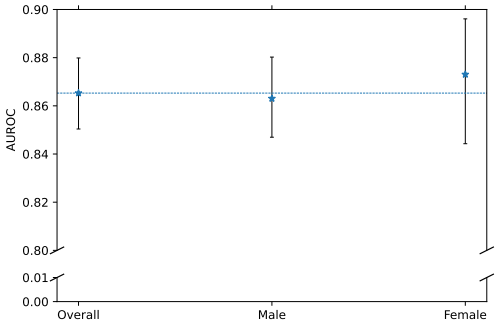


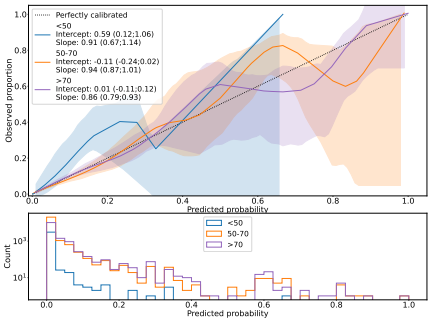

Figure 11: Clustermap of the correlation matrix of all included model predictors.



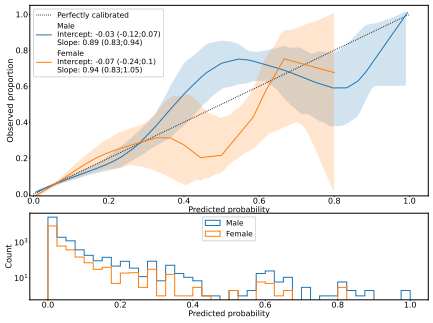

(a)



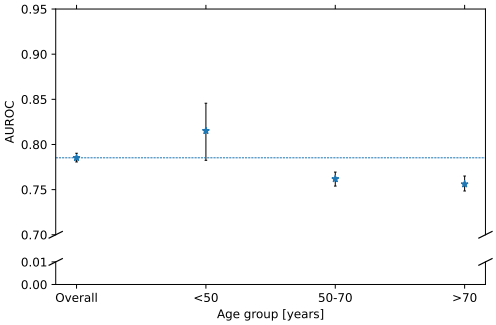

(b)



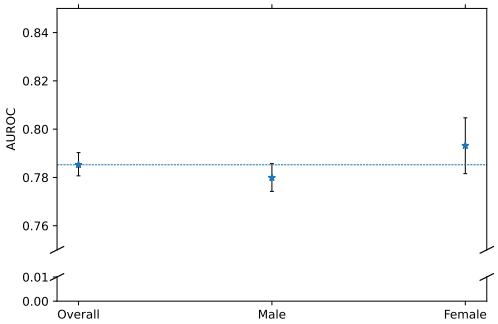

(c)



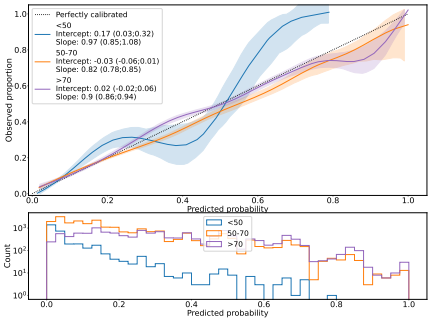

(d)



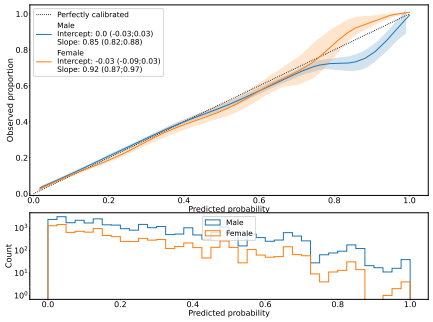

(e)




(f)

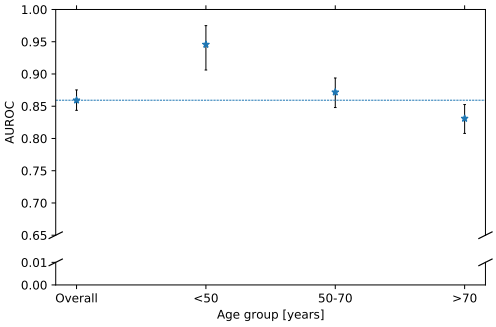



(g)

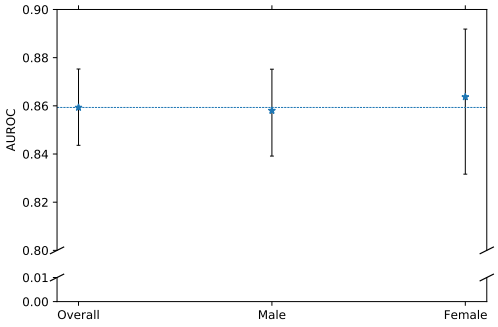



(h)

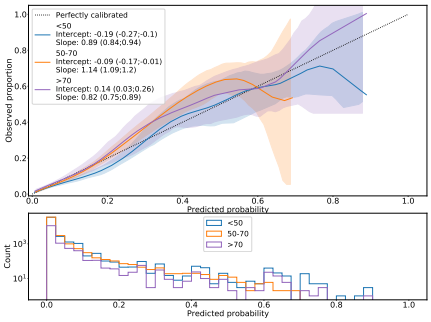
Figure 12: Model performance of the logistic regression (LR) models for near-term (24 hour) mortality prediction in different age groups and sexes in terms of discrimination (a,b) and calibration (c,d) and for long-term (in-ICU) mortality prediction in different age groups and sexes in terms of discrimination (e,f) and calibration (g,h)



