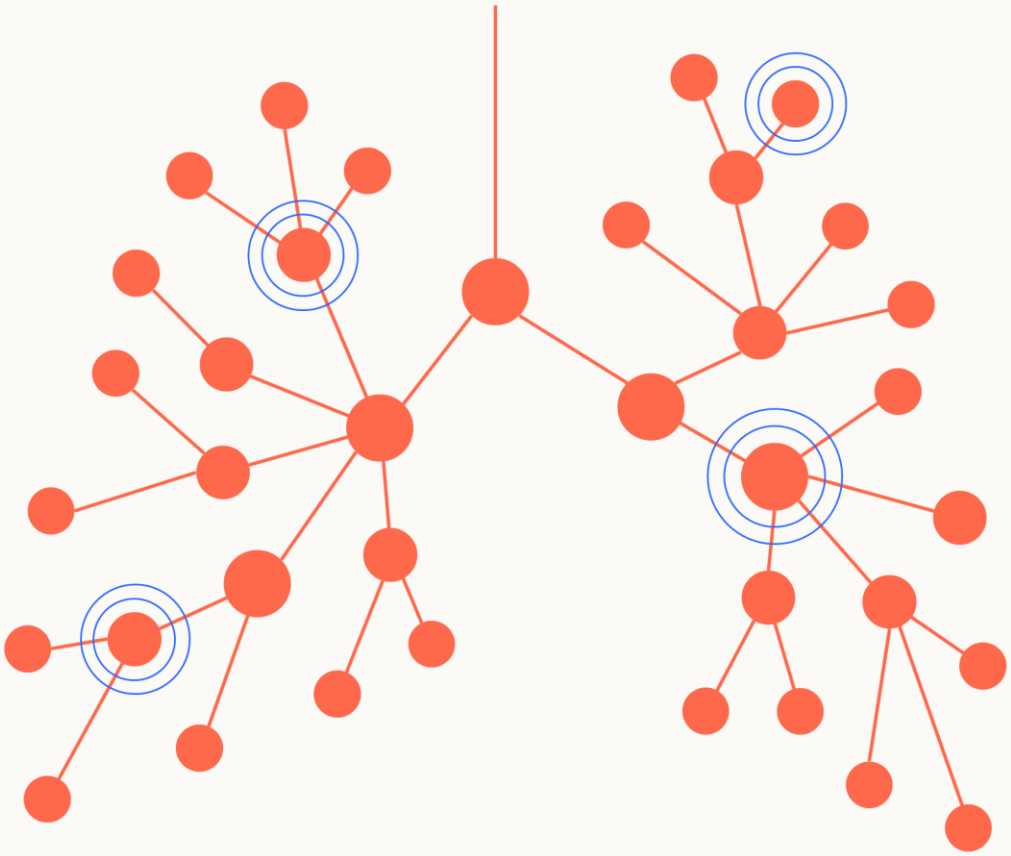
Developing a prediction model for respiratory deterioration in mechanically ventilated ICU patients



Emmelieve den Breejen

Master thesis Technical Medicine

February 2026

Developing a prediction model for respiratory deterioration in mechanically ventilated ICU patients

Emmelieve den Breejen

Student number: 4996690

26 February 2026

Thesis in partial fulfillment of the requirements for the joint degree of Master in Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Intensive Care Medicine, LUMC

May 2025 – February 2026

Supervisors:

Dr. A. Schoe, LUMC

Dr. D.M.J. Tax, TU Delft

Drs. F.E. Smits, LUMC

Thesis committee members:

Dr. A. Schoe, LUMC (chair)

Dr. D.M.J. Tax, TU Delft

Dr. Ir. H.J. Krijthe, TU Delft

Drs. F.E. Smits, LUMC

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Preface

Hierbij presenteer ik mijn masterthesis voor Technische Geneeskunde, waarin je leest hoe ik een dappere poging heb gewaagd om respiratoire verslechtering bij beademde IC patiënten te voorspellen, en of ik daarin geslaagd ben. Ik heb met veel plezier aan dit uitdagende project gewerkt, waarin ik me kon vastbijten in technische principes, telkens opnieuw naging hoe de methodes en uitkomsten zich verhouden tot IC patiënten en de klinische praktijk, en ook steeds moest uitleggen waar ik mee bezig was. Dat laatste is een van de dingen die ik het meest uitdagend en daarom ook het leukste vond om te doen.

Dit project is mede tot stand gekomen dankzij een geweldig team van begeleiders. Floor, David en Bram, heel erg bedankt voor jullie sturing, goede ideeën, vragen en discussies, en bemoedigende woorden. Ik reken me rijk met jullie. Jesse, dankjewel dat je de tijd neemt om mijn thesis te lezen en deel van mijn afstudeercommissie wil zijn.

Tot slot ben ik familie en vrienden dankbaar voor jullie trouwe steun en aanmoediging tijdens mijn hele studie. In het bijzonder mijn ouders, voor het rustige en liefdevolle thuis dat jullie mij geven. Dat heeft er, meer dan ik me altijd heb doen beseffen, aan bijgedragen dat ik nu bijna Technisch Geneeskundige geworden ben.

Ik wens je veel leesplezier!

Emmelieve
Sliedrecht, februari 2026

Table of Contents

1 Abstract

Objective

The primary aim of this study was to develop and validate a machine learning prediction model for respiratory deterioration in mechanically ventilated Intensive Care Unit (ICU) patients. The secondary aim was to identify physiological parameters associated with respiratory failure during mechanical ventilation.

Methods

Two distinct prediction models were developed using data from ICU patients admitted to the Leiden University Medical Centre (LUMC) between 2018 and 2023. Patients receiving invasive mechanical ventilation (IMV) for at least 48 hours with a PaO_2/FiO_2 ratio below 40 kPa were included and allocated to COVID training, COVID test, or non-COVID test sets. Model 1 predicts respiratory deterioration within six hours after switching from controlled to assisted ventilation. Model 2 is an hourly updating model predicting respiratory deterioration occurring more than six hours after this switch. XGBoost models were cross-validated on the COVID training set to identify the optimal observation windows and prediction horizons, after which feature selection and hyperparameter optimisation were performed. Model 1 was optimised for the area under the receiver operating characteristic (AUROC) and Model 2 for the area under the precision-recall curve (AUPRC). Discriminative performance, generalisability, and clinical utility were evaluated on the COVID and non-COVID test sets.

Results

A total of 296 patients were included in the COVID training set, 78 in the COVID test set, and 755 to the non-COVID test set. For Model 1, a one-hour observation window was selected. The most important features were the mean fraction of inspired oxygen (FiO_2), propofol infusion rate, and peripheral oxygen saturation (SpO_2). This model achieved an AUROC of 0.78 on the COVID test and 0.76 on the non-COVID test set. For model 2, a two-hour observation window and a six-hour prediction horizon were selected, with the SpO_2/FiO_2 ratio as the most important input feature. This model achieved an AUPRC of 0.05 on the COVID test set and 0.03 on the non-COVID test set.

Conclusion

Model 1 demonstrated moderate discriminative performance but limited clinical utility at relevant operating points. Model 2 showed very limited predictive value, primarily due to extreme class imbalance. Consequently, neither model is currently suitable for clinical implementation. With larger datasets and more advanced modelling techniques, Model 1 may have the potential to become a clinically useful decision support tool to support decisions on switching from controlled to assisted ventilation.

2 Introduction

Patients with respiratory failure are mechanically ventilated in the Intensive Care Unit (ICU). Initially, clinicians typically employ a controlled ventilation mode, without spontaneous respiratory activity of the patient. Generally, after 24 to 48 hours of controlled ventilation, a switch to assisted ventilation is attempted, allowing the patient to breathe spontaneously. Assisted ventilation is considered beneficial due to reduced sedation requirements, no need for neuromuscular blockade, less respiratory muscle atrophy, improved haemodynamic stability, better distal organ perfusion and lung protection (1–3). Moreover, early switching to assisted ventilation has been associated with shorter durations of invasive mechanical ventilation (IMV) and ICU stay (4).

However, in some cases, a deterioration in ventilatory parameters is observed during assisted ventilation without a clearly identifiable cause. This deterioration may necessitate a return to controlled ventilation, requiring neuromuscular blockade and increased sedation. These events were frequently observed during the COVID-19 pandemic (5). These switch failures are associated with worse outcomes, such as a higher 28-day mortality and less ventilator free days (5–8).

Potential causes for deterioration during assisted ventilation are both ventilator-induced lung injury (VILI) and patient self-inflicted lung injury (P-SILI). VILI comprises four mechanisms of lung injury. Barotrauma and volutrauma result from alveolar overdistension caused by high transpulmonary pressures and tidal volumes, while atelectrauma arises from cyclic opening and closing of alveoli due to insufficient positive end-expiratory pressure (PEEP). These mechanisms can trigger biotrauma, characterised by the release and systemic dissemination of inflammatory mediators from the alveolar space. Lung-protective ventilation strategies are effective in mitigating VILI (9).

More recently, P-SILI has been proposed, attributed to excessive patient breathing effort. The pathophysiology likely mirrors that of VILI, involving alveolar overdistension and atelectrauma. In P-SILI, a vicious cycle may ensue: lung injury worsens gas exchange, which increases respiratory drive and effort, thereby exacerbating the injury. In such cases, reinitiating controlled ventilation may become necessary as a therapeutic intervention (10,11).

If respiratory deterioration that necessitates a return to controlled ventilation could be predicted, clinicians could optimise the timing of initiating assisted ventilation, and, additionally, adapt ventilation strategy earlier in patients receiving assisted ventilation treatment to prevent P-SILI.

ICU patients are continuously monitored, generating large volumes of physiological data that can be harnessed to develop predictive models using machine learning. Such models have demonstrated potential to detect clinical deterioration earlier than clinicians and could support clinical decision-making (12–15). Additionally, identifying which physiological parameters are predictive of failure in assisted ventilation could inform future research into the underlying pathophysiology.

Therefore, the primary aim of this study is to develop and validate a machine learning model to predict transitions from assisted to controlled mechanical ventilation in ICU patients. The secondary aim is to identify physiological parameters associated with respiratory failure in mechanically ventilated ICU patients.

3 Background

3.1 Pathophysiology of lung injury in mechanical ventilated patients

3.1.1 Ventilator induced lung injury (VILI)

Positive-pressure mechanical ventilation differs substantially from physiological breathing, in which negative pressure generated by the respiratory muscles initiates inspiration. This counter-physiological mechanism may have several adverse effects, on both the lungs and peripheral organs (9). VILI encompasses various types of injury, commonly classified as barotrauma, volutrauma, atelectrauma, and biotrauma, as illustrated in Figure 1 (9,16).

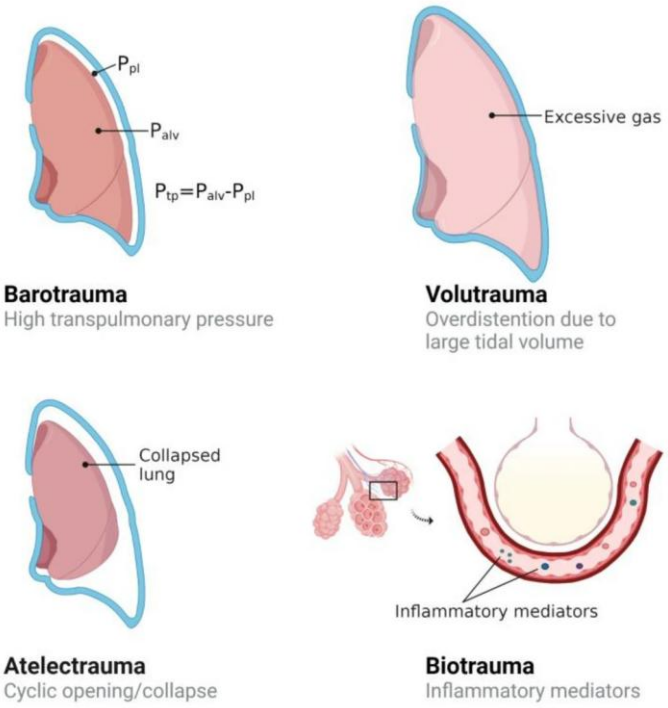


Figure 1. Types of injury induced by mechanical ventilation. P_{pl}: pleural pressure, P_{alv}: alveolar pressure, P_{tp}: transpulmonary pressure. Copied from Zou et al. 2024 (16)

Barotrauma and volutrauma are closely related phenomena. Barotrauma, primarily results from a high transpulmonary pressure, whereas volutrauma is caused by the administration of a large tidal volume. Despite these differences, both ultimately lead to excessive stress and strain within the lungs, either locally or globally. Strain is defined as the ratio of tidal volume to end-expiratory lung volume, while stress refers to the transpulmonary pressure, which is defined as the difference between the alveolar and pleural pressure. During positive-pressure ventilation, tidal volume is delivered by increasing airway pressure, creating compressive stress (9,10). High stress and strain lead to alveolar overdistension, deforming cells and their supporting matrix into abnormal shapes.

In severe cases, this may even cause alveolar rupture, allowing air to escape into surrounding tissues (9,16). Additionally, high tidal volumes have shown to induce pulmonary oedema (10).

Atelectrauma is caused by cyclic opening and closing of small airways and alveoli due to low tidal volumes and insufficient PEEP, resulting in abrasion of the epithelial lining (9,16).

Injured lung tissue and cyclic stretch trigger the release of injurious inflammatory mediators, a phenomenon known as biotrauma. Unlike the other mechanisms, this proinflammatory response is not confined to the lungs and may provoke a systemic inflammatory response (9,16).

3.1.2 Patient self-inflicted lung injury (P-SILI)

Assisted mechanical ventilation allows patients to breathe spontaneously using their respiratory muscles, while still providing ventilatory support with positive pressure. During assisted mechanical ventilation, lung injury may not only arise from the applied mechanical support, but also from the patient's increased spontaneous breathing effort, a phenomenon known as P-SILI. Similar to VILI, P-SILI is largely caused by increased stress and strain (10).

In spontaneous breathing, tidal volume is generated by creating negative pleural pressure rather than by increasing airway pressure, resulting in tensile stress. High inspiratory effort can lead to elevated levels of stress and strain, potentially inducing barotrauma and volutrauma (10).

Moreover, high inspiratory effort in combination with inhomogeneous distribution of transpulmonary pressure across the lung, can result in cyclic inflation in regions with high transpulmonary pressure variations, reproducing the mechanism of atelectrauma. Increased regional inhomogeneity may also cause pendelluft, in which gas shifts intrapulmonary from regions with low transpulmonary pressure variations to regions with high pressure variations. This process can result in local volutrauma, independent of the tidal volume (10,17).

During assisted ventilation, alveolar pressure can, in contrast to controlled ventilation, fall below the PEEP level due to high inspiratory effort. Such decreases in alveolar pressure may increase transvascular hydrostatic pressure, particularly in the presence of elevated airway resistance, potentially causing pulmonary oedema (10).

In addition, in a spontaneously breathing patient, patient-ventilator interactions can contribute to lung injury through several mechanisms. Over-assistance may increase transpulmonary pressure, increasing the risk of alveolar overdistension, whereas under-assistance may increase inspiratory effort, risking P-SILI. Besides, patient-ventilator asynchronies, such as double-triggering and reverse triggering, can cause breath stacking and large tidal volumes (10,17).

P-SILI can initiate a vicious cycle: lung injury impairs gas exchange, which increases respiratory drive and effort, further worsening the injury. In such cases, muscle relaxation and reinitiating controlled ventilation may be required as a therapeutic intervention (10,11).

3.2 Ventilatory parameters and modes

In mechanical ventilation, several ventilatory parameters are adjusted by the operator; these are referred to as control parameters. Additionally, the Hamilton ventilator measures and computes other parameters, referred to as output parameters. These, together with demographic, haemodynamic, blood gas parameters, and medication infusion rates, are used as predictors in the prediction models in this project. An overview of ventilatory parameters is provided in Figure 2 and Table 1.

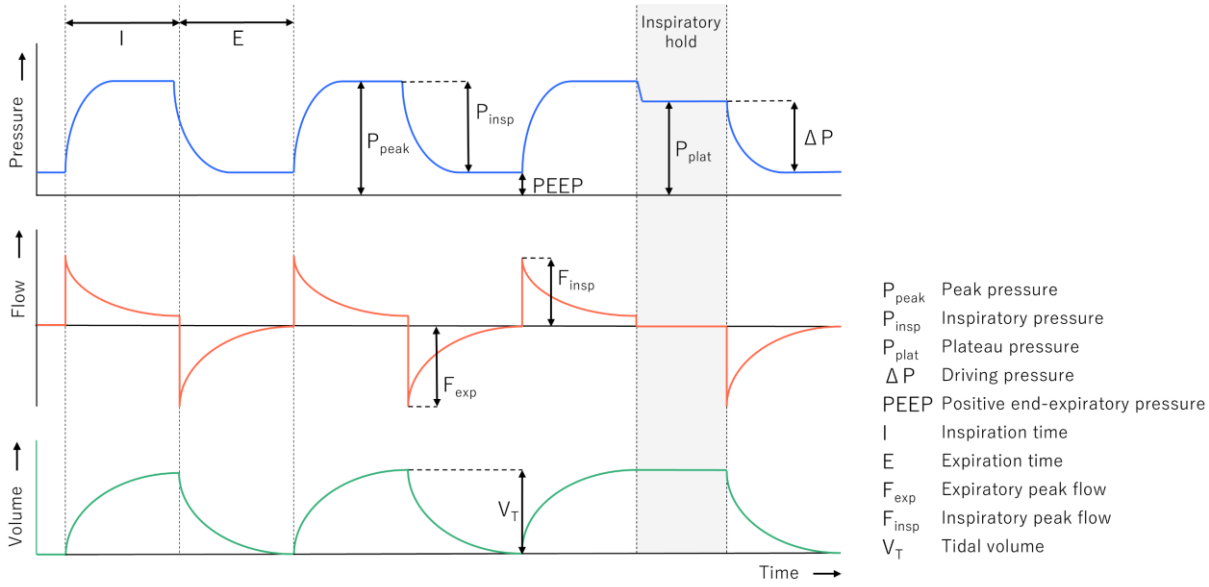


Figure 2. Pressure, flow and volume curves of pressure controlled ventilation, with ventilatory parameters indicated. During an inspiratory hold, the flow is set to zero and the plateau pressure is reached.

At the Intensive Care Unit of the Leiden University Medical Centre (LUMC), three distinct ventilation modes are used in clinical practice during IMV. For controlled ventilation, the Pressure-Controlled Mechanical Ventilation (P-CMV) mode is employed. In P-CMV, the primary control parameters are inspiratory pressure, PEEP, fraction of inspired oxygen (FiO_2), respiratory rate, and the inspiration-to-expiration time ratio (I:E ratio) (18).

For assisted ventilation, either the (Intellivent) Adaptive Support Ventilation (ASV) mode, or the spontaneous (SPONT) ventilation mode is used. In ASV, the main control parameters are target minute volume, PEEP, FiO_2, and expiratory trigger sensitivity (ETS). The expiratory trigger sensitivity is the percentage of the inspiratory maximum flow at which expiration is initiated. The ventilator determines the tidal volume and respiratory rate needed to achieve the target minute ventilation. When the patient is passive, ASV functions as pressure controlled ventilation. When the patient is active, the respiratory rate is controlled by the patient and the ventilator determines the inspiratory pressure needed to achieve the target minute volume (19). In Intellivent-ASV mode, target values for ventilation (end-tidal CO_2) and oxygenation (SpO_2) are set by the clinician, after which the ventilator automatically adjusts minute volume, PEEP, and FiO_2 to achieve these targets. All settings can be readily overridden by the clinician (20).

