TU Delft Master thesis

The ability of new prediction models to discriminate covid-19 patients at risk of unplanned intensive care unit admission and unexpected death

A multi-center retrospective and simulated prospective cohort study

J.M. Smit

July 1, 2021

Abstract

Background The covid-19 pandemic has overwhelmed hospitals worldwide and clinical prediction models may assist in timely identification of covid-19 patients at risk for clinical deterioration, i.e. 'early warning'. In this article, we report on the development and validation of a new early warning model that predicts unplanned ICU admission or unexpected death within 24 hours from the moment of prediction, specifically for covid-19 patients. We compared the performance with two well-known and widely used early warning scores (EWSs), i.e. the Modified Early Warning Score (MEWS) [2] and National Early Warning Score (NEWS) [3].

Methods We collected electronic medical record (EMR) data from covid-19 patients admitted to six Dutch hospitals between February 2020 and May 2021. We defined the clinical endpoint as a surrogate of unplanned ICU admission or unexpected death. To examine the added value of including non-linear predictor-outcome relations, we trained both a (linear) logistic regression (LR) and a (non-linear) random forest (RF) model. We included predictors based on patient demographics, vital signs and laboratory test results. We validated the models retrospectively in a 'leave-one-hospital-out' cross-validation (LOHO-CV) procedure. Furthermore, we simulated a prospective validation by splitting all included patients admitted before and after August 1 2020 and simulated as if the models would have been developed based on the data collected until August 2020 and implemented during the remaining study period. Additionally, we examined different strategies for monthly model updating. We evaluated model discrimination and calibration for the proposed models as well as the traditional EWSs, and performed a decision curve analysis [22]. Importance of individual predictors was quantified using SHAP values [13].

Findings In the retrospective validation, the LR model yielded a significant improvement in partial area under the receiver operating curve (pAUC) compared to the traditional EWSs in four of the six included hospitals, and in all hospitals by the RF model. In the simulated prospective validation, significant improvement was shown in two and four hospitals by the LR and RF models, respectively. Without any model updating, both model showed risk overestimation. We proposed a combination of monthly model retraining and hospital-specific re-calibration that could correct for this miscalibration effectively. In the decision curve analysis, the proposed models outperformed the traditional EWS in terms of net benefit (NB) over a wide range of clinically relevant model thresholds.

Interpretation We have derived and validated a new early warning model specifically for covid-19 patients that outperformed traditional EWSs and showed good generalizability over different Duch hospitals. Also, we introduced SpO2-to-O2 ratio as an important marker for disease severity in covid-19 patients. Finally, we showed the importance of repeated model updating when developing medical prediction models in the midst of the covid-19 pandemic and proposed an effective model updating strategy. Future research is needed to validate the model outside the Netherlands.

Funding No specific funding.

Contents

1 Introduction

Disease resulting from infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), or covid-19, has a high mortality rate with deaths predominantly caused by respiratory failure. The pandemic has continued to overwhelm hospital wards worldwide, and clinical prediction models may assist in timely identification of covid-19 patients at risk for clinical deterioration.

Long before the covid-19 outbreak, early warning scores (EWSs) had already been developed for this purpose functioning as the 'tracker' in track-and-trigger systems, also known as 'Rapid Response Systems' (RSSs) [1]. RSSs are widely implemented in every day health care, in which an EW score triggers either an urgent response or an emergency response (i.e., the 'trigger') if a specific threshold is reached (see supplementary figure 1). widely used EWSs are (variations of) the Modified Early Warning Score (MEWS) [2] and the National Early Warning Score (NEWS) [3], which were both designed for the general patient population. An EWS designed specifically for covid-19 patients may improve timely identification of deterioration in this patient group.

Many prognostic models specifically for covid-19 have recently been developed, although the vast majority of these were classified as being at a high risk of bias [4]. Also, while many of these models claim to serve the purpose of identifying patient deterioration in an early stage, most use relatively long or unspecified prediction horizons. For early warning models, a prediction horizon limited to a few days is recommended in the literature [5], as any signs linked to this outcome will probably not be seen for longer than this period. Moreover, a risk for near-term deterioration (e.g. within 24 hours) denotes a more precise prognosis compared to a long term risk of deterioration, making it more actionable for clinicians.

We report on the development and validation of a new EWS specifically for covid-19 patients. We trained a linear and non-linear prediction model using patient demographics, vital signs and laboratory test results, to predict unplanned ICU admission or unexpected death within 24 hours from the moment of prediction. We compared its performance with traditional EWSs retrospectively in six different Dutch hospitals. Additionally, we performed a simulation of prospective validation, comparing the performance of the developed model with traditional EWSs if it had been implemented between the first and second covid-19 'wave' in the Netherlands and examined the added value of monthly model updating.

2 Methods

2.1 Study design and setting

The study was performed in six hospitals across the Netherlands, consisting of two academic hospitals (Erasmus University Medical Center and Leiden University Medical Center) and four teaching hospitals (Maasstad, Haga, Albertschweitzer and Ikazia teaching Hospital). We collected electronic medical record (EMR) data from patients admitted to these hospitals who were hospitalized with covid-19 either proven by a polymerase chain reaction (PCR) test or diagnosed by their treating clinicians based on symptoms, lab and radiology. The period of data collection varied per hospital and ranges between February 2020 and May 2021. For model development and reporting, we followed the TRIPOD guidelines [8].

2.2 Data collection

We handled patients who switched from wards between different hospitals as separate admissions (as pseudonymization did not allow for patient matching). We handled patients who returned to the same hospital for covid-19 related matters, after being sent home first, as separate admissions as well. Patients who were admitted to the intensive care unit (ICU) straight from home or emergency department (ED) were excluded. We collected multiple observation sets, or 'samples', at different time points for every patient, starting at 24 hours after ward admission and adding one every 24 hours until either discharge, ICU admission, or death occurred. For patients who stayed shorter than 24 hours, we collected a sample halfway the stay. These samples served as inputs for model development and validation. Loss to follow-up occurred for patients who were transferred to other hospitals and those who were still admitted at the moment of data collection. In the first situation, we assumed that the clinical endpoint was not reached within 24 hours after hospital transfer, and censored at the moment of transfer. We censored patients who were still admitted at 24 hours before the final observed measurement, consequently excluding still admitted patients who stayed shorter than 24 hours.

2.3 Outcomes

We defined the clinical endpoint as a surrogate outcome of unplanned ICU admission or unexpected death. We classified ICU admission as unplanned if the admission could not have been postponed for longer than 12 hours

without any risk [9]. To define unexpected death, we first (re)defined different levels of limitation of medical treatment (LOMT) [10] as follows: Code A for 'full active care', Code B for 'do not perform cardiopulmonary resuscitation', Code C as 'do not admit to ICU' and Code D as 'only palliative care'. We classified death on ward as unexpected if Code A or B applied to the patient at the moment of death. As early warning is only useful for patient where ICU admission is a treatment option, we excluded patients with LOMT code C or D.

2.4 Independent predictor variables

To select predictors, we considered evidence in previous literature and availability. A priori, we selected a reduced set of candidate predictors which were identified as clinically important in covid-19 cohorts by Knight and colleagues [11]. To this list, we added supplementary oxygen (O2) and the SpO2-to-O2 ratio, as these are known to be good indicators for disease severity specifically for respiratory diseases like covid-19. O2 was added as a dichotomous (yes/no) and as a continuous (Liter/minute) predictor, where we only considered O2 measurements within the preceding eight hours of the moment of sampling. The SpO2-to-O2 ratios were constructed from the most recent pair of simultaneously measured values for SpO2 and O2. Based on other recent covid-19 literature, we added 8 extra laboratory measures (eosinophil count, monocyte count, red cell distribution width, D-Dimer, L-6, ferritin and neutrophil-to-lymphocyte ratio) as candidate predictors. To correct for time dependency of some included predictors, as well as to model the influence of the length of hospitalization on the disease severity, we added the current length of stay on the ward as a predictor. Finally, to model the effect of changes in frequently measured vital signs, we added the signed difference between the first and second most recently measured value for a selection of predictors (SpO2, HR, SBP, RR, Temp and SpO2/O2) in a 24-hour sliding window (before normalization). To give an overview of the frequency in which different predictors occurred in the EMR, we calculated daily entry densities (i.e., fractions of non-empty daily measurements) for each predictor and for all patients individually. To avoid the need for too much imputation, we excluded predictors with a total entry density ≤ 0.2. To handle missing values, we used an iterative imputation algorithm (Scikit-learn IterativeImputer [12]) with a single imputation round. Here, each missing predictor is estimated based on all the available predictors using Bayesian ridge regression in an iterative fashion. To normalize the samples, we centered and scaled each predictor by the standard deviation.

2.5 Model development

We labelled the samples as 'event samples' if unplanned ICU admission or unexpected death occurred within 24 hours from the moment of prediction, and 'non-event samples' otherwise. We trained classification models to discriminate between event and non-event samples. To examine the added value of including non-linear predictor-outcome relations, we trained both a (linear) logistic regression (LR) and a (non-linear) random forest (RF) model. For the LR model, we used l2 regularization. We optimized the regularization strength (λ) of the LR model and the 'maximum tree depth' and 'max features' hyperparameters of the RF model using an exhaustive gridsearch in a stratified 10-fold cross-validation procedure optimizing the area under the receiver operating curve (AUC). Supplementary table 1 shows the hyperparameter grids that were searched.

To obtain interpretability for the developed models, we calculated the impacts of individual predictors on risk output by SHAP values. A SHAP value is a model-agnostic representation of predictor importance, where the impact of each predictor is represented using Shapley values inspired by cooperative game theory [13]. We calculated SHAP values based on a LR and RF model trained on data from the comlete cohort.

2.6 Model validation

2.6.1 Retrospective validation

We retrospectively validated the models in a 'leave-one-hospital-out' (LOHO) cross-validation procedure. Here, in each round, patients from five of the six hospitals formed the development set. First, we fitted the imputation algorithm based on the development set and used it to impute all missing values. Then, we normalized all samples, optimized the model hyperparameters (as described in section 2.5) and trained the model using the development set. Finally, we validated the trained model on the left-out hospital (see figure 1a). We repeated this process until each hospital served as the validation set once, resulting in six LR and six RF models.

2.6.2 Simulated prospective validation

Additionally, we simulated the situation as if the model had been implemented halfway the covid-19 outbreak and implemented during the remaining study period. Therefore, we split all included patients into two cohorts:

patients admitted before and after August 1 2020. We refer to these cohorts as the 'wave 1 cohort' and 'wave 2 cohort', respectively, as these periods coincide with the first and second covid-19 'waves' in the Netherlands.

To simulate the situation if a model had been developed and implemented without any model updating, we trained both an LR and RF model on the wave 1 cohort and validated it on the wave 2 cohort, referred to as the 'baseline models'. We examined the effectiveness of different strategies for monthly model updating. Here, we applied different techniques for monthly model updating and hospital-specific re-calibration (and combinations of these). More details on this can be found in Appendix E. We selected a combination of monthly model retraining and hospital-specific re-calibration using isotonic regression [14] as the most effective strategy, referred to as 'monthly model updating' in the remainer of the article. Figure 1b visualizes this simulated prospective validation procedure.

2.6.3 Evaluation metrics

To evaluate model discrimination considering a clinically relevant range of model thresholds, we determined the partial area under the receiver operating curve (pAUC) [15] between a false positive rate (FPR) between 0 and 0.33 as a primary endpoint. The PPV (or precision) is suggested as a useful metric to estimate clinical workload when implementing EWSs [16] and therefore, we evaluated the area under the precision-recall curve (AUPRC) [17] as a secondary outcome, which is a single number summary of the PPVs for a range of model thresholds. Finally, to enable comparison of the proposed model with any other model in literature, we calculated the widely used (complete) area under the ROC curve (AUC).

To calculate uncertainties around the different metrics, we calculated bootstrap percentile confidence intervals [18] for the pAUC and AUC and binomial confidence intervals [17] for the AUPRCs. To test the significance of the improvements in discriminative performance compared to traditional EWSs, we used the bootstrapping procedure described in [15]. For more details on uncertainty calculation and significance testing, we refer to Appendix C and D.

We evaluated model calibration in the 'weak' and 'moderate' sense [19]. We evaluated model calibration in the weak sense by calculating the calibration intercept and slope [20] and in the moderate sense by plotting loess smoothed flexible calibration curves.

Additionally, we performed a decision curve analysis (DCA) [22] plotting the net benefit (NB) over a range of model thresholds. The NB is calculated by the proportion of true positives (i.e. finding patients who are deteriorating) and false positives (i.e. false alarms), where the latter is weighted by the 'exchange rate', defined as the odds ratio at a certain model threshold (see equation 1). We standardized the NB for validation in each hospital separately by dividing the NB by the proportion of event samples (which is the maximum NB).

$$\text{Net benefit} = \frac{\text{True positives}}{N} - \frac{\text{False positives}}{N} \times \frac{p_t}{1 - p_t} \tag{1}$$

The DCA should be performed in a clinically relevant range, which depends on how many false alarms one is willing to invest in order to find one case (true positive) when using the early warning model. As a physician most likely wants to trigger an emergency response if the probability of deterioration exceeds 10%, we chose to show the DCA results between 0 and 10%.

2.6.4 Comparison with existing Early Warning Scores

We compared the proposed models with two widely used EWSs: the modified early warning score (MEWS) [2] and the national early warning score (NEWS) [3]. These scores are calculated based on respiratory rate, systolic blood pressure, heart rate, temperature and level of consciousness using the AVPU system. The NEWS additionally requires information about oxygen saturation and supplemental oxygen. In case of missing values, we used the same imputed values as used for the development of the LR and RF models, except for the AVPU score. As patients on the ward are normally alert, we assumed an AVPU score A ('Alert') if no information was available.

As the traditional EWSs output a discrete score rather than a probability, these could not directly be compared to the proposed models in terms of model calibration and in the decision curve analysis (DCA). To evaluate calibration, we plotted a discrete calibration curve for the traditional EWSs. For the DCA, we fitted two extra logistic regression models with either the MEWS or NEWS score as the only predictor, and applied the output probabilities to calculate the net benefit (NB) for the MEWS and NEWS.

3 Results

We collected EMR data from cpvod-19 patients admitted to six Dutch hospitals between February 2020 until May 2020. Figure 2 shows the inclusion of admissions. After excluding patients who were admitted to the ICU immediately and patients with limitation-of-medical-treatment (LOMT) code C or D, we included 3 674 admissions. Table 1 shows the pathway and population characteristics for all admissions (tables for individual hospitals can be found in Appendix B). Unplanned ICU admission occurred in 605, unexpected death in three and hospital transfer in 538 admissions. To give an overview of the role of different predictors during the 24 hours preceding patient deterioration, supplementary figure 2 (Appendix A) shows the cumulative distributions for all candidate predictors based on samples taken within 24 hours before unplanned ICU admission or unexpected death (i.e., the 'event samples') compared to all other ('non-event') samples.

Supplementary figures 3 and 4 (Appendix A) visualize the daily data availability by boxplots showing the distributions of entry densities (i.e., fractions of non-empty daily measurements) for all candidate predictors in the complete cohort, as well as for individual hospitals. Based on availability (total entry density ≥ 0.2), we included two patient demographics, six clinical signs, four bedside investigations, twelve laboratory measures and the current length of stay on the ward in the model, resulting in a total of 32 predictors (see supplementary table 2, Appendix B). All predictors were continuous, except for one categorical predictor (AVPU) and two dichotomous predictors (sex and supplemental oxygen).

3.1 Retrospective validation

The pAUCs (95% CI) yielded by the traditional EWSs and the proposed models in the retrospective validation are visualized in figure 3a. As the NEWS yielded higher performances than the MEWS in all the hospitals, we tested the improvement in performance by the proposed models for significance only compared to the NEWS. In terms of pAUC, the LR model outperformed the NEWS significantly in five out of six hospitals and the RF model in all hospitals. The AUPRCs and AUCs yielded by the traditional EWSs and the proposed models are visualized in supplementary figure 5 (Appendix A). Also considering these metrics, the LR and RF models outperformed the NEWS significantly in most hospitals.

Combined predictions of the six LR models yielded a calibration intercept of -0.23 (-0.32;-0.14) and slope of 0.64 (0.59;0.69), suggesting risk overestimation and too extreme risk estimates. Combined predictions of the six RF models yielded a calibration intercept of -0.06 (-0.14;0.02, 95% CI) and slope of 1.60 (1.50;1.70, 95% CI), suggesting slight risk overestimation, but too moderate risk estimations. The corresponding loess smoothed calibration curves are plotted in figure 3c. Hospital-specific calibration curves can be found in supplementary figure 6 and discrete calibration curves yielded by the traditional EWSs can be found in supplementary figure 7 (Appendix A).

Figure 3b shows the results of the decision curve analysis (DCA). In five out of six hospitals, both LR and RF model show a clear improvement in net benefit (NB) compared to traditional EWSs over the entire clinically relevant probability range.

The SHAP values for interpretability of the LR and RF models are summarized in figure 3d and 3e, respectively. The SpO2-to-O2 ratio, respiratory rate and the ward length-of-stay were among the top five highest ranked predictors (based on mean SHAP magnitude) for both the LR and RF model.

3.2 Simulated prospective validation

The pAUCs (95% CI) yielded by the traditional EWSs, baseline models and the monthly updated models in the simulated prospective validation are visualized in figure 4a. Again, as the NEWS yielded higher performances than the MEWS in all the hospitals, we tested the improvement in performance by the monthly updated models for significance only compared to the NEWS. In terms of pAUC, the monthly updated LR model outperformed the NEWS significantly in two out of six hospitals and in four out of six hospitals by the monthy updated RF model. The AUPRCs and AUCs (95% CI) are visualized in supplementary figure 8 (Appendix A), showing similar improvements of the proposed models compared to the NEWS. We observed no significant difference in terms of pAUC, AUPRC or AUC between the baseline and monthly updated LR and RF models.

The baseline LR model yielded a calibration intercept of -0.15 (-0.21;-0.02) and slope of 0.66 (0.60;0.72), suggesting risk overestimation and too extreme risk estimates. With monthly model updating, this improved to a calibration intercept of -0.05 (-0.16;0.06) and slope of 0.83 (0.81;0.86). The baseline RF model yielded a calibration intercept of -0.35 (-0.45;-0.24) and slope of 1.65 (1.53;1.77), suggesting risk overestimation and too moderate risk estimates. With monthly model updating, this improved to a calibration intercept of -0.19 (-0.31;-0.07) and slope of 0.87 (0.80;0.94). The corresponding calibration curves of the baseline and monthly updated models are shown in figure 4c and 4d, respectively. Hospital-specific calibration curve (of both baseline

and monthly updated models) can found in Appendix E. The discrete calibration curves of the traditional EWSs can be found in supplementary figure 9.

Figure 4b shows the results of the decision curve analysis (DCA). Again, in five out of six hospitals both LR and RF model show a clear improvement in net benefit (NB) compared to traditional EWSs.

Figure 4e shows the receiver operating characteristic curves with 95% confidence intervals resulting from the predictions by the NEWS and the monthly updated RF model in all hospitals in the wave 2 cohort combined. We placed two landmarks for a NEW score of 5 and 7, which are recommended by the Royal College of Physicians [23] to be used as trigger thresholds for an urgent and emergency response, respectively. The vertical difference represents the potential improvement in model sensitivity, and the horizontal difference the potential reduction in false alarms, if the RF model had been implemented with monthly model updating during the wave 2 period.

4 Discussion

4.1 Principal findings

We have developed and validated a linear and non-linear early warning model specifically for covid-19 patients in a retrospective and simulated prospective validation cohort study, based on 3674 admissions from six different Dutch hospitals. In the retrospective validation, both models showed significant improvement in terms of pAUC compared to the traditional EWSs in the majority of the included hospitals, although the RF model showed moderate calibration (intercept$<$0, slope$<$1). In the decision curve analysis, both models showed improvement in net benefit compared to the traditional EWSs except for one of the hospitals, despite of an improvement in model discrimination. This can be explained by severe risk overestimation by both LR and RF model in this hospital (supplementary figure 13c). This overestimation may be explained by local differences in protocols for ICU admission.

In the simulated prospective validation, significant improvement in terms of pAUC compared to the traditional EWSs was shown in only two out of six by the LR model and in four out of six hospitals by the RF model (compared to four and six out of six hospitals in the retrospective validation). This may be partly explained by a better performance of the NEWS after than before August 2020 and by the fact that sample sizes of the validation sets were smaller, resulting in wider confidence intervals.

Both the baseline models show risk overestimation (calibration intercept$<$0), which may be explained by improvement of covid-19 care during the pandemic. For instance, the RECOVERY trial [24] published in July 2020 enabled physicians to treat covid-19 patients more effectively with the wide-spread use of Dexamethasone. Therefore, the a priori risk of deterioration for covid-19 patients lowered and could have caused the risk overestimation of baseline models (which were trained on data collected until July 2020). We showed that monthly model updating could correct for this miscalibration effectively.

In the decision curve analysis, both models showed improvement in net benefit compared to the traditional EWSs except for one of the hospitals, despite the correction for miscalibration. This can be explained by the lack of improvement in model discrimination compared to the NEWS, as shown in figure 4a.

The RF showed better performance than the LR model in terms of pAUC in both retrospective and simulated prospective validations. Although the RF model showed worse calibration compared to the LR model in the retrospective validation, the simulated prospective validation showed that this could be corrected for effectively by monthly model updating. The RF outperforming the LR model may be explained by its ability to model non-linear predictor-outcome relations, which is recommended for early warning models in the literature [5].

We added the SpO2-to-O2 ratio as a predictor, which was shown to be important in both the LR and RF model (see figure 3d and 3e). To our knowledge, this is the first prognostic model for covid-19 patients that includes this predictor. Another notably high ranked predictor was the ward length-of-stay (LOS). This can be explained by the relatively short stays on the ward among patients who experienced an unplanned ICU admission or unexpected death, which we observed consistently in the six included hospitals (see supplementary figure 2).

As shown in figure 4e, a NEW score of 5 yields a false positive rate (FPR) of approximately 20%. Given the extremely low prevalence of the clinical endpoint, this means that roughly every fifth risk assessment using the NEWS will trigger a false alarm, not surprisingly, ultimately leading to alarm fatigue [25]. The ROC curves show that the FPR could have been reduced significantly by at least a factor two, while maintaining the same sensitivity, had the RF model been implemented (and monthly updated) during the wave 2 period. For the emergency response, the model sensitivity could have been improved by roughly 10 percent at the same rate of false alarms.

4.2 Comparison with conventional EWSs

The proposed models in this study outperformed traditional EWSs (MEWS[2] and NEWS[3]) in both retrospective and simulated prospective validation in most included hospitals. This improvement may be explained by several factors. First, we validated the models for the early detection of unplanned ICU admission and unexpected death among covid-19 patients, while the traditional EWSs were designed to detect unplanned ICU admission, cardiac arrest or death in the general patient population on the ward. Second, we make use of continuous variables, while the traditional EWSs categorise predictors by binning the continuous variables and assigning a discrete score for each bin. Finally, the predictors included in the traditional EWSs consist of a small number of clinical signs. We also included predictors based on patient demographics, bedside investigations, laboratory measures and dynamics of clinical signs.

On the other hand, a big advantage of the traditional EWSs is their simplicity, enabling health care workers to calculate the score easily at the bedside. The clinical usage of a model like we presented in this study requires an app or implementation in the EMR, and (as we showed here) needs to be updated frequently. Whether these costs outweigh the potential decrease in false alarms or increase in model sensitivity as shown in figure 4e, remains open for discussion.

4.3 Clinically relevant model evaluation

Model discrimination for medical prediction models is typically quantified by the the area under the receiver operating curve (AUC). We chose to use the partial AUC (pAUC) between false positive rates of 0 and 0.33, as we argue that an FPR of 0.33 or higher would cause too many false alarms. Also, earlier large-scale retrospective studies for the NEWS [16, 26] have shown that the recommended trigger thresholds for an urgent or emergency response (i.e. a NEWS of 5 or 7 [23]) are located on the ROC curve at FPRs lower than 0.33. Thus, the clinically relevant region of the ROC curve for the NEWS is at FPRs < 0.33 and therefore, we argue that it should be evaluated in this region as well.

4.4 Comparison with other studies

Other recent studies [6, 7] reported models with similar endpoints and also use a prediction horizon of 24 hours. However, both models were only evaluated in a single center, did not evaluate model calibration and did not compare the performance of the proposed models to traditional EWSs.

As early warning models are not useful for patients for who ICU admission is not an option, i.e. those with a limitation of medical treatment (LOMT) code C or D, we chose to exclude these patients and evaluate the model for unexpected death. In the literature [27, 28], patients with LOMT code B (i.e., 'do not perform cardiopulmonary resuscitation') are often also excluded. We argue that a patient with code B (thus, where ICU admission was a treatment option), who deceased on the ward, should be classified as an unexpected death as well and therefore not be excluded in early warning validation studies.

4.5 Strengths and limitations of this study

The inclusion of six different hospitals allowed extensive external validation of the models, which enabled us to show model generalizability in the Netherlands.

Also, it is common to validate medical prediction models in a retrospective fashion, which provides no guarantee that the model will still perform well on patients admitted after model development. In this study, we simulated a prospective validation from August 2020 until May 2021, which provided temporal validation of the model.

Finally, we showed that a model developed in the midst of a covid-19 pandemic would have shown miscalibration when implemented during the remaining study period. This underwrites the importance of model updating for medical prediction models, especially in a rapidly changing situation like the covid-19 pandemic. Also, we proposed a monthly model updating strategy that could correct for the miscalibration effectively.

This study has several limitations. First, we had no estimate of the number of patients whose clinical course was positively influenced by any clinical intervention and who, as a result, were not admitted to the ICU or deceased. Therefore, we may have labelled some samples as 'non events' falsely because signs of deterioration were actually present.

Second, we made the assumption that patients who got transferred to other hospitals did not meet the clinical endpoint within 24 hours after transfer. Also, we censored patients who were transferred or still admitted. For still admitted patients, we can assume non-informative censoring. For transferred patients, informative censoring

may have introduced a bias. While several strategies are proposed in literature how to handle such competing risks situations [29], we chose not to implement these as we assumed the potential bias to be small.

Third, as the RF model is more complex than the LR model, it is also more prone to overfit. In our study, we show that the RF model generalizes well for different hospitals in the Netherlands. However, it is possible that the model is overfitting to Dutch covid-19 patients and/or practices in the Dutch health care system. Therefore, further external validation of the model is needed to check its generalizability elsewhere in the world.

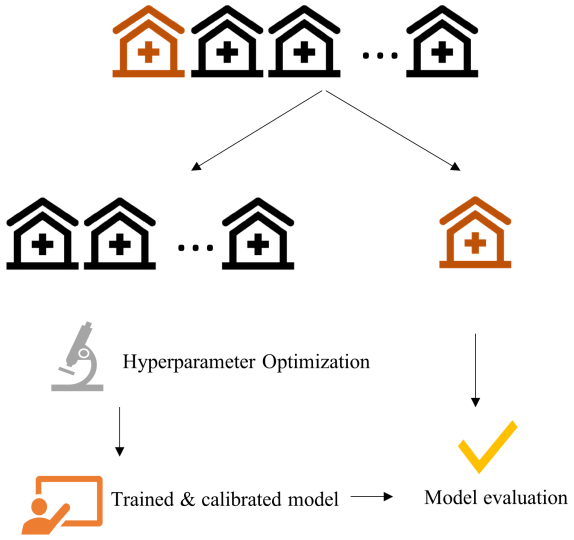
Fourth, we used repeated observations from individual patients in the analysis, resulting in dependency between the samples. As the methods we used to measure uncertainty around the performance metrics assume independent and identically distributed (IID) samples, this may have led to underestimation of the uncertainties. Therefore, the (significant) improvements found compared to the traditional EWSs may turn out to be too optimistic.

Finally, external validation is ideally performed by independent researchers, as choices that we made for the modelling could have been influenced by the data that was already available for us. For the same reason, the simulated prospective validation does not provide as much evidence as an actual prospective could have offered.