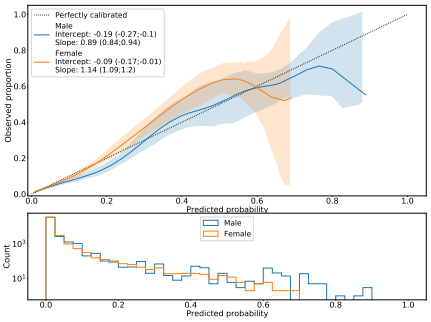

(a)



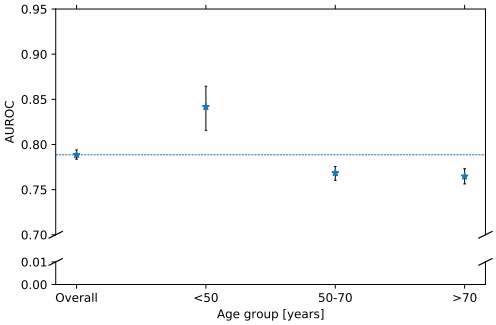

(b)



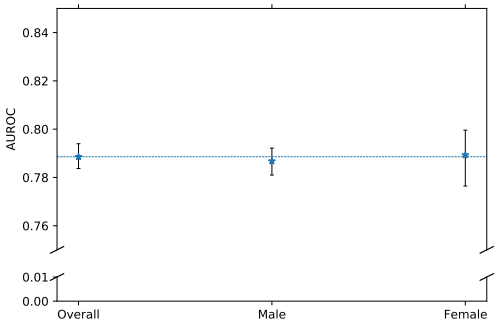

(c)



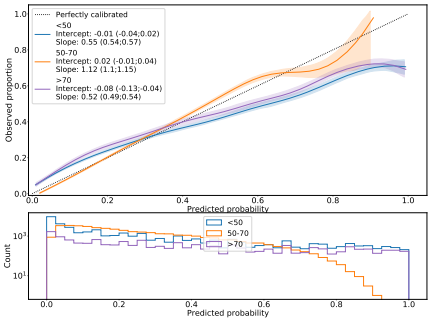

(d)



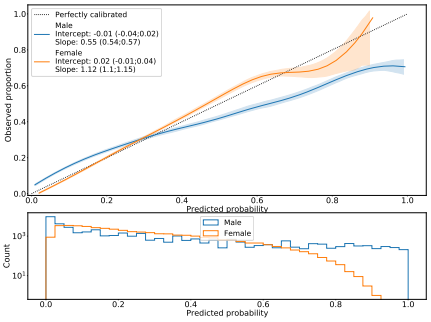

(e)




(f)




(g)




(h)

Figure 13: Model performance of the random forest (RF) models for near-term (24 hour) mortality prediction in different age groups and sexes in terms of discrimination (a,b) and calibration (c,d) and for long-term (in-ICU) mortality prediction in different age groups and sexes in terms of discrimination (e,f) and calibration (g,h)

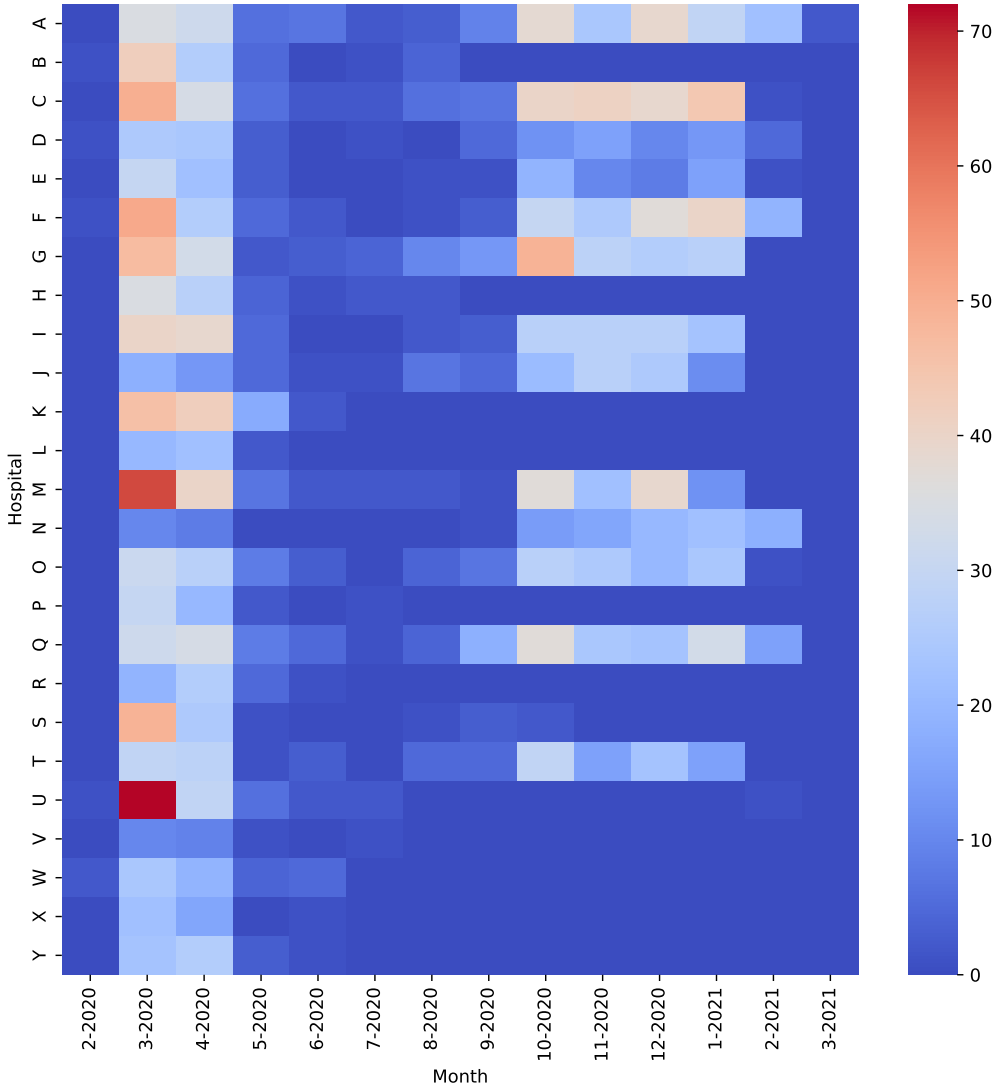



Figure 14: Number of ICU admissions per month among the 25 included hospitals. The number of patients peaks during two time-periods coinciding with the first (March-April 2020) and second (November 2020-January 2021) COVID-19 'waves' in the Netherlands