In spontaneous ventilation, the primary control parameters are inspiratory pressure, PEEP, FiO_2, and ETS, while the patient controls the respiratory rate and timing (18).

Table 1. Definitions of ventilatory parameters and type (control, output, adaptive) per ventilation mode.

Parameter	Definition	Control, output, adaptive
FiO_2	Fraction of oxygen in inspiration air	Control
Respiratory rate (RR)	Number of ventilation cycles per minute	P-CMV: control, SPONT: output, ASV: adaptive
Tidal volume (V_T)	Difference between the end inspiratory volume and end expiratory volume	P-CMV, SPONT: output ASV: adaptive
Minute ventilation	Ventilated volume per minute, product of tidal volume and respiratory rate	P-CMV, SPONT: output ASV: control
I:E ratio	Ratio between the inspiration time and expiration time	P-CMV: control ASV, SPONT: output
P_{insp}	Target airway pressure during inspiration	Control
P_{mean}	Mean airway pressure over one ventilation cycle	Output
P_{peak}	Maximum airway pressure during inspiration	Output
P_{plat}	Plateau pressure, airway pressure at the end of inspiration when flow is zero, measured by Hamilton at the end of inspiration when flow is close to zero and pressure is stable.	Output
PEEP	Positive end expiratory pressure, the airway pressure at the end of the expiration phase	Control
Auto PEEP	PEEP generated by the patient itself	Output
Driving pressure (ΔP)	Driving pressure, the difference between P_{plat} and PEEP.	Output
$Flow_{insp}$	Peak flow during inspiration	Output
$Flow_{exp}$	Peak flow during expiration	Output
R_{insp}	Difference between the P_{peak} and P_{plat} pressure divided by the inspiratory flow (21).	Output
Compliance	The elastic property of the respiratory system, ratio between V_T and ΔP (21).	Output
RC_{exp}	Expiratory time constant, describing the speed of change in volume after a change in pressure, the product of compliance and resistance measured at expiration (21).	Output
V_T/IBW	Ratio between tidal volume and ideal body weight	Output
End-tidal CO_2 ($EtCO_2$)	Partial pressure of CO_2 in the expiration air at the end of expiration	Output
RSBI	Rapid shallow breathing index, the ratio between the respiratory rate and tidal volume.	Output
PaO_2/FiO_2 (PF) ratio	Ratio between the partial pressure of oxygen in arterial blood (PaO_2) and the FiO_2	Output
SpO_2/FiO_2 (SF) ratio	Ratio between the oxygen saturation level in the blood (SpO_2) and the FiO_2	Output

3.3 Machine learning models

3.3.1 Logistic regression

The logistic regression model is a linear model that describes the relationship between predictor variables (X_i) and a binary outcome variable (Y; event or no event), expressed as the probability of the event $P(Y = 1)$ using the odds ratio (22). The odds ratio is defined as the ratio of the probability of an event occurring to the probability of an event not occurring:

$$odds\ P(Y = 1|X_1) = \frac{P(Y=1|X_1)}{1-P(Y=1|X_1)} \qquad \text{(Equation 1)}$$

$P(Y = 1|X_1)$: event probability given predictor variable X_1

An event probability below 0.5 results in an odds ratio between 0 and 1, whereas an event probability above 0.5 yields an odds ratio greater than 1, extending to infinity (Figure 3). To address this imbalanced scale, the logarithm of the odds ratio is used (23).

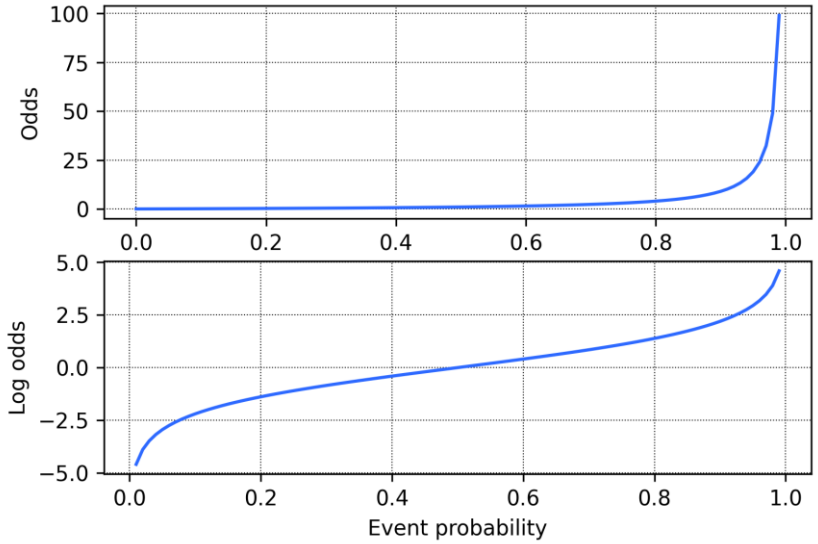


Figure 3. Relationship between event probability and the odds ratio (upper panel), and between event probability and the logarithm of the odds ratio (lower panel).

In logistic regression, it is assumed that the logarithmic odds ratio of the event probability is linearly related to the predictor variables (22). This relationship can be expressed as:

$$log\ odds(Y = 1|X_1) = \log\left(\frac{P(Y=1|X_1)}{1-P(Y=1|X_1)}\right) = \beta_0 + \beta_1 X_1 \qquad \text{(Equation 2)}$$

β_0: intercept

β_1: regression coefficient

The regression coefficient β_1 determines the rate of change in the outcome associated with predictor X_1, and represents the strength of the relationship between predictor X_1 and the outcome. The intercept β_0 represents the log-odds when X_1 equals zero. In models with multiple predictors, β_0 represents the log-odds when all predictors are zero, a situation that is usually not clinically meaningful (22).

From Equation 2, the logistic probability function can be written as:

$$P(Y = 1|X_1) = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}} \qquad \text{(Equation 3)}$$

The prediction probability function $P(Y = 1|X_1)$ has an S-shaped form, in which the predictor variable X_1 can range from -∞ to +∞, while the predicted probability P is constrained between 0 and 1 (Figure 4). In logistic regression, this S-shaped curve is fitted to map predictor values to event probabilities.

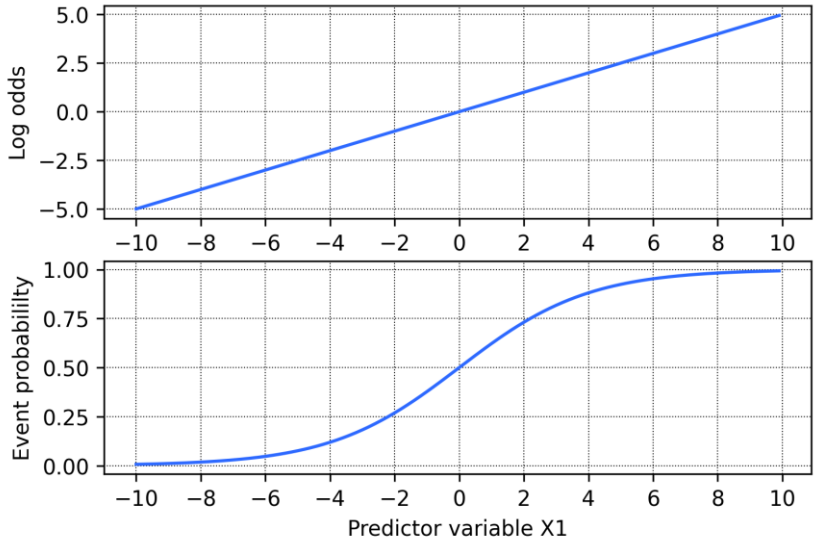


Figure 4. Linear relationship between the log-odds and predictor variable X_1 (upper panel), and the S-shaped relationship between event probability and predictor variable X_1 (lower panel). With $\beta_0 = 0$ and $\beta_1 = 0.5$.

In a model with multiple predictor variables, additivity is assumed, meaning that no interaction effects between predictors are included. The log-odds equation then becomes:

$$log\ odds(Y = 1|X_1, \ldots, X_n) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \qquad \text{(Equation 4)}$$

3.3.2 Extreme gradient boosting (XGBoost)

The XGBoost machine learning model is a decision tree ensemble, similar to a random forest, in which the prediction scores of individual trees are combined to produce the final prediction. The key difference between a random forest and an XGBoost model lies in the way the models are trained (24).

During XGBoost model training, trees are grown by optimising an objective function, which consists of a training loss component and a regularisation term to reduce overfitting. For binary classification, the loss function is typically logistic loss, measuring the difference between the predicted probability and the true outcome. In gradient boosting, trees are grown iteratively. At each boosting iteration, the previous tree is optimised by adding new splits to further reduce the loss (24,25). A schematic overview of the XGBoost training process is presented in Figure 5.

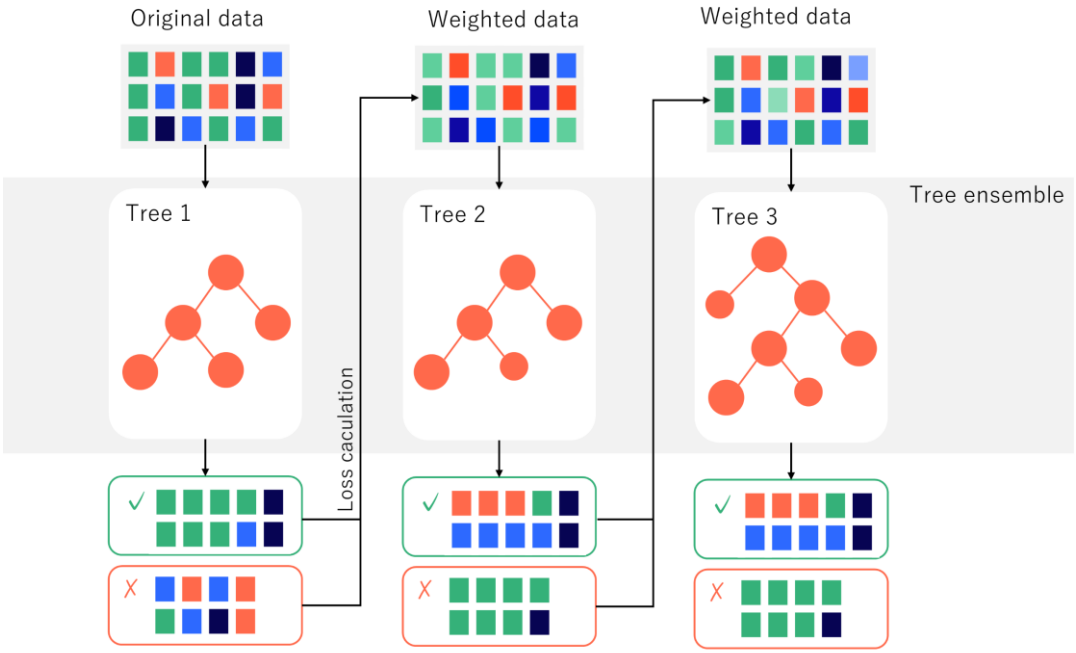


Figure 5. Schematic overview of the XGBoost training process with three boosting iterations. First, a decision tree is fitted to the original data and the training loss is computed, after which the data are reweighted. A new decision tree is then fitted, and the final prediction is obtained by combining all trees in the ensemble.

Several hyperparameters can be adjusted to optimise the XGBoost model, including those controlling the learning rate, model complexity, subsampling, and regularisation (25,26). An overview including definitions of the key hyperparameters is presented in *Supplement D.*

3.4 Evaluation methods in machine learning

3.4.1 Training quality

To evaluate the quality of model training, learning curves can be used. Similar to how humans improve at solving a problem through repeated exposure, machine learning models improve as they are trained on more data. A learning curve illustrates this behaviour by plotting model performance against the number of training samples (Figure 6).

Learning curves are typically estimated using k-fold cross-validation. The dataset is divided into k folds, and for each split a training and validation set are created. The size of the training set is varied, and for each training size the model is trained and evaluated on the corresponding validation set across all folds. This results in k performance estimates per training size, which are subsequently averaged (27).

Learning curves are assumed to converge to an asymptotic performance level. The point at which this level is reached is referred to as the saturation point. Before this point, models trained on smaller datasets show inferior performance, whereas beyond this point, increasing the training set size no longer yields performance gains. Observing a saturation point during model training indicates that sufficient data were available to train the model adequately (28).

To assess overfitting and underfitting, training performance can be plotted alongside validation performance. Similar to humans, memorisation performance decreases as the amount of training data increases. Consequently, training performance typically declines and converges towards the validation performance (27). When training performance remains high and validation performance stays low, the model is overfitting and additional training data or stronger regularisation may be required. Conversely, when both training and validation performance remain low, the model is underfitting and a more complex model may be needed.

In addition to varying the size of the training set, performance can be plotted against the number of optimisation iterations performed during model training (Figure 6). For XGBoost, this corresponds to the number of boosting iterations. Such iterative learning curves are also assumed to converge to a stable performance level (28). Identifying this saturation point allows the number of boosting to be limited, thereby reducing model complexity and mitigating overfitting.

A third type of learning curve is the feature curve, which shows model performance as a function of the number of features used for training (Figure 6). This curve can be used to assess the number of features required. In some cases, a peaking phenomenon can be observed, indicating that adding additional features may actually degrade model performance (27).

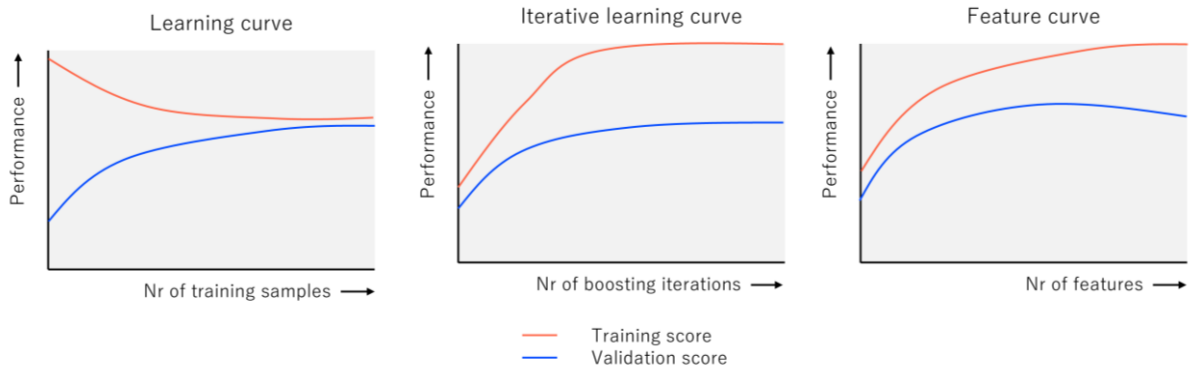


Figure 6. Examples of a learning curve based on the number of training samples (left), an iterative learning curve based on the number of boosting iterations (middle), and a feature curve based on the number of features used for training (right).

3.4.2 Classification and discriminative performance

Classification performance reflects how accurately samples are assigned to the correct class. Classification metrics are derived from the contingency table, or confusion matrix (Table 2). To calculate these metrics, a decision threshold must be applied to the predicted event probability to classify samples as low or high risk. Classification performance is optimal when all samples with an event have predicted probabilities above the threshold, and all individuals without an event have probabilities below it.

Table 2. Contingency table (confusion matrix) for a binary classification problem, showing true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN).

	Predicted high risk	Predicted low risk
Event	True positives (TP)	False negatives (FN)
Control	False positives (FP)	True negatives (TN)

Four basic classification measures can be distinguished: sensitivity (recall), specificity, positive predictive value (PPV; precision), and negative predictive value (NPV) (Table 3). Each addresses a different aspect of classification performance and is only relevant when reported together, as their values depend on the chosen classification threshold.

Table 3. Definitions of classification metrics: sensitivity (recall), specificity, positive predictive value (PPV; precision), and negative predictive value (NPV).

	Definition	Formula
Sensitivity or recall	The proportion of event samples that are classified as high risk.	$\frac{TP}{TP + FN}$
Specificity	The proportion of control samples that are classified as low risk.	$\frac{TN}{TN + FP}$
Positive predictive value (PPV) or precision	The proportion of samples classified as high risk that are an event sample.	$\frac{TP}{TP + FP}$
Negative predictive value (NPV)	The proportion of samples classified as low risk that are a control sample.	$\frac{TN}{TN + FN}$

Discriminative performance without selecting a classification threshold is commonly evaluated by using the area under the receiver operating characteristic curve (AUROC) (Figure 7). The receiver operating characteristic (ROC) curve plots sensitivity against 1-specificity. The AUROC a priori level is 0.5 and represents a model with no discriminative ability.

A precision-recall curve illustrates the trade-off between precision (PPV) and recall (sensitivity) (Figure 7). This is particularly relevant when a prediction model is used as an online alarm system, where both the predictive value of alarms (precision), and the ability to detect events (recall) are essential. The area under the precision-recall curve (AUPRC) summarises this trade-off. The a priori AUPRC depends on the ratio of event to control samples and equals the proportion of event samples in the dataset.

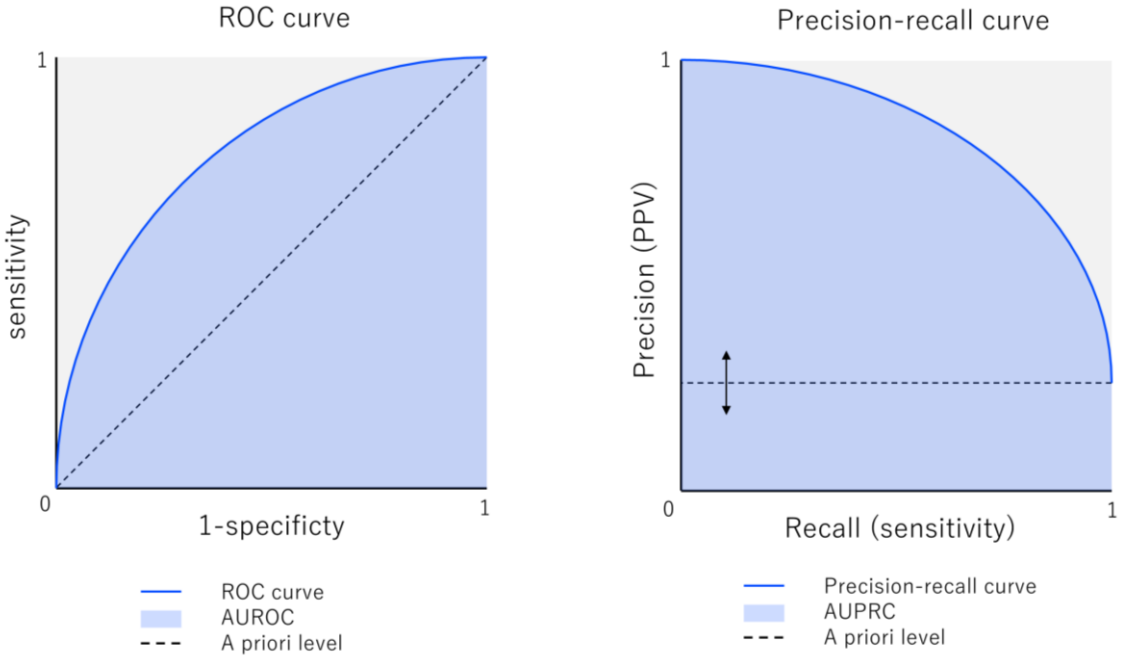


Figure 7. Receiver operating characteristic (ROC) curve (left) and precision-recall curve (right).