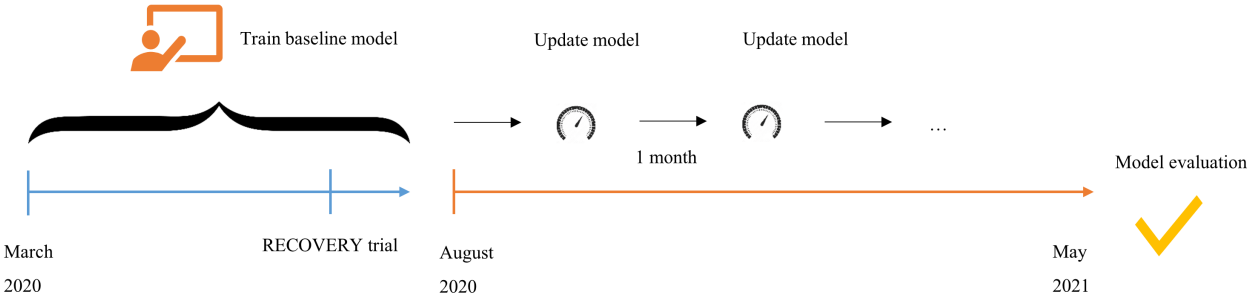
5 Conclusion

We have derived and validated a new early warning model specifically for covid-19 patients, which showed significant improvements compared to the traditional EWSs in a retrospective and simulated prospective validation study. Also, introduced a new important marker for disease severity in covid-19 patients, i.e. the SpO2-to-O2 ratio, and we show the importance of repeated model updating when dealing with a rapidly changing situation in the covid-19 pandemic. Future research is needed to validate the model outside the Netherlands.

6 Figures



(a) Leave-one-hospital-out (LOHO) cross-validation procedure for retrospective validation.



(b) Monthly model updating procedure for the simulated prospective validation. The RECOVERY trial [24], which initiated wide-spread use of Dexamethasone, was published on July 17, 2020

Figure 1: Model evaluation procedures.

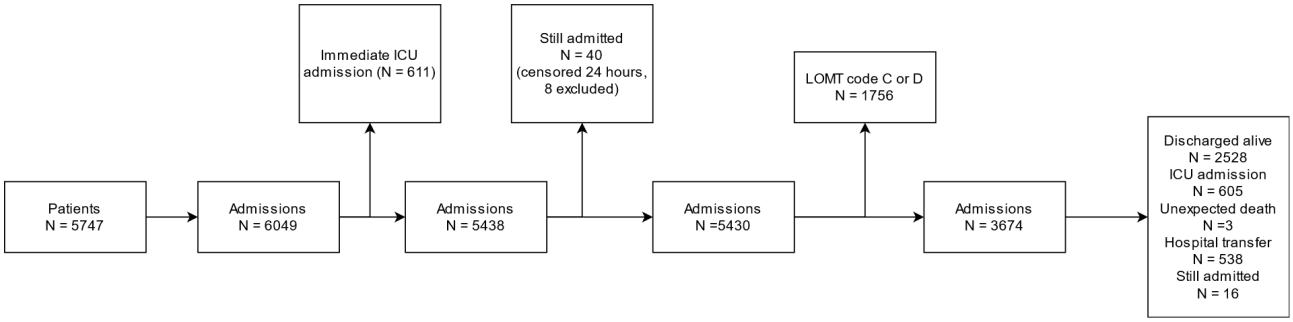
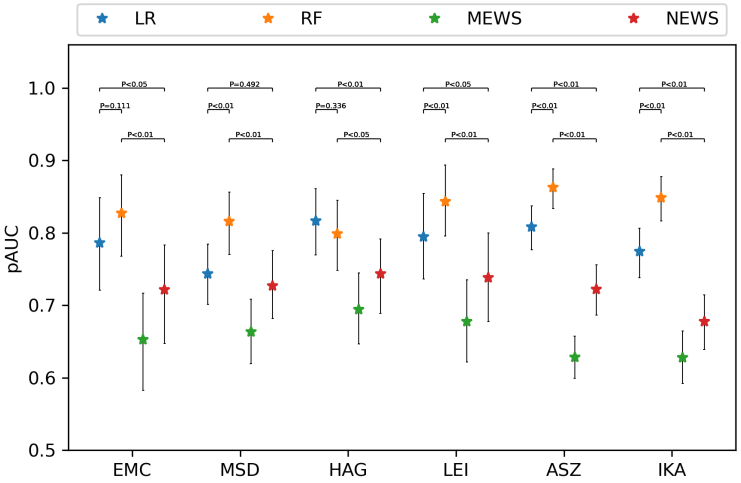
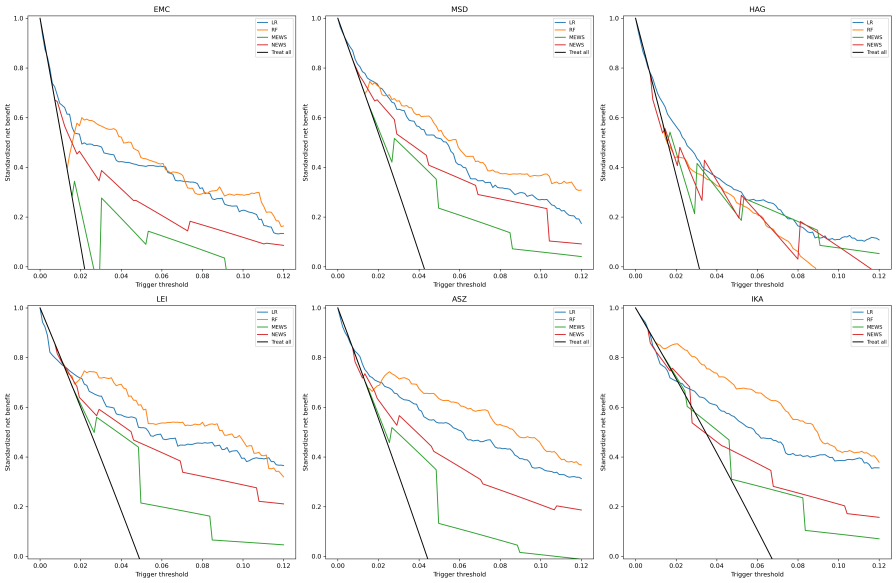


Figure 2: Flowchart of patient inclusion.

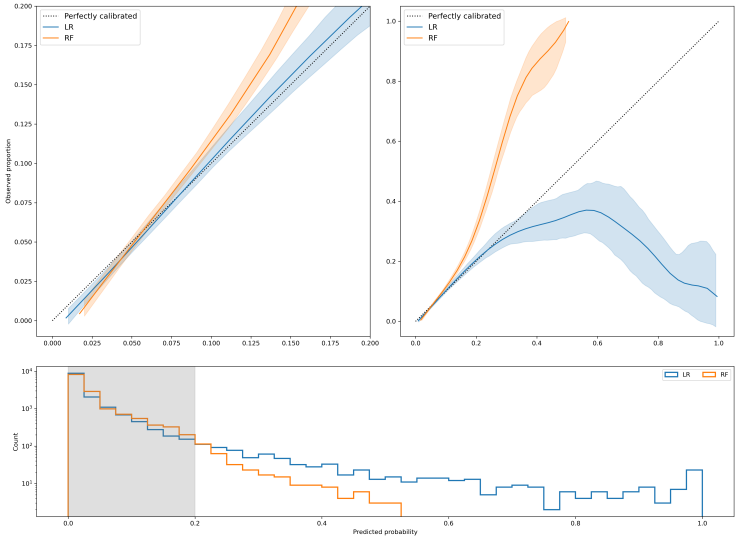


(a) Model discrimination in terms of partial area under the ROC curve (pAUC). P values are shown resulting from significance tests for difference between LR and NEWS (upper bar), LR and RF (middle bar) and between RF and NEWS (lower bar).

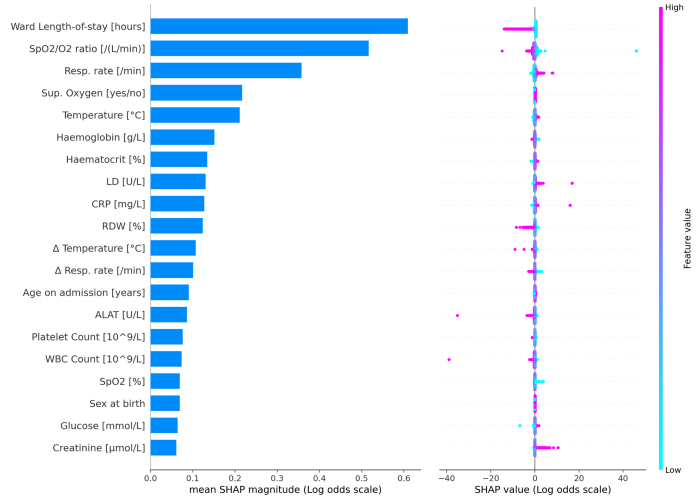


(b) Decision curve analysis (DCA) results. The standardized net benefit (NB) is plotted over a range of clinically relevant probability thresholds. The 'Treat all' line indicates the NB if an urgent or emergency response would always be triggered.

Figure 3: Figure continued on next page.

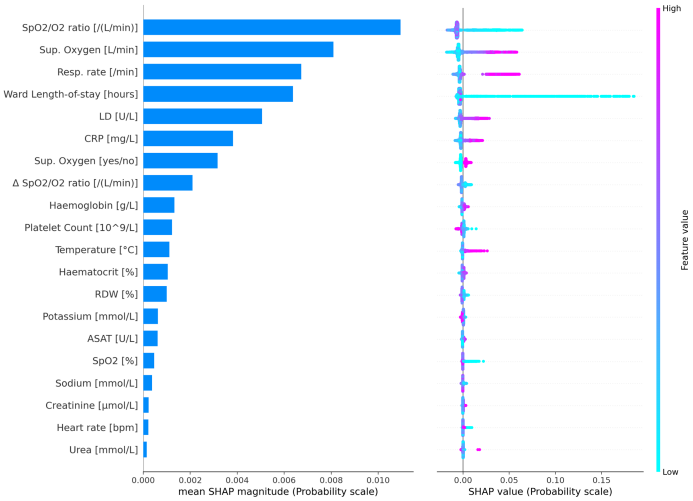


(c) Loess smoothed flexible calibration curves for the logistic regression (LR) and random forest (RF) models. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions. Shaded areas around the curves represent the 95%CIs.



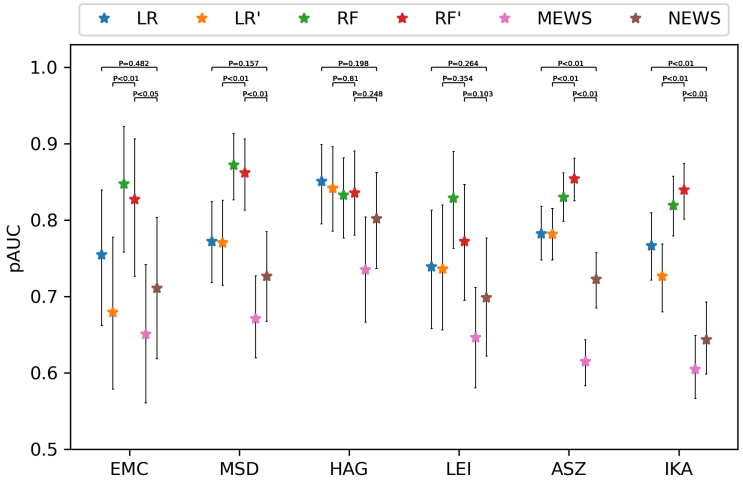
(d) (left) Bar plots for mean SHAP magnitude (log-odds scale) constructed from the logistic regression model. (right) Summary plot, where each SHAP value is represented by a single dot on each predictor row. LOS = ward length-of-stay, Δ=signed difference in a 24 hour sliding window. WBC = White Bloodcell Count, LD = Lactate dehydrogenase, CRP = C-reactive protein.

Figure 3: Figure continued on next page.



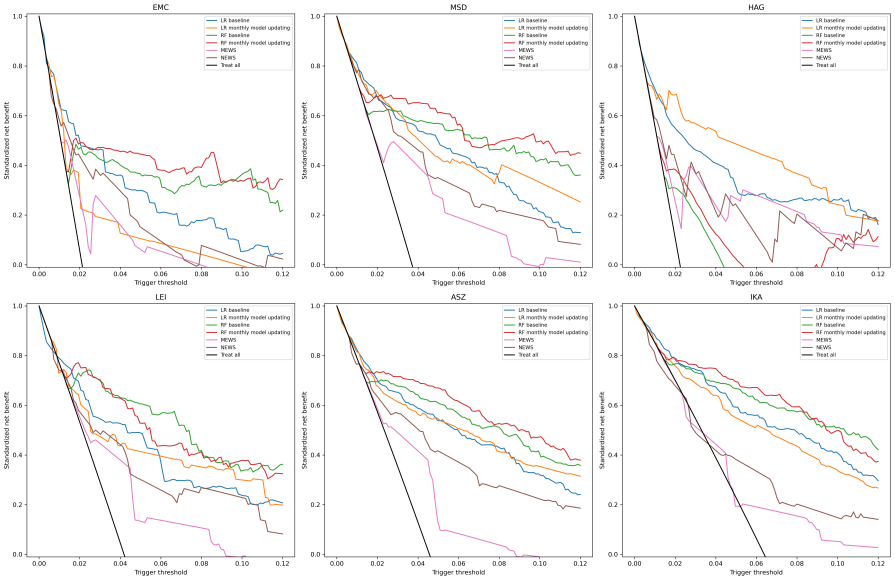
(e) (left) Bar plots for mean SHAP magnitude (probability scale) constructed from the random forest model. (right) Summary plot, where each SHAP value is represented by a single dot on each predictor row. LOS = ward length-of-stay, Δ=signed difference in a 24 hour sliding window. WBC = White Bloodcell Count, LD = Lactate dehydrogenase, CRP = C-reactive protein.

Figure 3: Results of the proposed models and traditional EWSs for retrospective validation. The SHAP values in figure d and e are based on the extra models trained on the complete cohort. EMC = Erasmus University medical Center, MSD = Maasstad Teaching hospital, HAG = Haga Teaching hospital, LEI = Leiden University Medical Center, ASZ = Albertschweitzer Teaching hospital, IKA = Ikazia Teaching hospital.

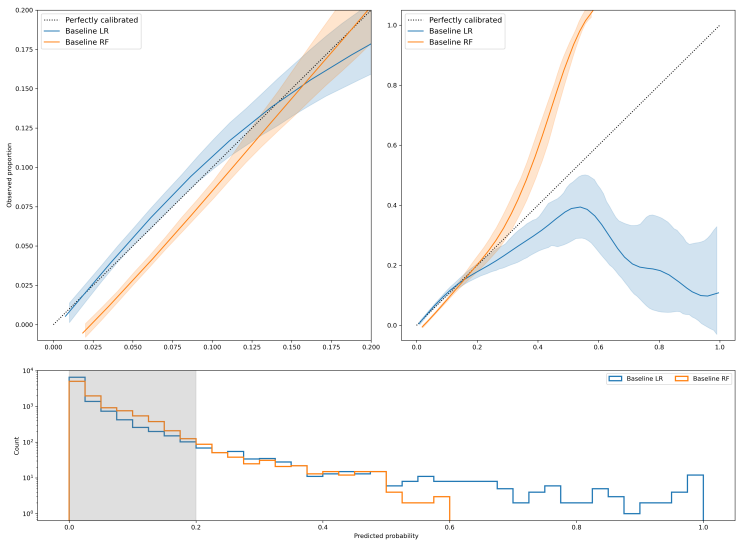


(a) Model discrimination in terms of partial area under the ROC curve (pAUC). LR' = LR model with monthly model updating, RF' = RF model with monthly model updating. P values are shown resulting from significance tests for difference between LR' and NEWS (upper bar), LR' and RF' (middle bar) and between RF' and NEWS (lower bar).

Figure 4: Figure continued on next page.

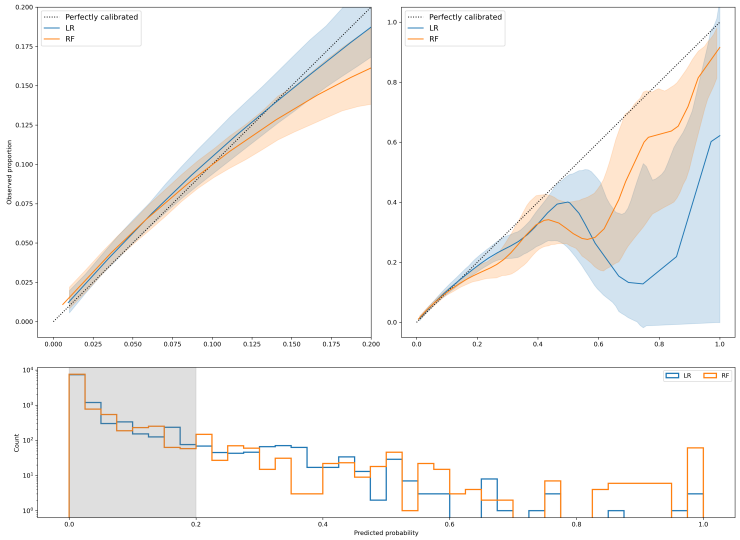


(b) Decision curve analysis (DCA) results. The standardized net benefit (NB) is plotted over a range of clinically relevant probability thresholds. The 'Treat all' line indicates the NB if an urgent or emergency response would always be triggered.

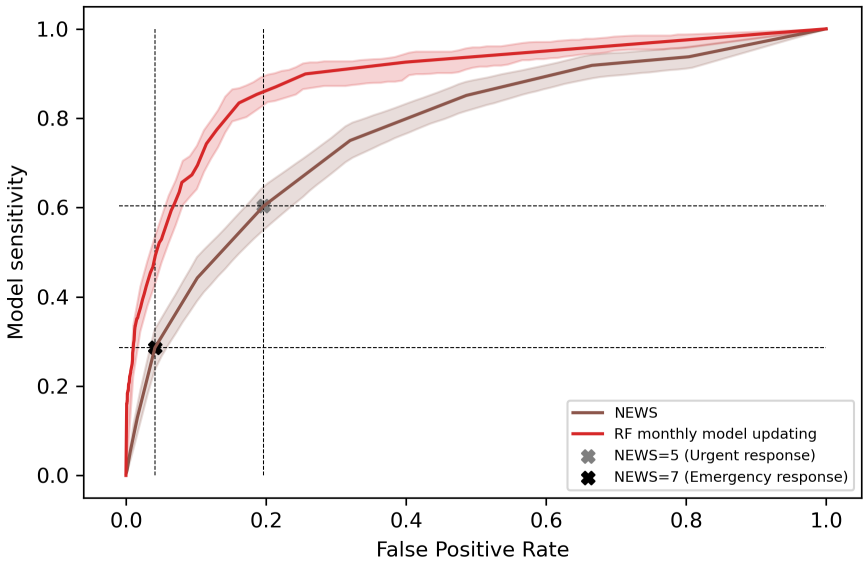


(c) Loess smoothed flexible calibration curves for the baseline logistic regression (LR) and random forest (RF) models. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions. Shaded areas around the curves represent the 95%CIs.

Figure 4: Figure continued on next page.



(d) Loess smoothed flexible calibration curves for the monthly updated logistic regression (LR) and random forest (RF) models. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions. Shaded areas around the curves represent the 95%CIs.



(e) Receiver operating characteristic curve (95% CI) of the NEWS and the monthly updated RF model.

Figure 4: Results of the proposed models and traditional EWSs for the simulated prosspective validation. Shaded areas around the curves represent the 95% bootstrap percentile confidence intervals, calculated in a 2000 sample bootstrapping procedure.

7 Tables

Table 1: Pathway and population characteristics of the patients with limitation of medical treatment (LOMT) code A or B and with LOMT code C or D.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=2528)	Unplanned ICU admission (N=605)	Unexpected death (N=3)	Hospital transfer (N=538)	Total (N=3658*)
Sex, male n(%)	1307.0 (54.8)	380.0 (65.6)	2.0 (66.7)	293.0 (56.1)	1982.0 (56.8)
Age, years					
med (IQR)	60.0 (51.0-70.0)	63.0 (55.0-70.0)	76.0 (74.5-77.5)	60.0 (54.0-69.0)	61.0 (52.0-70.0)
mean (SD)	59.3 (14.4)	61.2 (11.6)	76.0 (4.2)	59.9 (11.6)	59.7 (13.6)
Ward LOS, days					
med (IQR)	3.6 (1.9-6.3)	1.9 (0.8-3.6)	6.9 (3.6-7.6)	1.0 (0.7-2.0)	3.6 (1.7-7.6)
mean (SD)	5.2 (6.5)	3.0 (4.2)	5.1 (3.5)	1.8 (2.6)	6.8 (10.2)

References

[1] DeVita, M. A., Bellomo, R., Hillman, K., Kellum, J., Rotondi, A., Teres, D., Auerbach, A., Chen, W. J., Duncan, K., Kenward, G., Bell, M., Buist, M., Chen, J., Bion, J., Kirby, A., Lighthall, G., Ovreveit, J., Braithwaite, R. S., Gosbee, J., Milbrandt, E., Peberdy, M., Savitz, L., Young, L., and Galhotra, S. *Critical Care Medicine* **34**(9), 2463–2478 (2006).

[2] Subbe, C. P., Kruger, M., Rutherford, P., and Gemmel, L. *QJM - Monthly Journal of the Association of Physicians* **94**(10), 521–526 (2001).

[3] Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., and Featherstone, P. I. *Resuscitation* **84**(4), 465–470 (2013).

[4] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M., Damen, J. A., Debray, T. P., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Andaur Navarro, C. L., Reitsma, J. B., Sergeant, J. C., Shi, C., Skoetz, N., Smits, L. J., Snell, K. I., Sperrin, M., Spijker, R., Steyerberg, E. W., Takada, T., Van Kuijk, S. M., Van Royen, F. S., Wallisch, C., Hooft, L., Moons, K. G., and Van Smeden, M. *The BMJ* **369** apr (2020).

[5] Gerry, S., Bonnici, T., Birks, J., Kirtley, S., Virdee, P. S., Watkinson, P. J., and Collins, G. S. *BMJ (Clinical research ed.)* **369**, m1501 (2020).

[6] Cheng, F. Y., Joshi, H., Tandon, P., Freeman, R., Reich, D. L., Mazumdar, M., Kohli-seth, R., Levin, M., Timsina, P., and Kia, A. *J Clin Med* **9**(6) (2020).

[7] Douville, N. J., Douville, C. B., Mentz, G., Mathis, M. R., Pancaro, C., Tremper, K. K., and Engoren, M. *British Journal of Anaesthesia* (September) (2021).

[8] Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. *BMC Medicine* **13**(1), 1–10 (2015).

[9] Stichting-NICE. Technical report, (2021).

[10] Brunsveld-Reinders, A. H., Ludikhuize, J., Dijkgraaf, M. G., Arbous, M. S., de Jonge, E., van Putten, M. A., Adams, R., de Maaijer, P. F., de Rooij, S. E., Kerkhoven, C., Braber, A., Schoonderbeek, F. J., Kors, B. M., Sep, D. P., Vermeijden, J. W., Fikkers, B. G., Tangkau, P., van der Weijden, P. K., Koenders, S., Meertens, M., Brunsveld-Reinders, A. H., Hoeksema, M., and Smorenburg, S. M. *Critical Care* **20**(1), 1–7 (2016).

[11] Knight, S. R., Ho, A., and Pius, R. *BMJ* **2**(September) (2020).

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. *Journal of machine learning research* **12**(Oct), 2825–2830 (2011).

[13] Lundberg, S. M. and Lee, S.-i. (Section 2), 1–10 (2017).

[14] Niculescu-Mizil, A. and Caruana, R. *Proceedings of the 22nd international conference on Machine learning* (2005).

[15] McClish, D. K. *Medical Decision Making* **9**(3), 190–195 (1989).

[16] Haegdorens, F., Monsieurs, K. G., De Meester, K., and Van Bogaert, P. *Journal of Clinical Nursing* **29**(23-24), 4594–4603 (2020).

[17] Boyd, K., Eng, K. H., and Page, C. D. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science*, volume 8190. Springer, Berlin, Heidelberg, (2013).

[18] Qin, G. and Hotilovac, L. *Statistical Methods in Medical Research* **17**(2), 207–221 (2008).

[19] Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., MacAskill, P., McLernon, D. J., Moons, K. G., Steyerberg, E. W., Van Calster, B., Van Smeden, M., and Vickers, A. J. *BMC Medicine* **17**(1), 1–7 (2019).

[20] Cox, D. R. *Miscellanea* (1953), 562–565 (1958).

[21] van Calster, B., Nieboer, D., Vergouwe, Y., Cock, B. D., Pencina, M., and Steyerberg, E. *Journal of clinical epidemiology* **74**, 167–76 (2016).

[22] Vickers, A. J., van Calster, B., and Steyerberg, E. W. *Diagnostic and Prognostic Research* **3**(1), 1–8 (2019).

[23] RCOP. *National Early Warning Score (NEWS) - Standardising the assessment of acute-illness severity in the NHS. Report of a working party.* Number July. (2012).

[24] Horby, P., Lim, W. S., Emberson, J. R., and Mafham, M. *New England Journal of Medicine* , 1–11 (2020).

[25] Weeks, B. Y. K., Timalonis, J., and Donovan, L. *Nursing* **41**(5), 59–63 (2021).

[26] Smith, G. B., Prytherch, D. R., Jarvis, S., Kovacs, C., Meredith, P., Schmidt, P. E., and Briggs, J. *Critical Care Medicine* **44**(12), 2171–2181 (2016).

[27] Haegdorens, F., Van Bogaert, P., Roelant, E., De Meester, K., Misselyn, M., Wouters, K., and Monsieurs, K. G. *Resuscitation* **129**(March), 127–134 (2018).

[28] Hillman, K. *Lancet* **365**(9477), 2091–2097 (2005).

[29] van Geloven, N., Swanson, S. A., Ramspek, C. L., Luijken, K., van Diepen, M., Morris, T. P., Groenwold, R. H., van Houwelingen, H. C., Putter, H., and le Cessie, S. *European Journal of Epidemiology* **35**(7), 619–630 (2020).

The ability of new prediction models to discriminate covid-19 patients at risk of unplanned intensive care unit admission and unexpected death

Supplementary material

J.M. Smit

July 1, 2021

Contents

A Supplementary figures

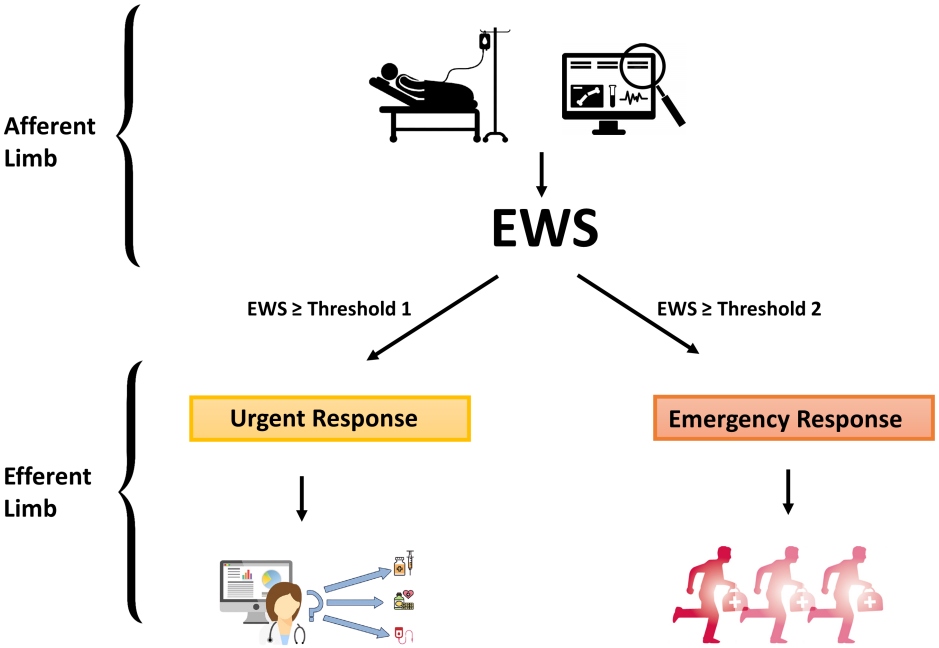


Figure 1: Visual representation of the Rapid Response System. The early warning model forms the afferent limb.