# B Supplementary tables

| | Unit | Included | Mean entry density | Median entry density |
|---|---|---|---|---|
| **Patient Demographics:** | | | | |
| Age on admission | years | ✓ | - | - |
| Sex at Birth | - | ✓ | - | - |
| | | | | |
| **Bedside investigations:** | | | | |
| GCS (eye) | - | ✓ | 0.73 | 0.96 |
| GCS (motor) | - | ✓ | 0.72 | 0.95 |
| | | | | |
| **Vital Signs:** | | | | |
| Respiratory Rate | breaths/min | ✓ | 0.70 | 1.00 |
| $SpO_2$ | % | ✓ | 0.88 | 1.00 |
| Systolic blood pressure | mmHg | ✓ | 0.83 | 1.00 |
| Temperature | °C | ✓ | 0.81 | 1.00 |
| Heart Rate | bpm | ✓ | 0.81 | 1.00 |
| | | | | |
| **Blood gasses:** | | | | |
| pH | - | ✓ | 0.77 | 1.00 |
| Base excess | mmol/L | ✓ | 0.91 | 1.00 |
| $PaO_2$ | mmHg | ✓ | 0.83 | 1.00 |
| $PaCO_2$ | mmHg | ✓ | 0.77 | 1.00 |
| $FiO_2$ | % | ✓ | 0.78 | |
| $PaO_2/FiO_2$ | - | ✓ | 0.75 | 1.00 |
| **Laboratory test results:** | | | | |
| Haemoglobin | mmol/L | ✓ | 0.92 | 1.00 |
| Haematocrit | L/L | ✓ | 0.77 | 1.00 |
| White cell count | $1 \times 10^9$/L | ✓ | 0.89 | 1.00 |
| Platelet count | $1 \times 10^9$/L | ✓ | 0.91 | 1.00 |
| Sodium | mmol/L | ✓ | 0.92 | 1.00 |
| Chloride | mmol/L | ✓ | 0.74 | 1.00 |
| Potassium | mmol/L | ✓ | 0.93 | 1.00 |
| Magnesium | mmol/L | ✓ | 0.69 | 1.00 |
| ALAT | IU/L | ✓ | 0.53 | 0.50 |
| ASAT | IU/L | ✓ | 0.44 | 0.38 |
| Albumin | g/L | ✓ | 0·55 | 0.50 |
| Lactate (arterial) | mmol/L | | 0·43 | 0.00 |
| Lactate dehydrogenase | IU/L | ✓ | 0·45 | 0.40 |
| Urea | mmol/L | ✓ | 0·80 | 1.00 |
| Creatinine | $\mu mol/L$ | ✓ | 0·89 | 1.00 |
| Urea to creatinine ratio | - | ✓ | 0·78 | 1.00 |
| C-reactive protein | mg/L | ✓ | 0·81 | 1.00 |
| Ionised Calcium | mmol/L | ✓ | 0·66 | 1.00 |
| Glucose | mmol/L | ✓ | 0·88 | 1.00 |
| Alkaline phosphatase | IU/L | ✓ | 0·43 | 0.38 |
| | | | | |
| **Extra** | | | | |
| Length of stay on ward | hours | ✓ | - | - |
| $SpO_2/FiO_2$ | - | ✓ | 0.72 | 1.00 |

Table 1: Candidate predictors evaluated for potential inclusion in the prediction model, based on evidence in literature and availability.

| Model | Hyperparameter | Search Space |
|---|---|---|
| LR / LASSO | $\lambda$ | $[10^{-4}, \ldots, 10^{4}]$ evenly spaced on log scale with 20 steps |
| RF | max features | [ $\sqrt{p}$ , $\log_2 p$ ] where p is the total number of predictors. |
| | max depth | [3, 4, 5, 6, 8, 10, 12, 15 ] |
| XGBoost | learning rate | [ 0.05, 0.10, 0.15, 0.20, 0.25, 0.30 ] |
| | max depth | [3, 4, 5, 6, 8, 10, 12, 15 ] |
| | min child weight | [1, 3, 5, 7 ] |
| | gamma | [0.0, 0.1, 0.2 , 0.3, 0.4 ] |
| | colsample bytree | [0.3, 0.4, 0.5 , 0.7 ] |

Table 2: Search spaces used in the grid-search for model hyperparameter optimization for the logistic regression models using L2 (LR) and L1 (LASSO) regularization, the random forest (RF) and the Gradient Boosting (XGBoost) models.