3.4.3 Clinical utility

To retrospectively assess whether a prediction model has the potential to improve current clinical care, a net benefit analysis can be used (Figure 8) (29). This analysis evaluates the net benefit across a range of threshold probabilities (P_t). The threshold probability represents the predicted risk at which a clinician would apply an intervention for an event. Net benefit quantifies the gain from true positive decisions while accounting for the harm of false positives and is expressed in units of true positives:

$$Net\ benefit = TP - FP \times exchange\ rate \qquad \text{(Equation 5)}$$

The gain from true positives and the harm from false positives are weighted by a value ratio, referred to as the exchange rate. This exchange rate depends on the chosen threshold probability and is equal to the corresponding odds ratio:

$$Exchange\ rate = \frac{P_t}{1-P_t} \qquad \text{(Equation 6)}$$

When performing a net benefit analysis, it is important to define the clinically relevant range of threshold probabilities in advance. At low threshold probabilities, clinicians prioritise avoiding missed events over the risk of unnecessary intervention. For example, at a threshold probability of 0.2, the odds ratio is 1:4, meaning that the gain of detecting one true positive is considered to outweigh the harm of four false positives. In contrast, at high threshold probabilities, clinicians place greater emphasis on avoiding unnecessary intervention rather than on missing events. At a threshold probability of 0.8, the odds ratio is 4:1, indicating that the harm of a false positive is weighted four times more heavily than the benefit of detecting a true positive (29). As the threshold probability can differ across patients and clinicians, a clinically relevant range should be defined.

In addition to the net benefit of the prediction model, the net benefits of an 'intervention for all' and 'intervention for none' strategy are shown in the graph (Figure 8). To evaluate whether a prediction model could be clinically useful, its net benefit should be compared with current practice, over the clinically relevant threshold range. In this study, current practice is best represented by the 'intervention for none' line, as no successful interventions were applied to prevent events in the dataset.

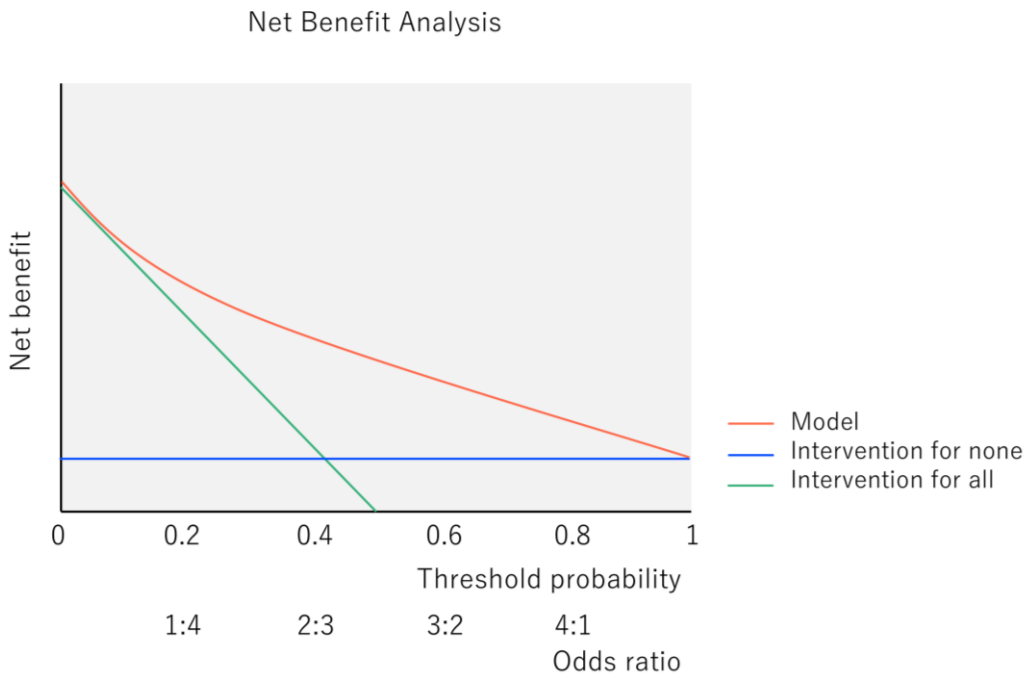


Figure 8. Net benefit analysis showing the net benefit across different threshold probabilities for a prediction model, an 'intervention for all' strategy, and an 'intervention for none' strategy, with corresponding odds ratios indicated.

4 Methods

The *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence* (TRIPOD+AI) guidelines (30) were used as a framework for model development and reporting. The completed TRIPOD+AI checklist is provided in *Supplement J*.

4.1 Data

The dataset used for model development and validation comprises detailed clinical information extracted from the Patient Data Management System (PDMS) MetaVision of ICU admissions at the LUMC, recorded between December 2018 and May 2023 as part of routine clinical care. The database includes ventilatory parameters and vital signs recorded at one-minute intervals, as well as laboratory results, administered medication, and demographic details.

4.2 Participants

Patients with a positive Sars-CoV-2 Polymerase Chain Reaction (PCR) test were allocated to the COVID group, and patients without a positive Sars-CoV-2 PCR test to the non-COVID group. Patients were excluded if they 1) did not receive IMV treatment, 2) had an IMV duration shorter than 48 hours, 3) were not ventilation with a Hamilton ventilator (C3 or C6), 4) did not have a PaO_2/FiO_2 (PF) ratio below 40 kPa, 5) received extracorporeal membrane oxygenation (ECMO) therapy, 6) were enrolled in the ICONIC trial, receiving a different PaO_2 target strategy (31), 7) were aged below eighteen, 8) or medication data was not available.

Inclusion and exclusion criteria were defined to obtain a homogeneous study population in which respiratory failure was the primary clinical problem (the COVID group), as well as a more heterogeneous non-COVID group with a major respiratory problem to serve as an external validation set to assess the model's generalisability.

Patient records in the COVID group were chronologically split into a training and test set based on the date of ICU admission. Patients admitted before November 2021 were assigned to the training set, whereas those admitted thereafter were assigned to the test set. Readmissions were allocated to the same subset as their initial admission to prevent data leakage.

4.3 Data preprocessing

Preprocessing was done to clearly differentiate between segments of controlled and assisted ventilation. First, time points with a spontaneous respiratory rate exceeding five breaths per minute were classified as assisted ventilation (Figure 9). A threshold of five was selected to balance between spontaneous irregular Cheyne-Stokes or opioid-induced breathing pattern and occasional spontaneous breaths occurring during controlled ventilation. Second, a median filter with a 31-minute window was applied to remove short segments of one ventilation mode lasting 15 minutes or less within a longer segment of the opposite mode. Finally, segments with missing data shorter than one hour were merged with the preceding ventilation mode.

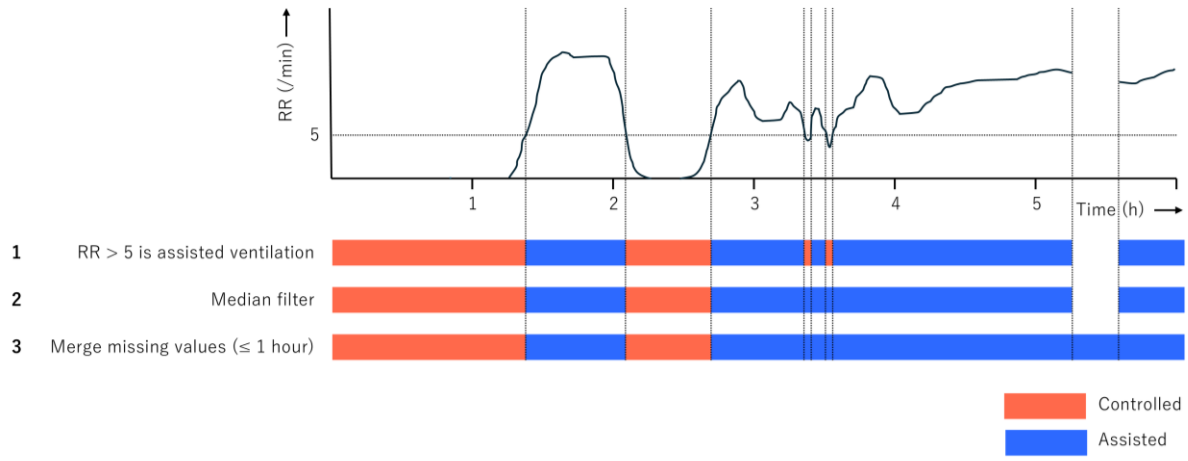


Figure 9. Data preprocessing steps for ventilation mode determination (controlled or assisted). 1) Time points with a spontaneous respiratory rate (RR) exceeding five are classified as assisted ventilation, 2) a median filter with a 31-minute window is applied, 3) segments with missing data shorter than one hour are merged with the preceding mode.

4.4 Event definition

The model outcome is the event probability. An event of respiratory deterioration, as defined by experienced ventilation specialists, is characterised by:

1) a transition from assisted to controlled ventilation, where
2) controlled ventilation persists for at least three hours, and
3) the FiO_2 was set to 40% or higher at least once within the time span of one hour before to one hour after the transition (Figure 10).

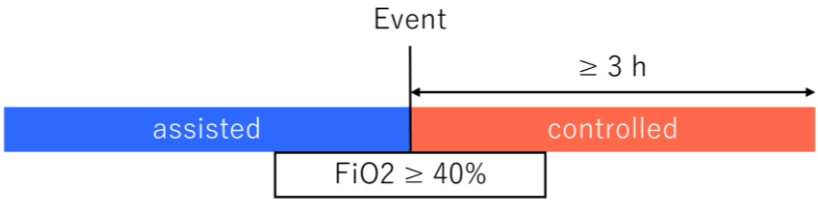


Figure 10. An event of respiratory deterioration is defined as a conversion from assisted to controlled ventilation, where controlled ventilation lasts at least three hours, the FiO_2 is set to 40% or higher in one hour before and after transition.

This outcome is proposed because it is an actionable target (32) during the course of mechanical ventilation treatment, rather than before intubation or during the weaning phase, contexts in which prediction models already exist (33–35). The restrictions applied in the event definition (≥ 3 hours of controlled ventilation and $FiO_2 \geq 40\%$) were introduced to minimise false-positive events resulting from sedation administered for procedural purposes. Transitions to controlled ventilation related to tracheostomy insertion were excluded as events.

As FiO_2 settings are operator-depended, this parameter serves as an imperfect marker of respiratory deterioration. To illustrate the influence of the FiO_2 threshold on the number of detected events, a histogram was generated for the number of events for different threshold values, in the COVID group. Events were automatically detected based on the event definition.

4.5 Input features

Initial predictors were selected based on availability, literature, and expert knowledge. Variables with less than 20% missing data and availability in at least 95% of records were included. A detailed overview of the variables used for input features is presented in *Supplement A.*

To capture temporal dependencies, summary statistics of ventilatory and haemodynamic parameters, including the mean, standard deviation, and trend over multiple windows, were used as input features (36). The trend is defined as the slope of the linear regression line. Additionally, the most recent arterial blood gas results (routinely measured every six hours) at the time of prediction, as well as age, sex, body mass index (BMI), total IMV duration, and duration of assisted ventilation, were included as input features. To guarantee reliability of feature values, summary statistics were only calculated if at least 50% of the datapoints in the selected window were available, otherwise a missing value was given.

Because aggregate features are employed, missing values are scarce and imputation of missing data is not required. Likewise, feature normalization is unnecessary for XGBoost, as decision trees split nodes according to the relative ordering of feature values rather than their absolute scale.

To fit a logistic regression model, samples with four or more missing feature values were first removed, after which features containing missing values were excluded. In addition, standard scaling was applied for the logistic regression model by subtracting the mean and scaling it to unit variance (Equation 7).

$$z = \frac{x-\mu}{\sigma} \qquad \text{(Equation 7)}$$

Z: standard value, x: feature value, μ: mean, σ: standard deviation

4.6 Model design

4.6.1 Model 1 | Unreadiness for assisted ventilation

The first model is designed to predict the probability of an event within six hours after switching from controlled to assisted ventilation. Therefore, the observation window is defined during the controlled ventilation phase, with predictions made at the point of transition from controlled to assisted ventilation (Figure 11). Events for this model are defined as occurrences within six hours after switching and are considered indicative that the patient was not yet ready for assisted ventilation. Control samples comprise cases in which patients did not fail or failed after a period longer than six hours after switching. Control samples within 24 hours prior to death were excluded.

In clinical practice, this model could be used as a decision-support tool to evaluate at a certain time-point whether a patient is ready to switch from controlled to assisted ventilation.

In addition, a variation to Model 1 with input features derived from both the last hour before transition to assisted ventilation and the first hour after transition was developed, results from this model are presented in *Supplement H.*

Furthermore, the clinical conditions and timing of all first attempts to switch from controlled to assisted ventilation were analysed and compared between patients who experienced an event within six or 72 hours and those who did not. The results of this analysis are presented in *Supplement I*.

4.6.2 Model 2 | Development of P-SILI

The second model is developed to predict the probability of an event over a 4-10-hour horizon during assisted ventilation. Therefore, the observation window is defined during assisted ventilation and predictions are made during assisted ventilation (Figure 11). Events for this model are defined as occurrences after a minimum of six hours of assisted ventilation, representing failure following a sustained period of assisted ventilation was attained, likely due to development of P-SILI.

The prediction window was set at 2 hours, therefore event samples were included three times (at 0, 1, and 2 hours after the prediction horizon). Control samples are drawn at a 1-hour interval from segments preceding true events or from segments without failure. Control samples within 24 hours before death were excluded.

This model could be used in clinical practice as a real-time alarm system, generating a prediction score every hour during assisted ventilation, to alert clinicians when the predicted probability of an event is high.

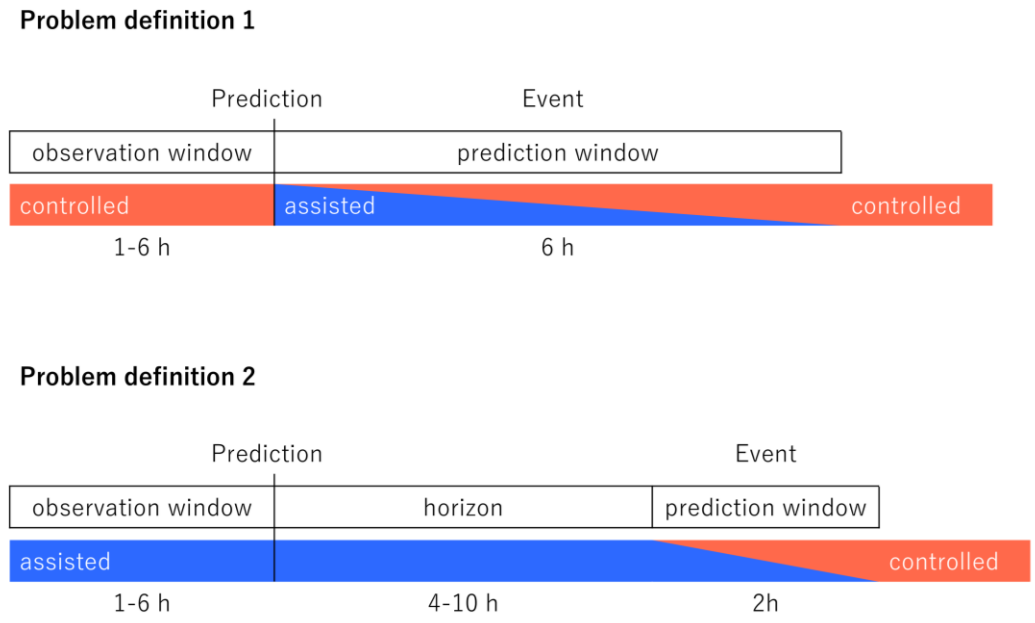


Figure 11. Overview of the observation window, prediction horizon, and prediction window aligned with ventilation modes for problem definition 1 (unreadiness) and problem definition 2 (P-SILI).

4.7 Model development

XGBoost models were employed because of their demonstrated high performance in comparable prediction tasks (35,37). During development, a series of experiments was conducted using stratified 10-fold cross-validation, grouped by patient ID, to maintain class distribution over different folds and prevent data leakage. First, the optimal observation window (1, 2, 4, or 6 hours) and horizon (4, 6, 8 or 10 hours) (Model 2) were determined. Second, a logistic regression model was trained to benchmark the performance of the XGBoost model. Subsequently, feature selection and hyperparameter optimisation was performed. Finally, performance obtained with grouped cross-validation was compared to ungrouped cross-validation, to assess the impact of learning patient-specific characteristics. An overview of all experiments conducted in this study is presented in Figure 12.

During model development, Model 1 was primarily optimised for the AUROC, reflecting the importance of discrimination in a decision-support tool, whereas Model 2 was primarily optimised for the AUPRC, reflecting the importance of predictive value in an alarm system. To compare the AUPRC between models during development, control samples were random under sampled obtaining a 1:3 ratio of event to control samples for Model 1 and 1:10 for Model 2. Differences in performance between models were evaluated using a one-sided Wilcoxon signed-rank test, with a significance level of 0.05.

Furthermore, learning curves were generated before and after feature selection and hyperparameter optimisation to evaluate model stability and to assess whether the sample size of training data was sufficient.

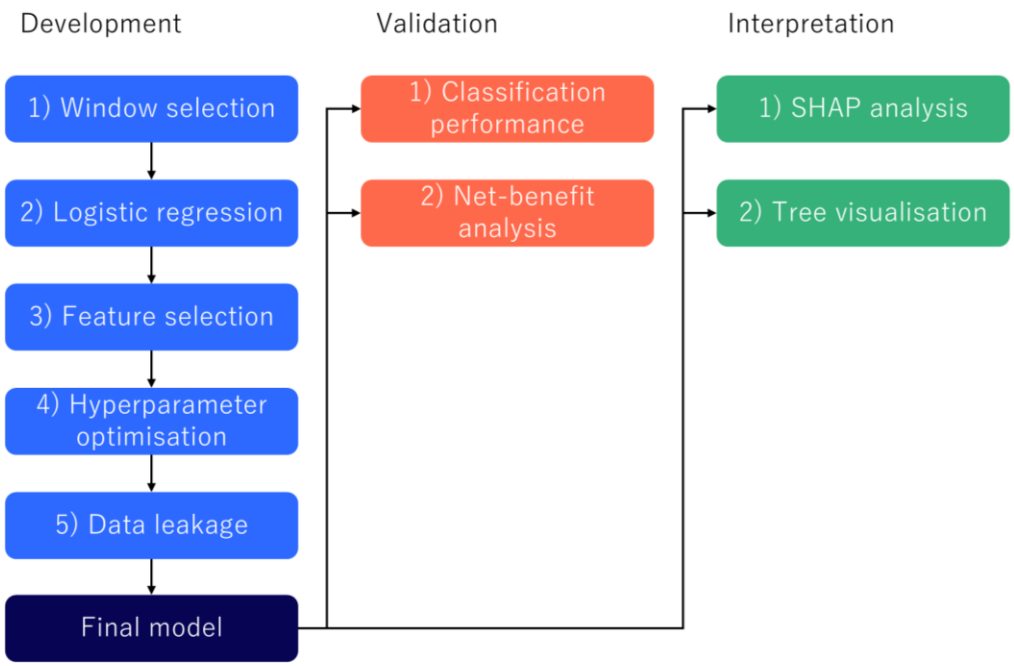


Figure 12. Overview of the experiments and analyses conducted for model development, validation and interpretation throughout the study.