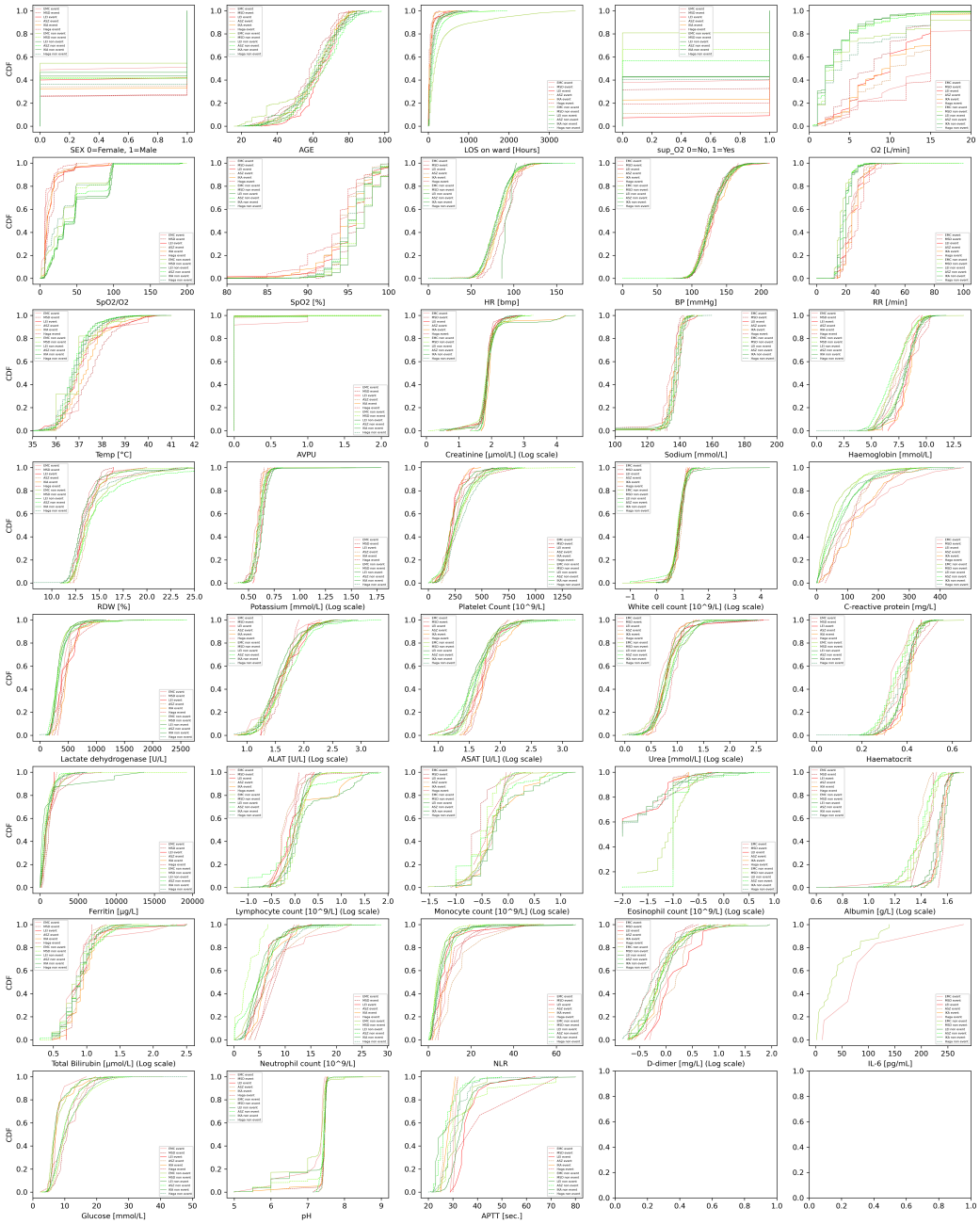


Figure 2: Cumulative density functions for all candidate predictors, separately for event and non-event samples. NLR = Neutrophile-to-lymphocyte ratio

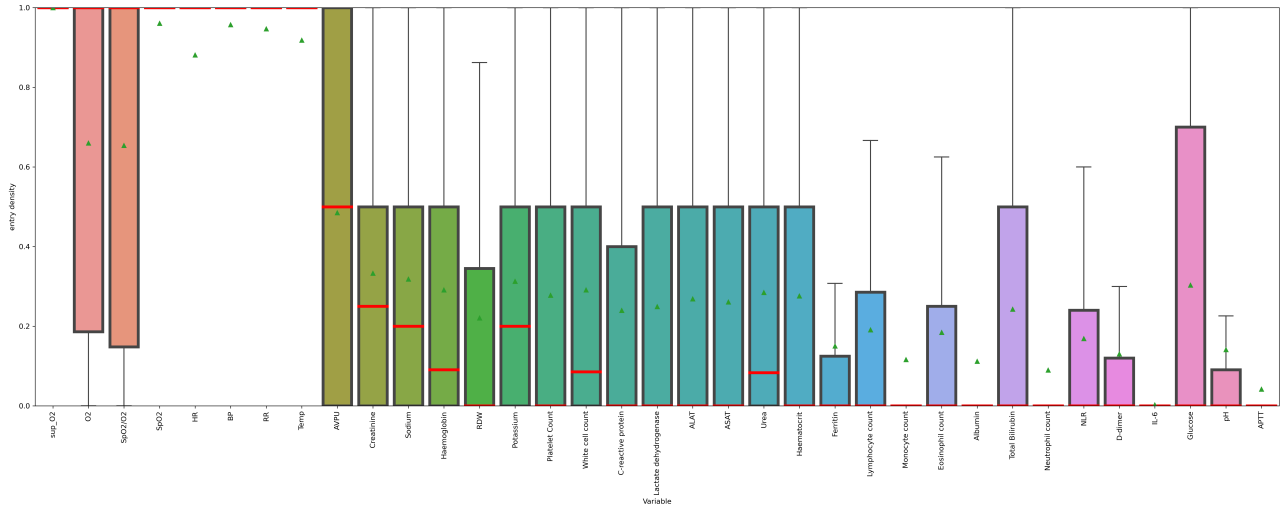
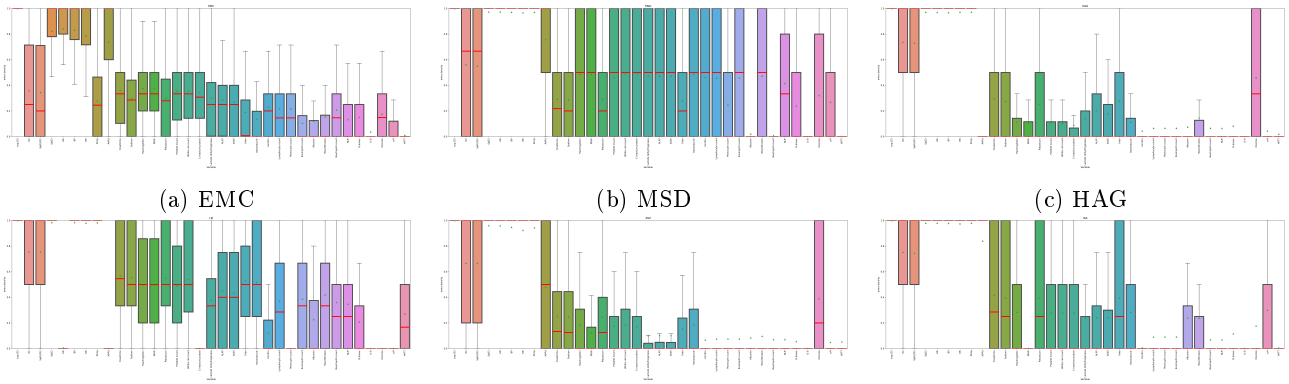


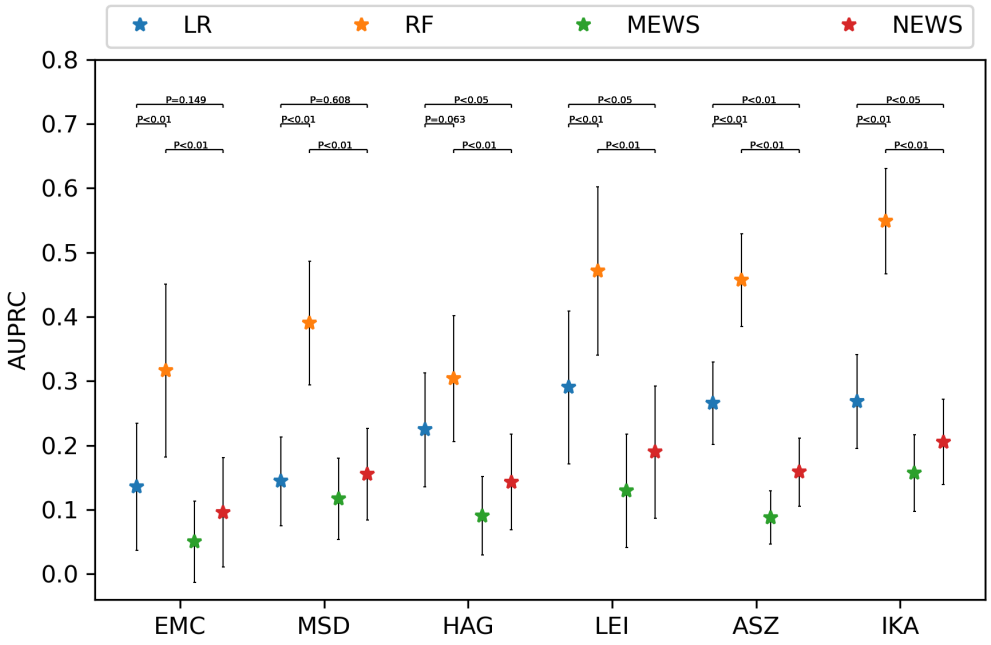
Figure 3: Distribution of daily entry densities (i.e., fractions of non-empty daily measurements) of all included patients for each candidate predictor.
Red line = median
Green triangle = mean
Box = 25-75 percentile



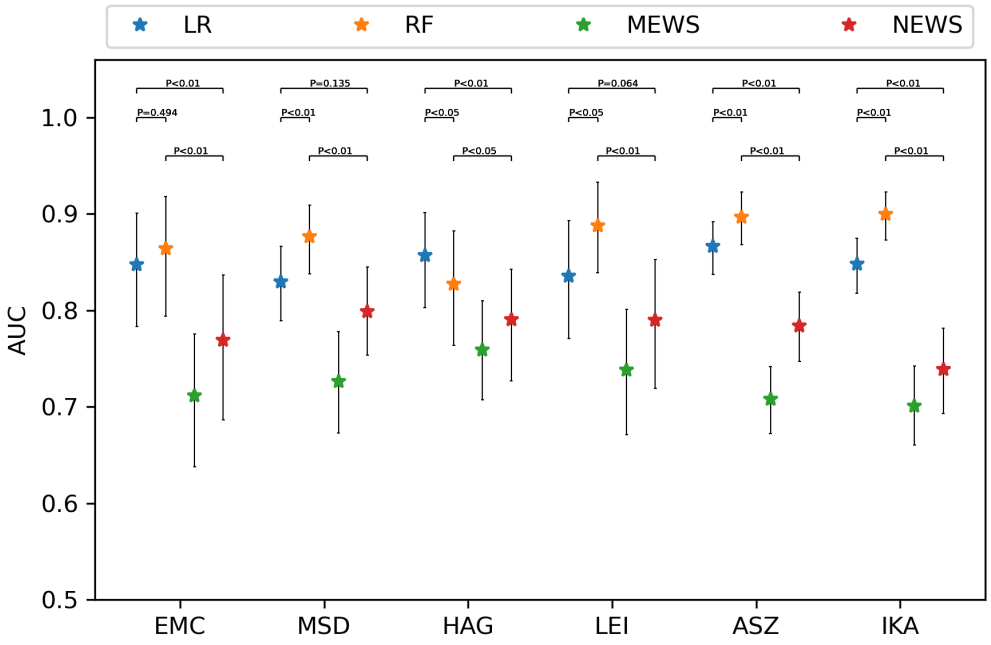
(a) EMC (b) MSD (c) HAG

(d) LEI (e) ASZ (f) IKA

Figure 4: Hospital-specific distribution of daily entry densities (i.e., fractions of non-empty daily measurements) of all included patients for each candidate predictor.
Red line = median
Green triangle = mean
Box = 25-75 percentile

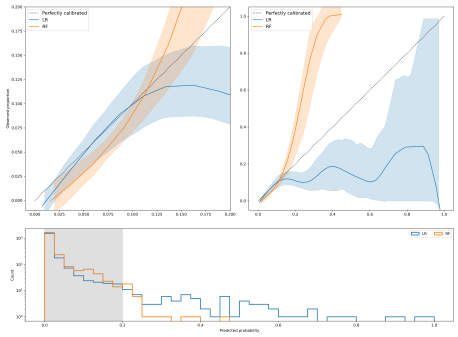


(a) AUPRC

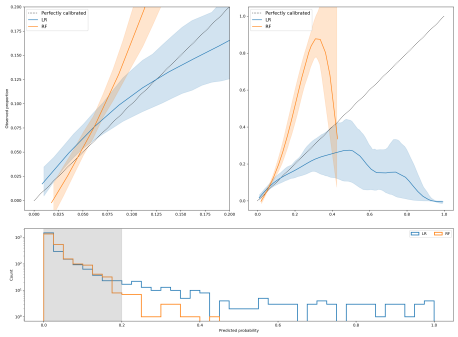


(b) AUC

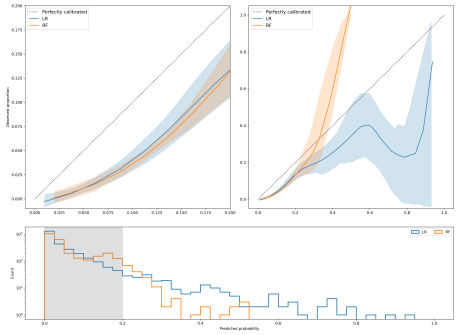
Figure 5: Area under the precision-recall curve (AUPRC) and area under the receiver operating curve (AUC) in the retrospective validation. P values are shown resulting from significance tests for difference between LR and NEWS (upper bar), RF and NEWS (middle bar) and between LR and RF (lower bar).



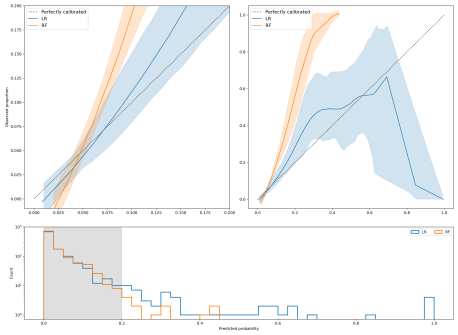
(a) EMC



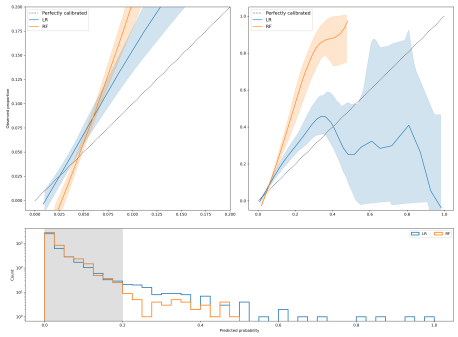
(b) MSD



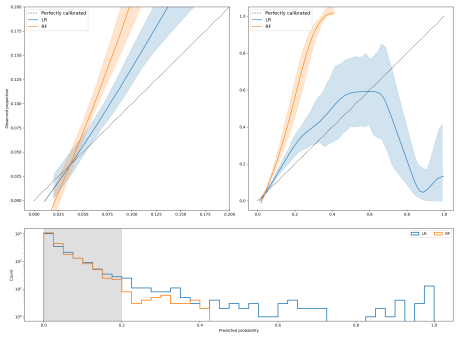
(c) HAG



(d) LEI

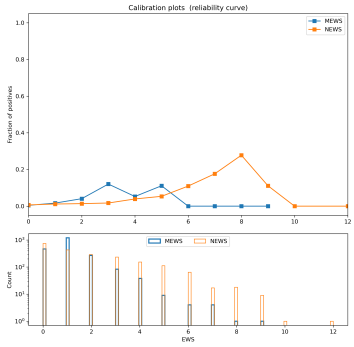


(e) ASZ

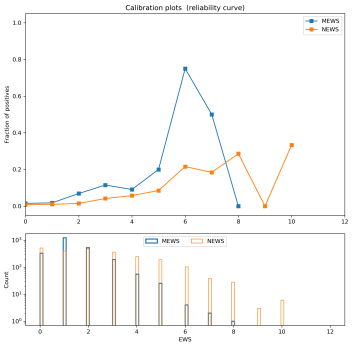


(f) IKA

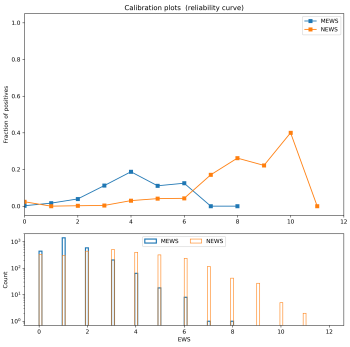
Figure 6: Loess smoothed flexible calibration curves yielded in the individual hospitals by the logistic regression (LR) and random forest (RF) models in the retrospective validation. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions in each hospital. Shaded areas around the curves represent the 95%CIs.



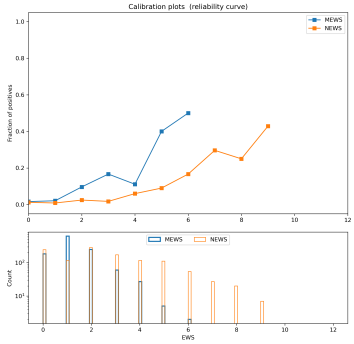
(a) EMC



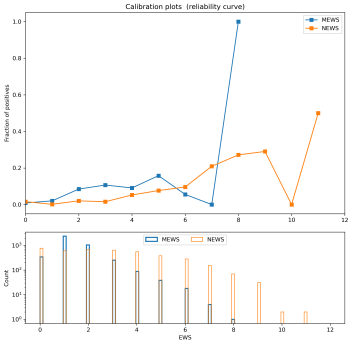
(b) MSD



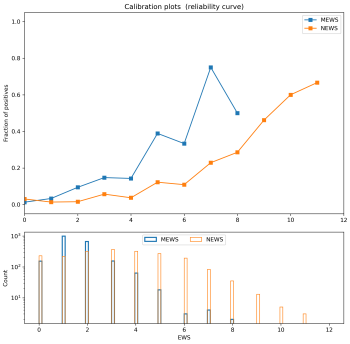
(c) HAG



(d) LEI

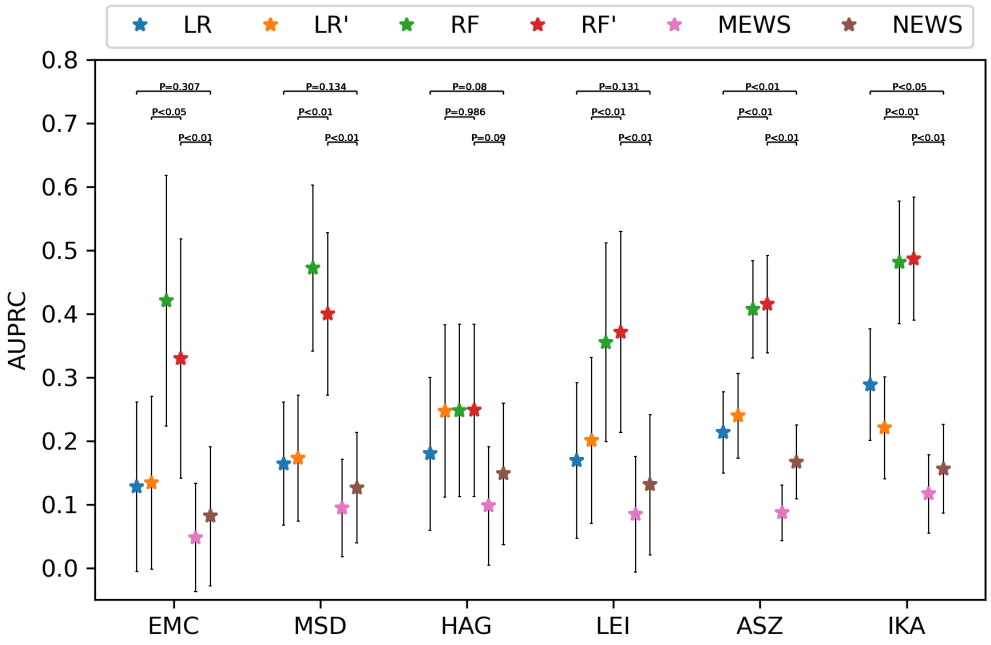


(e) ASZ

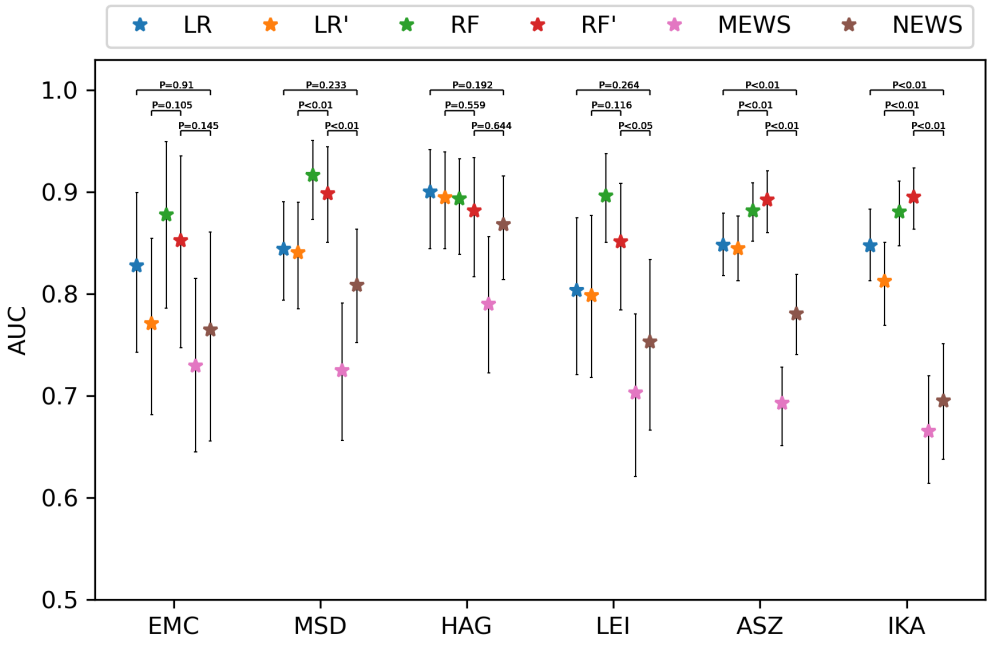


(f) IKA

Figure 7: Discrete calibration plots yielded by the traditional EWS in each individual hospital in the retrospective validation.

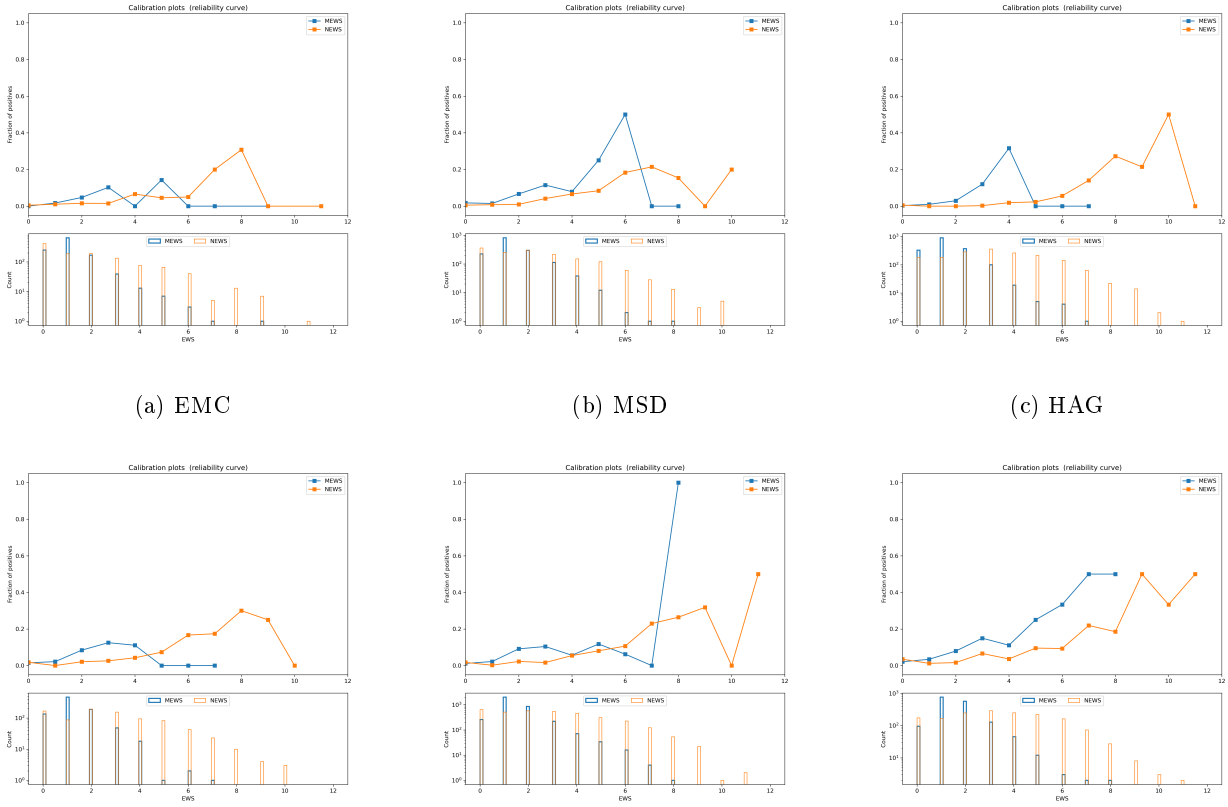


(a) AUPRC



(b) AUC

Figure 8: Area under the precision-recall curve (AUPRC) and area under the receiver operating curve (AUC) in the simulated prospective validation. LR' = LR model with monthly model updating, RF' = RF model with monthly model updating. P values are shown resulting from significance tests for difference between LR' and NEWS (upper bar), RF' and NEWS (middle bar) and between LR' and RF' (lower bar).



(a) EMC (b) MSD (c) HAG

(d) LEI (e) ASZ (f) IKA

Figure 9: Discrete calibration plots yielded by the traditional EWS in each individual hospital in the prospective validation.

B Supplementary tables

Table 1: Search spaces used in the grid-search for model hyperparameters optimization.