# C Multilayer Perceptron

We fitted a multilayer perceptron (MLP), i.e. a feedforward artificial neural network, using the Keras library in Python. The input the network is a vector containing the 36 predictors (after normalization and imputation). The network consists of two fully connected (hidden) layers sized 18 and 9 neurons with ReLu activation functions. After the input layer and both hidden layers, we implemented dropout (which randomly drops input neurons from the network during training to prevent overfitting) with a 20% rate. The output layer is a sigmoid function. We optimzed the corss-entropy using the Adam optimizer.[1] The code for importing the required packages and defining the model can be found below.

```
import tensorflow as tf
import keras

def nn_model(INPUT_DIM):
        clf = keras.Sequential([
        keras.layers.Dropout(.2, input_shape=(INPUT_DIM,)),
        keras.layers.Dense(18,input_dim=INPUT_DIM, activation=tf.nn.relu),
        keras.layers.Dropout(.2, input_shape=(18,)),
        keras.layers.Dense(9, activation=tf.nn.relu),
        keras.layers.Dropout(.2, input_shape=(9,)),
        keras.layers.Dense(1, activation=tf.nn.sigmoid)])

        clf.compile(
            optimizer='adam',
            loss='binary_crossentropy',
            metrics=[
                keras.metrics.AUC()])

        return clf
```

# D In-ICU mortality modeling

To compare modeling with near-term ($\leq$ 24 hours) and in-ICU ('long-term') mortality as a clinical endpoint in this study, we repeated the modeling development, re-calibration and validation procedure (as described in the main text) to predict in-ICU mortality. Figure 1 visualizes the differences between near-term and in-ICU mortality prediction.

## D.1 Methods

To model in-ICU mortality, we labeled all patient samples as 'event samples' the patient did not survive the ICU admission and as 'non-event samples' otherwise. Figure 2 visualizes the corresponding labeling strategies for near-term and in-ICU mortality modeling. Models were trained, re-calibrated and validated following the same procedure as performed for near-term mortality modeling (for which we refer to the Methods section in the main text), using both a (linear) logistic regression (LR) model and a (non-linear) random forest (RF) model, and benchmarked these with a logistic regression model with L1 regularization (LASSO), a gradient boosting (XGBoost) model and a multilayer perceptron (MLP). Also, we quantified predictor importance by SHAP values based on the models trained on the complete cohort (all 25 ICUs).

## D.2 Results

Overall, i.e. by combining the predictions of all iterations in the LOIO procedure, both the LR and RF models yielded an AUROC of 0·79 [0·78,0·79]. The LASSO, XGBoost and MLP models yielded similar or lower overall AUROCs compared to the LR model (figure 17). Point estimates of the AUROCs yielded in the individual ICUs are depicted in figure 15 and in table 3 shows the corresponding 95% CIs. We observed wide CIs for the models validated on ICUs with relatively small sample sizes (figure 16).

Figure 18 shows the flexible calibration curves yielded by the different models with and without re-calibration, including the corresponding calibration intercepts and slopes. Without re-calibration, the LR model overestimated the mortality risk (intercept<0) and the RF models yielded slightly too moderate predictions (slope>1). After re-calibration, both models show good calibration in the large, but slightly too extreme predictions, with a calibration slope of 0·87 [0·84,0·89] and 0·55 [0·54,0·57], respectively for the LR and RF model.

Table 4 shows the 20 most important predictors ranked based on the mean SHAP magnitude and figure 19 shows the corresponding summary plots for the SHAP values for the LR and RF model.

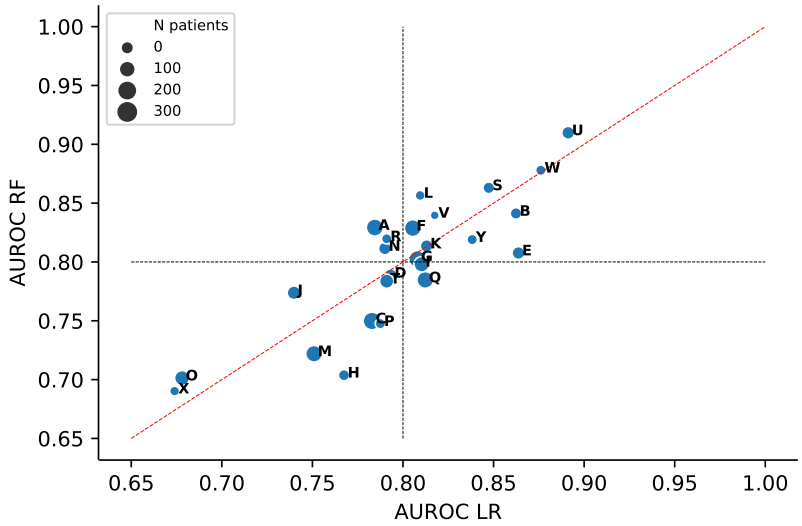



Figure 15: Results in-ICU mortality modeling: Areas under the receiver-operating-curve (AUROCs) for the logistic regression (LR) and random forest (RF) models validated on the different ICUs.