4.7.1 Feature engineering and selection

To reduce overfitting during feature selection the number of boosting iterations was reduced. The optimal number of estimators was assessed by plotting performance over each boosting iteration.

For each model, two distinct feature sets were employed for the selected observation windows. Feature set 1 comprised the mean, standard deviation, and trend of ventilatory and haemodynamic parameters calculated over the entire observation window. Feature set 2 consisted of mean values calculated over multiple shorter sub-intervals within the observation window. Both feature sets were supplemented with the most recent arterial blood gas results, age, BMI, sex, IMV duration, and assisted mechanical ventilation duration (Model 2).

For each feature set, mean feature importance was calculated across 10-fold cross-validation using the gain, defined as the average loss reduction resulting from tree splits in which a given feature is used. Subsequently, backward feature elimination was performed starting with the 25 features with the highest importance. After each feature elimination step, feature importance was re-evaluated and the least important feature was removed, until only a single feature remained. Feature curves showing performance as a function of the number of included features were used to determine the minimum number of features required to achieve a stable model performance, defined by high performance with a small interquartile range (IQR).

4.7.2 Hyperparameter optimisation

A stepwise approach for hyperparameter optimisation was employed, as outlined in Figure 13 (25,38). First, the number of boosting was determined using the default learning rate by identifying the number of iterations at which model performance stabilises. Subsequently, tree complexity, subsampling and regularisation parameters were optimised using a 10-fold cross-validated grid search. The parameter values yielding high validation and relatively low training performance were selected to reduce overfitting. Finally, the learning rate was lowered, and the final number of iterations was determined. The exact hyperparameter values used during the grid search are provided in *Supplement D.*

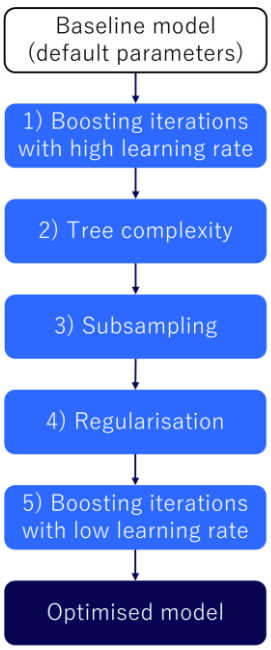


Figure 13. Stepwise hyperparameter optimisation approach. The number of boosting iterations is first determined, followed by optimisation of tree complexity, subsampling, and regularisation parameters. Finally, the learning rate is reduced and the final number of boosting iterations is selected.

4.8 Model validation

The optimised models were subsequently validated on the hold-out COVID and non-COVID test sets. Model performance was assessed using AUROC and AUPRC. In addition, the sensitivity, specificity, PPV, and NPV were calculated. For this purpose, the target specificity for Model 1 was set to 0.80, reflecting the importance of a highly specific test to prevent patients from unnecessarily prolonged controlled ventilation due to falsely high predicted probabilities of failure after switching to assisted ventilation. For Model 2, the target PPV was set to 0.80, reflecting the need for a highly precise alarm system that minimises false-positive alerts.

In addition to classification metrics, a net-benefit analysis was performed to assess clinical utility compared with the absence of a prediction model. To this end, the models were calibrated for each test set using a sigmoid function. Finally, probability distribution plots, showing the predicted probabilities for event and control samples, were generated to provide insight into overall model performance and behaviour (39).

4.9 Model interpretation

Model interpretability was achieved using Shapley additive explanations (SHAP) to quantify the contribution of individual features to model predictions and to interpret these contributions from a clinical perspective (40). In addition, detailed case descriptions of event and control samples with relatively high and low predicted probabilities were presented to explore the clinical conditions under which the model performs well or poorly. These case descriptions were complemented by local SHAP explanations to reveal the factors driving the model's predictions for these individual samples. Finally, the first decision tree of the XGBoost ensemble, which contributes most strongly to the prediction scores, was visualised to illustrate the underlying split criteria.

4.10 Software

Data processing, model training, evaluation, and interpretation were performed using Python 3.12.4, with the scikit-learn (v1.4.2), xgboost (v3.1.1), dcurves (v1.1.7), shap (v0.50.0), and dtreeviz (v2.2.2) libraries. Python scripts are available in the GitHub repository: *github.com/Emmelieve/TM3*.

4.11 Ethical approval

The study is approved by the Medical Ethics Committee Leiden The Hague Delft. Consent was waived as the data consists of routinely collected clinical information and it was not considered reasonable to request consent after an invasive ICU admission.

5 Results

5.1 Dataset characteristics

The LUMC ICU database comprises 11,922 records of patients admitted between December 2018 and May 2023. Of these, 640 tested positive for Sars-CoV2 by PCR. Of these, 266 records were excluded: 138 did not receive IMV, 106 were ventilated for less than 48 hours, five were not ventilated using a Hamilton ventilator, 14 received ECMO therapy, and three were enrolled in the ICONIC study (Figure 14). A total of 374 records were included in the COVID group and chronologically split into a training set (n = 296) and a test set (n = 78).

In the non-COVID group, 10,500 records were excluded: 4018 did not receive IMV, 5158 were ventilated for less than 48 hours, 1107 were not ventilated using a Hamilton ventilator, 69 received ECMO therapy, 131 were enrolled in the ICONIC study, three were aged under 18, and one record was excluded because medication data was not available. This resulted in a non-COVID test set of 755 records.

The COVID training set comprised 294 unreadiness events (failure within six hours) and 269 P-SILI events (failure after more than six hours). The COVID test set included 85 unreadiness events and 72 P-SILI events, while the non-COVID test set contained 483 unreadiness events and 317 P-SILI events. Patient characteristics for each dataset are summarised in Table 4.

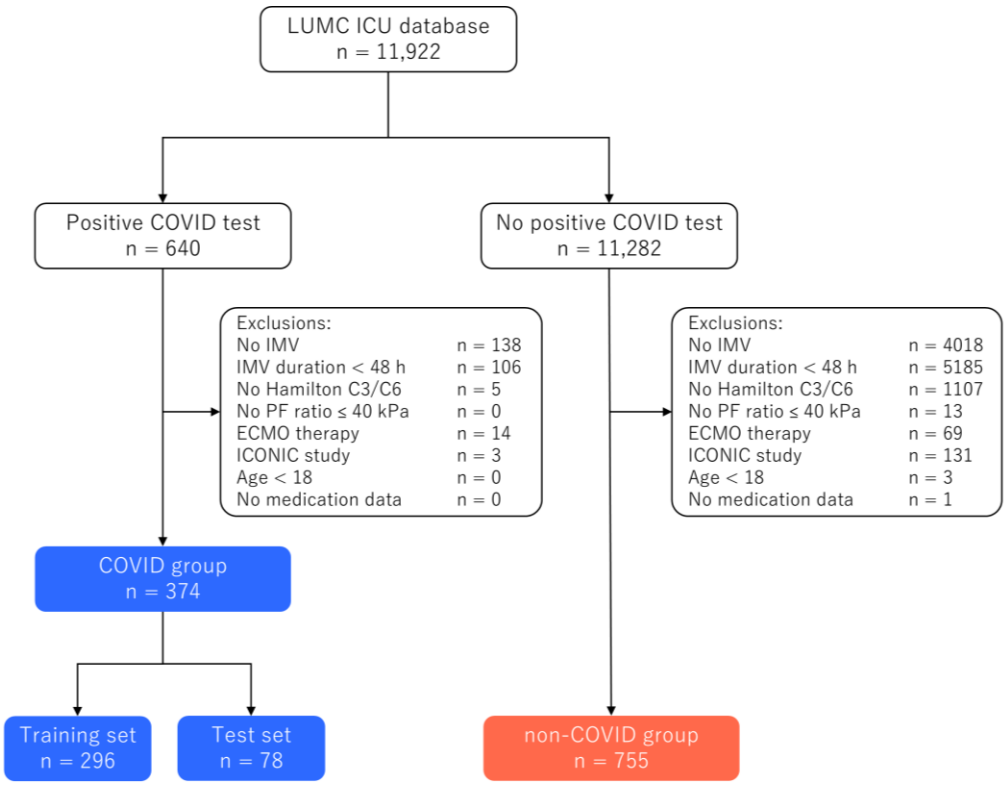


Figure 14. Flow of records included in the study, based on inclusion and exclusion criteria. IMV = invasive mechanical ventilation, PF = PaO_2/FiO_2, ECMO = extra-corporal membrane oxygenation

Table 4. Characteristics of patients included in the COVID training set, COVID test set, and non-COVID test set, including age, sex, BMI, ICU mortality, ICU length of stay, IMV duration, and the number of events.

		COVID training set	COVID test set	Non-COVID test set
Records	**n**	296	78	755
Age	**median (IQR)**	63.0 (56.8-70.0)	59.0 (49.0-66.8)	61.0 (50.0-70.0)
Male	**n (%)**	210 (71.0)	55 (70.5)	492 (65.2)
BMI	**median (IQR)**	29.31 (26.1-33.2)	27.15 (24.8-30.4)	25.95 (23.2-29.3)
ICU mortality	**n (%)**	80 (27.0)	25 (32.0)	207 (27.4)
ICU length of stay (days)	**median (IQR)**	13.75 (8.3-23.9])	13.31 (7.8-24.2)	9.33 (5.6-18.00
IMV duration (days)	**median (IQR)**	11.21 (6.0-20.4)	9.33 (5.5-17.8)	6.08 (3.6-11.9)
Unreadiness events	**n**	294	85	483
Records with unreadiness	**n (%)**	144 (48.7)	37 (47.4)	253 (33.5)
P-SILI events	**n**	269	72	317
Records with P-SILI events	**n (%)**	146 (49.3)	38 (48.7)	204 (27.0)

BMI = body mass index, IMV = invasive mechanical ventilation, IQR = interquartile range

5.2 Effect of the FiO_2 threshold on the number of events

The relation between number of detected events and the FiO_2 threshold in the event definition is shown in Figure 15. The number of detected events with FiO_2 threshold ranges from 30 to 80% is 436 to 54 for unreadiness events and 368 to 119 for P-SILI events in the COVID dataset.

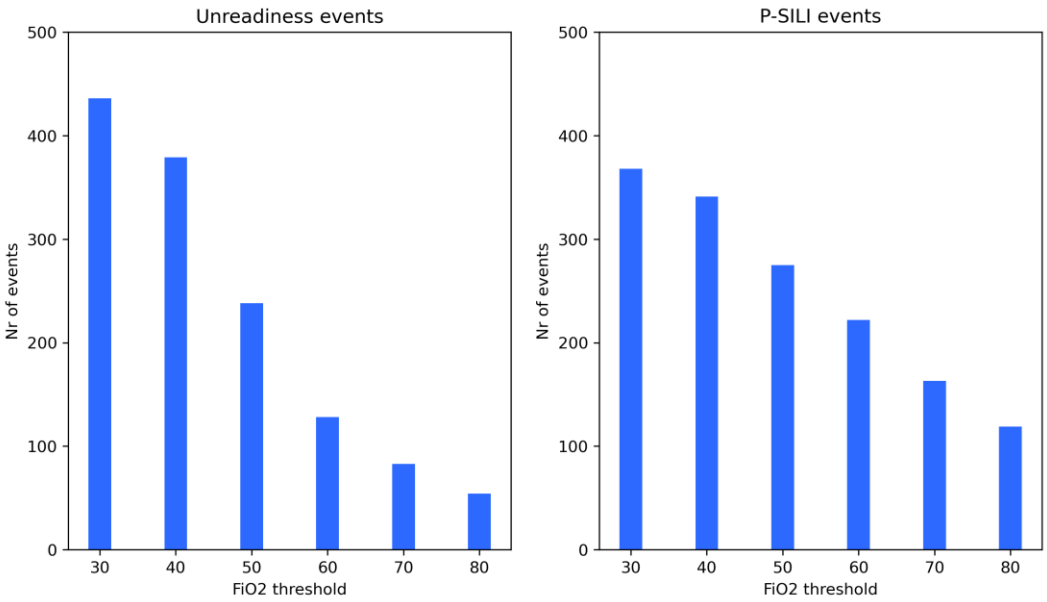


Figure 15. Relation between the number of detected unreadiness events and P-SILI events in the COVID dataset and the FiO_2 threshold in the event definition.

5.3 Input features

Summary statistics (mean, standard deviation, and trend) of six monitor parameters, 25 ventilatory parameters, and eight sedative and inotropic medication infusion rates were used as input features, together with the most recent value of 15 arterial blood gas parameter, as well as age, sex, and BMI. Details of the specific parameters, including missing data and availability per record in the COVID dataset, are provided in *Supplement A*.

5.4 Model 1 | Unreadiness for assisted ventilation

5.4.1 Observation window

Different observation windows, defined as the timeframe from which features are derived, were evaluated for Model 1, predicting unreadiness for assisted ventilation. The highest median AUROC values were obtained with 1-hour (AUROC 0.60, IQR 0.58-0.63) and 6-hour observation windows (AUROC 0.60, IQR 0.57-0.63). Models using 2-hour (AUROC 0.59, IQR 0.55-0.64) and 4-hour observation windows (AUROC 0.58, IQR 0.56-0.62) showed comparable performance, with no statistically significant differences ($p > 0.05$) (Table 5).

Table 5. Performance of Model 1 using different observation windows (1, 2, 4, and 6 hours), evaluated by the AUROC and AUPRC, p-values indicate differences in performance compared with the 1-hour observation window.

Observation window	1 h	2 h	4 h	6 h
Events	294	236	171	145
AUROC median (IQR)	0.60 (0.58-0.63)	0.59 (0.55-0.64)	0.58 (0.56-0.62)	0.60 (0.57-0.63)
p for AUROC	-	0.385	0.278	0.216
AUPRC median (IQR)	0.34 (0.29-0.36)	0.37 (0.29-0.41)	0.35 (0.26-0.38)	0.35 (0.33-0.40)
p for AUPRC	-	0.577	0.385	0.754

5.4.2 Logistic regression

The XGBoost model with a 1-hour observation window showed comparable performance (AUROC 0.63, IQR 0.57-0.66) compared with the logistic regression model (AUROC 0.61, IQR 0.57-0.66, p=0.784)(Table 6).

As logistic regression models do not handle missing values, 4 of 294 event samples and 15 of 117 features containing missing data were excluded for this analysis. Regression coefficients per feature and an analysis of the correlation between XGBoost feature importance and logistic regression coefficients are provided in *Supplement B.*

Table 6. Performance of the XGBoost and logistic regression model, both using a 1-hour observation window, evaluated by AUROC and AUPRC.

Model	XGBoost	Logistic regression	P
AUROC median (IQR)	0.63 (0.57-0.66)	0.61 (0.57-0.66)	0.784
AUPRC median (IQR)	0.35 (0.33-0.38)	0.35 (0.30-0.38)	0.688

5.4.3 Feature engineering and selection

To reduce overfitting during feature selection, the number of boosting iterations was set to 5; performance per iteration is shown in *Supplement D.* Four distinct feature sets were employed:

1. The mean, standard deviation, and trend over 1 hour, with a 1-hour observation window
2. The mean over each 20-minute interval, with a 1-hour observation window
3. The mean, standard deviation, and trend over 6 hours, with a 6-hour observation window
4. The mean over each 2-hour interval, with a 6-hour observation window

For each feature set, feature selection was performed using backward feature elimination. For feature set 1, nine features were selected, yielding a median AUROC of 0.68 (IQR 0.65-0.70) (Figure 16, Table 7). Feature set 2 selected 10 features, with AUROC 0.66 (IQR 0.63-0.68); feature set 3 selected 12 features, AUROC 0.64 (IQR 0.60-0.68); and feature set 4 selected nine features, AUROC 0.63 (IQR 0.60-0.66). Feature set 1 was selected for further optimisation based on the highest median AUROC. Feature importance graphs and feature curves for each feature set are provided in *Supplement C.*

Learning curves obtained before and after feature selection are shown in Figures 17 and 18.

Table 7. Overview of the different features sets, showing the selected features and performance as evaluated by AUROC and AUPRC.