Model	**Hyperparameter**	**Search Space**
Logistic Regression	λ	$[10^{-4}, \ldots, 10^{4}]$ evenly spaced on log scale with 20 steps
Random Forest	max features	[p, $\sqrt{p}$, $\log_2 p$] where p is the total number of predictors.
Random Forest	max three depth	[2,3]

Candidate predictors	Evidence for clinical importance	Included	Reason for Exclusion	Entry Density
Patient Demographics:				
Age on admission [years]	Knight et al.[1]	✓	-	-
Sex at Birth	Knight et al.[1]	✓	-	-
Clinical Signs:				
Respiratory Rate [breaths/min]	Knight et al.[1]	✓	-	0.88
Peripheral oxygen saturations [%]	Knight et al.[1]	✓	-	0.91
Systolic blood pressure [mmHg]	Knight et al.[1]	✓	-	0.91
Temperature [°C]	Knight et al.[1]	✓	-	0.85
Heart Rate [bpm]	Knight et al.[1]	✓	-	0.83
Glasgow Coma Score	Knight et al.[1]		Not available	-
AVPU	National Early Warning Score [2]	✓	-	0.45
SpO2/O2 [$\frac{1}{L/min}$]	-	✓	-	0.57
Bedside investigations:				
Supplemental Oxygen [yes/no]	National Early Warning Score [2]	✓	-	-
O2 [L/min]	-	✓	-	0.58
pH	Knight et al.[1]		Low entry density	0.11
Glucose [mmol/L]	Knight et al. [1]	✓	-	0.29
Infiltrates on chest radiograph	Knight et al. [1]		Not available	-
Laboratory measures:				
Haemoglobin [g/L]	Knight et al. [1]	✓	-	0.28
Haematocrit [%]	Knight et al. [1]	✓	-	0.26
White cell count [10^9/L]	Knight et al. [1]	✓	-	0.28
Neutrophil count [10^9/L]	Knight et al. [1]		Low entry density	0.09
Lymphocyte count [10^9/L]	Knight et al. [1]		Low entry density	0.16
Eosinophil count [10^9/L]	Xie et al. [3]		Low entry density	0.15
Monocyte count [10^9/L]	Linssen et al. [4]		Low entry density	0.11
Neutrophil-to-lymphocyte ratio	Liu et al. [5]		Low entry density	0.14
Platelet count [10^9/L]	Knight et al. [1]	✓	-	0.26
Prothrombin [seconds]	Knight et al. [1]		Not available	-
APTT [seconds]	Knight et al. [1]		Low entry density	0.03
Sodium [mmol/L]	Knight et al. [1]	✓	-	0.32
Potassium [mmol/L]	Knight et al. [1]	✓	-	0.21
Total Bilirubin [mg/dL]	Knight et al. [1]		Low entry density	0.19
ALAT [U/L]	Knight et al. [1]	✓	-	0.23
ASAT [U/L]	Knight et al.[1]	✓	-	0.22
Albumin [g/L]	Knight et al. [1]		Low entry density	0.09
Lactate dehydrogenase [U/L]	Knight et al. [1]	✓	-	0.21
Urea [mmol/L]	Knight et al. [1]	✓	-	0.26
Creatinine [$\mu mol/L$]	Knight et al. [1]	✓	-	0.33
C-reactive protein [mg/dL]	Knight et al.[1]	✓	-	0.22
RDW [%]	Foy et al. [6]	✓	-	0.23
D-dimer [mg/L]	Yu et al. [7]		Low entry density	0.12
IL-6 [pg/mL]	Coomes et al. [8]		Low entry density	0.002
Ferritin [$\mu g/L$]	Dahan et al. [9]		Low entry density	0.13
Dynamics of clinical signs				
Δ Respiratory Rate [breaths/min]	-	✓	-	0.82
Δ Peripheral oxygen saturation [%]	-	✓	-	0.86
Δ Systolic blood pressure [mmHg]	-	✓	-	0.86
Δ Temperature [°C]	-	✓	-	0.80
Δ Heart Rate [bpm]	-	✓	-	0.80
Δ SpO2/O2 [$\frac{1}{L/min}$]	-	✓	-	0.48
Other				
Length of stay on ward [hours]	-	✓	-	-

Table 2: Candidate predictors evaluated for potential inclusion in the prediction model, based on evidence in literature and availability. APPT = Activated Partial Thromboplastin Time. Δ = signed difference between first and second most recent measurement.

Table 3: Pathway and population characteristics of included patients from the Erasmus University Medical Center.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=207)	Unplanned ICU admission (N=46)	Unexpected death (N=0)	Hospital transfer (N=5)	Total (N=258*)
Sex, male n(%)	105.0 (50.7)	23.0 (50.0)	-	2.0 (40.0)	130.0 (50.4)
Age, years					
med (IQR)	60.0 (46.0-68.0)	66.0 (51.2-73.0)	-	59.5 (56.2-63.8)	60.0 (49.5-70.0)
mean (SD)	56.0 (16.5)	61.7 (14.1)	-	60.5 (6.6)	57.1 (16.1)
Ward LOS, days					
med (IQR)	6.6 (3.6-11.2)	1.8 (0.5-3.8)	-	5.1 (5.0-7.1)	7.5 (3.8-14.3)
mean (SD)	9.7 (13.4)	3.8 (6.6)	-	5.8 (1.4)	11.3 (13.4)

Table 4: Pathway and population characteristics of included patients from the Maasstad teaching hospital.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=556)	Unplanned ICU admission (N=98)	Unexpected death (N=1)	Hospital transfer (N=115)	Total (N=770*)
Sex, male n(%)	289.0 (52.0)	72.0 (73.5)	0.0 (0.0)	62.0 (53.9)	423.0 (54.9)
Age, years					
med (IQR)	59.0 (49.0-68.0)	58.0 (50.0-67.0)	-	59.0 (53.0-68.0)	59.0 (50.0-68.0)
mean (SD)	58.0 (13.7)	57.4 (13.0)	-	58.6 (11.8)	58.0 (13.3)
Ward LOS, days					
med (IQR)	2.8 (1.6-4.9)	1.3 (0.8-2.6)	0.3 (0.3-0.3)	1.1 (0.8-1.7)	2.8 (1.3-5.5)
mean (SD)	4.0 (4.6)	2.0 (2.0)	0.3 (0.0)	1.6 (2.9)	4.9 (7.2)

Table 5: Pathway and population characteristics of included patients from the Haga teaching hospital.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=529)	Unplanned ICU admission (N=85)	Unexpected death (N=0)	Hospital transfer (N=142)	Total (N=756*)
Sex, male n(%)	319.0 (60.4)	62.0 (72.9)	-	91.0 (64.1)	472.0 (62.5)
Age, years					
med (IQR)	59.0 (48.0-68.0)	63.0 (57.0-69.0)	-	60.0 (54.0-66.0)	60.0 (51.0-68.0)
mean (SD)	58.3 (14.8)	62.0 (10.0)	-	59.6 (11.0)	59.0 (13.7)
Ward LOS, days					
med (IQR)	3.0 (1.7-5.7)	2.8 (1.4-4.2)	-	0.9 (0.6-2.0)	2.9 (1.4-6.0)
mean (SD)	4.7 (5.1)	3.5 (3.6)	-	1.6 (1.6)	5.7 (8.3)

Table 6: Pathway and population characteristics of included patients from the Leiden University Medical Center.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=218)	Unplanned ICU admission (N=56)	Unexpected death (N=0)	Hospital transfer (N=27)	Total (N=301*)
Sex, male n(%)	131.0 (60.1)	33.0 (58.9)	-	11.0 (40.7)	175.0 (58.1)
Age, years					
med (IQR)	62.0 (53.0-71.0)	64.5 (58.8-72.0)	-	60.0 (53.5-71.5)	63.0 (54.0-71.0)
mean (SD)	61.2 (13.0)	64.3 (10.2)	-	61.3 (13.0)	61.8 (12.5)
Ward LOS, days					
med (IQR)	3.9 (2.4-6.0)	2.3 (1.2-3.9)	-	1.3 (1.0-2.0)	4.1 (2.3-7.8)
mean (SD)	4.8 (4.3)	3.6 (4.9)	-	1.8 (1.2)	8.6 (15.0)

Table 7: Pathway and population characteristics of included patients from the Albertschweitzer teaching hospital.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=588)	Unplanned ICU admission (N=181)	Unexpected death (N=2)	Hospital transfer (N=116)	Total (N=887*)
Sex, male n(%)	284.0 (54.6)	106.0 (63.9)	2.0 (100.0)	61.0 (56.5)	453.0 (56.9)
Age, years					
med (IQR)	63.0 (53.0-73.0)	64.0 (56.0-70.0)	76.0 (74.5-77.5)	60.0 (54.2-70.8)	63.0 (54.0-72.0)
mean (SD)	61.9 (14.4)	62.3 (10.8)	76.0 (4.2)	60.9 (11.6)	61.9 (13.3)
Ward LOS, days					
med (IQR)	4.4 (2.0-7.8)	2.1 (0.8-3.8)	7.6 (7.2-7.9)	1.3 (0.9-2.6)	4.2 (1.8-8.9)
mean (SD)	6.0 (6.8)	3.4 (5.1)	7.6 (0.7)	2.2 (2.7)	7.6 (10.4)

Table 8: Pathway and population characteristics of included patients from the Ikazia teaching hospital.
*This number is without patients still admitted.

LOMT code A or B	Discharged alive (N=414)	Unplanned ICU admission (N=139)	Unexpected death (N=0)	Hospital transfer (N=133)	Total (N=686*)
Sex, male n(%)	179.0 (50.3)	84.0 (65.6)	-	66.0 (52.8)	329.0 (54.0)
Age, years					
med (IQR)	61.0 (52.0-70.0)	62.0 (54.8-69.0)	-	61.0 (53.0-69.0)	61.0 (52.0-69.0)
mean (SD)	59.9 (13.6)	60.7 (11.7)	-	60.1 (12.1)	60.1 (12.9)
Ward LOS, days					
med (IQR)	3.4 (1.9-5.5)	1.7 (0.7-3.0)	-	0.9 (0.6-1.3)	3.3 (1.6-7.0)
mean (SD)	4.4 (3.8)	2.1 (2.4)	-	1.5 (3.1)	6.6 (10.1)

C Evaluation metrics

To quantify the performance of the different models, we use three different metrics:

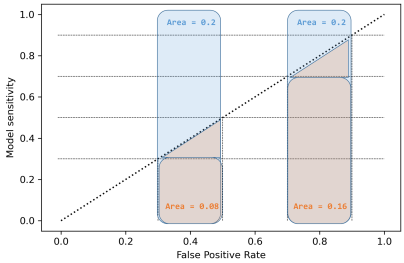
- The partial area under the ROC curve (pAUC) between a false positive rate (FPR) of 0 and 0.33.
- The area under the precision-recall (PR)-curve (AUPRC).
- The (complete) area under the ROC curve (AUC).

C.1 Partial and complete area under the ROC curve

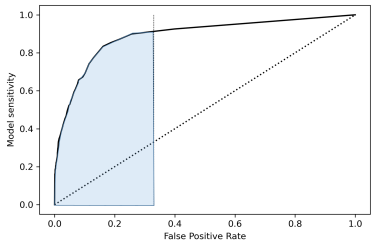
Like proposed in the original paper on partial areas under the ROC curve (pAUCs) [10], normalization of the pAUC is needed. For example, area underneath an ROC curve between false-positive rates (FPRs) 0.7 and 0.9 is 0.18. The maximum value this area could attain is 0.2 while, the minimum is 0.16. An area of 0.18 found FPRs 0.3 and 0.5 would have the same maximum of 0.2, but a minimum value of 0.08 (see figure 10a). As we do not want to value the two areas the same, the pAUC is normalized such that the minimum value is equal to 0.5 (just like the minimum value of the complete AUC):

$$A_{norm} = \frac{1}{2}[1 + \frac{A - A_{min}}{A_{max} - A_{min}}] \tag{1}$$

where A is the pAUC, A_{min} the minimum pAUC, and A_{max} the maximum pAUC in the specific FPR range. In this study, we considered the pAUC between 0 and 0.33 FPR (figure 10b).



(a) The maximum area underneath an ROC curve between false positive rates 0.7 and 0.9 and between 0.3 and 0.5 is both 0.2, whereas the minimum areas are 0.16 and 0.08, respectively.



(b) The shaded area denotes the partial area under the ROC curve between false positive rates 0 and 0.33, which we assume as the clinically relevant range.

As we will evaluate the portion of the receiver operating curve between 0 and 0.33 FPR, the maximum area is 0.33 and the minimum area is 0.05445. Therefore, we normalized the pAUC as:

$$A_{norm} = \frac{1}{2}[1 + \frac{A - 0.05445}{0.33 - 0.05445}] \tag{2}$$

The normalized pAUC was implemented using the 'roc auc score' function from Sklearn's metrics library, setting 'max fpr' to 0.33 (and the other parameters at default). For the complete area under the ROC curve (AUC), we used the same function, setting max fpr to 1.

For both pAUC and AUC, we calculated the Bootstrap percentile confidence intervals as described in [11], using 1000 stratified bootstrap samples.

C.2 Area under the precision-recall (PR)-curve.

The area under the precision-recall (PR) curve (AUPRC) is a useful performance metric for imbalanced data in a problem setting where finding the true positives is important. This is true for the setting in this study, as the prevalence of the primary clinical endpoint (unplanned ICU admission or unexpected death) is very low and the sampling strategy makes this imbalance problem even bigger.

In their work on the AUPRC, Boyd and colleagues [12] recommends the average precision (AP) as the point estimate for the AUPRC. Using the AP is preferred over computing the AUPRC with the trapezoidal rule as the AP is not interpolated, while the linear interpolation used in the trapezoidal rule can be too optimistic.

The Average Precision (AP) summarizes a PR-curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. It is defined in as:

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{3}$$

where R_n and P_n are the recall (or sensitivity) and precision (or PPV) at the n^{th} threshold. We used the 'average precision score' function from Sklearn's metrics library with default parameters.

To calculate uncertainty around this metric, we used the binomial confidence interval as it has shown to be valid in previous work [12] on this metric and does not require any additional calculations compared to, e.g., a bootstrapping method. The binmoial interval for 95% coverage is defined as follows:

$$AP \pm 1.96\sqrt{\frac{AP(1-AP)}{n}} \tag{4}$$

where n is the number of positive samples, not the total sample size, as the number of positive samples specify the maximum number of unique recall (or sensitivity) values.

C.3 Net Benefit

The net benefit (NB) is a tool for evaluating the clinical implications of models, markers, and tests. A model gives a predicted probability directly for a certain adverse event, e.g. unplanned ICU admission or unexpected death within 24 hours. The metric is defined by [13] as follows:

$$NB = \text{benefit} - \text{harm} \times \text{exchange rate} \tag{5}$$

where the benefit is defined as the number of true positives (as a fraction of the total observations) and the cost as the number of false positives (as a fraction of the total observations). The exchange rate is a clinical judgement of the relative value of benefits (finding cases) and harms (causing false alarms). The exchange rate can be derived by the maximum number of triggers (or alarms) one is willing to invest to find one case. For example, a physician may say that, find one patient who is deteriorating, no more than 20 false alarms should go off. This implies that the harm of missing a patient who is deteriorating is nineteen times greater than that of a false alarm. So, we want to 'weight' finding a case as nine times more important than avoiding one false alarm. This, the 'harm' is weighted by the odds ratio corresponding to the probability threshold (assuming that the model is well calibrated).

So, the NB can be calculated by:

$$NB = \frac{\text{True positives}}{N} - \frac{\text{False positives}}{N} \times \frac{p_t}{1-p_t} \tag{6}$$

where N is the total number of samples and p_t the probability threshold used to trigger a certain action. In the context of early warning, this could be either an urgent clinical response or emergency clinical response (see figure 1).

D Comparing performances

As we want to compare the performances yielded by the two proposed models to the performance of the conventional EWSs, we need to compare the different metrics. We chose the same definition for the test statistic to compare all three types of areas (pAUC, AUPRC and AUC) as described in [10]:

$$Z = \frac{A_1 - A_2}{\sqrt{Var(A_1 - A_2)}} \tag{7}$$

where A_1 and A_2 are the two areas to be compared, and $Var(A_1 - A_2)$ is approximated from the bootstrap differences in a 2000 sample stratified bootstrapping procedure. Z is then compared to the normal distribution to determine the corresponding P value.

E Miscalibration correction

E.1 Baseline situation

In the simulated prospective validation, we simulated the situation as if a new early warning model had been developed based on the data gathered in the six included hospitals until July 2020, i.e. the 'wave 1 cohort', and used to predict deterioration for the patients admitted from August 2020 until May 2021, i.e. the 'wave 2 cohort'. In the figure below, this is marked with the 'Train baseline model' icon.

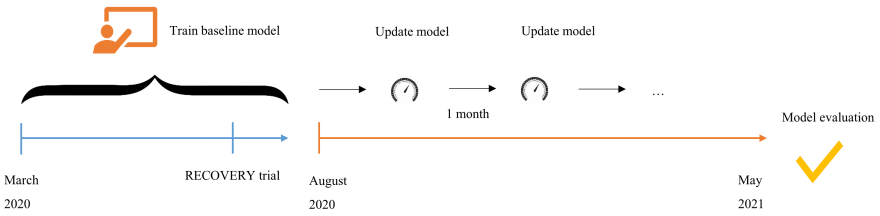


Figure 11: Simulation of a prospective validation of the proposed models.

We plotted the flexible calibration curves resulting from the predictions made in all the hospitals combined. These plots suggest that the logistic regression (LR) model calibrated relatively well but its predictions are too extreme (calibration intercept −0.15, and slope 0.66), while the random forest (RF) predictions show risk overestimation and predictions that are too moderate (calibration intercept −0.35 and slope 1.65).

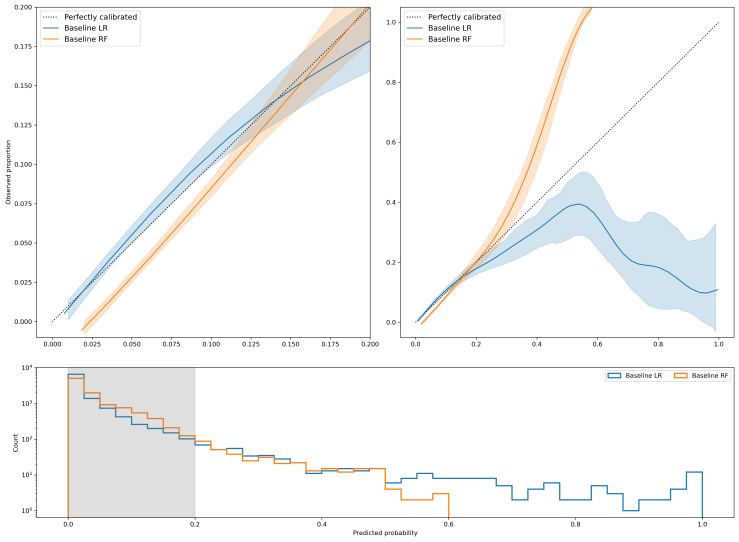
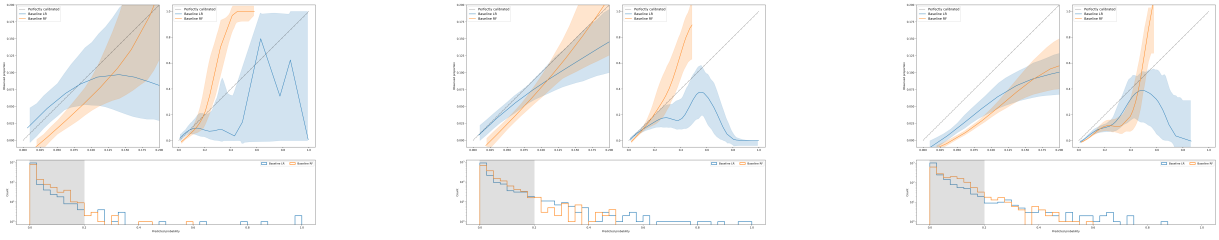
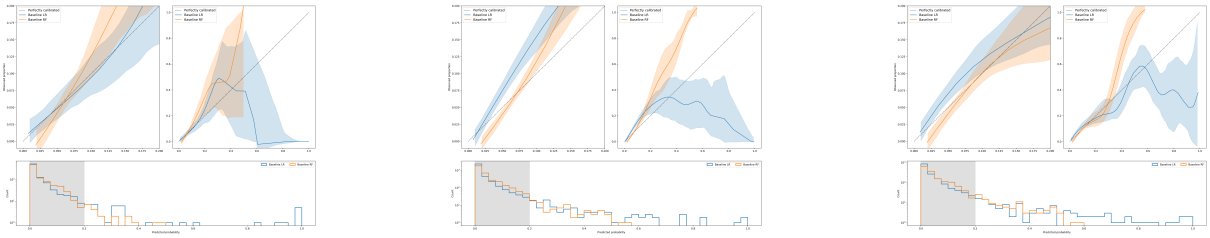


Figure 12: Loess smoothed flexible calibration curves for the baseline logistic regression (LR) and random forest (RF) models in the simulated prospective validation. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions. Shaded areas around the curves represent the 95%CIs.

In the figures below, the flexible calibration curves of the baseline models based on predictions in every hospital separately are plotted.



(a) EMC (b) MSD (c) HAG



(d) LEI (e) ASZ (f) IKA

Figure 13: Hospital-specific loess smoothed flexible calibration curves for the logistic regression (LR) and random forest (RF) models in the simulated prospective validation. Left plots show zoom-ins of the right plots in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions in each hospital. Shaded areas around the curves represent the 95%CIs.)
EMC = Erasmus University medical Center, MSD = Maasstad Teaching hospital, HAG = Haga Teaching hospital, LEI = Leiden University Medical Center, ASZ = Albertschweitzer Teaching hospital, IKA = Ikazia Teaching hospital.

Here, we notice a couple of things:

- The number of predictions for some hospitals is very low, which results in large uncertainty (i.e. wide CIs), which makes the evaluation of model calibration in these hospitals harder.
- While the baseline LR model seems to calibrate well when we examine the calibration curve based on predictions from all hospitals combined (figure 12), evaluation in individual hospitals does clear show miscalibration of the LR model in some hospitals.
- The severity and direction of miscalibration differs between hospitals. For example, both LR and RF model show underestimation in the Haga Teaching hospital, but underestimation in the Albertschweitzer Teaching hospital.

To correct for this miscalibration, we examined different specific strategies. We applied each of the specific strategies to simulate as if one corrected for miscalibration in the models on a mothly basis during the wave 2 period, as visualized in figure 11.

We constructed the specific strategies based on three main strategies:

- Update the classifier, i.e. 'model updating'.
- Apply an extra mapping function (a calibrator) on top of the model, i.e. 're-calibration'.
- Use a combination of model updating and re-calibration.

E.2 Main strategies

E.2.1 Model updating

The logistic regession (LR) model is defined as:

$$logit[Pr(y = 1)] = \alpha + \beta_1 x_1 + ... + \beta_n x_n = \alpha + \mathbf{x}^T \boldsymbol{\beta} \tag{8}$$

Prior shift adjustment (PSA)

During the pandemic, the a priori risk to reach the clinical endpoint (unplanned ICU admission or unexpected death) may have changed over time, and could have been different for every hospital. The changes over time may be explained by the improvement of care for covid-19 patients during this period, for example caused by the wide introduction of dexamethasone after the RECOVERY trial [14]. Differences between hospitals may be explained by differences in policies considering ICU admission in covid-19 patients. Under the assumption that, in a new situation (e.g. later during the pandemic or in another hospital), only the a priori risk has changed and the class conditional distribution ($p(x|y)$) for the different predictors included in the model remained the same compared to the situation where the model was trained on, this would result in the prior probability shift situation as described in [15].

To correct for this shift, we can update the posterior probability from an old situation (referred to with the subscript 'o') in the new situation, e.g. later during the pandemic or in another hospital (referred to with the subscript 'n'), as follows:

$$p_n(y|x) = \frac{p(x|y)p_n(y)}{\sum_y p(x|y)p_n(y)} = \frac{\frac{p_o(y|x)p_o(x)}{p_o(y)}p_n(y)}{\sum_y \frac{p_o(y|x)p_o(x)}{p_o(y)}p_n(y)} = \frac{\frac{p_n(y)}{p_o(y)}p_o(y|x)}{\sum_y \frac{p_n(y)}{p_o(y)}p_o(y|x)} \tag{9}$$

Thus, to update the old posterior to the new situation, we reweigh by $\frac{p_n(y)}{p_o(y)}$ and re-normalize these probabilities.

Recalling eq. 8, the LR model is defined as:

$$log\left[\frac{p_o(y=1|x)}{p_o(y=0|x)}\right] = \alpha + \mathbf{x}^T\boldsymbol{\beta} \tag{10}$$

We can update the posterior:

$$log\left[\frac{p_n(y=1|x)}{p_n(y=0|x)}\right] = log\left[\frac{\frac{p_n(y=1)}{p_o(y=1)}p_o(y=1|x)Z}{\frac{p_n(y=0)}{p_o(y=0)}p_o(y=0|x)Z}\right] = log\left[\frac{p_o(y=1|x)}{p_o(y=0|x)}\right] + log\left[\frac{p_n(y=1)}{p_n(y=0)}\right] - log\left[\frac{p_o(y=1)}{p_o(y=0)}\right] \tag{11}$$

where $Z = \left[\sum_y \frac{p_n(y)}{p_o(y)}p_o(y|x)\right]^{-1}$ (normalization constant) and $log\left[\frac{p(y=1|x)}{p(y=0|x)}\right] = \mathbf{x}^T\boldsymbol{\beta} + \alpha$.

Thus,

$$log\left[\frac{p_o(y=1|x)}{p_o(y=0|x)}\right] + log\left[\frac{p_n(y=1)}{p_n(y=0)}\right] - log\left[\frac{p_o(y=1)}{p_o(y=0)}\right] = \alpha_n + \mathbf{x}^T\boldsymbol{\beta} \tag{12}$$

where

$$\alpha_n = \alpha + log\left[\frac{p_n(y=1)}{p_n(y=0)}\right] - log\left[\frac{p_o(y=1)}{p_o(y=0)}\right] \tag{13}$$

Therefore, if you know the prior probability in the new situation ($p_n(y)$) and the prior probability in the old situation ($p_o(y)$), one can update the α in the logistic regression model directly.

In the Random Forest model (RF), the definition of the posterior differs from the LR model. It is computed as the mean of the posteriors of all trees in the forest, while the posterior in a single tree is the fraction of positive samples in a leaf. We did not define a method to correct for a prior shift in an RF model.

Full model retraining (FMR)

For full model retraining, the model is simply refitted to the complete dataset available at the moment of refitting, i.e. full model refit (FMR).

E.3 Model Re-calibration

Instead of updating an already fitted model, one may try to recalibrate the model's predictions in the probability scale by learning a mapping from the predicted probabilities to updated (calibrated) probabilities. Here, one regresses the observed binary outcome directly to the predictions of the classifier. Two commonly used functions for this are logistic regression function (called Platt Scaling) and isotonic regression [16]. These methods may apply to both the LR and RF classifier. We did not report on the findings of Platt scaling, as these results showed to be inferior to isotonic regression.

Isotonic regression (IR)

In isotonic regression, it is assumed that originally trained classifier ranks examples correctly (i.e., good discrimination), which means that mapping the original predictions into the observed probabilities is non-decreasing and thus can be learned by an isotonic regressor. Given the posterior estimates E from a model and the true targets y_i, the basic assumption in isotonic regression is:

$$y_i = m(E_i) + \epsilon_i \tag{14}$$

Then, given a train set (f_i, y_i), the isotonic regression problem is finding the isotonic function m such that

$$m = argmin_z \sum (y_i - z(E_i))^2 \tag{15}$$

A commonly used algorithm that finds a stepwise constant function for z is the pair-adjacent violators (PAV) algorithm (described in [16]) and is used in the Sklearn function we used ('CalibratedClassifierCV') as well.

E.4 Different strategies for miscalibration correction

As shown in figure 14, the miscalibration of the baseline models is different for every hospital. To correct for this hospital-specific miscalibration effectively, we applied the different techniques for model updating or re-calibration to every hospital separately, based on data collected in that specific hospital.

We examined the effect of prior shift adjustment (PSA) only for the logistic regression model (as we did not define this for the random forest model), and examined the effect of isotonic regression (IR) and full model retraining (FMR) for both the LR and RF model. Additionally, we examined the effect of combining FMR and IR (and FMR and PSA, for logistic regression).

Under the hypothesis that the a priori probability for deterioration changes over time for the individual hospitals, better calibration may be achieved by using only recently collected data for re-calibration or model updating. Therefore, we carried out every re-calibration and model updating strategy in a cumulative fashion and using a sliding window of 4 months.