| ICU | N patients | Prevalence in-ICU mortality | LR AUROC [95% CI] | RF AUROC [95% CI] |
|---|---|---|---|---|
| V | 21 | 0·33 | 0·82 [0·76,0·86] | 0·84 [0·79,0·88] |
| X | 39 | 0·41 | 0·67 [0·62,0·72] | 0·69 [0·64,0·74] |
| L | 44 | 0·20 | 0·81 [0·75,0·87] | 0·86 [0·81,0·90] |
| R | 51 | 0·31 | 0·79 [0·76,0·82] | 0·82 [0·79,0·85] |
| Y | 53 | 0·13 | 0·84 [0·78,0·89] | 0·82 [0·76,0·87] |
| P | 53 | 0·25 | 0·79 [0·75,0·82] | 0·75 [0·71,0·78] |
| W | 54 | 0·07 | 0·88 [0·83,0·92] | 0·88 [0·83,0·92] |
| H | 71 | 0·14 | 0·77 [0·72,0·81] | 0·70 [0·65,0·75] |
| B | 79 | 0·30 | 0·86 [0·83,0·89] | 0·84 [0·81,0·87] |
| S | 81 | 0·35 | 0·85 [0·82,0·87] | 0·86 [0·84,0·88] |
| K | 107 | 0·11 | 0·81 [0·78,0·84] | 0·81 [0·78,0·84] |
| N | 109 | 0·18 | 0·79 [0·75,0·83] | 0·81 [0·77,0·84] |
| E | 110 | 0·19 | 0·86 [0·85,0·88] | 0·81 [0·79,0·83] |
| U | 113 | 0·18 | 0·89 [0·87,0·91] | 0·91 [0·89,0·93] |
| D | 114 | 0·23 | 0·79 [0·77,0·82] | 0·79 [0·76,0·81] |
| J | 134 | 0·14 | 0·74 [0·71,0·77] | 0·77 [0·75,0·80] |
| T | 153 | 0·29 | 0·79 [0·77,0·81] | 0·78 [0·76,0·80] |
| O | 177 | 0·18 | 0·68 [0·65,0·71] | 0·70 [0·68,0·73] |
| I | 193 | 0·33 | 0·81 [0·79,0·83] | 0·80 [0·78,0·81] |
| M | 230 | 0·10 | 0·75 [0·72,0·78] | 0·72 [0·70,0·75] |
| Q | 234 | 0·14 | 0·81 [0·79,0·83] | 0·78 [0·76,0·81] |
| F | 240 | 0·30 | 0·81 [0·79,0·82] | 0·83 [0·81,0·84] |
| G | 242 | 0·18 | 0·81 [0·79,0·83] | 0·80 [0·78,0·82] |
| A | 248 | 0·25 | 0·78 [0·76,0·80] | 0·83 [0·81,0·84] |
| C | 272 | 0·16 | 0·78 [0·76,0·80] | 0·75 [0·73,0·77] |

Table 3: Results for in-ICU mortality: AUROCs with 95% CI yielded by the logistic regression (LR) and random forest (RF) models in the different left-out ICUs (sorted by sample size). Prevalence is the fraction of patients who experience in-ICU mortality per ICU.

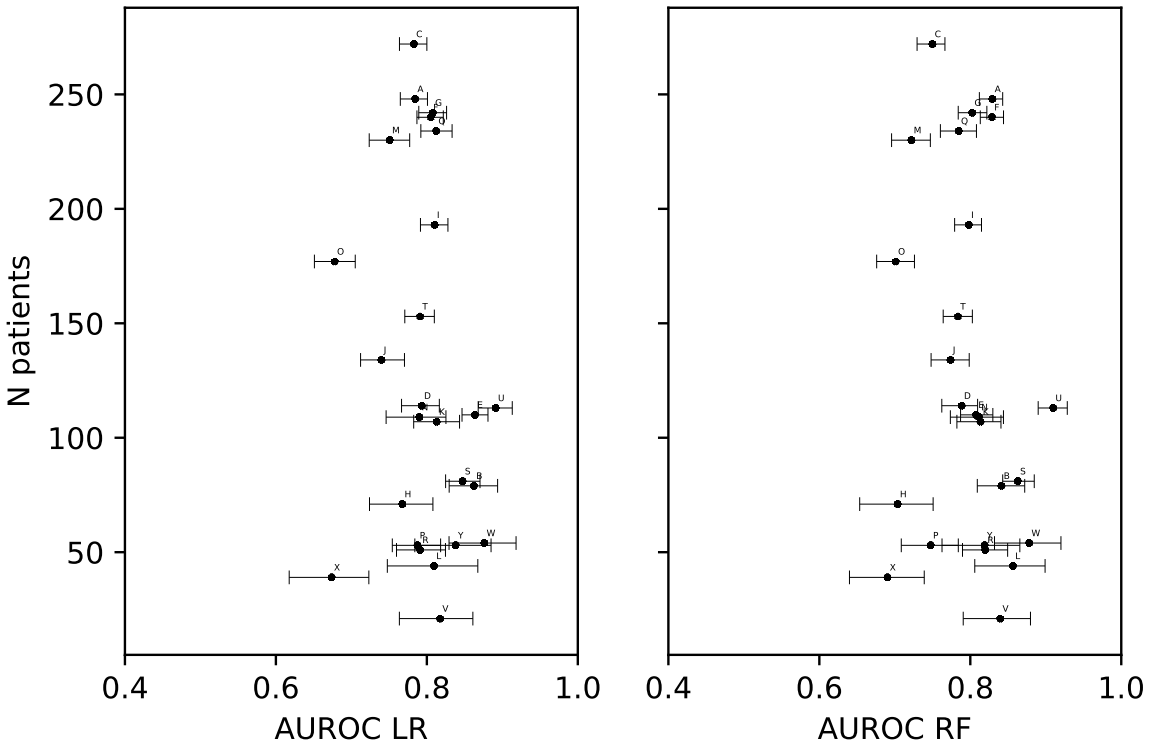



Figure 16: Results in-ICU mortality modeling: areas under the receiver-operating-curve (AUROCs) with 95% CIs for the logistic regression (LR) and random forest (RF) model, validated on the different ICUs, sorted by sample size.