	Set 1	Set 2	Set 3	Set 4
Observation window	1 hour	1 hour	6 hours	6 hours
Features	Mean, std, trend over 1 hour	Mean over 20 min	Mean, std, trend over 6 hours	Mean over 2 hours
Selected features	Mean FiO_2 Mean SF Mean PEEP Mean mida. Mean SpO_2 Mean V_T/IBW Std of PFI Std of $EtCO_2$ Mean prop.	Mean FiO_2 40-60 min Mean SF 40-60 min Mean PEEP 40-60 min Mean SpO_2 40-60 min Mean SF 20-40 min Mean SpO_2 20-40 min Mean V_T/IBW 20-40 min Mean SF 0-20 min Mean compl. 40-60 min Mean mida. 0-20 min	Mean FiO_2 Std of MAP Trend of MAP Mean SF Trend of V_T/IBW Std of P_{insp} Mean RC_{exp} Std of dia. ABP Trend of SpO_2 Trend of sys. ABP Mean dia. ABP Mean prop.	Mean FiO_2 4-6 h Mean suf. 0-2 h Mean SF 4-6 h Mean sys. ABP 2-4 h Mean $V_{T,insp}$ 4-6 h Mean MAP 4-6 h Mean RC_{exp} 2-4 h Mean MAP 0-2 h Mean prop. 2-4 h
AUROC median (IQR)	0.68 (0.65-0.70)	0.66 (0.63-0.68)	0.64 (0.60-0.68)	0.63 (0.60-0.66)
p for AUROC	-	0.216	0.065	0.024
AUPRC median (IQR)	0.43 (0.38-0.47)	0.39 (0.36-0.41)	0.37 (0.33-0.49)	0.38 (0.29-0.44)
p for AUPRC	-	0.138	0.188	0.246

FiO_2 = fraction of inspired oxygen, SpO_2 = peripheral oxygen saturation level, SF = SpO2/FiO2 ratio, PEEP = positive end-expiratory pressure, V_T = tidal volume, IBW = ideal body weight, PFI = peripheral flow index, compl. = compliance, $EtCO_2$ = end-tidal CO_2, prop. = propofol infusion rate, compl. = compliance, mida. = midazolam infusion rate, MAP = mean arterial blood pressure, P_{insp} = inspiratory pressure, RC_{exp} = expiratory time constant, dia. ABP = diastolic arterial blood pressure, sys. ABP = systolic arterial blood pressure, suf. = sufentanil infusion rate

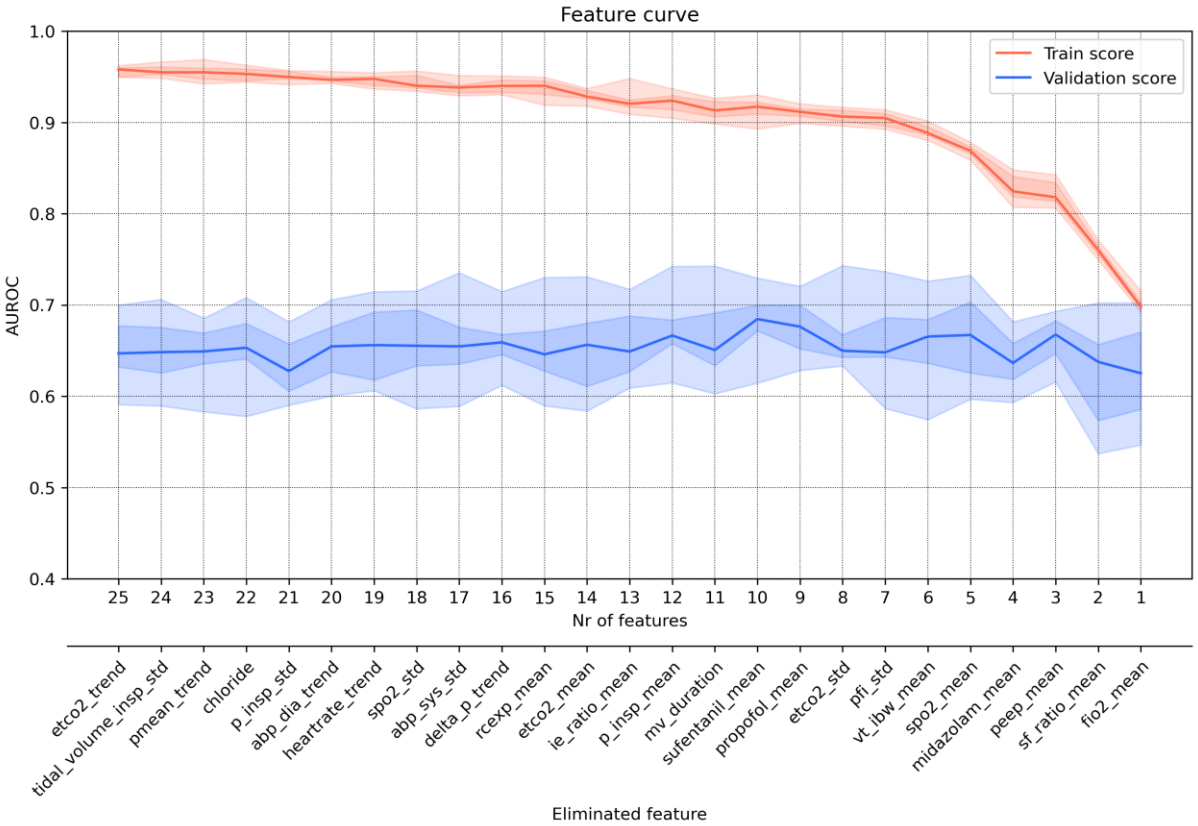


Figure 16. Feature curve illustrating the backward feature elimination process for Model 1 with feature set 1 (mean, standard deviation, and trend over 1 hour). The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

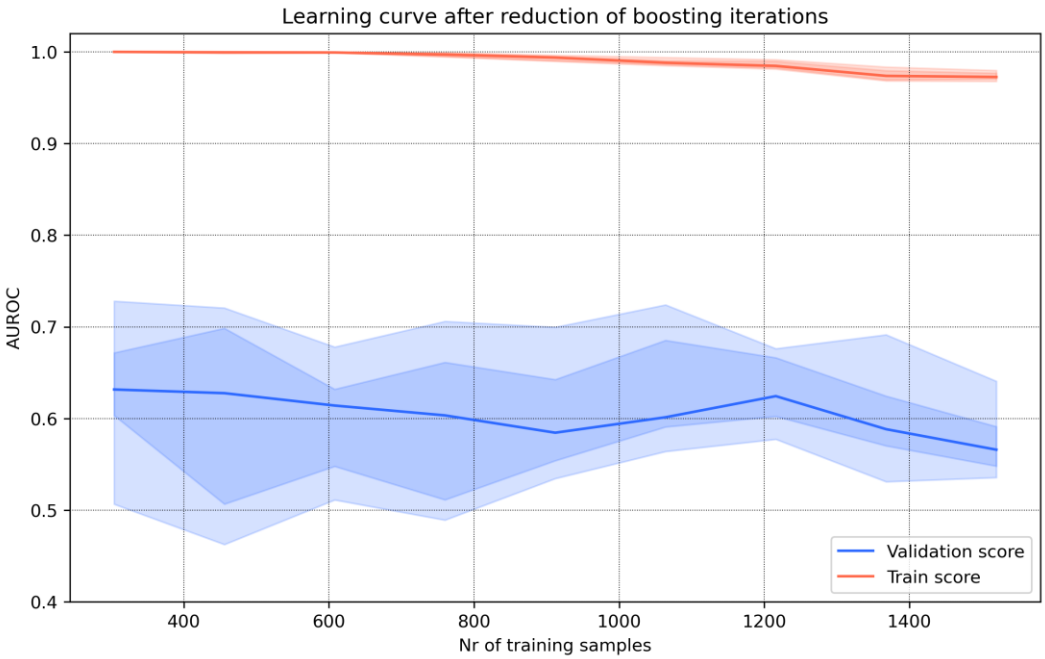


Figure 17. Learning curve, illustrating the performance over the number of training samples, after reduction of boosting iterations, prior to feature selection. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

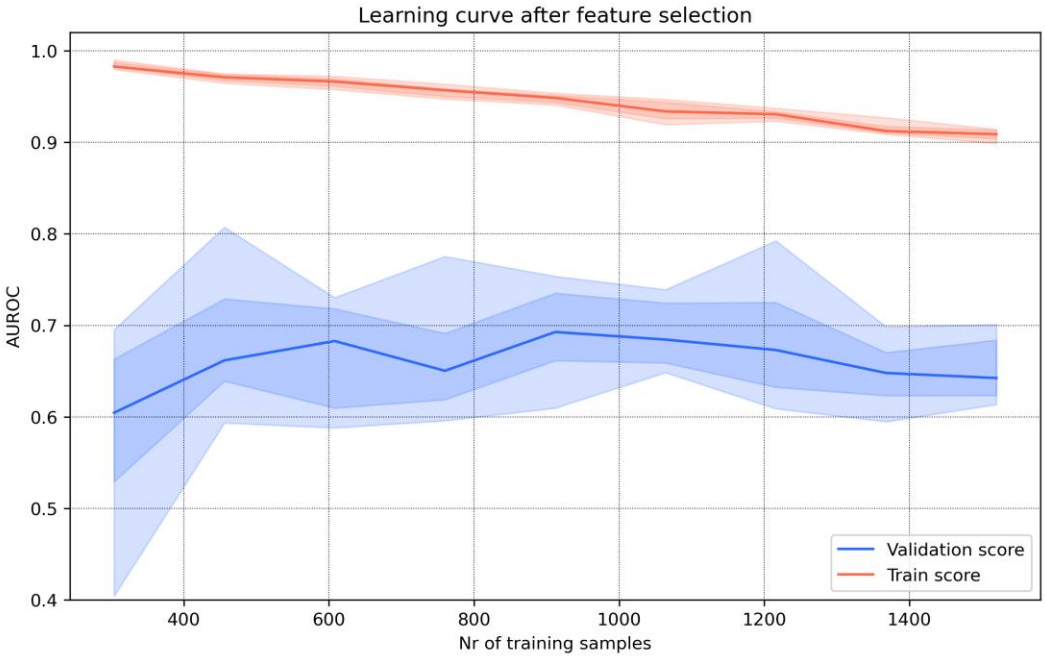


Figure 18. Learning curve of Model 1, illustrating the performance over the number of training samples, obtained after feature selection. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

5.4.4 Hyperparameter optimisation

The AUROC and AUPRC after hyperparameter optimisation were respectively, 0.67 (IQR 0.64-0.69) and AUPRC 0.41 (IQR 0.35-0.46). Details on optimised hyperparameter settings are provided in *Supplement D*. A learning curve obtained after hyperparameter optimisation is presented in Figure 19.

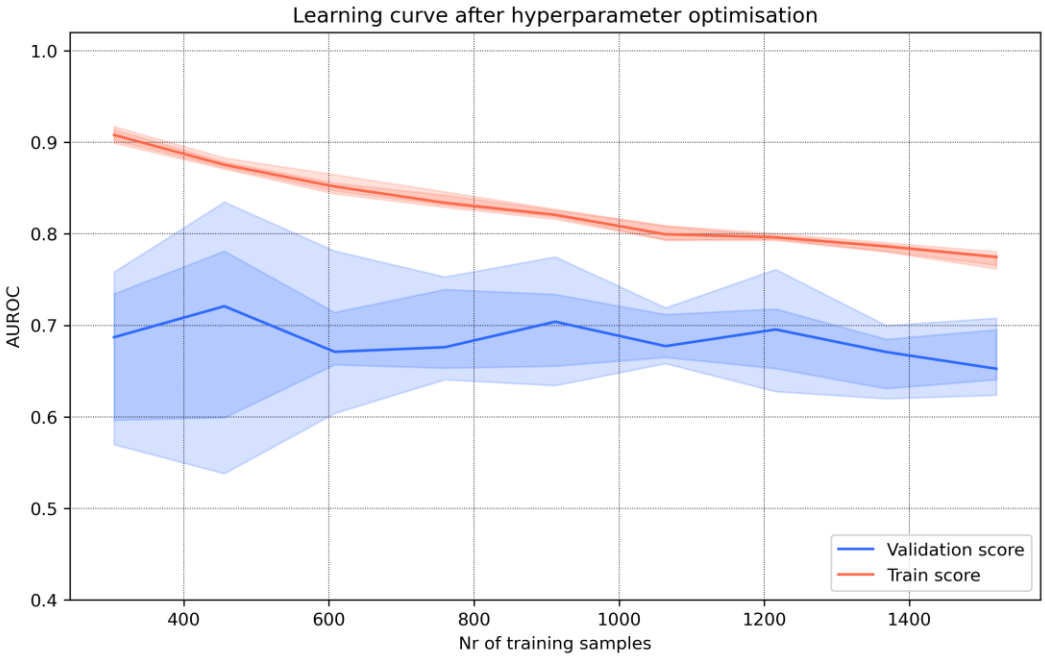


Figure 19. Learning curve of Model 1, illustrating the performance over the number of training samples, obtained after feature selection and hyperparameter optimisation. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

5.4.5 Training with data leakage

Ten-fold cross-validation with data grouped by patient ID yielded a median AUROC of 0.67 (IQR 0.64-0.69), whereas cross-validation without grouping, which introduces data-leakage, showed similar performance with a median AUROC of 0.66 (IQR 0.61-0.68, p=0.161) (Table 8).

Table 8. Performance of Model 1 evaluated using 10-fold cross-validation with and without grouping by patient ID, reported as AUROC and AUPRC.

	Grouped cross-validation	Ungrouped cross-validation	p
AUROC median (IQR)	0.67 (0.64-0.69)	0.66 (0.61-0.68)	0.161
AUPRC median (IQR)	0.41 (0.35-0.46)	0.38 (0.33-0.42)	0.161

5.4.6 Model validation

Validation on the test sets yielded an AUROC of 0.78 and an AUPRC of 0.38 for the COVID test set, and an AUROC of 0.76 and an AUPRC of 0.27 for the non-COVID test set. ROC and precision-recall curves are shown in Figures 20 and 21. For the COVID test set, setting specificity at 0.80 resulted in a sensitivity of 0.56, PPV of 0.39, and NPV of 0.89 (Table 9). For the non-COVID test set, setting specificity at 0.80 resulted in a sensitivity of 0.49, PPV of 0.25, and NPV of 0.92.

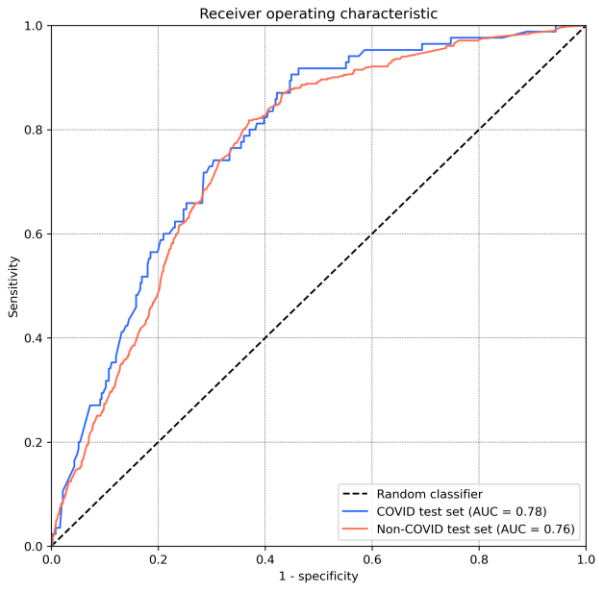


Figure 20. Receiver operating characteristic curve of validation on the COVID test dataset (AUC = 0.78) and non-COVID test dataset (AUC = 0.76).

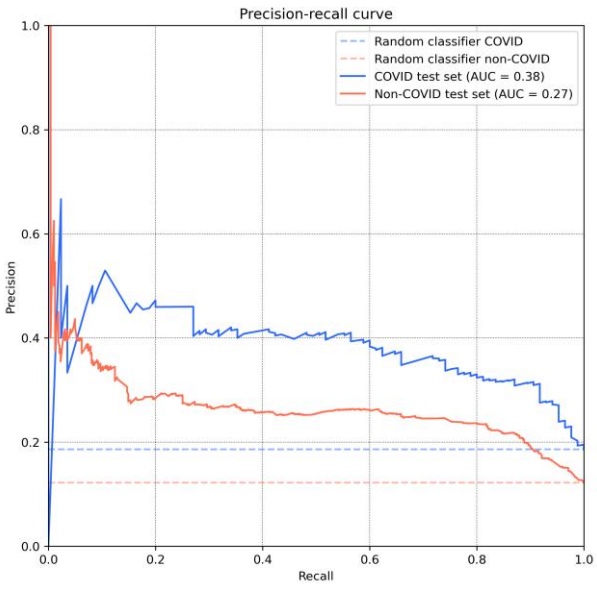


Figure 21. Precision-recall curve of validation on the COVID test dataset (AUC = 0.38) and non-COVID test dataset (AUC = 0.27).

Table 9. Overview of classification metrics for the COVID and non-COVID test set for Model 1.

	COVID test set	Non-COVID test set
Events	85	483
Controls	372	3483
AUROC	0.78	0.76
AUPRC	0.38	0.27
Sensitivity (recall)	0.56	0.49
Specificity	0.80	0.80
Positive predictive value (PPV) (precision)	0.39	0.25
Negative predictive value (NPV)	0.89	0.92

After calibrating the model on the test sets, net benefit analyses were performed (Figure 22). A superior net benefit compared with a 'treat all as unready' or 'treat all as ready' strategy was observed for threshold probabilities ranging from 0 to 0.5 for the COVID group and 0 to 0.3 for the non-COVID group. Calibration curves are provided in *Supplement E.* The distribution of predicted probabilities per event class after model calibration is shown in Figure 23.

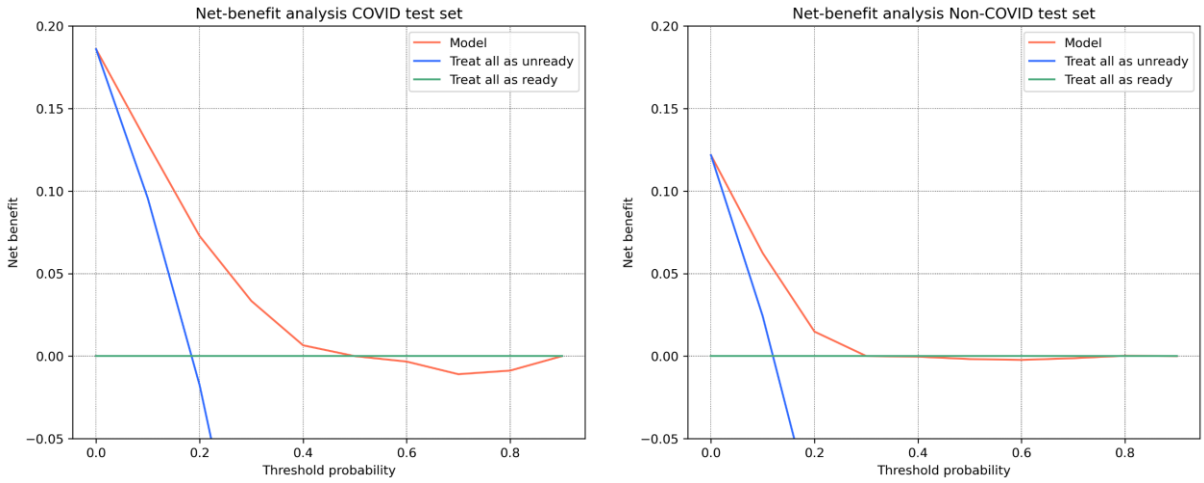


Figure 22. Net-benefit analysis for Model 1 on the COVID and non-COVID test sets, showing the threshold probability on the X-axis and the net benefit on Y-axis.

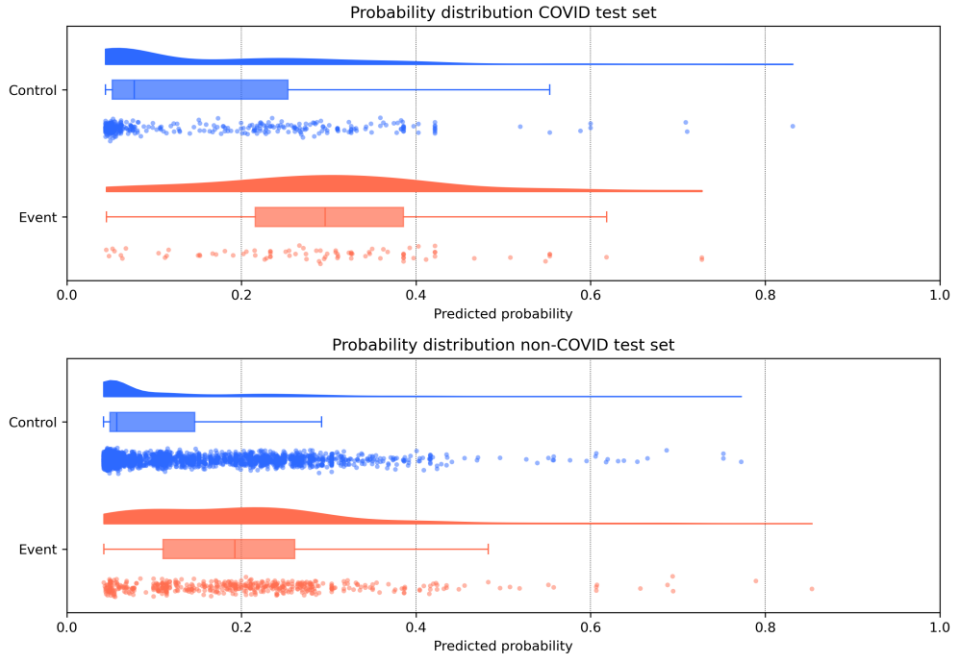


Figure 23. Predicted probability distributions for the COVID and non-COVID test sets obtained with Model 1 after calibration, shown as violin plots (top), boxplots (middle), and scatterplots (bottom).