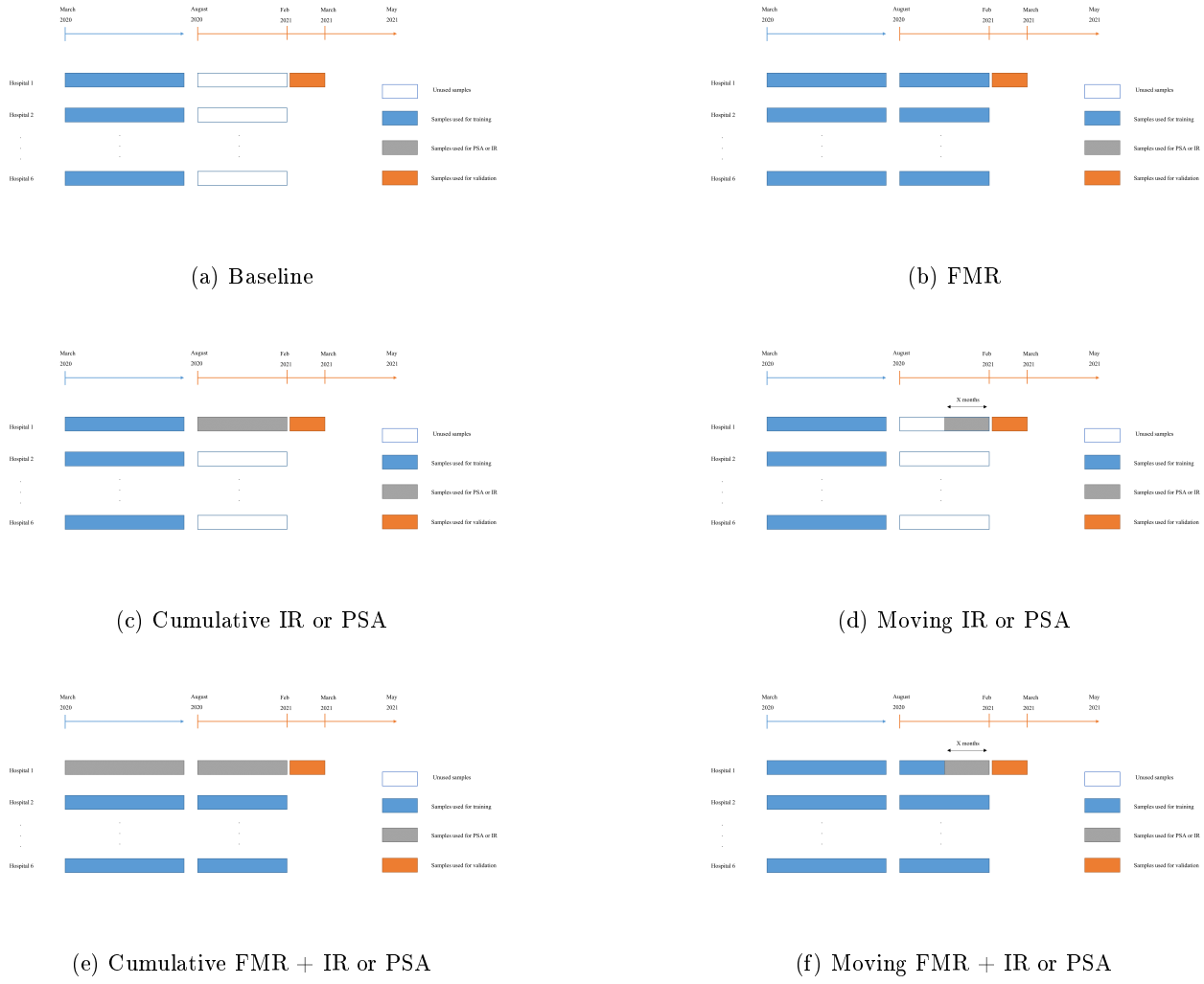
In total, we examined the effectiveness of 9 different strategies to correct for miscalibration of the LR model and 5 different strategies for the RF model:

- **FMR**: Full model retraining.
  Here, cumulatively collected data is used to retrained the models.
- **Cumulative IR**: Model recalibration using isotonic regression.
  Here, hospital-specific cumulatively collected data is used to train an isotonic regressor to re-map the predictions of the baseline model.
- **Cumulative PSA**: Model updating using prior shift adjustment (only for Logistic regression)
  Here, hospital-specific cumulatively collected data is used to determine the prior shift compared to data used for the baseline models. Then, the baseline models are updated based on this shift.
- **Moving IR**: Model recalibration using isotonic regression. Here, hospital-specific collected data from the most recent 4 months is used to train an isotonic regressor to re-map the predictions of the baseline model every month.
- **Moving PSA**: Model updating using prior shift adjustment (only for Logistic regression)
  Here, hospital-specific collected data from the most recent 4 months is used to determine the prior shift. Then, the baseline models are updated based on this shift.
- **Cumulative FMR + IR**: Combining full model retraining and recalibration using isotonic regression.
  Here, hospital-specific cumulatively collected data is used to train an isotonic regressor to re-map the predictions of a model that is retrained on all remaining data.
- **Cumulative FMR + PSA**: Combining full model retraining and model updating using prior shift adjustment.
  Here, hospital-specific cumulatively collected data is used to determine the prior shift. Then, the models, which are first retrained on the remaining data, are updated based on this shift.
- **Moving FMR + IR**: Combining full model retraining and recalibration using isotonic regression.
  Here, hospital-specific collected data from the most recent 4 months is used to train an isotonic regressor to re-map the predictions of a model that is retrained on all remaining data.

- **Moving FMR + PSA**: Combining full model retraining and model updating using prior shift adjustment. Here, hospital-specific collected data from the most recent 4 months is used to determine the prior shift. Then, the models, which are first retrained on the remaining data, are updated based on this shift.

We compared the model calibration achieved by the different strategies with the calibration of the baseline models, which were trained on the wave 1 cohort and eveluated on the wave 2 cohort without any re-calibration or model updating.

The figure below visualizes which data is used to train the model and to update or re-calibrate the model to make predictions on data collected between February and March 2021 for one of the six hospitals, according to the different strategies. For every month in the wace 2 cohort (August 2020 until May 2021), this process is repeated for every hospital.



(a) Baseline

(b) FMR

(c) Cumulative IR or PSA

(d) Moving IR or PSA

(e) Cumulative FMR + IR or PSA

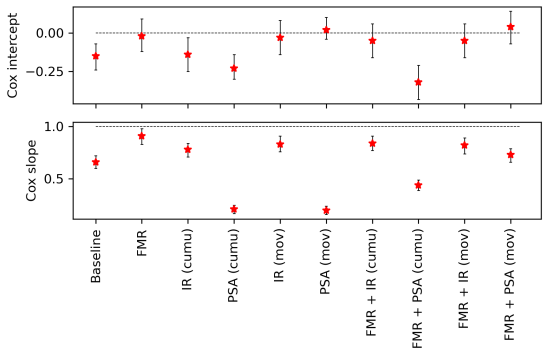
(f) Moving FMR + IR or PSA

Figure 14: Schematic representations of the different strategies to correct for miscalibration. FMR = full model retraining, IR = Isotonic regression, PSA = prior shift adjustment.

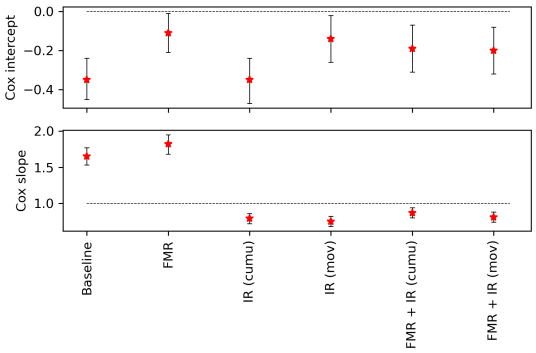
We evaluated the effectiveness of every strategy by evaluating the resulting calibration in the 'weak' sense, calculating the calibration slope and intercept [17], and in the 'moderate' sense by plotting the flexible calibration curves [18]. As calibration in individual hospitals is hard due to large uncertainties resulting from small sample sizes, we evaluated calibration first for the predictions in all hospitals combined. After selecting the most effective strategy, we also show the correction for miscalibration resulting from this strategy in the individual hospitals.

E.5 Results

Figure 15 shows the calibration intercepts and slopes (95% CI) yielded by the different strategies for miscalibration correction in the LR and RF model based on the predictions in the six hospitls combined.



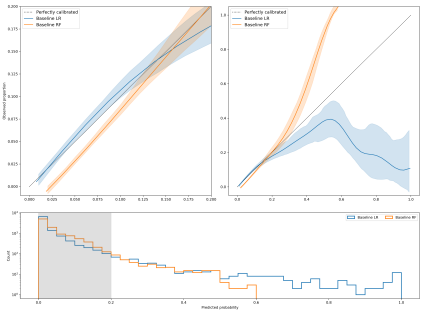
(a) LR



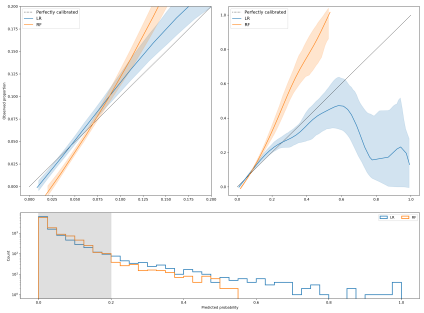
(b) RF

Figure 15: Calibration intercepts and slopes yielded by the different miscalibration correction strategies for the logistic regression (LR) and random forest (RF) model. cumu = cumulative, mov = using a 4 month sliding window

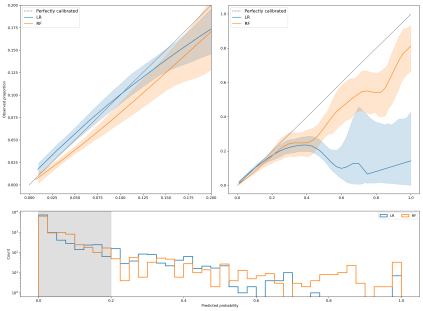
Figure below shows the corresponding flexible calibration curves.



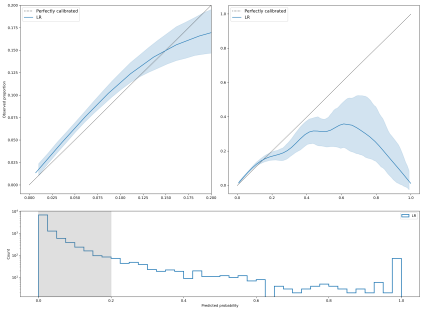
(a) Baseline



(b) FMR

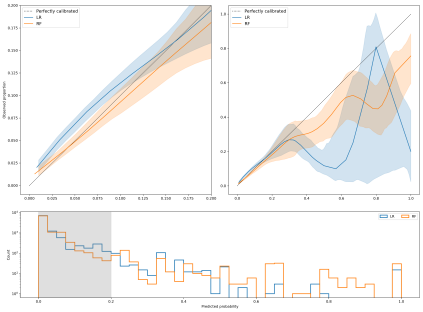


(c) Cumulative IR

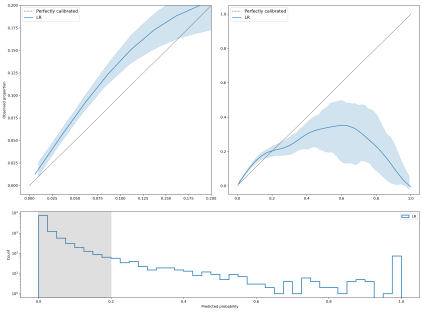


(d) Cumulative PSA

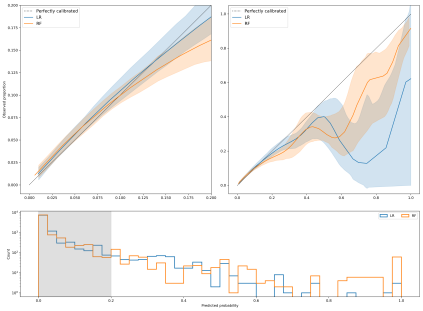
Figure 16: Figure continued on next page.



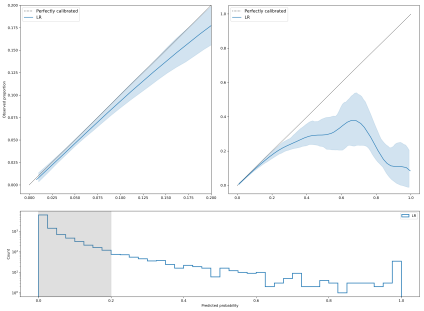
(e) Moving IR.



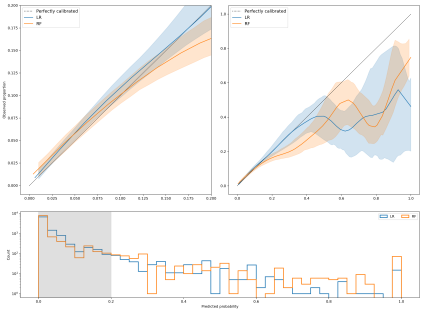
(f) Moving PSA.



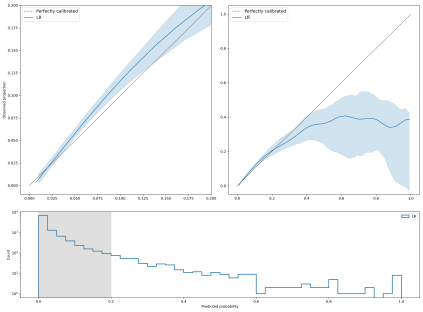
(g) Cumulative FMR + IR



(h) Cumulative FMR + PSA



(i) Moving FMR + IR.



(j) Moving FMR + PSA.

Figure 16: Loess smoothed flexible calibration curves yielded by the different strategies to correct for miscalibration in the baseline logistic regression (LR) and random forest (RF) models in the simulated prospective validation. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions. Shaded areas around the curves represent the 95%CIs.

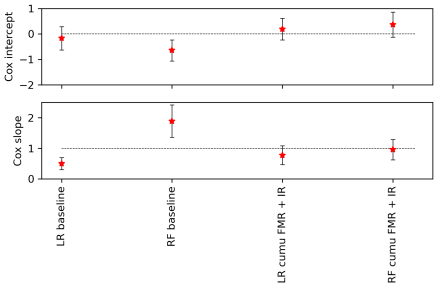
Based on the flexible calibration curves, we observed similar effective correction for miscalibration (i.e., bringing the calibration curve closer to 0 intercept and unit slope) in both the LR and RF model by two strategies:

- cumulative FMR + IR
- moving FMR + IR

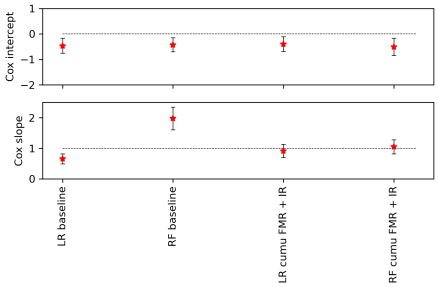
As shown in figure 15, both these strategies show good correction for calibration-in-the-large (moving the calibration slope towards 0) for both the LR model en RF model. To correct for the spread of predictions (moving the calibration slope towards 1), the cumlative FMR + IR method is slightly more effective in the RF model (see figure 15b). Also, as the isotonic regressor may start overfitting when trained on a small sample size, training it on as much hospital-specific data as possible (thus cumulatively) is reasonable. Therefore, we selected the cumulative FMR + IR strategy as the most effective to correct for miscalibration in the baseline models.

N.B., the calibration intercepts and slopes in figure 15 are **not** the intercept and slopes of calibration curves in the probability domain. They were first introduced by Cox in 1958 [17], and result from regressing the predictions in the log-odds domain to the observed posteriors.

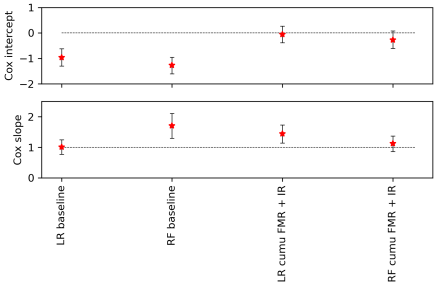
In figure 17, the calibration slopes and intercepts (95% CI) are visualized yielded by the baseline models and by the models after correcting for miscalibration using cumulative FMR + IR for each hospital separately. The corresponding flexible calibration curves are shown in figure 18, which should be compared to the baseline curves in figure 14.



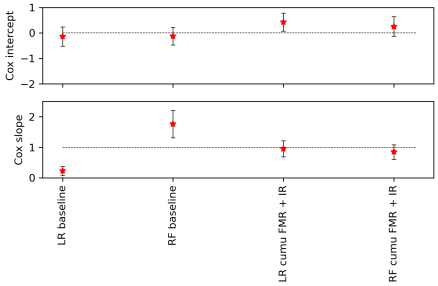
(a) EMC



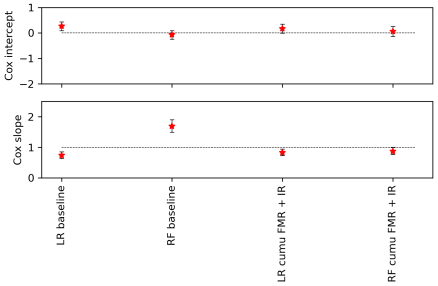
(b) MSD



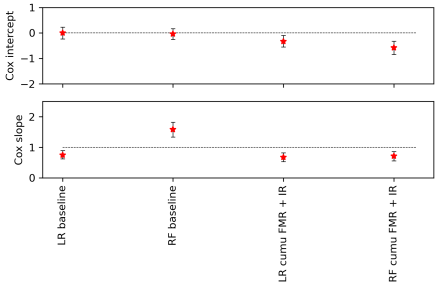
(c) HAG



(d) LEI

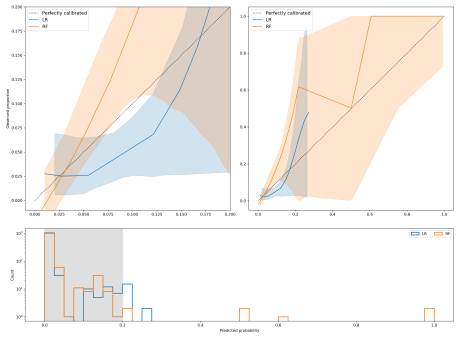


(e) ASZ

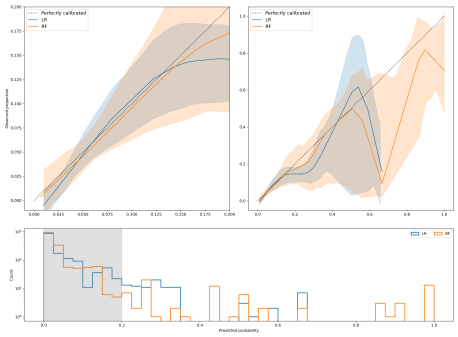


(f) IKA

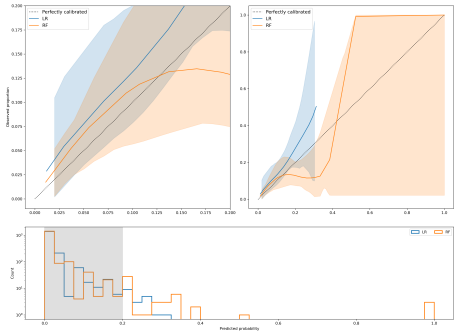
Figure 17: Calibration slopes and intercepts (95% CI) for the logistic regression (LR) and random forest (RF) model over the different hospitals, for the baseline model and after correcting for miscalibration with the cumulative FMR + IR strategy.



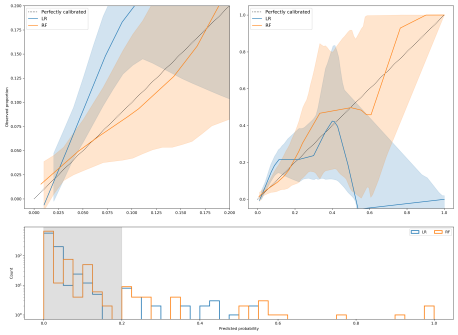
(a) EMC



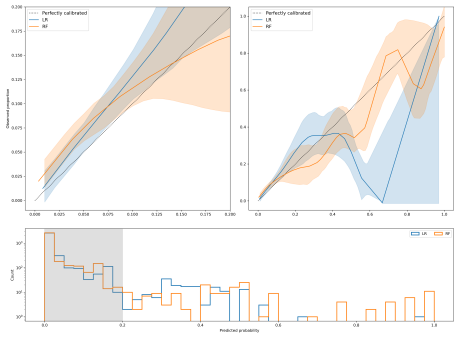
(b) MSD



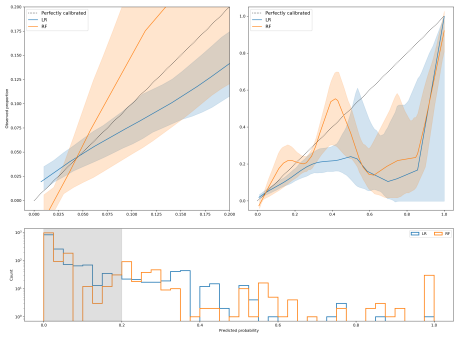
(c) HAG



(d) LEI



(e) ASZ



(f) IKA

Figure 18: Loess smoothed flexible calibration curves yielded in the individual hospitals by the baseline logistic regression (LR) and random forest (RF) models after correcting for miscalibration with the cumulative FMR + IR strategy. Left plot shows a zoom-in of the right plot in the probability range between 0 and 0.2 (grey area), which covers >95% of the predictions in each hospital.
Shaded areas around the curves represent the 95%CIs.

E.6 Conclusion

In this analysis we showed that the logistic regression (LR) and especially the random forest (RF) model showed miscalibration if these had been developed and implemented between the wave 1 and wave 2 period (i.e. August 2020) without any model updating. After examining different strategies, we showed that a combination of monthly model retraining and hospital-specific re-calibration using isotonic regression could have corrected for this miscalibration effectively. The results of this strategy (cumulative FMR + IR) are presented in the main paper, referred to as 'monthly model updating'.

References

[1] Knight, S. R., Ho, A., and Pius, R. *BMJ* **2**(September) (2020).

[2] Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., and Featherstone, P. I. *Resuscitation* **84**(4), 465–470 (2013).

[3] Xie, G., Ding, F., Han, L., Yin, D., Lu, H., and Zhang, M. *Allergy: European Journal of Allergy and Clinical Immunology* **76**(2), 471–482 (2021).

[4] Linssen, J., Ermens, A., Berrevoets, M., Seghezzi, M., Previtali, G., van der Sar-van der Brugge, S., Russcher, H., Verbon, A., Gillis, J. M., Riedl, J., de Jongh, E., Saker, J., Münster, M., Munnix, I. C., Dofferhof, A., Scharnhorst, V., Ammerlaan, H., Deiteren, K., Bakker, S. J., Van Pelt, L. J., Kluiters-de Hingh, Y., Leers, M. P., and van der Ven, A. J. *eLife* **9**, 1–28 (2020).

[5] Liu, F., Li, L., Xu, M., Wu, J., Luo, D., Zhu, Y., Li, B., and Song, X. (January) (2020).

[6] Foy, B. H., Carlson, J. C., Reinertsen, E., Padros I Valls, R., Pallares Lopez, R., Palanques-Tost, E., Mow, C., Westover, M. B., Aguirre, A. D., and Higgins, J. M. *JAMA network open* **3**(9), e2022058 (2020).

[7] Yu, F., Yan, L., Wang, N., Yang, S., Wang, L., Tang, Y., Gao, G., Wang, S., Ma, C., Xie, R., Wang, F., Tan, C., Zhu, L., Guo, Y., and Zhang, F. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **71**, 793–798 (2020).

[8] Coomes, E. A. and Haghbayan, H. *Reviews in Medical Virology* **30**(6), 1–9 (2020).

[9] Dahan, S. M. *Orphanet Journal of Rare Diseases* **21**(1), 1–9 (2020).

[10] McClish, D. K. *Medical Decision Making* **9**(3), 190–195 (1989).

[11] Qin, G. and Hotilovac, L. *Statistical Methods in Medical Research* **17**(2), 207–221 (2008).

[12] Boyd, K., Eng, K. H., and Page, C. D. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science*, volume 8190. Springer, Berlin, Heidelberg, (2013).

[13] Vickers, A. J., van Calster, B., and Steyerberg, E. W. *Diagnostic and Prognostic Research* **3**(1), 1–8 (2019).

[14] Horby, P., Lim, W. S., Emberson, J. R., and Mafham, M. *New England Journal of Medicine* , 1–11 (2020).

[15] Storkey, A. J. In *In Dataset Shift in Machine Learning*, 3–28. MIT Press, (2009).

[16] Niculescu-Mizil, A. and Caruana, R. *Proceedings of the 22nd international conference on Machine learning* (2005).

[17] Cox, D. R. *Miscellanea* (1953), 562–565 (1958).

[18] Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., MacAskill, P., McLernon, D. J., Moons, K. G., Steyerberg, E. W., Van Calster, B., Van Smeden, M., and Vickers, A. J. *BMC Medicine* **17**(1), 1–7 (2019).

Dynamic prediction of mortality in covid-19 patients in the intensive care unit: a retrospective multi-center cohort study

J.M. Smit

July 1, 2021

Abstract

Background The covid-19 pandemic has overwhelmed intensive care units (ICUs) worldwide. Improved prediction of a covid-19 patient's risk of dying may assist decision making in the intensive care unit (ICU) setting. In contrast to traditional mortality models like APACHE II [1] and SAPS II [2], dynamic mortality models allow for repeated risk stratification of patients throughout the ICU stay. Earlier works [3, 4, 5] in dynamic mortality modelling show promising results, although most of these works propose models for the general (non-covid) ICU population and use relatively long or unspecified prediction horizons. In this study, we report on the development and retrospective validation of a model for dynamic, near-term mortality for critically ill covid-19 patients.

Methods We collected EMR data from 3 481 ICU admissions with a covid-19 infection from the Dutch Data Warehouse (DDW) [6], coming from 25 different ICUs in the Netherlands. We extracted daily samples of each patient and trained both a linear (logistic regression) and non-linear (random forest) model to predict in-ICU mortality within 24 hours from the moment of prediction. Isotonic regression was used to re-calibrate the predictions of the trained models. We evaluated the models in a leave-one-ICU-out (LOIO) cross-validation procedure.

Findings Validation in 21 out of 25 and 18 out of 25 ICUs yielded an area under the receiver operating characteristic curve (AUROC) >0.80 for the logistic regression and random forest model, respectively. In the four hospitals that yielded an AUROC<0.8, local differences in protocols concerning discontinuation of treatment may have played a role. The re-calibrated model estimations showed good calibration for both models (calibration intercept $= -0.12$, slope $= 0.87$ for logistic regression and intercept $= -0.05$, slope $= 0.82$ for random forest).

Interpretation This study is different from previous works on dynamic mortality prediction in the ICU as we presented a model specifically for covid-19 patients and introduced near-term mortality predictions (compared to long-term or in-hospital mortality). The predictions were calculated based on a mixture of static information (e.g. age and sex) and dynamic information (e.g., vital signs and laboratory values) and the importance of individual predictors was quantified using SHAP values [7], where we found FiO2, oxygen saturation and pH to be important predictors. The potential clinical utility of dynamic mortality models such as guidance in resource allocation and real-time patient benchmarking could be topics for future research.

Funding No specific funding.

Contents

1 Research in context

1.1 Evidence before this study

An existing systematic review evaluated prediction models for covid-19 indexed in PubMed, Embase, and Ovid up to 1 July 2020 and arXiv, medRxiv, and bioRxiv up to 5 May, 2020. Most of the multivariate models were not recommended for clinical implementation due to high or unclear risk of bias, whereas only one prognostic model was identified as promising. This model is a 'static' model, predicting in-hospital mortality for covid-19 patients based on measurements from the day of admission. Similar and well-known static mortality models like APACHE II and SAPS II have been widely implemented in clinical practise long before the covid-19 pandemic emerged. Dynamic mortality modelling can provide real-time patient prognostication based on the most up-to-date patient status. Several recent studies reported promising results on dynamic mortality models for the critically ill, although all of these were developed before the covid-19 pandemic. Furthermore, most of the methods predict in-hospital mortality or long-term (e.g. 90-day) mortality.

1.2 Added value of this study

We developed a dynamic, near-term mortality model for critically ill covid-19 patients in intensive care units. This model offers real-time predictions of a patient's prognosis, based on updated measurements collected throughout the ICU admission. By predicting mortality within 24 hours rather than, e.g., in-hospital mortality, we aim to model an acute risk for mortality, rather than identify patients who are generally more likely to die after ICU admission. To our knowledge, this is the first dynamic mortality model specifically developed for critically ill covid-19 patients, as well as the first model to predict near-term mortality in the ICU.

1.3 Implications of all the available evidence

This study shows that it is possible to a develop a dynamic, near-term mortality model for critically ill covid-19 patients based on a mixture of static information (e.g. age and sex) and dynamic information (e.g., vital signs and laboratory values), using relatively simple classification models. We offered insight in the model predictions by calculating individual contributions of the different predictors to the predicted risk. The clinical utility of dynamic prediction models with mortality as clinical endpoint remains an open discussion. Further research is needed to examine its possible applications, such as guidance in resource allocation and real-time patient benchmarking. Finally, the presented mortality models may serve as a guidance in the development of causal models, which may support decision making in the ICU.

2 Introduction

The covid-19 pandemic has continued to overwhelm intensive care units (ICUs) worldwide. Improved prediction of a covid-19 patient's risk of dying may assist decision making in the intensive care unit (ICU) setting. Well-known scoring systems like APACHE II [1] and SAPS II [2] provide static predictions for hospital mortality among the general ICU population based on measurements obtained during the first 24 hours of admission in the ICU. These static prediction models leave events that may occur later during ICU admission, potentially influencing the prognosis, unconsidered.