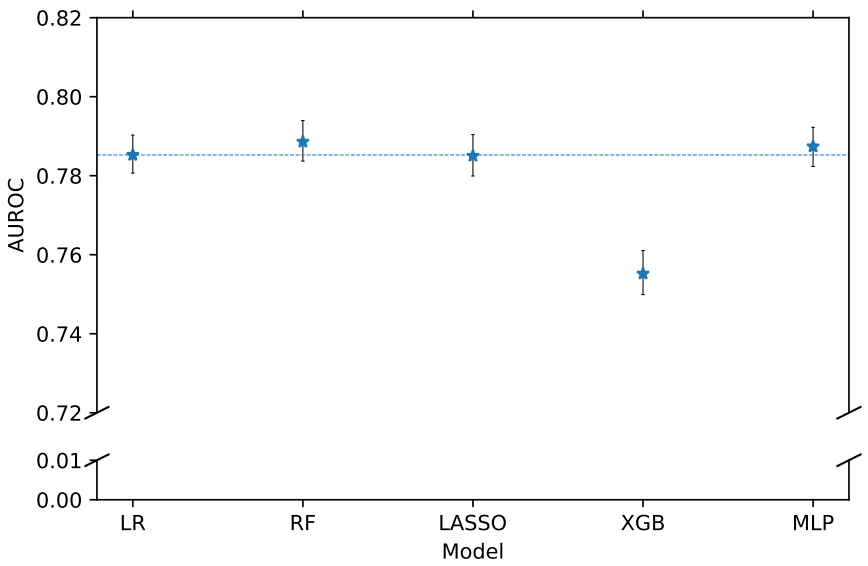
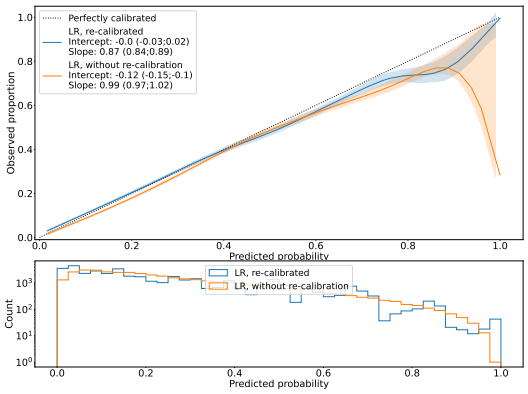
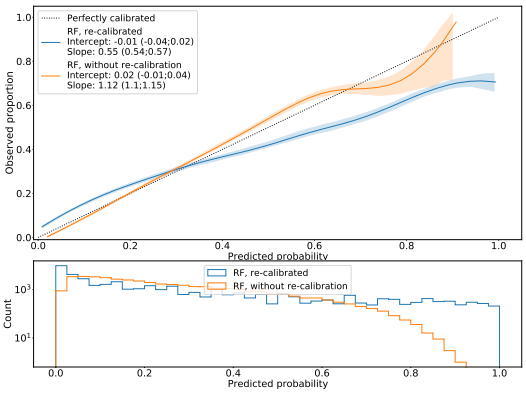


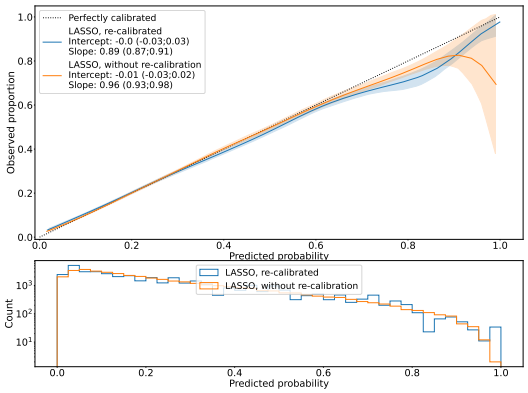

Figure 17: Overall areas under the receiver operating characteristic curves (AUROCs) yielded by the different models for long-term (in-ICU) mortality prediction