5.4.7 Model interpretation

A global SHAP analysis is presented in Figure 24. This analysis shows that high FiO_2, propofol infusion rates, and PEEP values contribute most strongly to high predicted probabilities, whereas low a FiO_2, PEEP, and V_T/IBW, together with a high SpO_2 and SpO_2/FiO_2 ratio, are associated with low prediction scores.

Local SHAP explanations for samples with high or low predicted probabilities (*Supplement F),* as well as visualisation of the first decision tree (*Supplement G),* further indicate that model output is predominantly driven by FiO_2 and propofol infusion rate. Detailed clinical case descriptions are provided in *Supplement F.*

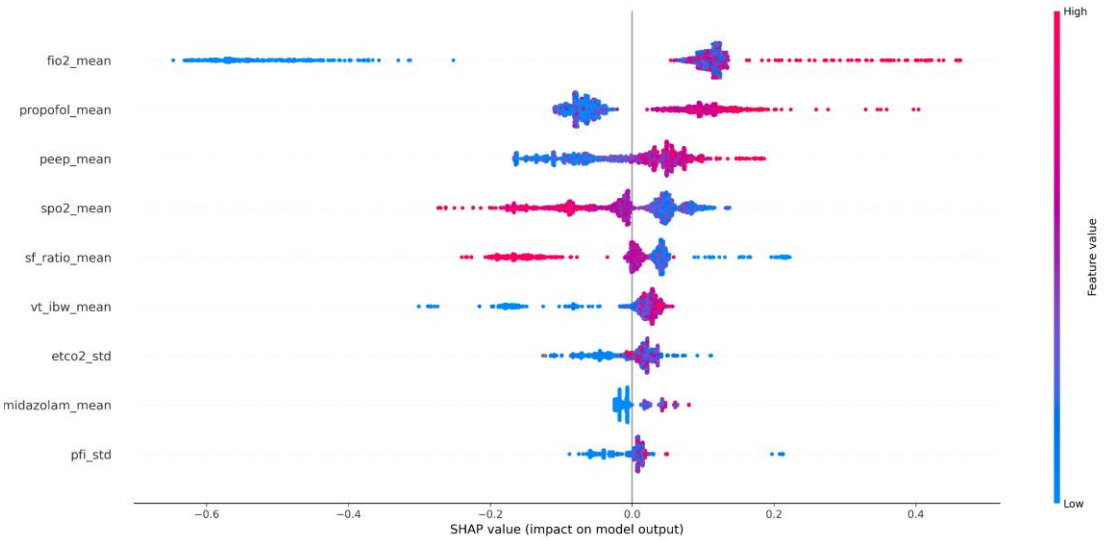


Figure 24. Global SHAP analysis of Model 1 showing the impact of feature values on the model output by aggregating local SHAP values across all training samples.

5.5 Model 2 | Development of P-SILI

5.5.1 Observation window

Different observation windows were evaluated for Model 2, which predicts respiratory deterioration due to P-SILI. The highest median AUPRC values were obtained with 2-hour (AUPRC 0.22, IQR 0.20-0.24) and 4-hour observation windows (AUPRC 0.22, IQR 0.20-0.24). Models using 1-hour (AUPRC 0.21, IQR 0.19-0.23) and 6-hour observation windows (AUPRC 0.18, IQR 0.16-0.21) showed comparable performance, with no statistically significant differences ($p > 0.05$) (Table 10).

Table 10. Performance of Model 2 using different observation windows (1, 2, 4, and 6 hours), evaluated by AUROC and AUPRC, p-values indicate differences in performance compared with the 4-hour observation window.

Observation window	1 h	2 h	4 h	6 h
Events	718	681	621	571
AUPRC median (IQR)	0.21 (0.19-0.23)	0.22 (0.20-0.24)	0.22 (0.20-0.24)	0.18 (0.16-0.25)
p for AUPRC	0.188	0.312	-	0.116
AUROC median (IQR)	0.69 (0.67-0.70)	0.70 (0.63-0.73)	0.70 (0.68-0.72)	0.67 (0.63-0.73)
p for AUROC	0.246	0.615	-	0.080

5.5.2 Prediction horizon

Varying prediction horizons, defined as the time between the prediction and the occurrence of an event, were evaluated in combination with a 4-hour observation window. The model with a 4-hour horizon obtained a median AUROC of 0.20 (IQR 0.17-0.27). A 6-hour horizon resulted in a median AUROC 0.22 (IQR 0.20-0.24), an 8-hour horizon in 0.20 (IQR 0.18-0.23), and a 10-hour horizon in 0.23 (IQR 0.16-0.26). A 6-hour horizon was selected for further model optimisation because it combines a relatively high median AUPRC with a narrow IQR (Table 11).

Table 11. Performance of Model 2 using different prediction horizons (4, 6, 8, and 10 hours), evaluated by AUROC and AUPRC, p-values indicate differences in performance compared with the 6-hour horizon.

Horizon	4 h	6 h	8 h	10 h
Events	727	621	604	516
AUPRC median (IQR)	0.20 (0.17-0.27)	0.22 (0.20-0.24)	0.20 (0.18-0.23)	0.23 (0.16-0.26)
p for AUPRC	0.216	-	0.116	0.313
AUROC median (IQR)	0.70 (0.66-0.74)	0.70 (0.68-0.72)	0.68 (0.66-0.69)	0.69 (0.63-0.71)
p for AUROC	0.385	-	0.097	0.161

5.5.3 Logistic regression

The logistic regression model with a 4-hour observation window and a 6-hour horizon showed comparable performance (AUPRC 0.23, IQR 0.18-0.26) compared with the XGBoost model (AUPRC 0.21, IQR 0.18-0.24, p=0.138) (Table 12).

As logistic regression models do not handle missing values, 24 of 621 event samples and 11 of 127 features containing missing data were excluded for this analysis. Regression coefficients per feature are provided in *Supplement B.*

Table 12. Performance of XGBoost and logistic regression for Model 2, both using a 2-hour observation window and a 4-hour horizon.

Model	XGBoost	Logistic regression	p
AUPRC median (IQR)	0.21 (0.18-0.24)	0.23 (0.18-0.26)	0.138
AUROC median (IQR)	0.71 (0.66-0.74)	0.71 (0.63-0.78)	0.784

5.5.4 Feature engineering and selection

To reduce overfitting during feature selection, the number of boosting iterations was set to 5; performance per iteration is shown in *Supplement D.* Four distinct feature sets were employed:

1. The mean, standard deviation, and trend over 2 hours, with a 2-hour observation window
2. The mean over each 30-minute interval, with a 2-hour observation window
3. The mean, standard deviation, and trend over 4 hours, with a 4-hour observation window
4. The mean over each 1-hour interval, with a 4-hour observation window

For each feature set, feature selection was performed using backward feature elimination. For feature set 1, seven features were selected, yielding an AUPRC of 0.27 (IQR 0.25-0.35) (Table 13). For feature set 2, nine features were selected, resulting in AUPRC 0.32 (IQR 0.26-0.35) (Figure 25). For feature set 3 and 4, 10 features were selected, resulting in AUPRC 0.26 (IQR 0.23-0.27) and 0.27 (IQR 0.24-0.27), respectively. Feature set 2 was selected for further optimisation based on the highest median AUPRC score. Feature importance graphs and feature curves for each feature set are provided in *Supplement C.*

Learning curves obtained before and after feature selection are shown in Figures 26 and 27.

Table 13. Overview of the different feature sets for Model 2, showing the selected features and performances as evaluated by AUPRC and AUROC. P-values indicate differences in performance compared with feature set 2.

	Set 1	Set 2	Set 3	Set 4
Observation window	2 hour	2 hour	4 hours	4 hours
Features	Mean, std, trend over 2 hours	Mean over each 30 min	Mean, std, trend over 4 hours	Mean over each 1 hour
Selected features	Mean SF IMV duration Mean nor Std of P_{insp} Assisted duration Art. base excess Mean V_T/IBW	Mean SF 90-120 min Mean SF 60-90 min Mean nor 60-90 min IMV duration Art. base excess Mean V_T/IBW 60-90 min Mean SF 0-30 min Glucose	Mean SF Mean prop. Mean nor. Mean $V_{T,exp}$ IMV duration Art. base excess Glucose Chloride $PaCO_2$ Mean PEEP	Mean SF 3-4 h Mean prop. 3-4 h Mean nor. 3-4 h Mean $V_{T,exp}$ IMV duration Art. base excess Glucose Chloride $PaCO_2$ Mean V_T/IBW 3-4 h
AUPRC median (IQR)	0.27 (0.25-0.35)	0.32 (0.26-0.35)	0.26 (0.23-0.27)	0.27 (0.24-0.27)
p for AUPRC	0.539	-	0.065	0.042
AUROC median (IQR)	0.74 (0.68-0.79)	0.76 (0.68-0.78)	0.73 (0.71-0.80)	0.73 (0.68-0.76)
p for AUROC	0.348	-	0.652	0.461

SF = SpO_2/FiO_2 ratio, V_T = tidal volume, IWB = ideal body weight, nor = noradrenaline infusion rate, RC_{exp} = expiratory time constant

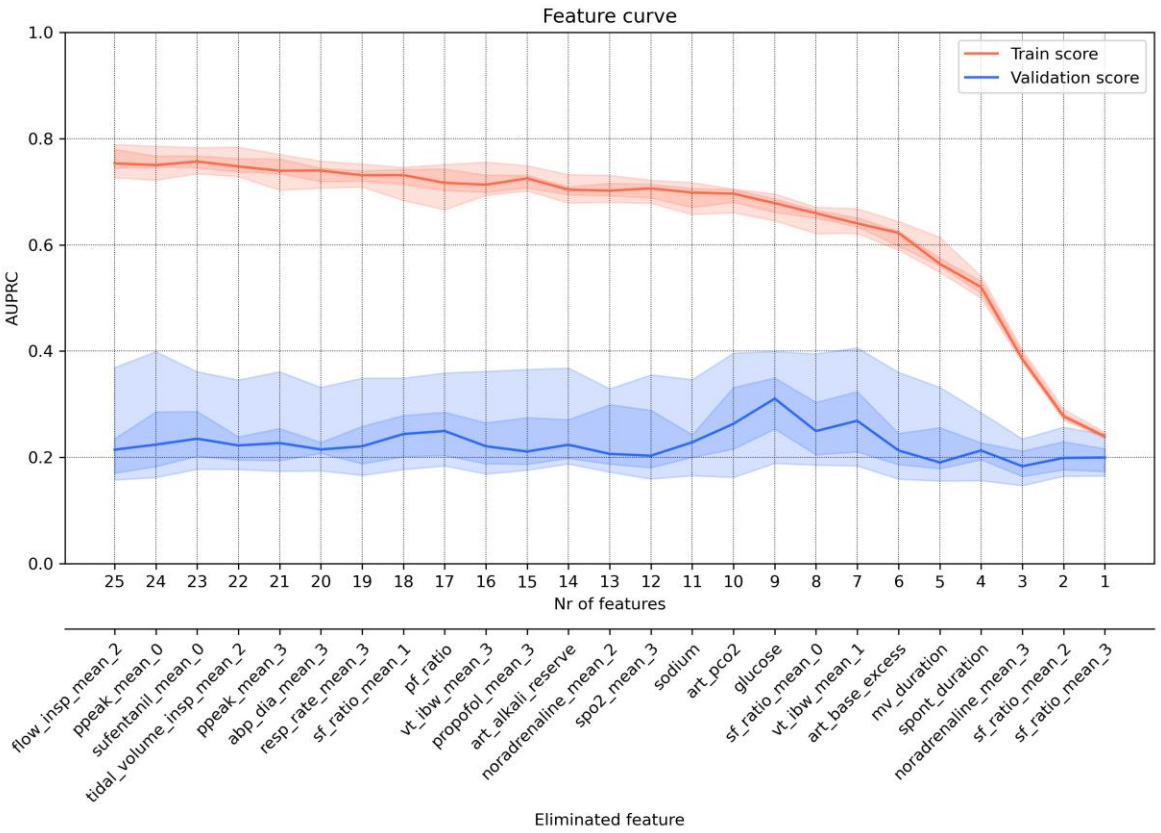


Figure 25. Feature performance curve illustrating the backward feature elimination process for Model 2 with feature set 2 (mean over each 30 minutes). 0 denotes minutes 0-30, 1 denotes minutes 30-60, 2 denotes minutes 60-90, and 3 denotes minutes 90-120. The 5th 25th, 50th, 75th and 95th percentiles are indicated.

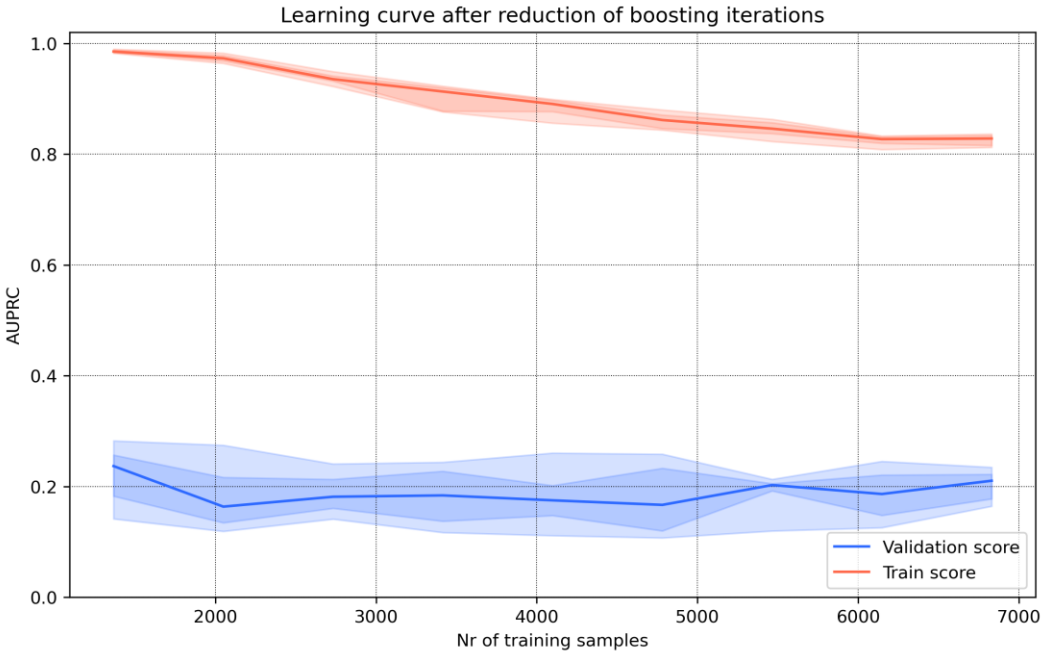


Figure 26. Learning curve of Model 2, illustrating the performance over the number of training samples, after reduction of boosting iterations, prior to feature selection. The, 5th, 25th,50th, 75th and 95th percentiles are indicated.

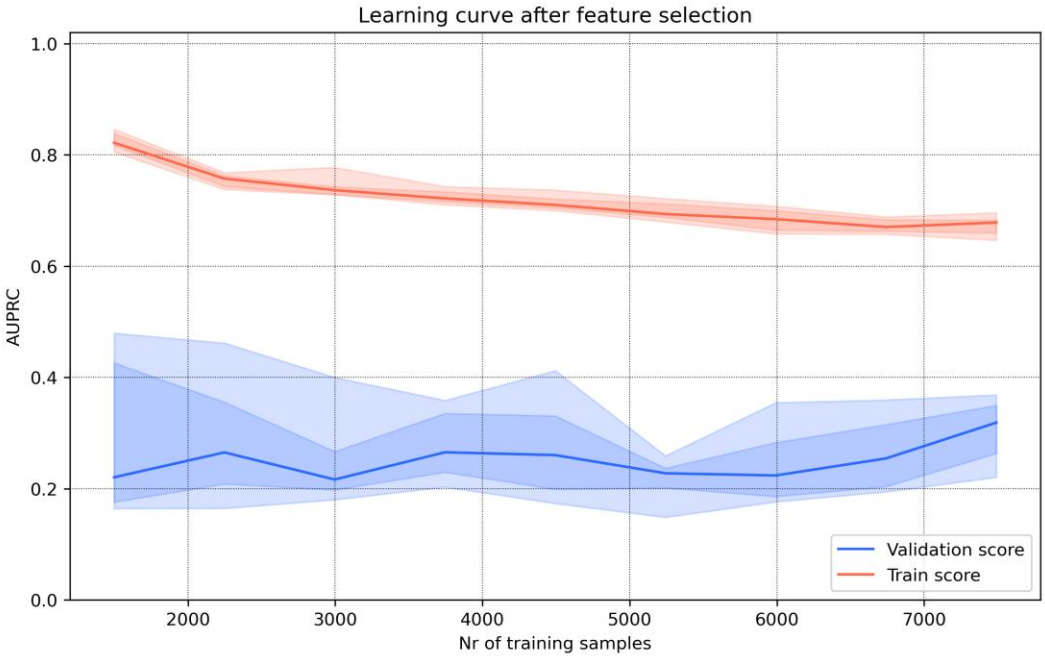


Figure 27. Learning curve of Model 2, illustrating the performance over the number of training samples, after feature selection. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

5.5.5 Hyperparameter optimisation

The AUPRC and AUROC after hyperparameter optimisation were respectively, 0.29 (IQR 0.23-0.35) and 0.76 (IQR 0.70-0.79). Details on optimised hyperparameter settings are provided in *Supplement D.* A learning curve obtained after hyperparameter optimisation is presented in Figure 28.

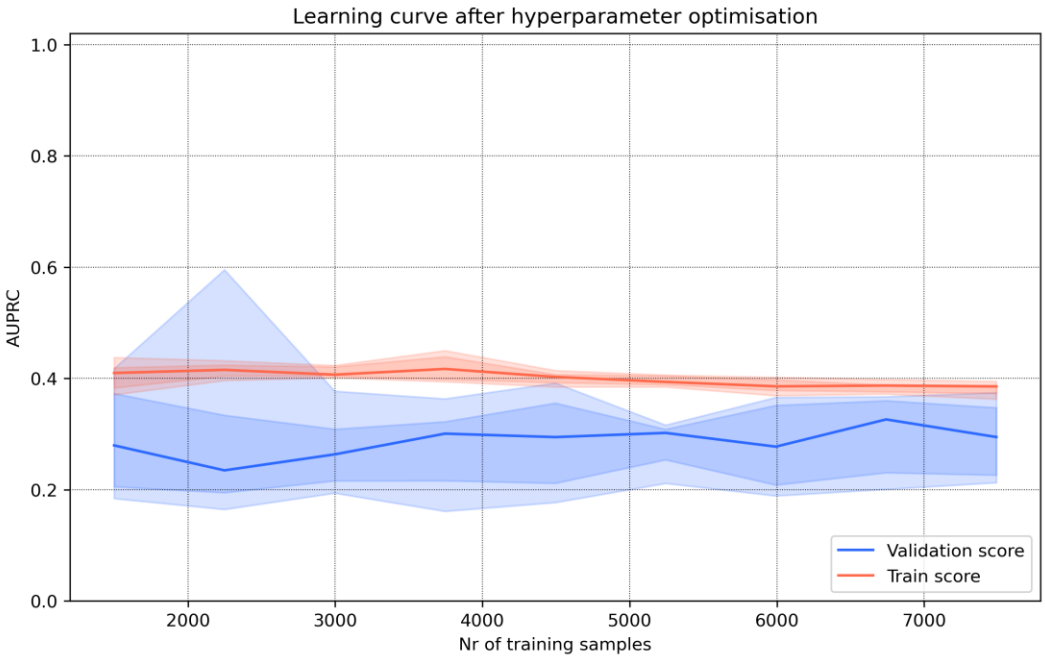


Figure 28. Learning curve of Model 2, illustrating the performance over the number of training samples, obtained after feature selection and hyperparameter optimisation. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

5.5.6 Training with data leakage

Ten-fold cross-validation, grouped by patient ID, yielded a median AUPRC of 0.29 (IQR 0.23-0.35), whereas cross-validation without grouping, which introduces data-leakage, resulted in a median AUPRC of 0.24 (IQR 0.26-0.31, p=0.116) (Table 14).