In contrast, dynamic mortality modelling enables real-time mortality predictions throughout the ICU admission. Recent works on dynamic mortality prediction in the ICU [4, 3, 5] have shown promising results, however, all were developed before the covid-19 pandemic. The patient population in the ICU is very heterogeneous and therefore, improved mortality prediction may be achieved by focusing on specific patient subgroups. We developed a dynamic mortality model specifically for critically ill covid-19 patients.

Furthermore, we use a prediction horizon of 24 hours, that is, the model predicts the patient's mortality within 24 hours from the moment of prediction. In contrast, dynamic mortality models in literature predict in-hospital mortality (leaving the prediction horizon unspecified) or use relatively long prediction horizons, e.g. 90-day mortality [4]. We hypothesize that predictions for near-term mortality serve as a better surrogate for a patient's current disease severity than predictions of long-term (or in-ICU) mortality, because long-term mortality models tend to identify patients who are generally more likely to die after ICU admission and as a result, put too much emphasis on variables physicians cannot manipulate (such as admission type or age). To compare near-term and long-term mortality predictions in our setting, we developed an extra model for in-ICU mortality based on the same data and the same model development procedure.

Summarizing, we report on the development of a dynamic, near-term mortality model for critically ill covid-19 patients admitted to the ICU. The model predictors formed a mixture of static information (e.g. age and sex)

and dynamic information (e.g., vital signs and laboratory values). We tested model generalizability by extensive 'leave-one-ICU-out' (LOIO) cross validation and examine the performance of both a linear and non-linear model.

3 Methods

3.1 Data sources

We used retrospectively collected data from the Dutch Data Warehouse (DDW), a large-scale ICU data sharing collaboration in the Netherlands initiated during the covid-19 pandemic [6]. This database includes data from patients with proven covid-19 infection from 25 different ICUs in the Netherlands, admitted between February 2020 and March 2021. We extracted demographic information, vital signs, laboratory test results and blood gasses, where the vital signs (SpO2, Respiratory rate, heart rate, systolic blood pressure, temperature and Fio2) were down-sampled by extracting one value every 30 minutes. Supplemetary figure 2 gives an overview of all included predictors. We collected data from the moment of ICU admission until either discharge or death occurred. Where possible, we matched patients who were transferred between different ICUs. We handled patients who were re-admitted to the ICU (after being sent home) as separate patient episodes. Loss to follow-up occurred for patients who were transferred to ICUs that were not included in the DDW, and for patients who were still admitted at moment of data collection. In the first situation, we assumed that death did not occur within 24 hours from the moment of transfer and censored at the moment of transfer. For the latter situation, we censored at 24 hours before the final observed measurement, assuming non-informative censoring. We collected multiple observation sets, or 'samples', at different time points during each admission, starting at 24 hours after admission and adding one every 24 hours until either discharge or death occurred.

3.2 Predictors

We made a selection of predictors based on availability. To give an overview of the frequency in which different predictors occurred in the EMR, we calculated daily entry densities (i.e., fractions of non-empty daily measurements) for each predictor and for all patients individually. Given the respiratory nature of disease caused by covid-19 indection, we included an extra candidate predictor similar to the PO2/FiO2 ratio (or P/F ratio), the SpO2-FiO2 ratio, from simultaneously measured values for SpO2 and Fio2. The P/F ratio is known as an important marker for disease severity in covid-19, and the SpO2/FiO2 ratio may add extra information. To correct for time dependency of some included predictors, as well as to model the influence of the duration of ICU admission on mortality risk, we added the current length of ICU admission as a predictor.

For every predictor, the last observation was carried forward. If there was no observation available at all (i.e. data missingness), we used a K-Nearest-Neighbour (KNN) imputation algorithm. Here, missing predictors were imputed using values from the five nearest neighbours (i.e., the shortest euclidean distance regarding the remaining predictors) that have a value for the predictor, averaging these uniformly. We trained the imputation algorithm using development data and used it for imputation in both development and validation data. Predictors were centered and scaled by the standard deviation, based on the distributions of the individual predictors in the development cohort.

3.3 Model development

To model near-term mortality, we chose a prediction horizon of 24 hours. Therefore, we labelled samples as 'event samples' if death occurred within 24 hours from the time of sampling and as 'non-event samples' otherwise. We trained classification models to discriminate between these samples.

Additionally, we trained a model to predict in-ICU mortality, for which we labelled samples as 'event samples' if in-ICU death occurred and as 'non-event samples' otherwise. Supplemental figure 2 visualizes the modelling procedures and corresponding labelling strategies.

To examine the added value of modelling non-linear dependencies in the data, we trained both a linear (logistic regression) and a non-linear model (random forest) for both near-term mortality and in-ICU mortality. We trained the logistic regression (LR) model using l2 regularization and optimized the regularization strength (λ). We set the 'maximum tree depth' for the Random Forest (RF) model to three to limit model over-fitting and optimized the 'max features' hyperparameter. Model hyperparameters were optimized using using an exhaustive gridsearch in a stratified 5-fold cross-validation procedure optimizing the area under the receiver operating curve (AUROC). Supplementary table 1 shows the hyperparameter grids which were searched here.

3.4 Model re-calibration

To improve the calibration of predictions, we re-calibrated the original model predictions by isotonic regression [8]. Here, model estimates are transformed by passing the predictions through a calibrator function (a monotonically increasing step-function), which results from fitting an isotonic regressor on a left-out set of samples. To fit the calibrator function based on samples disjoint from the samples used for fitting the classification model, we made an extra split in the development cohort. Here, we randomly assigned one third of the samples to the 'calibration fold' and two thirds to the 'training fold'. First, we trained the imputation algorithm, optimized the model hyperparameters (as descibed in section 3.3) and trained the logistic regression or random forest classifier using the samples in the train fold. Then, we fitted the calibrators using the predictions by the trained classifiers and the actual labels of the samples in the calibration fold.

3.5 Model performance

We evaluated the models in a leave-one-ICU-out (LOIO) cross-validation procedure. For every iteration, patient samples from one ICU formed the validation set which we used to evaluate the models that were trained (and re-calibrated with the calibrators trained) on the patient samples from the 24 remaining ICUs, forming the development set. Thus, both for near-term mortality and in-ICU mortality, we trained 25 logistic regression (LR) and and 25 random forest (RF) models and evaluated these on the unseen data from the left-out ICU. This process is visualized in supplementary figure 1.

To evaluate model discrimination, we determined the area under the receiver operating characteristic curve (AUROC) for each LR and RF model individually. We estimated the uncertainty around this metric by calculating logit-transformation (LT)-based 95% confidence intervals (CIs) [9]. We chose this method to calculate the CIs as the model evaluation in some ICUs resulted in relatively small sample sizes and the LT-based CIs have shown good small sample performance in previous work [10].

To evaluate model calibration, we used the predictions on the left-out hospitals by all 25 LR and 25 RF models combined. We evaluated model calibration in the 'weak' and 'moderate' sense [11]. For calibration in the weak sense, we determined the calibration intercept and slope [12]. Here, an intercept of 0 and slope of 1 indicate perfect calibration. For calibration in the moderate sense, we plotted loess smoothed flexible calibration curves [11], in which deviations of points from a diagonal line with unit slope indicate lack of calibration.

3.6 Explainable predictions

To gain insight in the influence of different predictors on the model predictions of near-term mortality and in-ICU mortality, we assessed the importance of the individual predictors by training an extra LR and RF model trained on the complete cohort (all 25 ICUs). We applied the Shapley additive explanations (SHAP) algorithm [13] to obtain a surrogate for predictor importance. The SHAP value can be interpreted as the change in risk for in-ICU death in the expected model prediction when conditioning on that predictor (and in case of non-linear models, averaging these changes in risk across all possible predictor orderings). For the LR model, the SHAP value is approximated directly from the model's weight coefficients. For a more detailed description of the SHAP calculation in the RF model, we refer to [13]. The SHAP values for the included predictors were calculated based on predictions on every patient sample. Global importances of the individual predictors were obtained by averaging the magnitudes across all, i.e. the mean SHAP magnitude.

4 Results

We collected data from 3 481 ICU admissions of patients with proven covid-19 infection, among which 710 died in the ICU. Table 1 shows the baseline characteristics of the included patients. The mean age was higher for the patients who died during ICU admission (68·6 vs 61·3 years). The majority of the patients were male (72·2%), and the percentage of male patients was higher for the patients who died inside the ICU (77·2 vs 70·9 %). Patients who died during ICU admission showed relatively long ICU stays more often than the patients who survived. Figure 1 shows the number of patients who were discharged alive, died inside the ICU, were transferred to another ICU or were still admitted at the moment of data collection among the individual ICUs. The prevalence of in-ICU mortality varied between 0·06 and 0·41, with an average value of 0.20.

To give an overview of the role of different predictors during the 24 hours preceding patient death, supplementary figure 3 shows the cumulative distributions for the different predictors of based on samples taken within 24 hours before death ('event samples') compared to all other ('non-event') samples. Supplementary figure 4 visualizes the daily data availability by boxplots showing the distributions of daily entry densities (i.e.

fractions of non-empty daily measurements) across all patient samples for the candidate predictors. We judged the entry density for all candidate predictors as sufficient, and therefore we included all candidate predictors in the models.

First, we report on the results of near-term (≤24h) mortality modelling. Table 2 shows the number of patients and events for all 25 included ICUs, as well as the areas under the receiver operating characteristic curve (AUROCs, 95% CI) yielded by the logistic regression (LR) and random forest (RF), when validated on the corresponding ICU. Figure 2 visualizes the AUROCs yielded by the LR and RF models for the different ICUs. Generally, the LR models yielded a slightly higher AUROC than the RF models. Validation of the LR models yielded an AUROC >0·80 in 21 out of the 25 ICUs and in 18 out of 25 ICUs for validation of the RF models, suggesting good model generalizability in Dutch ICUs. As shown in table 2, and visualized in supplementary figure 5, AUROCs of models validated on ICUs with relatively low sample sizes yielded notably wide CIs.

As shown in figure 2, both LR and RF models validated on ICUs O, P, R and X yielded a notably low AUROC (<0.80). To check for notable deviations for any predictors in patients from these ICUs compared to the remaining ICUs, we examined the cumulative distributions for all predictors based on the samples taken within 24 hours before death ('event samples') of patients from ICU O, P, R and X (see supplementary figure 6). The Fio2 distributions in these ICUs appear notably low, as shown in figure 3. Based on a two-sided Kolmogorov-Smirnov (KS) test, we found the FiO2 distribution to be significantly lower compared to the complete distribution of event samples in ICU O (KS-statistic = 0.31, P=0.011), R (KS-statistic = 0.44, P=0.0028) and P (KS-statistic = 0.42, P=0.015), but not in ICU X (KS-statistic = 0.32, P=0.060).

Figure 4 shows the flexible calibration curves for both models with and without re-calibration, including the corresponding calibration intercepts and slopes. Without re-calibration, both models overestimated the mortality risk (calibration intercept<0) and yielded too moderate predictions (calibration slope>1). After re-calibration, both models show good calibration in the large, with a calibration intercept of −0·12 (−0·20,−0·04) and −0·05 (−0·13,0·04) and good spread of predictions, with a calibration slope of 0·87 (0·82,0·92) and 0·82 (0·78,0·94), for the LR and RF model respectively.

Table 3 shows the 20 most important predictors ranked on the mean SHAP magnitude and figure 5 shows the corresponding summary plots for the SHAP values for the LR and RF model.

For in-ICU mortality modelling, we observed similar results for model discimination and calibration, which can be found in supplementary table 3 and supplementary figures 7-9. Supplementary table 4 shows the 20 most important predictors ranked on the mean SHAP magnitude and supplementary figure 10 shows the corresponding summary plots for the SHAP values for the LR and RF model. Here, we observed changes in the predictor importances compared to near-term mortality modelling, as age became the most important predictor (in terms of mean SHAP magnitude) for the LR model, and climbed in the ranking of predictor importances for the RF model as well.

5 Discussion

5.1 Principal findings

In the majority of the included ICUs, both logistic regression (LR) and random forest (RF) models yielded good discrimination (AUROC>0.80) for near-term (≤24h) mortality. Therefore, we have shown that the models generalized well over different ICUs in the Netherlands. In the majority of the ICUs, the AUROC of the LR model was slightly better compared to the RF model and therefore, modelling non-linearities for the task of dynamic mortality prediction did not show to be advantageous in this study. Without re-calibration, both models show overestimation of the mortality, which could be explained by the class imbalance. The overall in-ICU mortality rate was relatively low (20%) and the class imbalance is aggravated by the sampling strategy (as only one event sample is taken per patient who died in the ICU). Re-calibrating the model predictions using isotonic regression showed to be an effective way to correct for this.

While we evaluated the models for discrimination separately for every ICU, we chose to evaluate model calibration based on the combined predictions of all 25 LR and 25 RF models, which were trained on different datasets. We did not evaluate calibration of individual models, as the sample sizes of most individual ICUs were too small to enable good judgement of model calibration. Also, since the 25 models are trained on similar datasets (yielding similar models), we argue that the joint validation for model calibration is reasonable.

The low FiO2 distributions we observed in samples taken within 24 hours before death ('event samples') of patients from ICU O, P, R and X compared to the event samples from the complete cohort may have influenced the predictive performance of the models validated on these ICUs. As the FiO2 is a setting for mechanical ventilation set by a physician, this observation may be explained by local differences in protocols concerning

discontinuation of treatment. However, the distributions in ICU O, P, R and X are based on relatively small sample sizes and therefore, care has to be taken in interpreting these findings.

As expected in a respiratory disease caused by covid-19 infection, predictors concerning oxygenation such as FiO2, oxygen saturation (SpO2) and arterial pH appeared in the top 10 most important predictors for both the LR and RF model. Table 3 includes two unexpected electrolytes for the LR model: magnesium and sodium. Considering the cumulative distributions of the event and non-event samples (supplementary figure 4), sodium is not expected to be an important predictor as the distributions for event and non-event samples are very similar. For magnesium, the impact (quantified in SHAP values) on mortality risk is expected to be the opposite to what we observed in the LR model (figure 5), as the event samples show a higher magnesium distribution than the non-event samples (see supplementary figure 2). These discrepancies may be explained by collinearity among the predictors and/or the way the model is regularized, as magnesium shows relatively strong correlations with several other predictors (see supplementary figure 11).

5.2 Comparisons with other studies

The model we presented in this study has some important differences compared to related work on ICU mortality prediction. First, both traditional 'static' mortality models [1, 2] and most recent works on dynamic mortality models [5, 4] focus on the general ICU population. In contrast, we presented a model specifically developed for covid-19 patients. Given the heterogeneity among ICU patients, improved mortality prediction may be achieved by focusing on sub-populations. This study therefore serves as a proof of concept to move from 'one-size-fits-all' modelling towards modelling for subgroups in the ICU population.

Second, we presented a model that predicts near-term mortality (i.e. within 24 hours from the moment of prediction), whereas most published works aim to predict in-ICU mortality or long-term mortality. We hypothesized that predictions of a dynamic mortality model serve as a better surrogate for a patient's current disease severity than predictions of long-term (or in-ICU) mortality. In recent work by Thorsen-Meyer and colleagues [4], who presented a model for long-term (90-day) mortality, age was found as the most important predictor and admission type (i.e., whether a patient was admitted to the ICU after scheduled surgery) was ranked third in terms of mean SHAP magnitude. These are both variables that cannot be manipulated by a physician, but merely indicate the generally higher risk of dying. In our study, the importance of age was found higher for in-ICU mortality modelling compared to near-term mortality modelling, which suggests that predictions of a near-term mortality model better reflect a patient's current disease state. However, age was also ranked relatively high in near-term mortality modelling (2^{nd} in the LR model) and we did not include admission type and therefore do not know the importance this variable would have had for either near-term or in-ICU mortality modelling in this setting.

Third, where other studies on dynamic mortality modelling only performed internal validation [3] or external validation in a single center [4, 5], we externally validated the model over 25 different ICUs. However, all included ICUs are located in the Netherlands and thus, the model may generalize only for the Dutch ICUs. Further external validation is needed to test the model's generalizability in other countries.

Finally, in our sampling approach using logistic regression or random forest for binary classification, each sample is (falsely) assumed to be independent of previous and next ones (IID assumption). As we are dealing with time-series data, subsequent samples from the same patient actually have high dependency. Therefore, it would be interesting to examine the added value of modelling these dependencies, e.g. by using recurrent neural networks (RNNs), which have been used in most other studies on dynamic mortality modelling [3, 4, 5]. On the other hand, an RNN is more complex than logistic regression (LR), and we doubt the added value of more complex modelling, as the LR outperformed the more complex random forest in the majority of the ICUs.

5.3 Clinical implications

The clinical utility of dynamic mortality models in the ICU remains an open discussion, several suggestions have been made in the literature. Meyer and colleagues [3] note that the mortality predictions do not target a specific pathological entity, but suggest that these may serve to draw attention of the care team, such that subtle changes that could develop into a critical state will not be missed. However, we argue that in most cases a patient dies inside the ICU, this is the result of a well-advised clinical decision, rather than a sudden event that could have been avoided by drawing more attention. Therefore, we doubt the clinical utility of mortality models when implemented as a 'red flag model', e.g. triggering an alarm for high mortality risk. Thorsen-Meyer and colleagues [4] question the clinical utility of their presented mortality model mainly due to its lack of causality, which is true for the model presented in this study as well. Based on the prediction of a model which lacks a causal structure, one cannot know if any action based on this will change the outcome. However, non-causal

mortality models like those presented here may serve as a guidance in the development of models with more causal structure.

Despite of the lack of causality of the presented dynamic mortality model, we foresee two potential clinical utilities. First, model prediction may serve as a guidance for resource allocation in the ICU, e.g. by assigning more nurses per patient for those with high risk of mortality, although further research is needed to examine whether mortality risk is indeed a good surrogate for clinical workload. Second, where static mortality models like SAPS II [2] and APACHE II [1] are widely used for benchmarking purposes, a dynamic mortality model enables benchmarking of patients throughout the whole ICU admission. As static mortality models are based on measurements from the admission day, they represent disease severity before a patient receives any care in the ICU. Therefore, they serve as a good benchmark for patients when they enter the ICU. Predictions by the dynamic model enable benchmarking of patients at any moment during admission. Since all clinical events and/or interventions occurring during the admission influence the model predictions, care has to be taken in interpreting these.

5.4 Study limitations

First, several potentially relevant predictors, such as comorbidities or medical history, were not included in the models as we did not collect these. The inclusion of these predictors may have improved the predictive performance and enabled correction for potential confounding.

Second, as supplementary figure 4 shows, not all included predictors were daily available for all patients. Missingness was especially high for certain laboratory test results (such as albumin). As demonstrated in previous work [14] on in-ICU sepsis prediction, the frequency of occurrence of predictors may be associated with the event of interest (as more lab may be requested for deteriorating patients). Thus, as low entry densities for some predictors may have been informative for mortality risk, not including predictors derived from missingness may have introduced a bias to the predictions.

Third, the included patients were admitted over a wide range of time and this differed between individual ICUs. Supplementary figure 12 shows an overview of the number of admitted patients per month for the included ICUs. This number peaks during two periods of the complete study period: the first half of 2020 and the final months of 2020 until the first months of 2021. These periods coincide with the first and second covid-19 'waves' in the Netherlands. All the included ICUs contain admissions during the first wave, but roughly half of the included ICUs contain none (or very few) admissions during the second wave, as this data was simply not collected in the DDW. Advances in covid-19 research have improved the patient care during the pandemic, for instance the start of wide spread usage of dexamethasone [15] in July 2020. Therefore, models evaluated on ICUs that only contain patients admitted during the first wave may have resulted in predictions with lower mortality risks compared to models evaluated on ICUs that contain admissions during both waves.

Fourth, we drew repeated observations (samples) on the same patient at different points in time, resulting in highly correlated sample clusters. As stated before, the binary classification methods we use, as well as the methods to estimate uncertainty of the performance, falsely assume these samples to be IID. As a consequence, the width of the calculated confidence intervals may be underestimated. A solution to this could be the use of a better model which handles the entire time-series data of an individual patient as an independent sample (e.g. an RNN).

5.5 Conclusion

In this study, we developed dynamic mortality models for covid-19 patients admitted to the ICU from a dataset of 3 460 admissions from 25 different ICUs. The models have shown good discrimination and calibration, and showed to generalize well over 25 different Dutch ICUs. The model contributes to traditional mortality models [1, 2] and more recently published dynamic mortality models [3, 4, 5] by focussing on a patient sub-population (i.e. covid-19 patients) and by introducing near-term mortality predictions instead long-term or in-ICU mortality prediction. The clinical utility of dynamic mortality models in the ICU remains an open discussion. Further research is required to examine its possible applications, such as guidance in resource allocation and real-time benchmarking. Finally, interpretable mortality models may pave the way for the development ICU models with more causal structure, which may provide actionable advice about patient treatment in the future.

6 Tables

Table 1: Baseline characteristics of the included patient episodes. SA=still admitted.

	In-ICU mortality (N=710)	**Non In-ICU mortality (N=2 771)**	**All (N=3 481)**
Age, years: mean (sd)	68·6 (9·4)	61·3 (12·5)	62·8 (12·3)
Sex, male: n (%)	548 (77·2)	1 964 (70·9)	2 512 (72·2)
Length-of-stay: n (%)			
0-24 hrs	37 (5·2)	358 (12·9)	395 (11·3)
1-7 days	171 (24·1)	942 (34·0)	1 113 (32·0)
7-14 days	191 (26·9)	550 (19·8)	741 (21·3)
14-21 days	152 (21·4)	246 (8·9)	398 (11·4)
>21 days	159 (22·4)	471 (17·0)	630 (18·1)
SA	0 (0)	196 (7·1)	196 (5·6)

Table 2: AUROCs with 95% CI for all models validated on the left-out ICU. Prevalence is the fraction of patients who experience in-ICU mortality per ICU (sorted by sample size).
LR = logistic regression
RF = random forest

ICU	**N patients**	**Prevalence in-ICU death**	**LR AUROC [95% CI]**	**RF AUROC [95% CI]**
V	21	0·33	0·93 [0·85,0·97]	0·85 [0·67,0·94]
X	39	0·41	0·71 [0·54,0·84]	0·68 [0·53,0·80]
L	44	0·20	0·94 [0·87,0·97]	0·88 [0·72,0·95]
R	51	0·31	0·80 [0·65,0·89]	0·78 [0·66,0·87]
P	53	0·25	0·74 [0·58,0·86]	0·69 [0·53,0·82]
Y	53	0·11	0·82 [0·52,0·95]	0·85 [0·62,0·95]
W	53	0·08	0·93 [0·65,0·99]	0·91 [0·72,0·98]
H	71	0·14	0·91 [0·82,0·95]	0·90 [0·79,0·95]
S	81	0·33	0·89 [0·82,0·94]	0·88 [0·81,0·93]
K	109	0·06	0·89 [0·79,0·94]	0·78 [0·61,0·90]
N	109	0·18	0·90 [0·77,0·96]	0·89 [0·78,0·95]
E	110	0·19	0·88 [0·79,0·94]	0·79 [0·66,0·88]
U	113	0·17	0·87 [0·81,0·92]	0·86 [0·80,0·91]
D	114	0·23	0·95 [0·87,0·98]	0·91 [0·83,0·95]
J	134	0·14	0·90 [0·85,0·93]	0·90 [0·84,0·93]
T	153	0·29	0·89 [0·82,0·93]	0·87 [0·81,0·92]
O	177	0·18	0·79 [0·70,0·86]	0·74 [0·66,0·81]
I	192	0·33	0·89 [0·84,0·93]	0·86 [0·81,0·90]
M	225	0·09	0·85 [0·76,0·91]	0·78 [0·67,0·86]
Q	233	0·14	0·92 [0·87,0·95]	0·90 [0·82,0·94]
F	239	0·30	0·86 [0·81,0·90]	0·81 [0·76,0·86]
G	242	0·18	0·87 [0·82,0·91]	0·85 [0·78,0·90]
A	249	0·25	0·91 [0·86,0·94]	0·89 [0·85,0·93]
C	268	0·16	0·85 [0·78,0·91]	0·82 [0·75,0·87]
B	346	0·22	0·86 [0·75,0·93]	0·88 [0·79,0·94]

Table 3: Global importances of the top 20 most important predictors for the Logistic Regression and Random Forest model, ranked on mean SHAP magnitude. The predictors in **bold** are in the top 20 predictors for both models.

Logistic regression		**Random Forest**	
Predictor	mean SHAP magnitude	Predictor	mean SHAP magnitude
fio2 [%]	0·294	SpO2/FiO2 ratio	0·035
Age [y]	0·275	**fio2 [%]**	0·028
SpO2 [%]	0·186	**pH (arterial)**	0·024
pH (arterial)	0·166	**GCS-score (motor)**	0·015
HR [bpm]	0·131	**PaCO2 [mmHg]**	0·014
WBC [10^9/L]	0·124	**GCS-score (eye)**	0·014
Potassium [mmol/L]	0·123	**SpO2 [%]**	0·014
RR [/min]	0·109	PaO2/FiO2 ratio	0·011
SBP [mmHg]	0·108	**Age [y]**	0·010
GCS-score (motor)	0·103	Creatinine [μ mol/L]	0·010
GCS-score (eye)	0·098	**Potassium [mmol/L]**	0·007
Platelet Count [10^9/L]	0·095	Urea [mmol/L]	0·005
Sodium [mmol/L]	0·094	**SBP [mmHg]**	0·004
PaCO2 [mmHg]	0·087	**Base excess [mmol/L]**	0·003
PaO2 [mmHg]	0·083	**WBC [10^9/L]**	0·003
Magnesium [mmol/L]	0·075	CRP [mg/L]	0·003
CRP [mg/L]	0·058	**HR [bpm]**	0·002
Haemoglobin [mmol/L]	0·055	**Platelet Count [10^9/L]**	0·002
Base excess [mmol/L]	0·054	LD [U/L]	0·002
Lactate (arterial) [mmol/L]	0·053	**PaO2 [mmHg]**	0·002

7 Figures

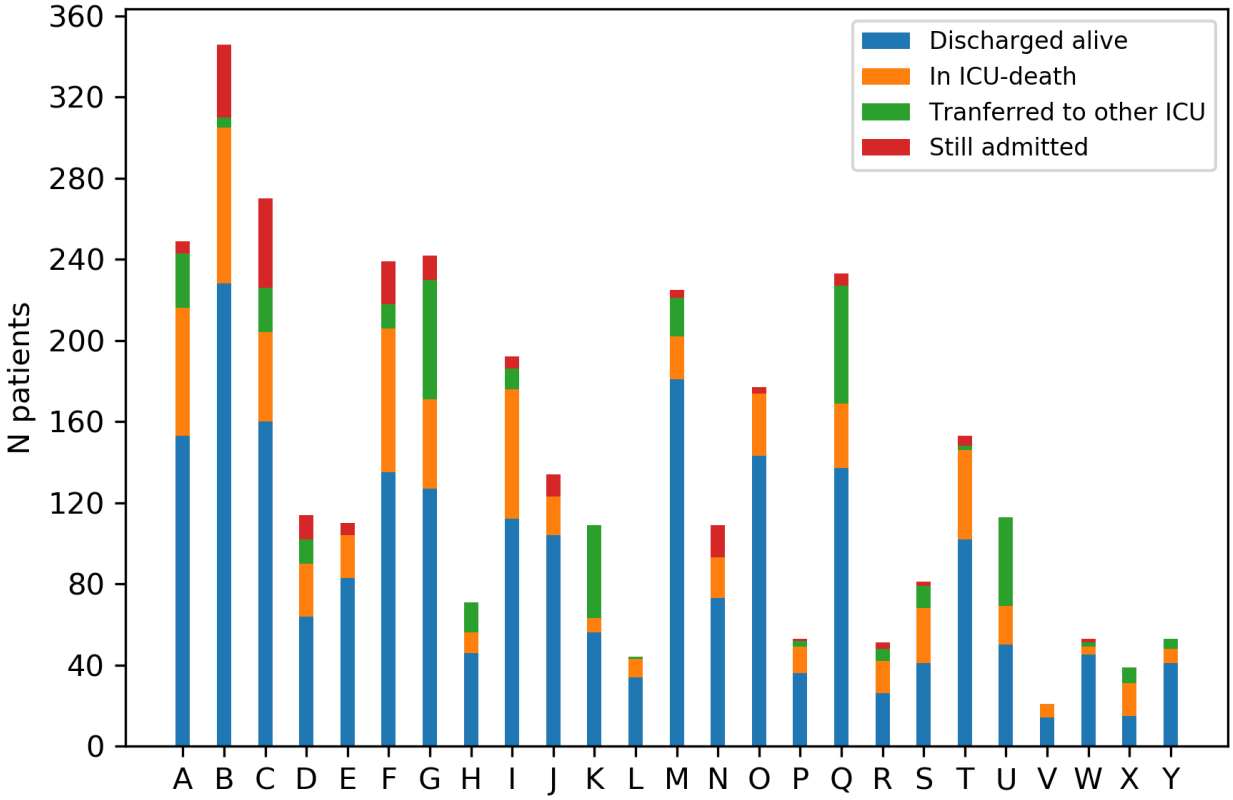


Figure 1: Number of patients in different groups (in-ICU death, discharged alive, transported to other ICU and still admitted) among the included ICUs.