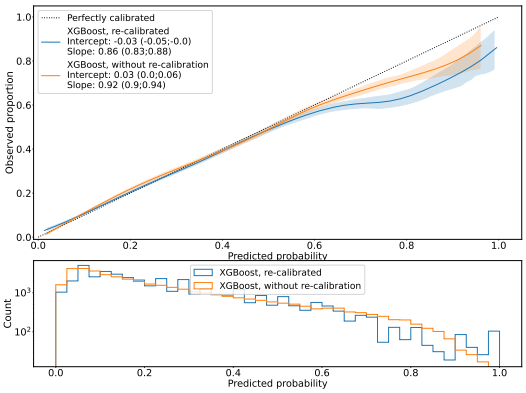

(a) LR



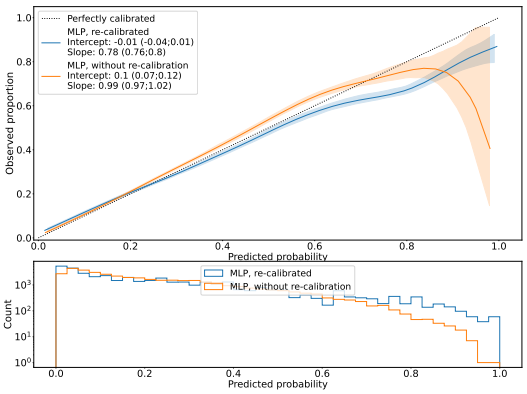

(b) RF




(c) LASSO




(d) XGBoost

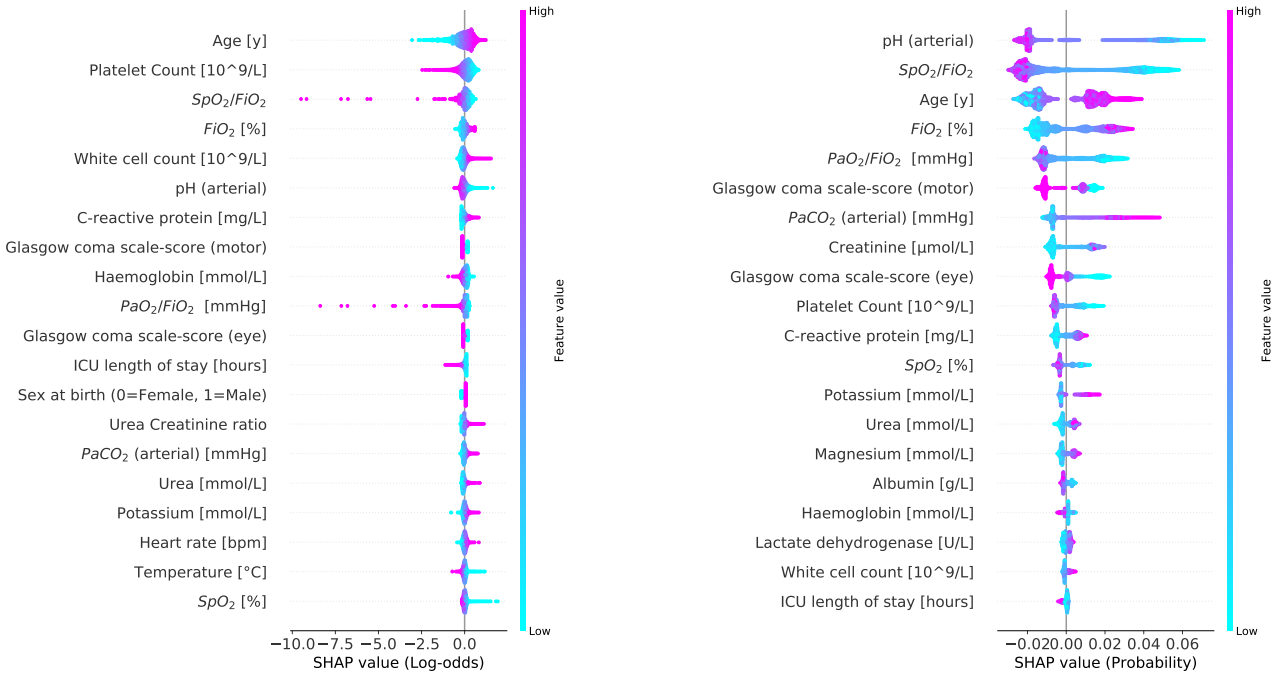



(e) MLP

Figure 18: Results in-ICU mortality modeling: smoothed flexible calibration curves for the (a) logistic regression (LR), (b) random forest, (c) logistic regression with L1 regularization (LASSO), (d) Gradient Boosting (XGBoost) and (e) multilayer perceptron (MLP) models, with and without re-calibration using isotonic regression. Shaded areas around the curves represent the 95%CIs. In the bottom plots, histograms of the predictions are shown.

| Predictor | Predictor importance LR model (mean \|SHAP\|, log-odds scale) | Predictor | Predictor importance RF model (mean \|SHAP\|, probability scale) |
|---|---|---|---|
| Age [y] | 0.373 | pH (arterial) | 0.0276 |
| Platelet Count [$1 \times 10^9$/L] | 0.258 | $SpO_2/FiO_2$ | 0.0248 |
| $SpO_2/FiO_2$ | 0.183 | Age [y] | 0.0164 |
| $FiO_2$ [%] | 0.141 | $FiO_2$ [%] | 0.0148 |
| White cell count [$1 \times 10^9$/L] | 0.134 | $PaO_2/FiO_2$ [mmHg] | 0.0129 |
| pH (arterial) | 0.132 | Glasgow coma scale-score (motor) | 0.0111 |
| C-reactive protein [mg/L] | 0.108 | $PaCO_2$ (arterial) [mmHg] | 0.0108 |
| Glasgow coma scale-score (motor) | 0.101 | Creatinine [mol/L] | 0.0089 |
| Haemoglobin [mmol/L] | 0.099 | Glasgow coma scale-score (eye) | 0.0079 |
| $PaO_2/FiO_2$ [mmHg] | 0.097 | Platelet Count [10^9/L] | 0.0063 |
| Glasgow coma scale-score (eye) | 0.087 | C-reactive protein [mg/L] | 0.0048 |
| ICU length of stay [hours] | 0.086 | $SpO_2$ [%] | 0.0043 |
| Sex at birth (0=Female, 1=Male) | 0.078 | Potassium [mmol/L] | 0.0042 |
| Urea Creatinine ratio | 0.076 | Urea [mmol/L] | 0.0029 |
| $PaCO_2$ (arterial) [mmHg] | 0.075 | Magnesium [mmol/L] | 0.0028 |
| Urea [mmol/L] | 0.075 | Albumin [g/L] | 0.0017 |
| Potassium [mmol/L] | 0.074 | Haemoglobin [mmol/L] | 0.0015 |
| Heart rate [bpm] | 0.069 | Lactate dehydrogenase [U/L] | 0.0012 |
| Temperature [°C] | 0.061 | White cell count [$1 \times 10^9$/L] | 0.0011 |
| $SpO_2$ [%] | 0.061 | ICU length of stay [hours] | 0.0008 |

Table 4: Results for in-ICU mortality: Global importances of the top 20 most important predictors for the logistic regression (LR) and random forest (RF) model trained for in-ICU mortality, ranked based on mean SHAP magnitude.




(a) (b)

Figure 19: Results in-ICU mortality modeling: summary plots for the SHAP values constructed from both logistic regression (left) and random forest model (right). Each SHAP value is represented by a single dot on each predictor row. Color is used to display the corresponding value of the predictor. Predictors are ordered by the mean SHAP magnitude.