Table 14. Performance of Model 2 evaluated using 10-fold cross-validation with and without grouping by patient ID, reported as AUPRC and AUROC.

	Grouped cross-validation	Ungrouped cross-validation	P
AUPRC median (IQR)	0.29 (0.23-0.35)	0.24 (0.26-0.31)	0.116
AUROC median (IQR)	0.76 (0.70-0.79)	0.70 (0.66-0.77)	0.020

5.5.7 Model validation

Validation on the test sets yielded an AUPRC of 0.05 and an AUROC of 0.69 for the COVID test set, and an AUPRC of 0.03 and an AUROC of 0.66 for the non-COVID test set. Precision-recall and ROC curves are shown in Figures 29 and 30. For both test sets the target PPV of 0.80 was not achieved, and the sensitivity was extremely low (≤ 0.02) (Table 15). The distribution of predicted probabilities per event class after model calibration is shown in Figure 31.

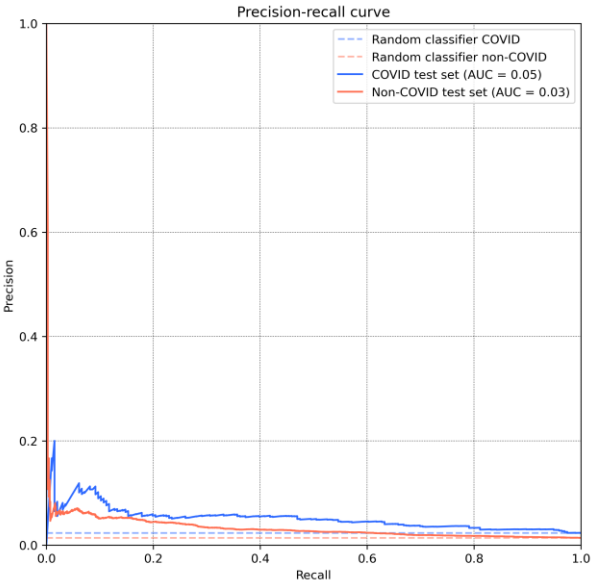


Figure 29. Precision-recall curve of validation of Model 2 on the COVID test dataset (AUC = 0.05) and non-COVID test dataset (AUC = 0.03).

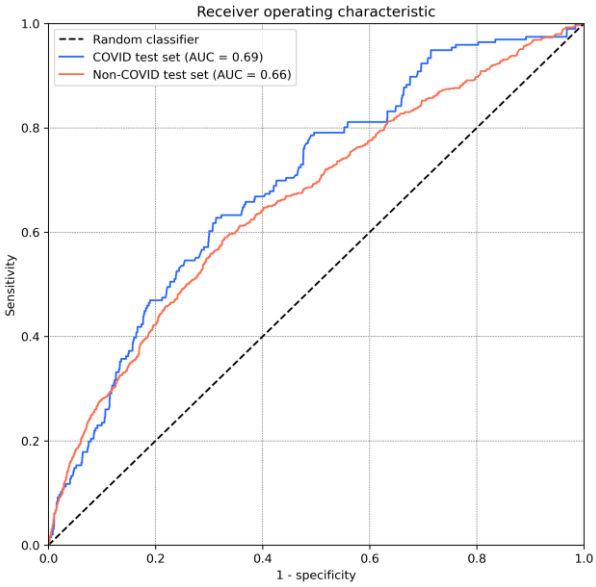


Figure 30. Receiver operating characteristic of validation of Model 2 on the COVID test dataset (AUC = 0.69) and non-COVID test dataset (AUC = 0.66).

Table 15. Overview of the classification metrics for the COVID and non-COVID test set obtained with Model 2.

	COVID test set	Non-COVID test set
Events	169	1083
Controls	8237	76619
AUPRC	0.05	0.03
AUROC	0.69	0.66
Sensitivity (recall)	0.02	0.00
Specificity	1.00	1.00
Positive predictive value (PPV) (precision)	0.20	0.60
Negative predictive value (NPV)	0.98	0.99

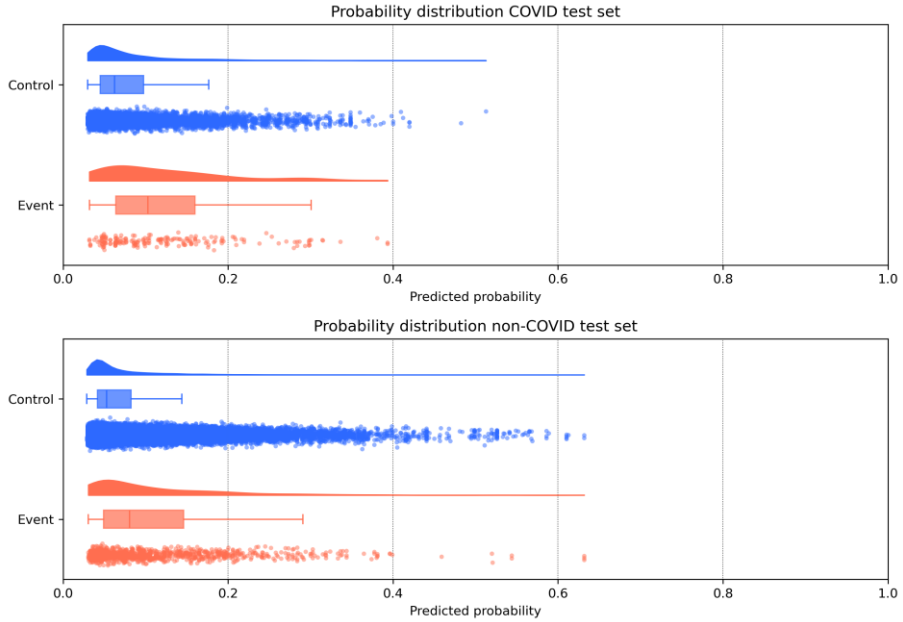


Figure 31. Predicted probability distributions for the COVID and non-COVID test sets obtained with Model 2, shown as violin plots (top), boxplots (middle), and scatterplots (bottom).

5.5.8 Model interpretation

A global SHAP analysis is presented in Figure 32. This analysis shows that low SpO_2/FiO_2 ratios, a long duration of assisted mechanical ventilation in most samples, a long total duration of IMV in a subset of samples, and high noradrenaline infusion rates contribute most strongly to high predicted probabilities. In contrast, a high SpO_2/FiO_2 ratio and V_T/IBW, are associated with low prediction scores.

Local SHAP explanations for event samples with high and low predicted probabilities (*Supplement F),* together with a visualisation of the first decision tree (*Supplement G),* further demonstrate that model output is predominantly driven by the SpO_2/FiO_2 ratio.

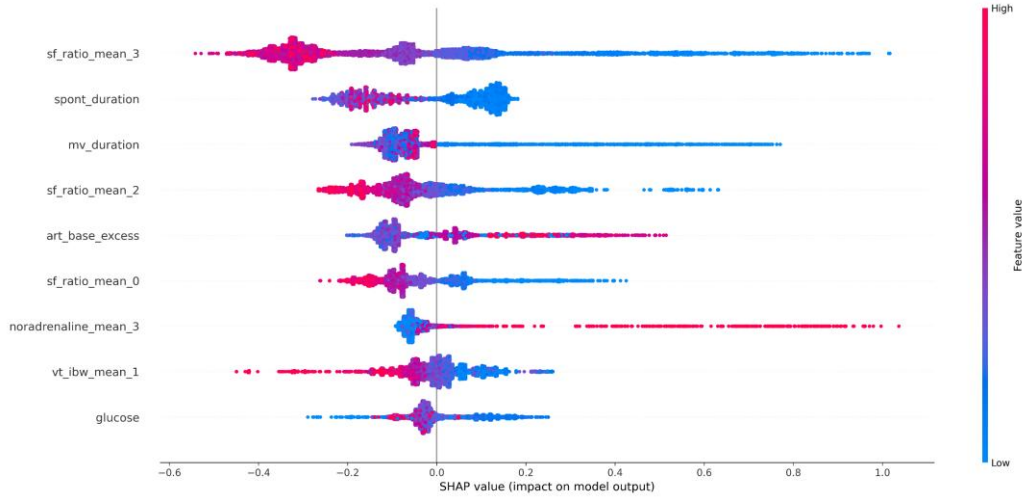


Figure 32. Global SHAP analysis of Model 2 showing the impact of feature values on the model output by aggregating local SHAP values across all training samples. 0 denotes minutes 0-30, 1 denotes minutes 30-60, 2 denotes minutes 60-90, and 3 denotes minutes 90-120.

6 Discussion

In this study, two types of prediction models for mechanically ventilated ICU patients were developed. The first model was aimed to predict early respiratory deterioration (≤ 6 hours) following a premature switch from controlled to assisted ventilation. This model demonstrated moderate discriminative performance, with an AUROC of 0.78 for COVID patients and 0.76 for non-COVID patients. However, its clinical value appears limited, as performance was poor within clinically relevant probability ranges. The most important predictors for unreadiness were the FiO_2 and propofol infusion rate.

The second model was designed as a real-time alarm system to predict delayed respiratory deterioration (> 6 hours after the switch to assisted ventilation) due to the development of P-SILI. The SpO_2/FiO_2 ratio emerged as the most relevant predictor. Nevertheless, the selected features were insufficient to capture the relatively rare respiratory deterioration events 6-8 hours in advance, resulting in an extremely low AUPRC (≤ 0.05).

6.1 Event definition

A critical component of any prediction model is a well-defined event definition. In this retrospective study, establishing such a definition was challenging because event labelling depended on historical clinical decisions. The timing of recognition and intervention varied between clinicians and patients, resulting in events that differed in both nature and severity.

The event definition used in this study was carefully discussed and established by experienced ventilation specialist and subsequently refined through iterative review of detected and missed events in a subsample of patients. However, due to time constraints, it was not feasible to assess the sensitivity and specificity of this definition in a large validation sample. Consequently, uncertainty remains regarding the accuracy of the event labels, which may have affected model performance.

To improve specificity, additional restrictions were applied to reduce false-positive event labels. A minimum duration of three hours of controlled ventilation following the event was required, thereby excluding brief interventions, requiring sedation, not related to respiratory deterioration. In addition, a minimum FiO_2 threshold of 40% in the hour before and after the transition was imposed, increasing the likelihood that events reflected true respiratory deterioration demanding augmented oxygen support. Nevertheless, FiO_2 settings are clinician dependent, and no clear threshold was identified at which most events occurred (Figure 15).

Based on reviewed samples, the only events not adequately distinguished by these restrictions were gastroscopies and bronchoscopies, which are typically accompanied by sedation and elevated FiO_2 levels. Whereas bronchoscopies are often associated with respiratory deterioration, gastroscopies are generally not, likely resulting in a small number of falsely labelled positive events. Furthermore, control samples within 24 hours prior to death were excluded. However, this time window should

likely have been extended. As illustrated by a control sample presented in *Supplement F (1.4)*, a patient in a severely deteriorated condition 30 hours before death was incorrectly labelled as a control sample. Nevertheless, the model appropriately assigned a high predicted risk of respiratory deterioration to this sample.

Despite these limitations, the event definition used in this study seems to be more specific than those employed in comparable studies, for example by Smit et al. (6), where switch failure was defined solely by a minimum duration of one hour of controlled ventilation following the event.

Furthermore, during review and discussion of detected events, two distinct event patterns emerged: early failures occurring shortly after the switch to assisted ventilation, and delayed deteriorations following a period of initial stability with spontaneous breathing. The first group likely reflects patients who were not yet ready to resume spontaneous inspiratory effort, as the acute phase of respiratory failure had not fully resolved. In contrast, the second group appears to represent secondary deterioration during assisted ventilation, plausibly driven by P-SILI. Given the fundamentally different underlying pathophysiology, supported by previously observed differences in ventilatory parameters immediately prior to the transition to assisted ventilation (6), and the distinct clinical implications of these event types, two separate prediction models were developed to address them.

6.2 Model 1 | Predicting respiratory deterioration due to unreadiness for assisted ventilation

For predicting readiness for assisted ventilation, the length of the observation window used to derive aggregate features did not significantly influence model performance. Moreover, no specific period within the 0-6 hour window was identified as containing substantially more predictive information. Feature importance analyses and feature performance curves showed that only a limited number of features contributed meaningfully to model performance (*Supplement C)*. SHAP analyses demonstrated that high FiO_2, propofol infusion rate, and PEEP values were particularly predictive for subsequent respiratory deterioration, whereas low FiO_2, PEEP, and tidal volume per kg ideal body weight, and high SpO_2 and SpO_2/FiO_2 values were particularly predictive of the absence of respiratory deterioration (Figure 24). These findings are consistent with clinical experience, as most of these parameters are key indicators of a patient's oxygenation status. In addition, previous prediction models have also identified PEEP and FiO_2 as important predictors of respiratory status in ICU patients (41–43). Together, this concordance with clinical knowledge and existing literature supports the face validity of the model and reduces the likelihood that its performance is driven by noise in the dataset.

The logistic regression model demonstrated performance comparable to that of the XGBoost model. Notably, the regression coefficients of individual features did not fully align with the feature importance scores derived from the XGBoost model, as detailed in *Supplement B*. This discrepancy is likely attributable to differences in the underlying model assumptions, as well as to overfitting effects in both models.

Comparison of learning curves across different stages of model development shows that feature reduction and hyperparameter optimisation, which substantially reduced model complexity, markedly decreased overfitting and slightly improved overall performance (Figures 17-19). For the final model, the training and validation curves converge to a stable plateau, suggesting that the available sample size was adequate for training this low-complexity XGBoost model.

Introducing data leakage during model training did not result in improved performance (Table 6), arguing against the notion that the model first has to 'learn' patient-specific characteristics during training to make accurate predictions (44,45). In theory, this would allow data from the first hours of ICU admission to be used to calibrate the model to individual patients. However, in this setting, such an approach is unlikely to provide additional benefit.

The model achieved an AUROC of 0.76-0.78 on the test sets, indicating a moderate discriminative and outperforming the AUROC of 0.58 reported by Smit et al. (6) for a similar task with a 72-hour prediction window. However, inspection of the ROC curves shows that this performance is largely driven by good discrimination at high sensitivity and low specificity, corresponding to low probability thresholds (Figure 20). This observation is consistent with the probability distribution plots (Figure 23) and calibration curves (*Supplement E),* which demonstrate that the model is poorly capable of identifying samples with a high event probability. As a result, the model only provides added value in detecting events of respiratory deterioration when applied at low threshold probabilities (Figure 22). In clinical practice, however, this is undesirable, as it would likely lead to unnecessary prolongation of controlled ventilation in many patients, exposing them to increased risks of complications such as respiratory muscle weakness (1). At the clinically relevant operating point with a target specificity of 0.80, sensitivity is very low (0.49-0.56), and comparable to a no-model strategy. Consequently, in its current form and performance level, this model does not appear to have clinical utility.

The model demonstrated comparable performance on the COVID and non-COVID test sets, suggesting good generalisability to the broader ICU population at the LUMC with respiratory failure (PaO_2/FiO_2 ratio ≤ 40).

6.3 Model 2 | Predicting respiratory deterioration due to development of P-SILI

For predicting respiratory deterioration due to P-SILI, a real-time alarm system with hourly updates was envisaged. Accordingly, hourly samples were extracted from periods of assisted ventilation and labelled as control or event samples. This design, combined with the relatively low event incidence, resulted in a highly imbalanced dataset, with an event-to-non-event ratio of approximately 1:400 in COVID patients and even lower in non-COVID patients, yielding an extremely low a priori AUPRC. During model development, under sampling was applied to facilitate model comparison and reduce computational burden. However, during validation on the test sets, the original class proportions were retained to reflect real-world performance.

For this model, a two-hour observation window in combination with a six-hour horizon yielded the best performance. Although shorter prediction horizons might have improved predictive performance, this would offer limited clinical value, as clinicians are generally able to recognise respiratory deterioration within four hours before escalation to controlled ventilation becomes necessary.

Within the XGBoost model, the SpO_2/FiO_2 ratio consistently emerged as the most relevant predictor. This parameter is indeed a key marker of lung function and oxygenation and is included in the global Acute Respiratory Distress Syndrome (ARDS) criteria (46,47). However, its value is strongly influenced by FiO_2 settings, which are adjusted by clinicians. Consequently, increases in FiO_2 initiated by clinicians in response to deterioration may have been captured by the model, suggesting that clinical recognition of deterioration could have preceded model detection. SHAP analysis indicates that a high predicted event probability was particularly influenced by a low SpO_2/FiO_2 ratio (especially during the last 30 minutes of the observation window), a high noradrenaline infusion rate and a short duration of both assisted and total IMV duration, although these effects were not entirely consistent (Figure 32). Conversely, a low predicted event probability was primarily driven by high SpO_2/FiO_2 ratios and tidal volume per kg ideal body weight.

Using these features, the model was able to achieve an AUPRC significantly higher than the a priori baseline. Nevertheless, the absolute performance remained far below a level that would be clinically useful, and the predefined target precision could not be reached. Given the extreme imbalance between events and non-events and the hourly sampling design, achieving a clinically acceptable precision appears infeasible in this population and model setup.

As observed for the first model, logistic regression showed similar performance as XGBoost. In addition, introducing data leakage did not improve performance.

6.4 Strengths and limitations

This study has several limitations. First, the event definition was inherently imperfect, introducing uncertainty into the prediction targets and, consequently, the model outputs. Second, the sample size was limited. Although learning curves indicated that sufficient data were available to train an XGBoost model, both prediction tasks ultimately showed limited predictive performance. This may be attributable to high inter-patient variability, severe class imbalance, and the possibility that the underlying events are intrinsically difficult to predict from routinely collected ICU data. This latter explanation is supported by the low performance (AUROC 0.58-0.70) observed in a variation of Model 1 that used data from the first hour after the switch to assisted ventilation to predict early failure (*Supplement H)*. Even with post-switch data, switch failure could not be reliably predicted. Furthermore, it has been shown that pre-switch characteristics for successful and failed first switch attempts are generally very similar (6), which was also the case in the LUMC dataset (*Supplement I)*. Third, no separate validation dataset was reserved for model development and optimisation. Instead, 10-fold cross-validation was employed, which is a widely accepted and appropriate alternative given the limited sample size (48). However, during feature selection, a small degree of data leakage was introduced by determining the feature elimination order using information from

all ten folds. Fourth, only two different types of machine learning models, XGBoost and logistic regression, were evaluated. Although other machine learning techniques might have yielded improved performance, previous studies suggest that other simple models, suitable for this amount of data, often achieve comparable or lower performances than XGBoost (42,49,50).