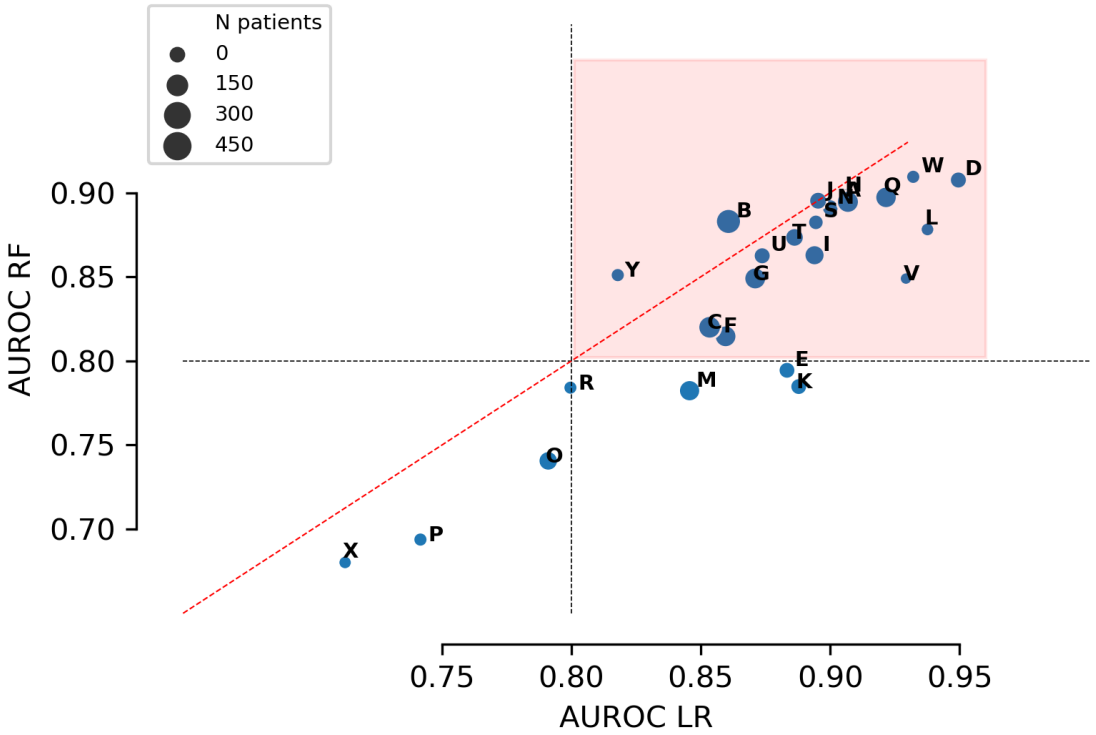


Figure 2: Areas under the receiver-operating-curve (AUROCs) for the logistic regression (LR) and random forest (RF) models validated on the different ICUs. In 18 of the 25 ICUs validated, both the LR and the RF model yielded an AUROC>0.80 (red shaded area).

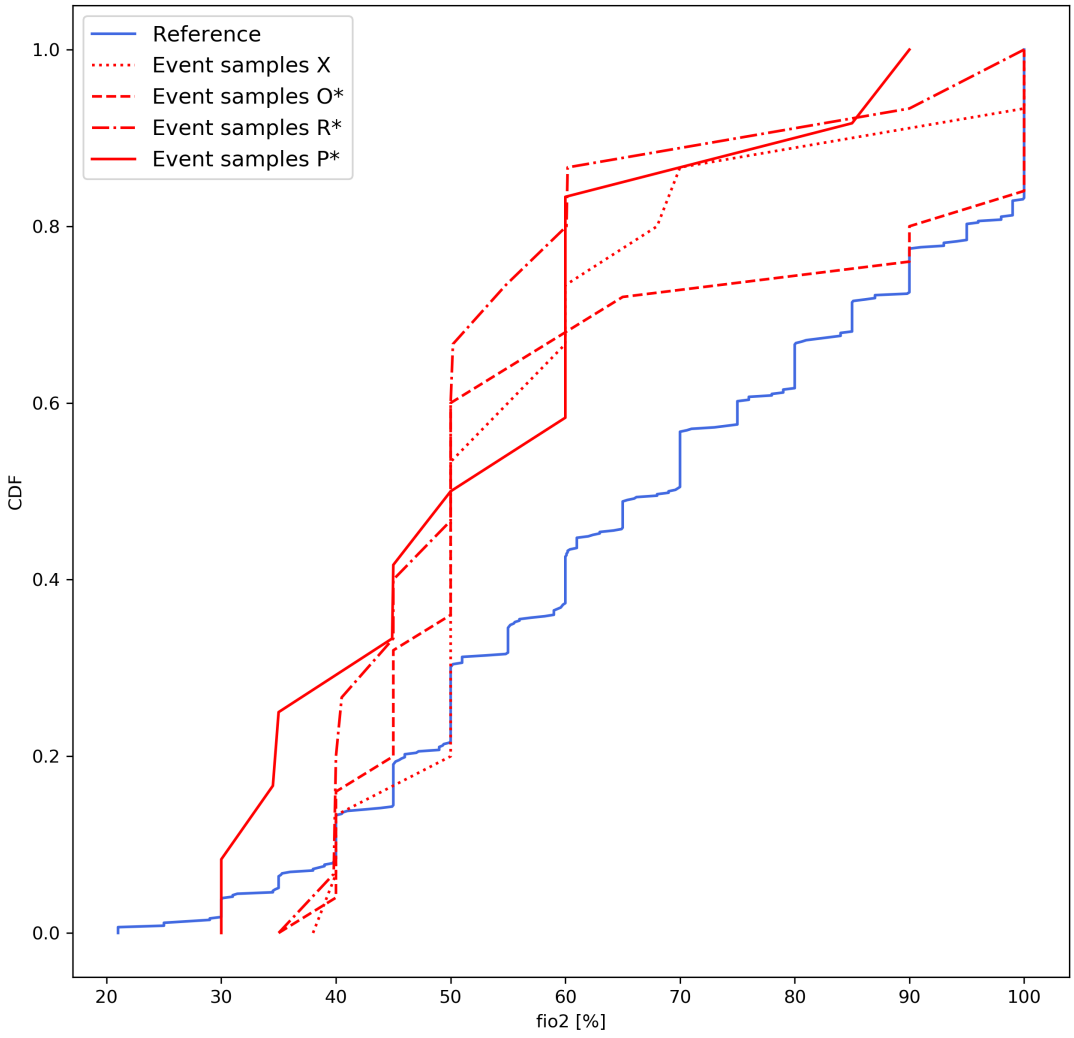
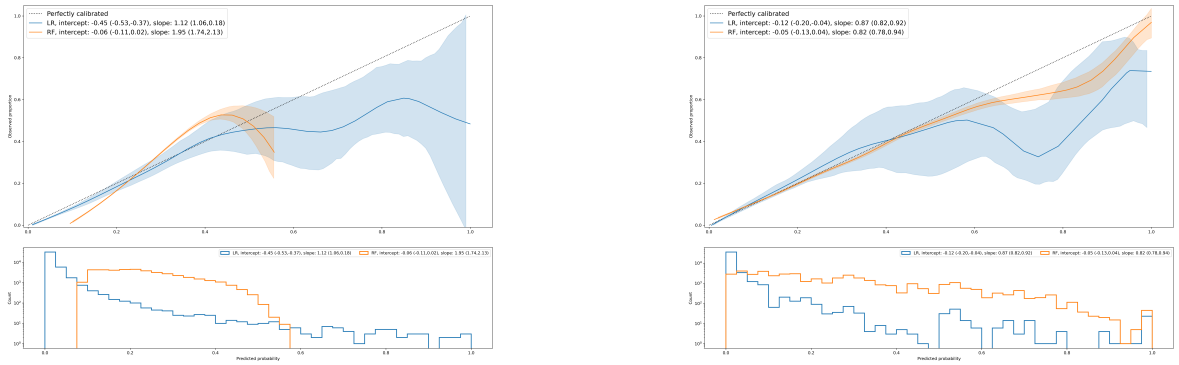


Figure 3: Cumulative distributions for Fio2 based on the samples taken within 24 hours before death ('event samples') of patients from ICU O(N=31), P(N=13), R(N=16) and X(N=16). The cumulative distributions based on event samples of patients from all ICUs (N=709) are plotted as references. *Distributions were found significantly different ($P<0.05$) from the reference based on a two-sided Kolmogorov-Smirnov test.



(a) Without re-calibration.

(b) Re-calibration by isotonic regression.

Figure 4: Loess smoothed flexible calibration curves for the logistic regression (LR) and random forest (RF) models, without re-calibration (a) and with re-calibration using isotonic regression (b). Shaded areas around the curves represent the 95%CIs.

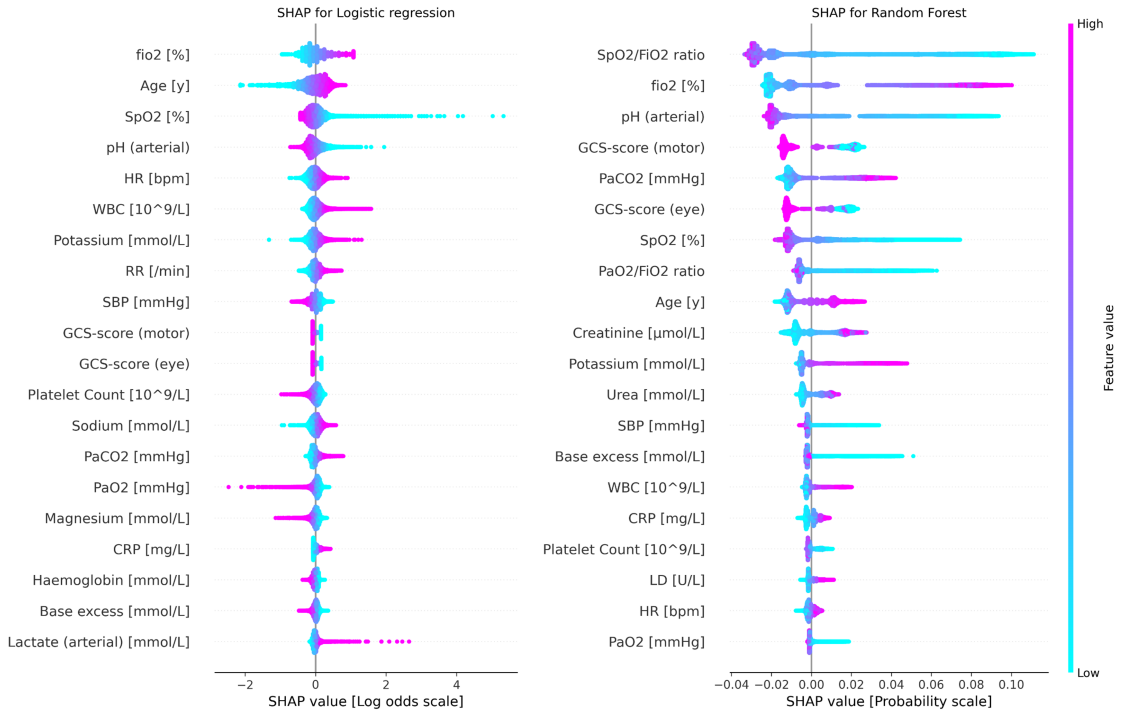


Figure 5: Summary plots for the SHAP values constructed from both Logistic regression (left) and Random Forest model (right). Each SHAP value is represented by a single dot on each feature row. Color is used to display the corresponding value of the predictor. Predictors are ordered by the average SHAP magnitude.

References

[1] Knaus, W., Draper, E., Wagner, D., and Zimmerman, J. *Crit Care Med* **13**(10), 818–828 (1985).

[2] Le Gall, J. R., Lemeshow, S., and Saulnier, F. *JAMA - Journal of the American Medical Association* **270**(24), 2957–2963 (1993).

[3] Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S. H., Stamm, C., Hofmann, T., Falk, V., and Eickhoff, C. *The Lancet Respiratory Medicine* **6**(12), 905–914 (2018).

[4] Thorsen-Meyer, H. C., Nielsen, A. B., Nielsen, A. P., Kaas-Hansen, B. S., Toft, P., Schierbeck, J., Strøm, T., Chmura, P. J., Heimann, M., Dybdahl, L., Spangsege, L., Hulsen, P., Belling, K., Brunak, S., and Perner, A. *The Lancet Digital Health* **2**(4), 179–191 (2020).

[5] Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., and Rashidi, P. *Scientific Reports* **9**(1), 1–12 (2019).

[6] Fleuren, L. M., de Bruin, D. P., Tonutti, M., Lalisang, R. C., Elbers, P. W., Gommers, D., Cremer, O. L., Bosman, R. J., Vonk, S. J., Fornasa, M., Machado, T., Dam, T., de Keizer, N. F., Raeissi, M., van der Meer, N. J., Rigter, S., Wils, E. J., Frenzel, T., Dongelmans, D. A., de Jong, R., Peters, M., Kamps, M. J., Ramnarain, D., Nowitzky, R., Nooteboom, F. G., de Ruijter, W., Urlings-Strop, L. C., Smit, E. G., Mehagnoul-Schipper, J., Dormans, T., Houwert, T., Hovenkamp, H., Londono, R. N., Quintarelli, D., Scholtemeijer, M. G., de Beer, A. A., Ercole, A., van der Schaar, M., Beudel, M., Hoogendoorn, M., Girbes, A. R., Herter, W. E., Thoral, P. J., Roggeveen, L., van Diggelen, F., el Hassouni, A., Guzman, D. R., Bhulai, S., Ouweneel, D., Driessen, R., Peppink, J., de Grooth, H. J., Zijlstra, G. J., van Tienhoven, A. J., van der Heiden, E., Spijkstra, J. J., van der Spoel, H., de Man, A., Klausch, T., de Vries, H., de Neree tot Babberich, M., Thijssens, O., Wagemakers, L., Berend, J., Silva, V. C., Kullberg, B., Heunks, L., Juffermans, N., Slooter, A., Rettig, T. C., Reuland, M. C., van Manen, L., Montenij, L., van Bommel, J., van den Berg, R., van Geest, E., Hana, A., Simsek, S., van den Bogaard, B., Pickkers, P., van der Heiden, P., van Gemeren, C., Meinders, A. J., de Bruin, M., Rademaker, E., van Osch, F., de Kruif, M., Hendriks, S. H., Schroten, N., Boelens, A. D., Arnold, K. S., Karakus, A., Fijen, J. W., Festen-Spanjer, B., Achterberg, S., Lens, J., van Koesveld, J., van den Tempel, W., Simons, K. S., de Jager, C. P., Oostdijk, E., Labout, J., van der Gaauw, B., Reidinga, A. C., Koetsier, P., Kuiper, M., Cornet, A. D., Beishuizen, A., de Jong, P., Geutjes, D., Faber, H. J., Lutisan, J., Brunnekreef, G., van Gemert, A. W., Entjes, R., van den Akker, R., Simons, B., Rijkeboer, A. A., Arbous, S., Aries, M., van den Oever, N. C., and van Tellingen, M. *Intensive Care Medicine* , 478–481 (2021).

[7] Lundberg, S. M. and Lee, S.-i. (Section 2), 1–10 (2017).

[8] Niculescu-Mizil, A. and Caruana, R. *Proceedings of the 22nd international conference on Machine learning* (2005).

[9] Qin, G. and Zhou, X. H. *Biometrics* **62**(2), 613–622 (2006).

[10] Qin, G. and Hotilovac, L. *Statistical Methods in Medical Research* **17**(2), 207–221 (2008).

[11] Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., MacAskill, P., McLernon, D. J., Moons, K. G., Steyerberg, E. W., Van Calster, B., Van Smeden, M., and Vickers, A. J. *BMC Medicine* **17**(1), 1–7 (2019).

[12] Cox, D. R. *Miscellanea* (1953), 562–565 (1958).

[13] Lundberg, S. and Lee, S.-I. (2017).

[14] Yang, M., Liu, C., Wang, X., Li, Y., Gao, H., Liu, X., and Li, J. *Critical Care Medicine* , E1091–E1096 (2020).

[15] The RECOVERY Collaborative Group. *New England Journal of Medicine* , 1–11 (2020).

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

J.M. Smit
July 1, 2021

Contents

1 Supplementary figures

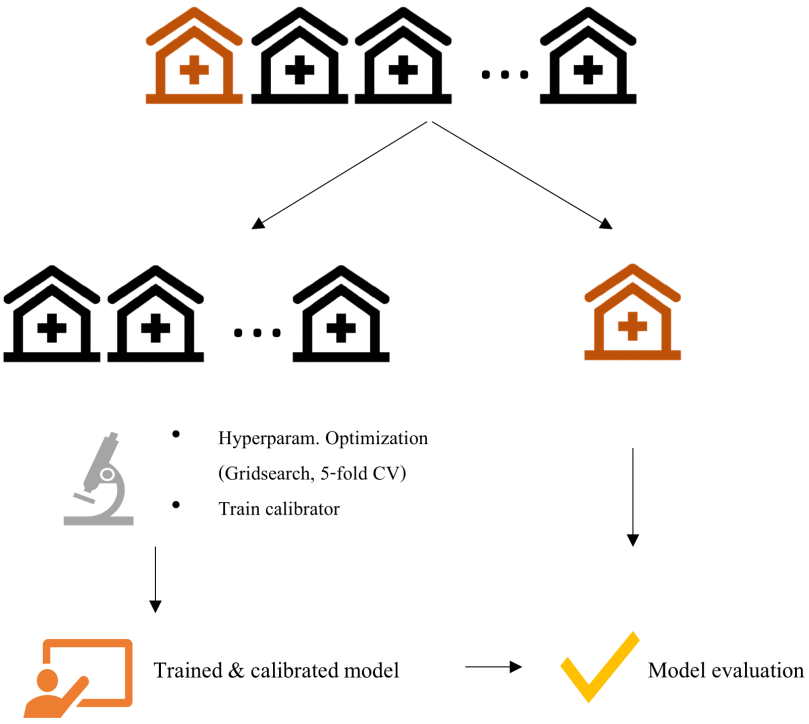


Figure 1: Leave-one-ICU-out (LOIO) cross-validation procedure.

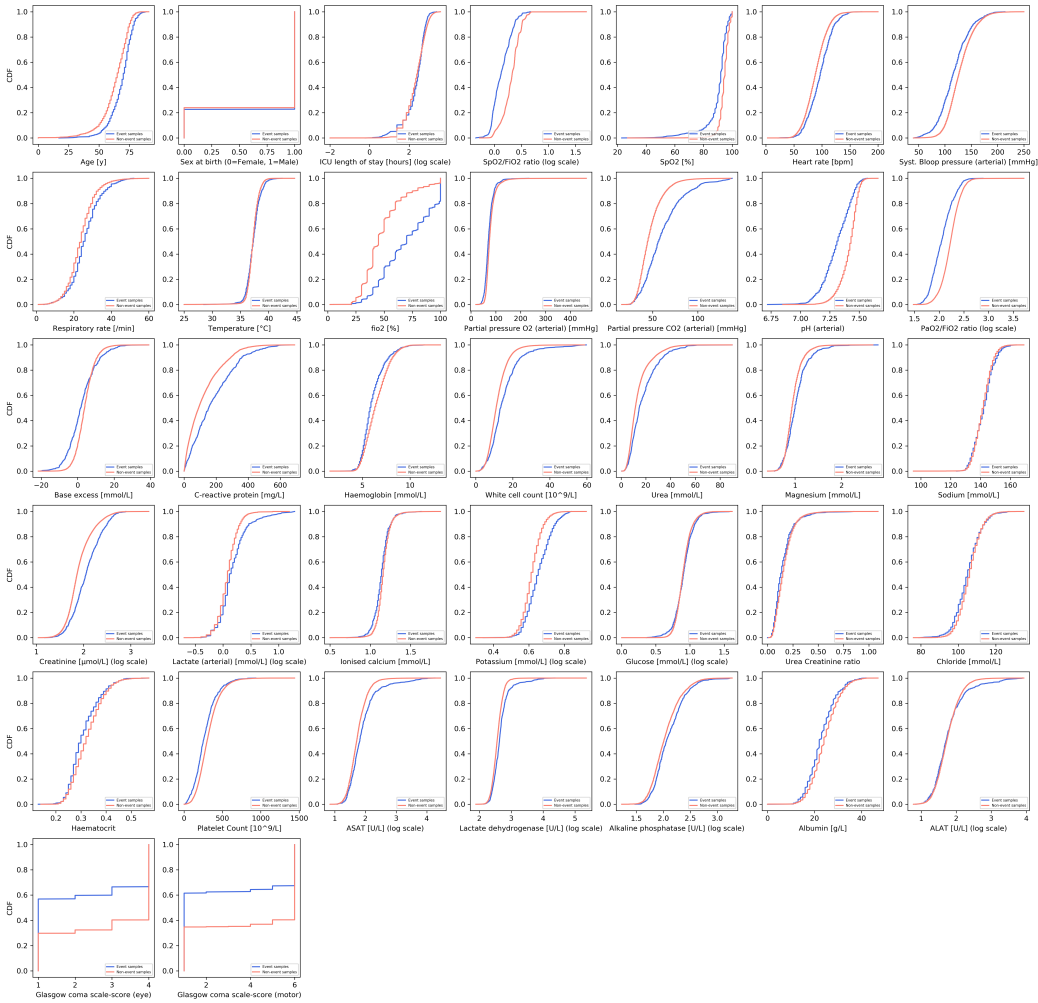
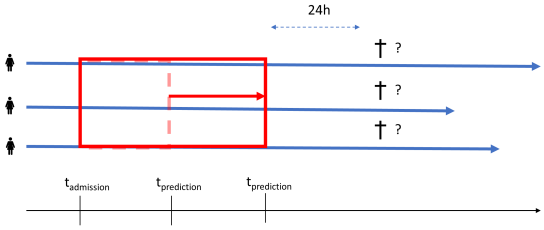
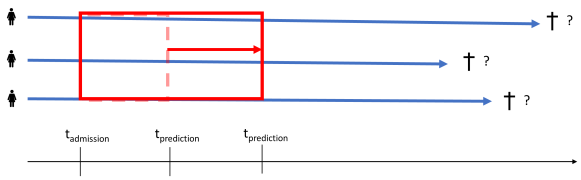


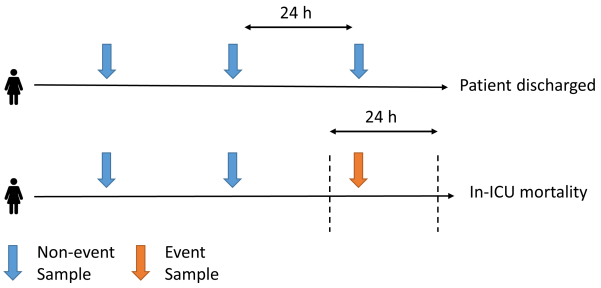
Figure 2: Cumulative distributions of different predictors based on samples taken within 24 hours of ICU death ('event samples') and all other ('non-event') samples from all included patients.



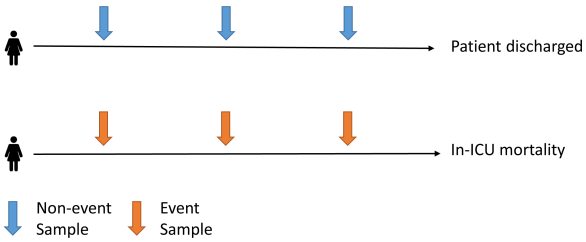
(a) Near-term mortality ($\leq$ 24 hours) modelling.



(b) In-ICU mortality modelling.



(c) Labelling strategy for near-term mortality ($\leq$ 24 hours) modelling.



(d) Labelling strategy for in-ICU mortality modelling.

Figure 3: Visual representation of near-term mortality modelling (a) and in-ICU mortality modelling (b) and the corresponding sampling strategies (c& d).

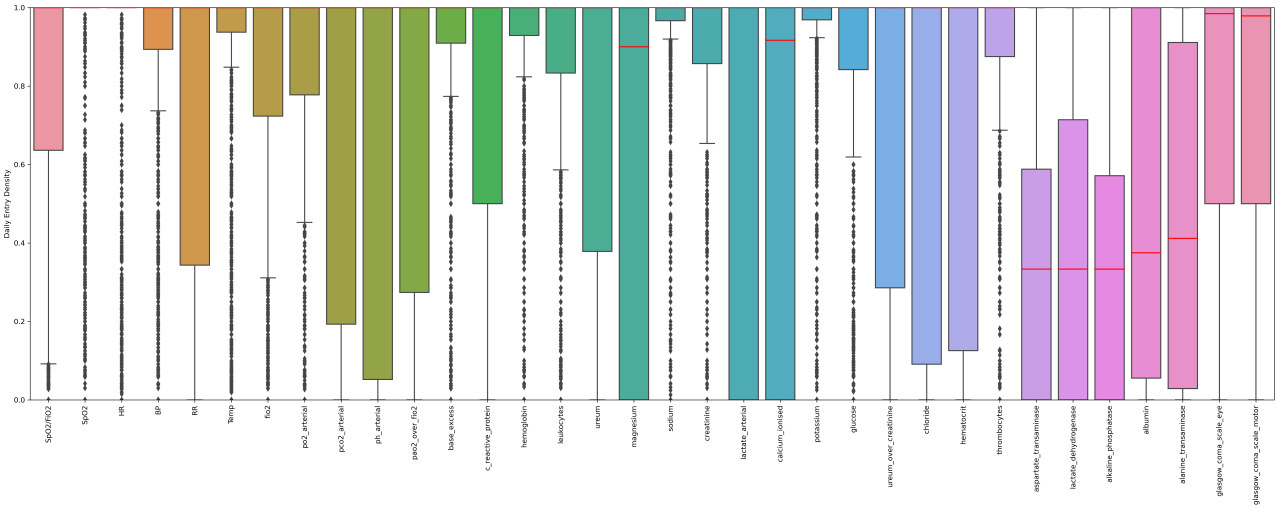


Figure 4: Boxplots of the daily entry density (i.e., fractions of non-empty daily measurements) distributions for each predictor.