# E Data missingness pattern

## E.1 Method

We analyzed the patterns in data missingness by first transforming the predictor matrix (i.e., the 33 predictors with misingness for at least one sample) in a binary matrix, inserting a 1 for missing and a 0 for non-missing. Hence, each column (i.e. each predictor) in this matrix now consists of a binary vector representing whether a value is missing (1) or available (0). Then, we calculated the Jaccard similarity coefficient score for each predictor pair among all these predictors. Given that:

- $M_{1,1}$ = total numbers of attributes, for which both A and B have 1
- $M_{0,1}$ = total numbers of attributes, for which A has 0 and B has 1
- $M_{1,0}$ = total numbers of attributes, for which A has 1 and B has 0
- $M_{0,0}$ = total numbers of attributes, for which both A and B have 0,

the Jaccard similarity coefficient score ($J$) between two vectors A and B is defined as follows:

$$J = \frac{M_{1,1}}{M_{1,0} + M_{0,1} + M_{1,1}} \tag{1}$$

Hence, two predictor columns that yield a high $J$ can be interpreted as predictors which are often missing in the same samples (thus at the same time points) and therefore, have a similar missingness pattern. We calculated the $J$s using the 'metrics jaccard score' function offered by scikit-learn in Python.[2]

## E.2 Results and interpretation

Figure 20 shows a heatmap of the $J$'s for each possible predictor pair. We observe several clusters (i.e. predictors with a similar missingness pattern), e.g. for blood gasses (pH, $PaO_2$,$PaCO_2$ and $PaO_2/SpO_2$), for a collection of laboratory test results (CRP, glucose, sodium, potassium, haemoglobin and creatinine) and the liver enzymes (ASAT, ALAT and alkaline phosphatase). The top 20 highest ranked predictors in importance (based on mean SHAP magnitude resulting from the LR model fitted using the full cohort) appear in many different clusters. These most important predictors are not strongly concentrated within these clusters, making it unlikely that the pattern in the importances we observed are due to the missingness pattern. Whether different imputation techniques, such as multiple imputation, would lead to even better model performance was beyond the scope of the current analysis.

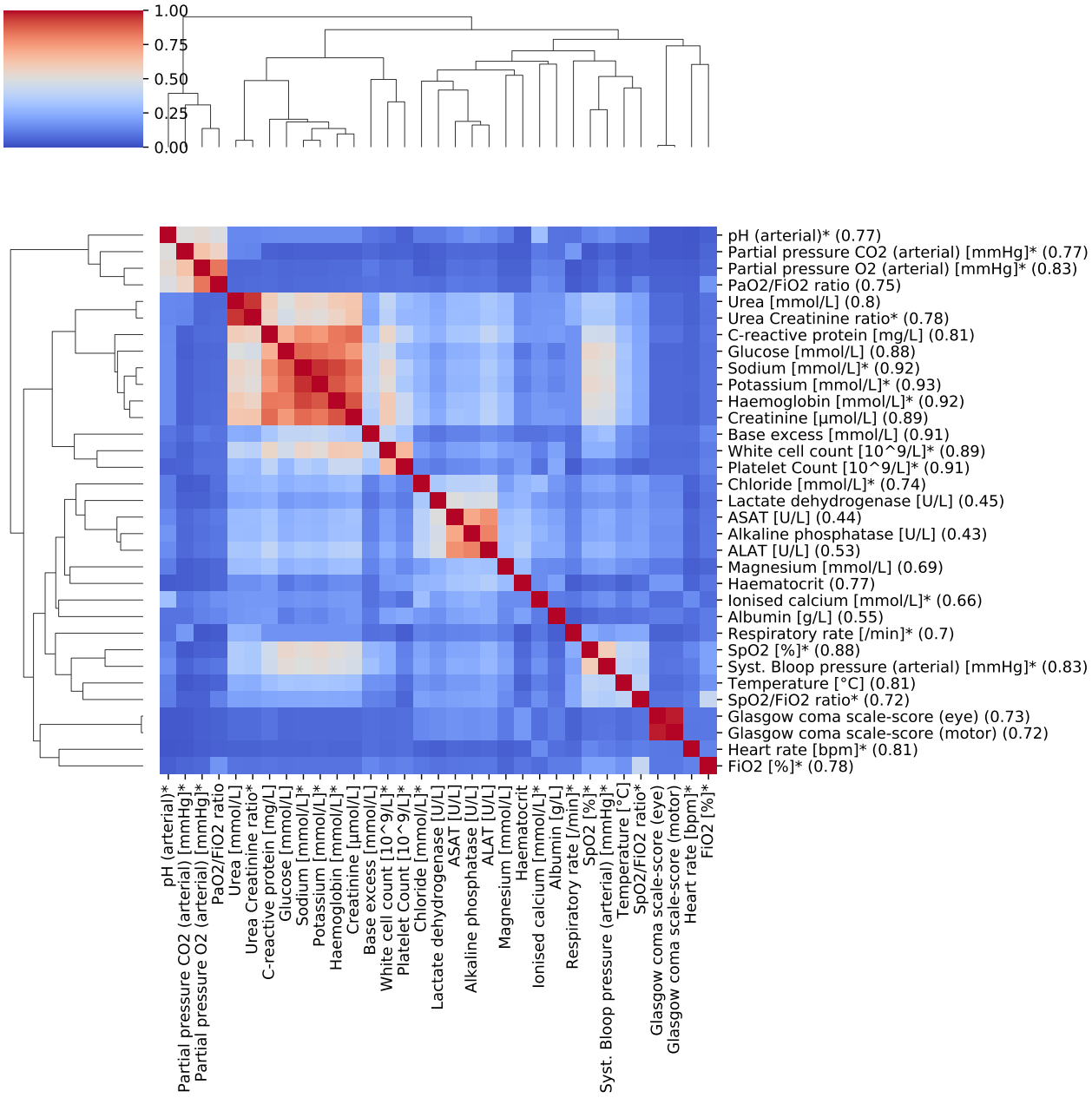



Figure 20: Missingness pattern of the included predictors: heatmap of Jaccard similarity coefficient scores ($Js$) of all possible predictor pairs (for all predictors with at least one missing value). For each predictor, the mean entry density (table 1) is given between brackets. Predictors highlighted with an asterisk are in the top 20 most important predictors ranked based on mean SHAP magnitudes resulting from the LR model fitted using the full cohort.

# References

[1] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y, editors. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings; 2015. Available from: http://arxiv.org/abs/1412.6980.

[2] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.