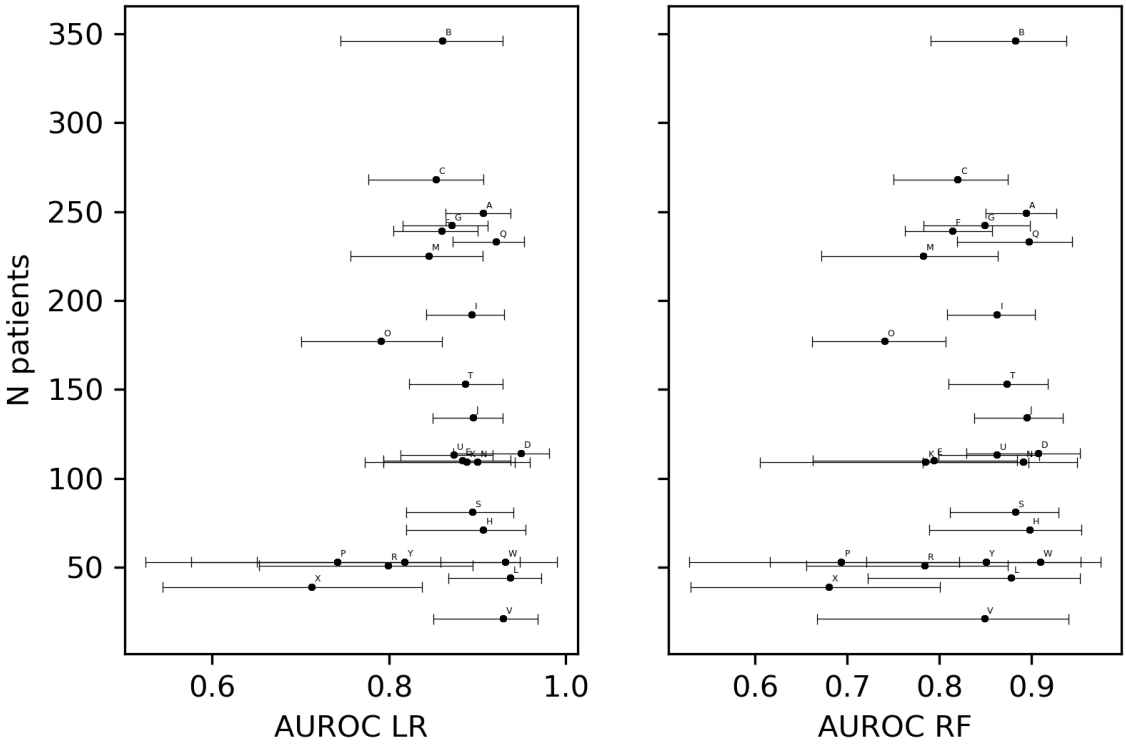


Figure 5: Near-term mortality: areas under the receiver-operating-curve (AUROCs) with logit-transformation (LT)-based 95% CIs (as described in [?]) for the logistic regression (LR) and random forest (RF) model, validated on the different ICUs sorted by sample size.

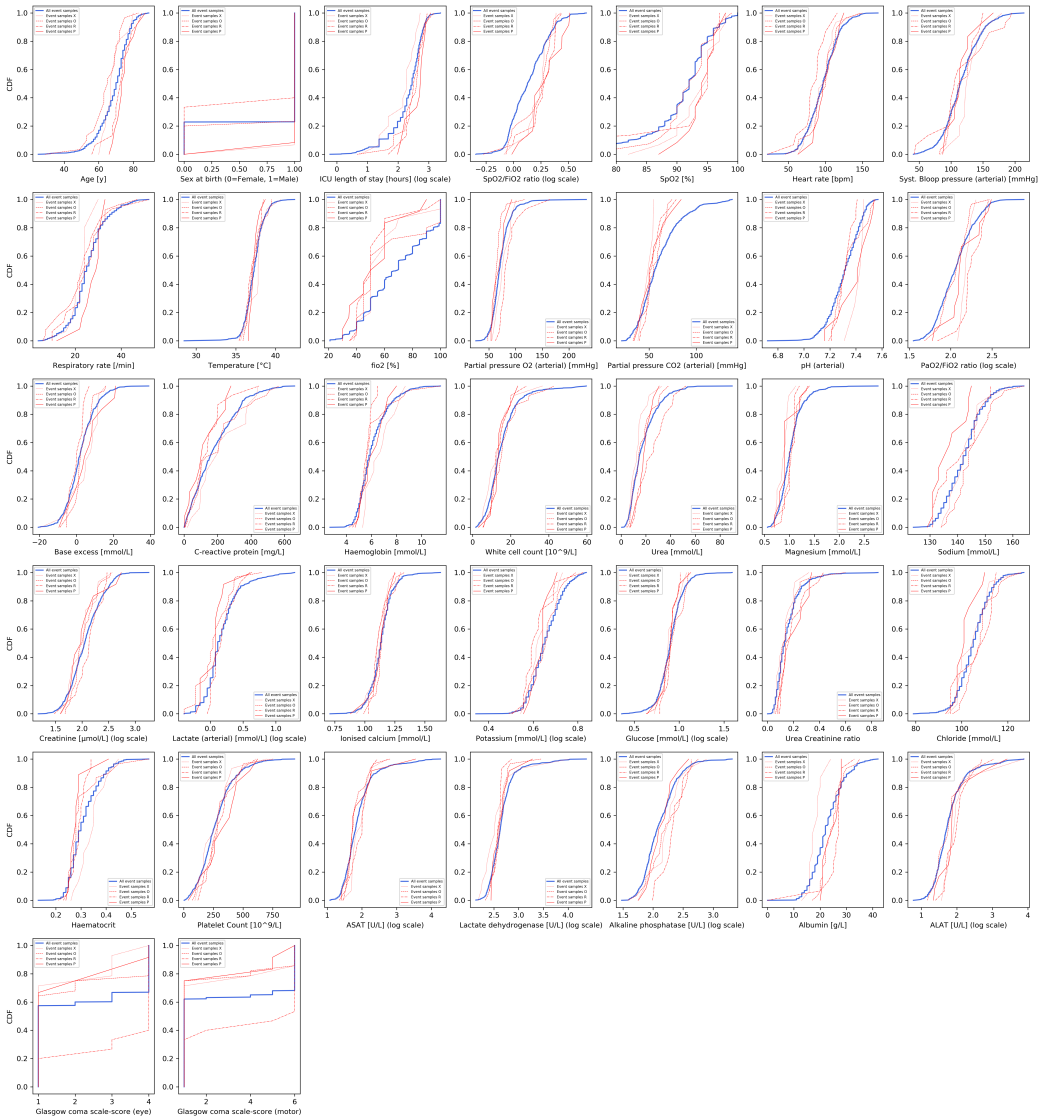


Figure 6: Cumulative distributions for all predictors based on the samples taken within 24 hours of ICU death ('event samples') of patients from ICU O(N=31), P(N=13), R(N=16) and X(N=16). The cumulative distributions based on event samples of patients from all ICUs (N=709) are plotted as references.

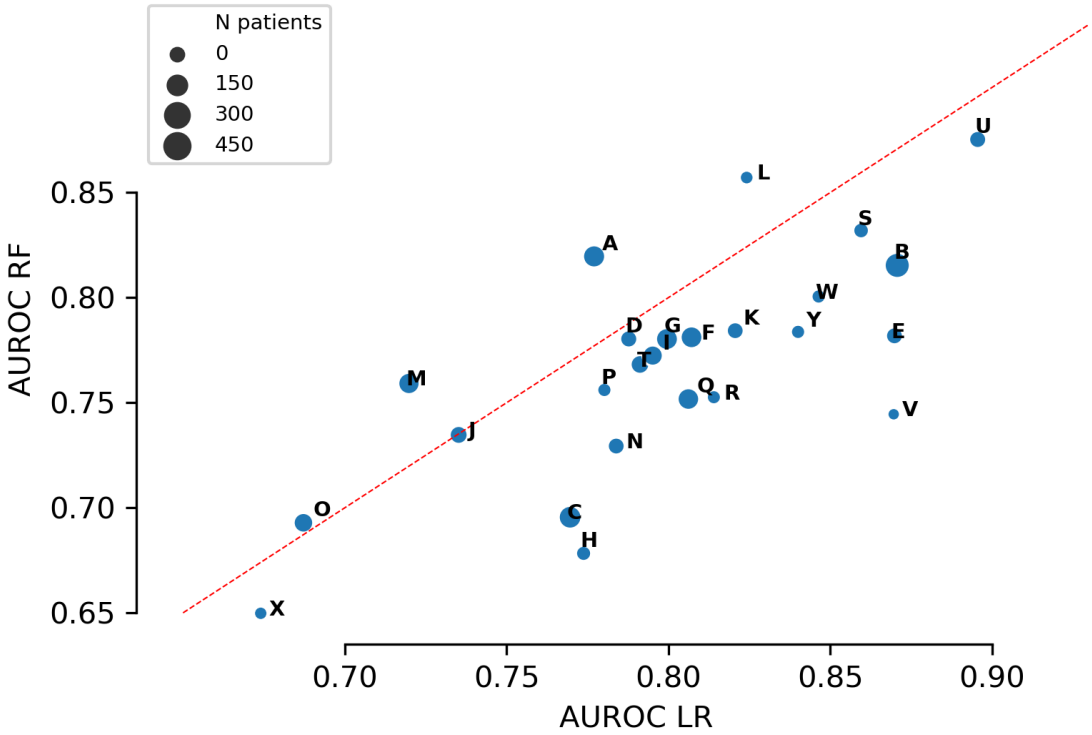


Figure 7: Results in-ICU mortality modelling: Areas under the receiver-operating-curve (AUROCs) for the logistic regression (LR) and random forest (RF) models validated on the different ICUs.

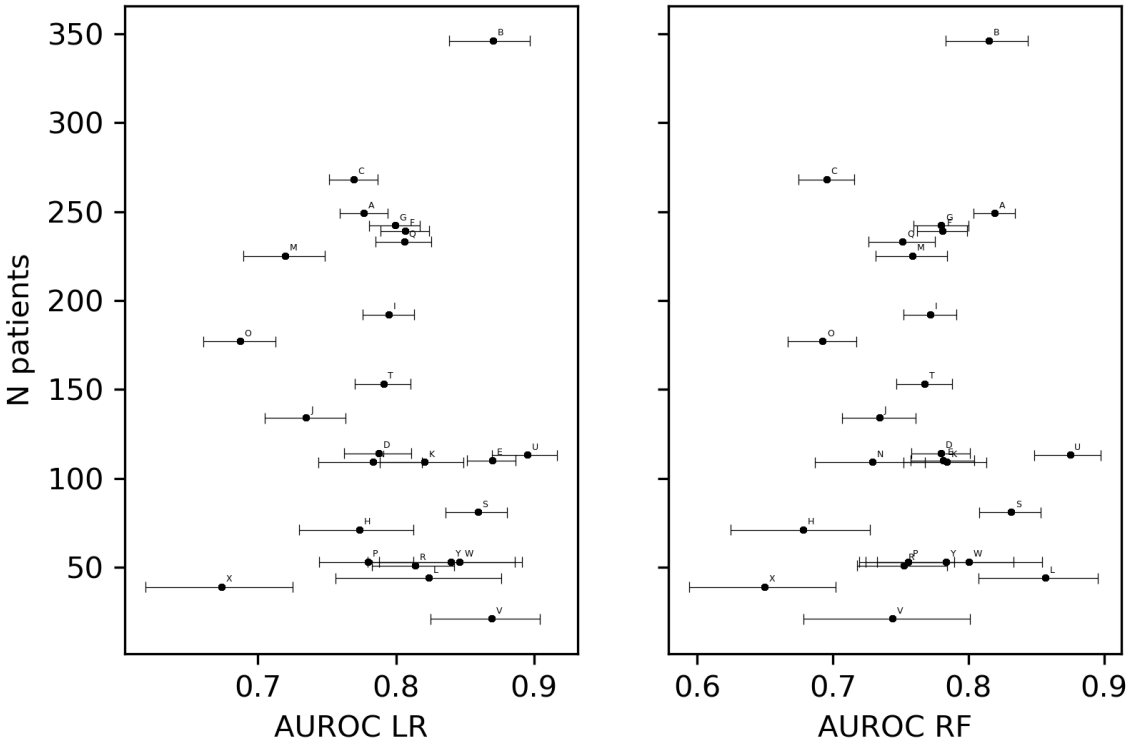
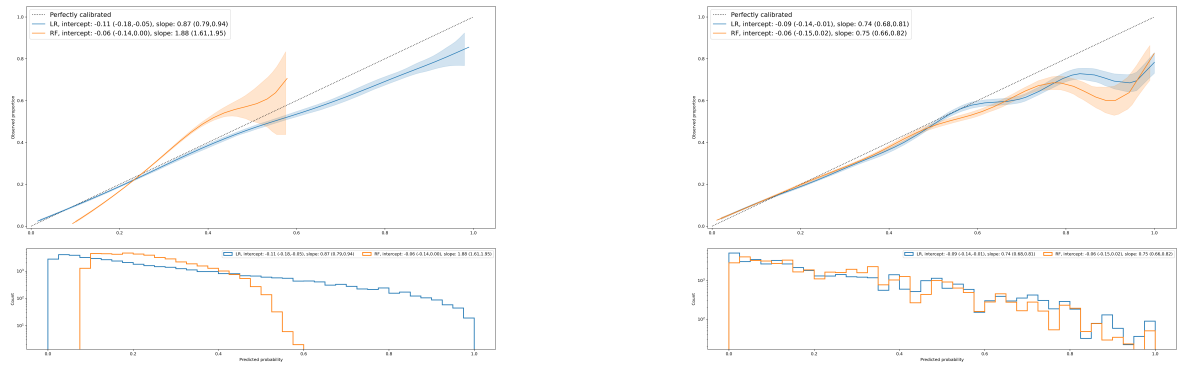


Figure 8: Results in-ICU mortality modelling: areas under the receiver-operating-curve (AUROCs) with logit-transformation (LT)-based 95% CIs (as described in [?]) for the logistic regression (LR) and random forest (RF) model, validated on the different ICUs sorted by sample size.



(a) Without re-calibration.

(b) Re-calibration by isotonic regression.

Figure 9: Results in-ICU mortality modelling: loess smoothed flexible calibration curves for the logistic regression (LR) and random forest (RF) models, without re-calibration (a) and with re-calibration using isotonic regression (b). Shaded areas around the curves represent the 95%CIs.

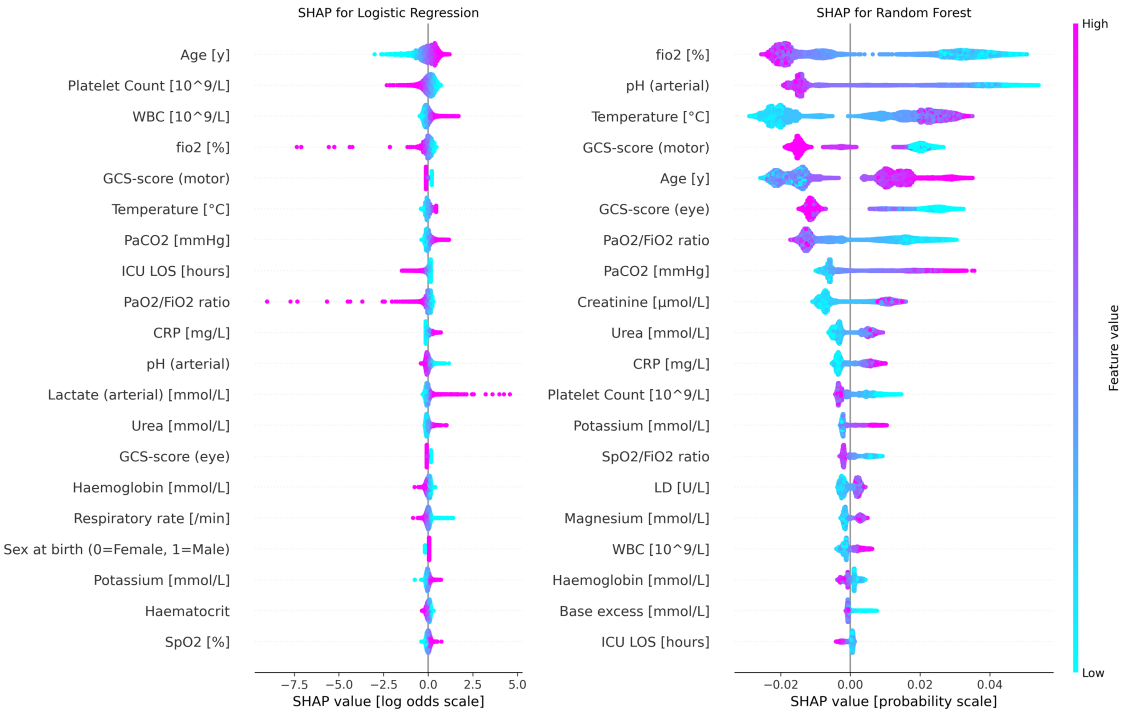


Figure 10: Results in-ICU mortality modelling: summary plots for the SHAP values constructed from both Logistic regression (left) and Random Forest model (right). Each SHAP value is represented by a single dot on each feature row. Color is used to display the corresponding value of the predictor. Predictors are ordered by the average SHAP magnitude.

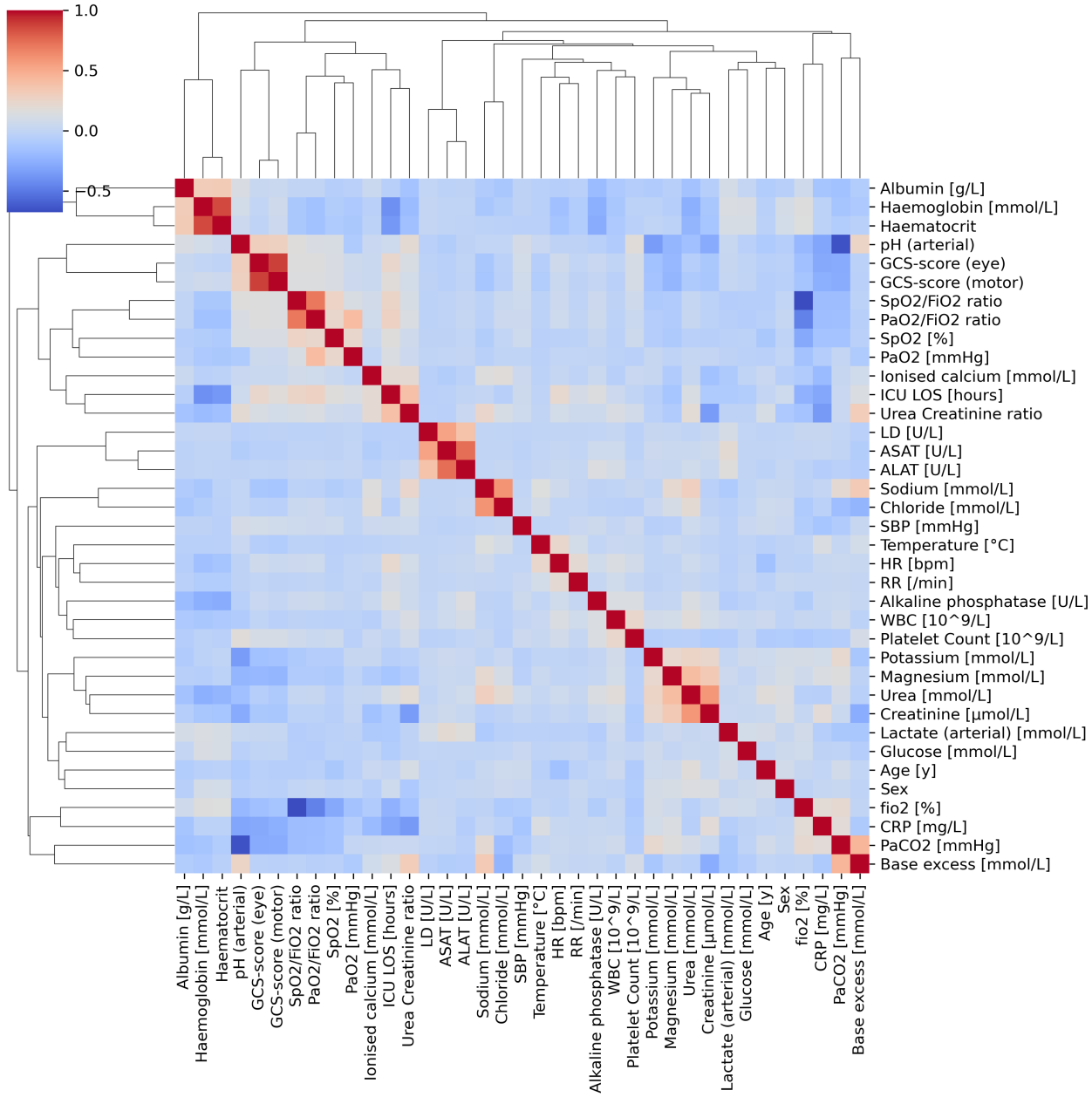


Figure 11: Clustermap of the correlation matrix of all included model predictors.

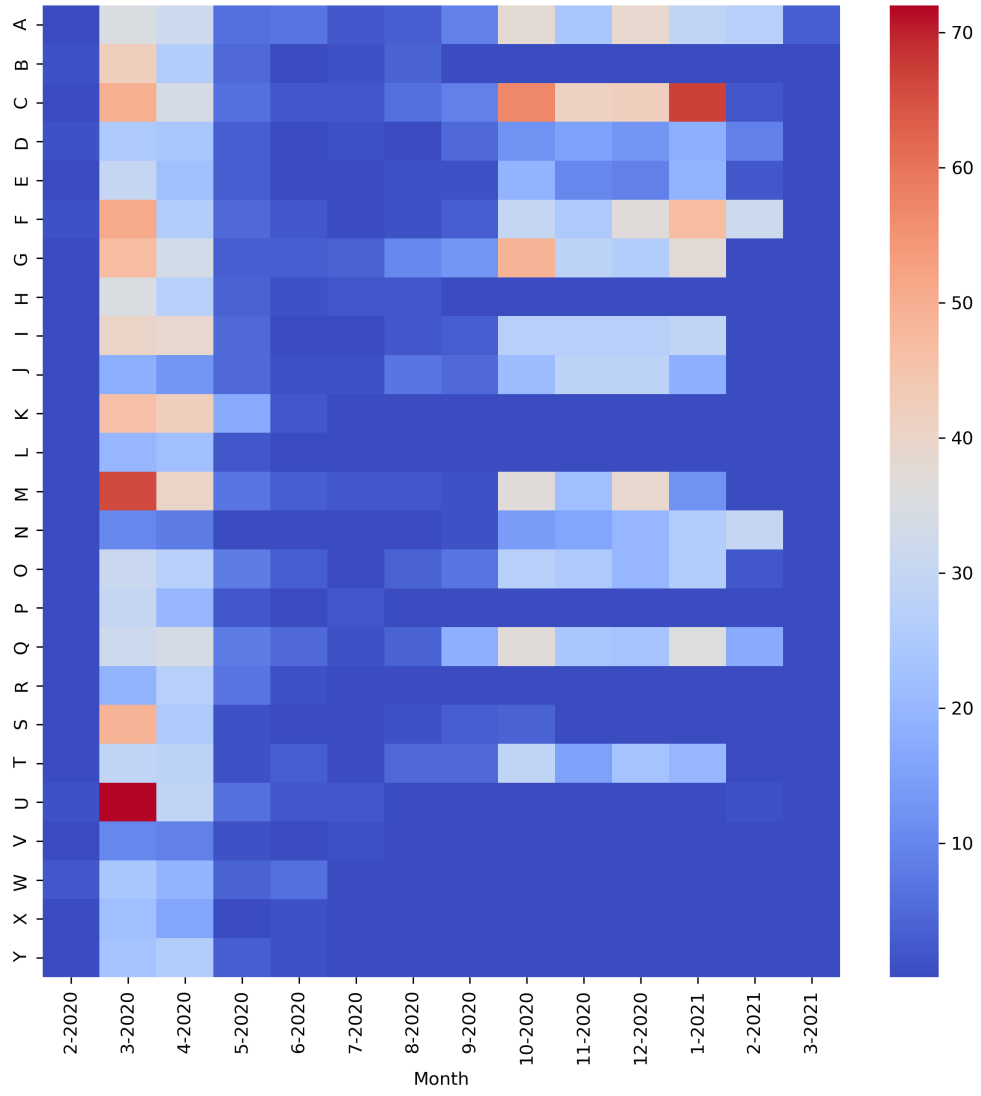


Figure 12: Number of ICU admissions per month among the 25 included hospitals. The number of patients peaks during two sub-periods of the complete study period, namely during the first half of 2020 ('wave 1') and during the final months of 2020 and first months of 2021 ('wave 2').

2 Supplementary tables

Table 1: Search spaces used in the grid-search for model hyperparameters optimization.

Model	**Hyperparameter**	**Search Space**
Logistic Regression	λ	$[10^{-4}, \ldots, 10^{4}]$ evenly spaced on log scale with 20 steps
Random Forest	max features	[p, $\sqrt{p}$, $\log_2 p$] where p is the total number of predictors.

Table 2: Results for in-ICU mortality: AUROCs with 95% CI for all models validated on the left-out ICU. Prevalence is the fraction of patients who experience in-ICU mortality per ICU (sorted by sample size) yielded by the models trained on in-ICU mortality.
LR = logistic regression
RF = random forest

code	N patients	prev	AUROC LR range	AUROC RF range
V	21	0·33	0·87 [0·82,0·9]	0·74 [0·68,0·8]
X	39	0·41	0·67 [0·62,0·73]	0·65 [0·59,0·7]
L	44	0·2	0·82 [0·76,0·88]	0·86 [0·81,0·9]
R	51	0·31	0·81 [0·78,0·84]	0·75 [0·72,0·78]
P	53	0·25	0·78 [0·74,0·81]	0·76 [0·72,0·79]
Y	53	0·11	0·84 [0·78,0·89]	0·78 [0·72,0·83]
W	53	0·08	0·85 [0·79,0·89]	0·8 [0·73,0·85]
H	71	0·14	0·77 [0·73,0·81]	0·68 [0·62,0·73]
S	81	0·33	0·86 [0·84,0·88]	0·83 [0·81,0·85]
K	109	0·06	0·82 [0·79,0·85]	0·78 [0·75,0·81]
N	109	0·18	0·78 [0·74,0·82]	0·73 [0·69,0·77]
E	110	0·19	0·87 [0·85,0·89]	0·78 [0·76,0·8]
U	113	0·17	0·9 [0·87,0·92]	0·88 [0·85,0·9]
D	114	0·23	0·79 [0·76,0·81]	0·78 [0·76,0·8]
J	134	0·14	0·74 [0·71,0·76]	0·73 [0·71,0·76]
T	153	0·29	0·79 [0·77,0·81]	0·77 [0·75,0·79]
O	177	0·18	0·69 [0·66,0·71]	0·69 [0·67,0·72]
I	192	0·33	0·8 [0·78,0·81]	0·77 [0·75,0·79]
M	225	0·09	0·72 [0·69,0·75]	0·76 [0·73,0·78]
Q	233	0·14	0·81 [0·79,0·83]	0·75 [0·73,0·78]
F	239	0·3	0·81 [0·79,0·82]	0·78 [0·76,0·8]
G	242	0·18	0·8 [0·78,0·82]	0·78 [0·76,0·8]
A	249	0·25	0·78 [0·76,0·79]	0·82 [0·8,0·83]
C	268	0·16	0·77 [0·75,0·79]	0·7 [0·67,0·72]
B	346	0·22	0·87 [0·84,0·9]	0·82 [0·78,0·84]

Table 3: Results for in-ICU mortality: Global importances of the top 20 most important predictors for the Logistic Regression and Random Forest model trained for in-ICU mortality, ranked on mean SHAP magnitude. The predictors in **bold** are in the top 20 predictors for both models.

Logistic regression		**Random Forest**	
Predictor	mean SHAP magnitude	Predictor	mean SHAP magnitude
Age [y]	0·370	**fio2 [%]**	0·023
Platelet Count [10^9/L]	0·240	**pH (arterial)**	0·021
White cell count [10^9/L]	0·156	**Temperature [°C]**	0·020
fio2 [%]	0·150	**GSC-score (motor)**	0·016
GCS-score (motor)	0·141	**Age [y]**	0·015
Temperature [°C]	0·123	**GSC-score (eye)**	0·015
PaCO2 [mmHg]	0·120	**PaO2/FiO2 ratio**	0·013
ICU length of stay [hours]	0·117	**PaCO2 [mmHg]**	0·009
PaO2/FiO2 ratio	0·112	Creatinine [μmol/L]	0·008
C-reactive protein [mg/L]	0·106	**Urea [mmol/L]**	0·004
pH (arterial)	0·099	**C-reactive protein [mg/L]**	0·004
Lactate (arterial) [mmol/L]	0·098	**Platelet Count [10^9/L]**	0·004
Urea [mmol/L]	0·093	**Potassium [mmol/L]**	0·003
GCS-score (eye)	0·092	SpO2/FiO2 ratio	0·003
Haemoglobin [mmol/L]	0·083	Lactate dehydrogenase [U/L]	0·002
Respiratory rate [/min]	0·083	Magnesium [mmol/L]	0·002
Sex at birth (0=Female, 1=Male)	0·080	**White cell count [10^9/L]**	0·002
Potassium [mmol/L]	0·074	**Haemoglobin [mmol/L]**	0·002
Haematocrit	0·071	Base excess [mmol/L]	0·001
SpO2 [%]	0·069	ICU length of stay [hours]	0·001

References

[1] Qin, G. and Hotilovac, L. *Statistical Methods in Medical Research* **17**(2), 207–221 (2008).

[2] Cox, D. R. *Miscellanea* (1953), 562–565 (1958).