Despite these limitations, the study also has notable strengths. First, the event definition was carefully developed through expert consultation, and respiratory deterioration events were stratified into two clinically distinct types. This distinction allowed for the development of separate models tailored to their respective clinical implications. Second, the LUMC ICU dataset is a unique, highly granular dataset with minimal missing data, providing access to a broad range of haemodynamic and ventilatory parameters. Third, extensive feature exploration and selection, and stepwise hyperparameter optimisation were performed. This approach effectively reduced overfitting. Finally, comprehensive model evaluation was conducted using performance measures aligned with the intended clinical applications. Validation on unseen data from distinct patient populations further enabled a meaningful assessment of generalisability.

6.5 Clinical implications

With the current performance levels, neither model is suitable for clinical use. If future retrospective performance improves to a clinically acceptable level, prospective validation would be required to assess effects on treatment outcome and obtain user feedback. In addition, appropriate clinical interventions corresponding to different predicted probability thresholds would need to be defined.

Careful consideration of implementation is essential, as key input variables such as FiO_2 and PEEP are clinician-dependent and may change once clinicians are aware of the model's use. Moreover, users must be aware that the models were trained and validated exclusively on LUMC ICU patients ventilated for at least 48 hours with a PaO_2/FiO_2 ratio ≤ 40 kPa and are therefore not applicable to all ICU patients. Finally, technical feasibility should be explored, ideally integrating the models into the PDMS, where predictions could be accessed or activated at the clinician's discretion.

6.6 Future directions

Future research should primarily focus on improving model performance. This may be achieved by using larger datasets in combination with more advanced modelling approaches. In particular, recurrent neural networks could directly leverage high-frequency ICU data rather than aggregated features, enabling more effective capture of temporal dynamics in ventilatory and haemodynamic parameters. Such models have demonstrated promising performance in related prediction tasks for ICU patients (51).

For the hourly updated P-SILI prediction model, an alternative design should be considered. The current approach inherently results in extreme class imbalance, making it unlikely to achieve clinically useful performance. Potential solutions include increasing the update interval or focusing on patient populations with a higher event density.

In addition, given the limited number of predictive features, especially for P-SILI, more fundamental research is needed to identify key pathophysiological drivers that could serve as meaningful predictors.

Once satisfactory performance is achieved, retrospective validation on external hospital datasets should be performed, followed by prospective validation to assess clinical impact and effects on patient outcomes.

7 Conclusion

In this study, two prediction models were developed to support clinical decision-making during the transition from controlled to assisted ventilation, and to alert clinicians to respiratory deterioration during assisted mechanical ventilation. The first model, aimed at predicting early respiratory deterioration due to premature switching, demonstrated moderate discriminative performance but lacked sufficient clinical utility at relevant operating points. The second model, designed as a real-time alarm system to predict delayed deterioration due to P-SILI, showed very limited predictive value, largely due to extreme class imbalance and the scarcity of informative predictors.

Although extensive exploration of optimal observation windows, prediction horizons, and predictive features, as well as careful model optimisation, were performed, both models remain unsuitable for clinical implementation in their current form. These findings highlight the complexity of predicting respiratory deterioration in mechanically ventilated ICU patients and underscore the challenges posed by retrospective event labelling, and clinician-dependent parameters.

Nevertheless, this study provides a careful and clinically relevant model design and establishes a transparent and reproducible framework for future work. With larger datasets and advanced modelling techniques capable of leveraging high-frequency ICU data, clinically useful prediction models may become feasible.

8 References

1. Goligher EC, Dres M, Patel BK, Sahetya SK, Beitler JR, Telias I, et al. Lung- and Diaphragm-Protective Ventilation. Am J Respir Crit Care Med. 2020 Oct 1;202(7):950–61.

2. Saddy F, Sutherasan Y, Rocco PRM, Pelosi P. Ventilator-associated lung injury during assisted mechanical ventilation. Semin Respir Crit Care Med. 2014 Aug;35(4):409–17.

3. Wongtangman K, Grabitz SD, Hammer M, Wachtendorf LJ, Xu X, Schaefer MS, et al. Optimal Sedation in Patients Who Receive Neuromuscular Blocking Agent Infusions for Treatment of Acute Respiratory Distress Syndrome—A Retrospective Cohort Study From a New England Health Care Network*. Crit Care Med. 2021 Jul;49(7):1137.

4. Reep CAT, Wils EJ, Fleuren LM, Breskin A, Bellani G, Laffey JG, et al. Early versus Delayed Switching from Controlled to Assisted Ventilation: A Target Trial Emulation. Am J Respir Crit Care Med. 2025 Jun;211(6):975–83.

5. Pérez J, Accoce M, Dorado JH, Gilgado DI, Navarro E, Cardoso GP, et al. Failure of First Transition to Pressure Support Ventilation After Spontaneous Awakening Trials in Hypoxemic Respiratory Failure: Influence of COVID-19. Crit Care Explor. 2023 Sep;5(9):e0968.

6. Smit JM, Van Bommel J, Gommers DAMPJ, Reinders MJT, Van Genderen ME, Krijthe JH, et al. Switching from controlled to assisted mechanical ventilation: a multi-center retrospective study (SWITCH). Intensive Care Med Exp. 2025 Jul 16;13(1):73.

7. Polo Friz M, Rezoagli E, Safaee Fakhr B, Florio G, Carlesso E, Giudici R, et al. Successful Versus Failed Transition From Controlled Ventilation to Pressure Support Ventilation in COVID-19 Patients: A Retrospective Cohort Study. Crit Care Explor. 2024 Feb;6(2):e1039.

8. Haudebourg AF, Chantelot L, Nemlaghi S, Haudebourg L, Labedade P, Boujelben MA, et al. Factors influencing the transition phase in acute respiratory distress syndrome: an observational cohort study. Ann Intensive Care. 2025 Jan 1;15(1):71.

9. Silva PL, Ball L, Rocco PRM, Pelosi P. Physiological and Pathophysiological Consequences of Mechanical Ventilation. Semin Respir Crit Care Med. 2022 Jun;43(3):321–34.

10. Carteaux G, Parfait M, Combet M, Haudebourg AF, Tuffet S, Mekontso Dessap A. Patient-Self Inflicted Lung Injury: A Practical Review. J Clin Med. 2021 Jun 21;10(12):2738.

11. Brochard L, Slutsky A, Pesenti A. Mechanical Ventilation to Minimize Progression of Lung Injury in Acute Respiratory Failure. Am J Respir Crit Care Med. 2017 Feb 15;195(4):438–42.

12. Hu W, Jin T, Pan Z, Xu H, Yu L, Chen T, et al. An interpretable ensemble learning model facilitates early risk stratification of ischemic stroke in intensive care unit: Development and external validation of ICU-ISPM. Comput Biol Med. 2023 Nov;166:107577.

13. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Crit Care Med. 2018 Apr;46(4):547–53.

14. Hong N, Liu C, Gao J, Han L, Chang F, Gong M, et al. State of the Art of Machine Learning-Enabled Clinical Decision Support in Intensive Care Units: Literature Review. JMIR Med Inform. 2022 Mar 3;10(3):e28781.

15. Rooney SR, Clermont G. Forecasting algorithms in the ICU. J Electrocardiol. 2023;81:253–7.

16. Zou Y, Liu Z, Miao Q, Wu J. A review of intraoperative protective ventilation. Anesthesiol Perioper Sci. 2024 Feb 6;2(1):10.

17. Jonkman AH, de Vries HJ, Heunks LMA. Physiology of the Respiratory Drive in ICU Patients: Implications for Diagnosis and Treatment. Crit Care. 2020 Mar 24;24:104.

18. Ventilation modes [Internet]. Hamilton Medical AG; [cited 2025 Dec 22]. Available from: https://www.hamilton-medical.com/en_NL/Products/Compare/Compare-ventilation-modes.html

19. ASV - Adaptive Support Ventilation [Internet]. Hamilton Medical AG; [cited 2025 Dec 22]. Available from: https://www.hamilton-medical.com/en_NL/Products/Technologies/ASV.html

20. INTELLiVENT-ASV [Internet]. Hamilton Medical AG; [cited 2026 Jan 6]. Available from: https://www.hamilton-medical.com/tr/Products/Technologies/INTELLiVENT-ASV.html

21. Arnal JM. Monitoring respiratory mechanics in mechaniscally ventilated patients [Internet]. Hamilton Medical AG; 2018. Available from: https://www.hamilton-medical.com/en_NL/Article-page~knowledge-base~6e39d4bb-1ab7-4c46-bc18-83f3e77897f9~Monitoring-respiratory-mechanics-in-mechanically-ventilated-patients~.html

22. Nick TG, Campbell KM. Logistic Regression. In: Topics in Biostatistics [Internet]. Totowa: Humana Press; 2007 [cited 2026 Jan 11]. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-59745-530-5_14

23. Lee F. What is logistic regression? [Internet]. IBM; [cited 2026 Jan 11]. Available from: https://www.ibm.com/think/topics/logistic-regression

24. Introduction to Boosted Trees [Internet]. xgboost developers; [cited 2025 Dec 22]. Available from: https://xgboost.readthedocs.io/en/stable/tutorials/model.html

25. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2025 Dec 19]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

26. XGBoost Parameters [Internet]. xgboost developers; [cited 2025 Dec 22]. Available from: https://xgboost.readthedocs.io/en/stable/parameter.html

27. Viering T, Loog M. The Shape of Learning Curves: A Review. IEEE Trans Pattern Anal Mach Intell. 2023 Jun;45(6):7799–819.

28. Mohr F, Rijn JN van. Learning Curves for Decision Making in Supervised Machine Learning: A Survey. Mach Learn. 2024 Dec;113(11–12):8371–425.

29. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019 Oct 4;3(1):18.

30. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378.

31. van der Wal LI, Grim CCA, del Prado MR, van Westerloo DJ, Boerma EC, Rijnhart-de Jong HG, et al. Conservative versus Liberal Oxygenation Targets in Intensive Care Unit Patients (ICONIC): A Randomized Clinical Trial. Am J Respir Crit Care Med. 2023 Oct;208(7):770–9.

32. Blythe R, Naicker S, White N, Donovan R, Scott IA, McKelliget A, et al. Clinician perspectives and recommendations regarding design of clinical prediction models for deteriorating patients in acute care. BMC Med Inform Decis Mak. 2024 Sep 2;24:241.

33. Yang HC, Hao ATH, Liu SC, Chang YC, Tsai YT, Weng SJ, et al. Prediction of Spontaneous Breathing Trial Outcome in Critically Ill-Ventilated Patients Using Deep Learning: Development and Verification Study. JMIR Med Inform. 2025 May 21;13:e64592.

34. Liu J, Duan X, Duan M, Jiang Y, Mao W, Wang L, et al. Development and external validation of an interpretable machine learning model for the prediction of intubation in the intensive care unit. Sci Rep. 2024 Nov 8;14(1):27174.

35. Bendavid I, Statlender L, Shvartser L, Teppler S, Azullay R, Sapir R, et al. A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19. Sci Rep. 2022 Jun 22;12(1):10573.

36. Shafiuzzaman M, Safayet Islam M, Rubaith Bashar TM, Munem M, Nahiduzzaman M, Ahsan M, et al. Enhanced very short-term load forecasting with multi-lag feature engineering and prophet-XGBoost-CatBoost architecture. Energy. 2025 Oct 30;335:137981.

37. Wang Z, Zhang L, Huang T, Yang R, Cheng H, Wang H, et al. Developing an explainable machine learning model to predict the mechanical ventilation duration of patients with ARDS in intensive care units. Heart Lung J Crit Care. 2023 Apr;58:74–81.

38. A Structured Approach to Tuning. In: Getting Started with Gradient Boosting Algorithms [Internet]. ApX Machine Learning; 2025 [cited 2025 Dec 19]. Available from: https://apxml.com/courses/getting-started-with-gradient-boosting-algorithms/chapter-6-hyperparameter-tuning-and-optimization/structured-approach-to-tuning

39. Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digit Health. 2025 Dec;7(12):100916.

40. Lundberg S, Lee SI. SHAP [Internet]. SHAP; 2025 [cited 2025 Oct 16]. Available from: https://github.com/shap/shap

41. Otaguro T, Tanaka H, Igarashi Y, Tagami T, Masuno T, Yokobori S, et al. Machine Learning for Prediction of Successful Extubation of Mechanical Ventilated Patients in an Intensive Care Unit: A Retrospective Observational Study. J Nippon Med Sch Nippon Ika Daigaku Zasshi. 2021 Nov 17;88(5):408–17.

42. Liu CF, Hung CM, Ko SC, Cheng KC, Chao CM, Sung MI, et al. An artificial intelligence system to predict the optimal timing for mechanical ventilation weaning for intensive care unit patients: A two-stage prediction approach. Front Med. 2022;9:935366.

43. Jia Y, Kaul C, Lawton T, Murray-Smith R, Habli I. Prediction of weaning from mechanical ventilation using Convolutional Neural Networks. Artif Intell Med. 2021 Jul;117:102087.

44. Visweswaran S, Angus DC, Hsieh M, Weissfeld L, Yealy D, Cooper GF. Learning patient-specific predictive models from clinical data. J Biomed Inform. 2010 Oct;43(5):669–85.

45. Liu K, Zhang X, Chen W, Yu ASL, Kellum JA, Matheny ME, et al. Development and Validation of a Personalized Model With Transfer Learning for Acute Kidney Injury Risk Estimation Using Electronic Health Records. JAMA Netw Open. 2022 Jul 1;5(7):e2219776.

46. Rangappa R. A Game Changer for ARDS? Unraveling the Potential of the SF Ratio. Indian J Crit Care Med Peer-Rev Off Publ Indian Soc Crit Care Med. 2024 Mar;28(3):191–2.

47. Matthay MA, Arabi Y, Arroliga AC, Bernard G, Bersten AD, Brochard LJ, et al. A New Global Definition of Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med. 2024 Jan;209(1):37–47.

48. Eertink JJ, Heymans MW, Zwezerijnen GJC, Zijlstra JM, de Vet HCW, Boellaard R. External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients. EJNMMI Res. 2022 Sep 11;12:58.

49. Lyu G, Nakayama M. Prediction of respiratory failure risk in patients with pneumonia in the ICU using ensemble learning models. PloS One. 2023;18(9):e0291711.

50. Xu H, Ma Y, Zhuang Y, Zheng Y, Du Z, Zhou X. Machine learning-based risk prediction model construction of difficult weaning in ICU patients with mechanical ventilation. Sci Rep. 2024 Sep 6;14(1):20875.

51. Chen D, Wang R, Jiang Y, Xing Z, Sheng Q, Liu X, et al. Application of artificial neural network in daily prediction of bleeding in ICU patients treated with anti-thrombotic therapy. BMC Med Inform Decis Mak. 2023 Aug 31;23(1):171.

Supplementary Materials

A Input variables

Table A1. Overview of variables used as input features specified for Model 1 and 2, with percentages of missing data and availability per record in the COVID dataset.

Variable	Type	Metric	% missing data	Nr records available	Model 1	Model 2
Heartrate	Monitor	Mean, std, trend	0.9	374	✓	✓
Mean arterial blood pressure	Monitor	Mean, std, trend	1.3	374	✓	✓
Systolic arterial blood pressure	Monitor	Mean, std, trend	1.3	374	✓	✓
Diastolic arterial blood pressure	Monitor	Mean, std, trend	1.4	374	✓	✓
Perfusion flow index	Monitor	Mean, std, trend	2.0	372	✓	✓
SpO_2	Monitor	Mean, std, trend	2.0	372	✓	✓
FiO_2	Ventilator	Mean, std, trend	4.3	374	✓	✓
Respiratory rate	Ventilator	Mean, std, trend	4.3	374	✓	✓
Spontaneous respiratory rate	Ventilator	Mean, std, trend	4.4	374		✓
Expiratory tidal volume	Ventilator	Mean, std, trend	4.6	374	✓	✓
Inspiratory tidal volume	Ventilator	Mean, std, trend	4.6	374	✓	✓
I:E ratio	Ventilator	Mean	5.0	374	✓	✓
End-tidal CO_2	Ventilator	Mean, std, trend	4.5	373	✓	✓
Spontaneous minute ventilation	Ventilator	Mean, std, trend	5.1	374		✓
Rinsp	Ventilator	Mean, std, trend	8.8	374	✓	✓
V_T/IBW	Ventilator	Mean, std, trend	5.0	374	✓	✓
Pinsp	Ventilator	Mean, std, trend	5.4	374	✓	✓
Minute ventilation	Ventilator	Mean, std, trend	4.5	374	✓	✓
Pmean	Ventilator	Mean, std, trend	4.5	374	✓	✓
RCexp	Ventilator	Mean, std, trend	4.5	374	✓	✓
PEEP	Ventilator	Mean, std, trend	4.5	374	✓	✓
Auto PEEP	Ventilator	Mean, std, trend	4.7	374	✓	✓
Ppeak	Ventilator	Mean, std, trend	4.9	374	✓	✓
$Flow_{insp}$	Ventilator	Mean, std, trend	4.9	374	✓	✓
$Flow_{exp}$	Ventilator	Mean, std, trend	5.0	374	✓	✓
Compliance	Ventilator	Mean, std, trend	5.7	374	✓	✓
Delta P	Ventilator	Mean, std, trend	5.4	374	✓	✓
PF ratio	Ventilator	Mean, std, trend	15.8	374	✓	✓
RSBI	Ventilator	Mean, std, trend	4.6	374		✓
SF ratio	Ventilator	Mean, std, trend	4.7	374	✓	✓
IMV duration	Ventilator		0	374	✓	✓
Assisted IMV duration	Ventilator		0	374		✓

Variable	Type	Metric	% missing data	Nr records available	Model 1	Model 2
Sufentanil	Medication	Mean	0	374	✓	✓
Noradrenaline	Medication	Mean	0	374	✓	✓
Propofol	Medication	Mean	0	374	✓	✓
Rocuronium	Medication	Mean	0	374	✓	✓
Midazolam	Medication	Mean	0	374	✓	✓
Dexmedetomidine	Medication	Mean	0	374	✓	✓
Remifentanil	Medication	Mean	0	374	✓	✓
Dobutamine	Medication	Mean	0	374	✓	✓
Glucose	Lab	Last value	6.8	374	✓	✓
Potassium	Lab	Last value	8.1	374	✓	✓
Chloride	Lab	Last value	8.4	374	✓	✓
Free calcium	Lab	Last value	10.0	374	✓	✓
Sodium	Lab	Last value	8.8	374	✓	✓
Arterial pCO_2	Lab	Last value	8.8	374	✓	✓
Arterial pH	Lab	Last value	8.9	374	✓	✓
Arterial pO_2	Lab	Last value	8.9	374	✓	✓
Arterial Alkali Reserve	Lab	Last value	8.9	374	✓	✓
Arterial Base Excess	Lab	Last value	8.9	374	✓	✓
Lactate	Lab	Last value	9.1	374	✓	✓
Arterial O_2 saturation	Lab	Last value	10.3	374	✓	✓
Arterial Ht	Lab	Last value	10.4	374	✓	✓
Age	Other		0	374	✓	✓
Sex	Other		0	374	✓	✓
BMI	Other		0	374	✓	✓

B Logistic regression coefficients

1 Model 1

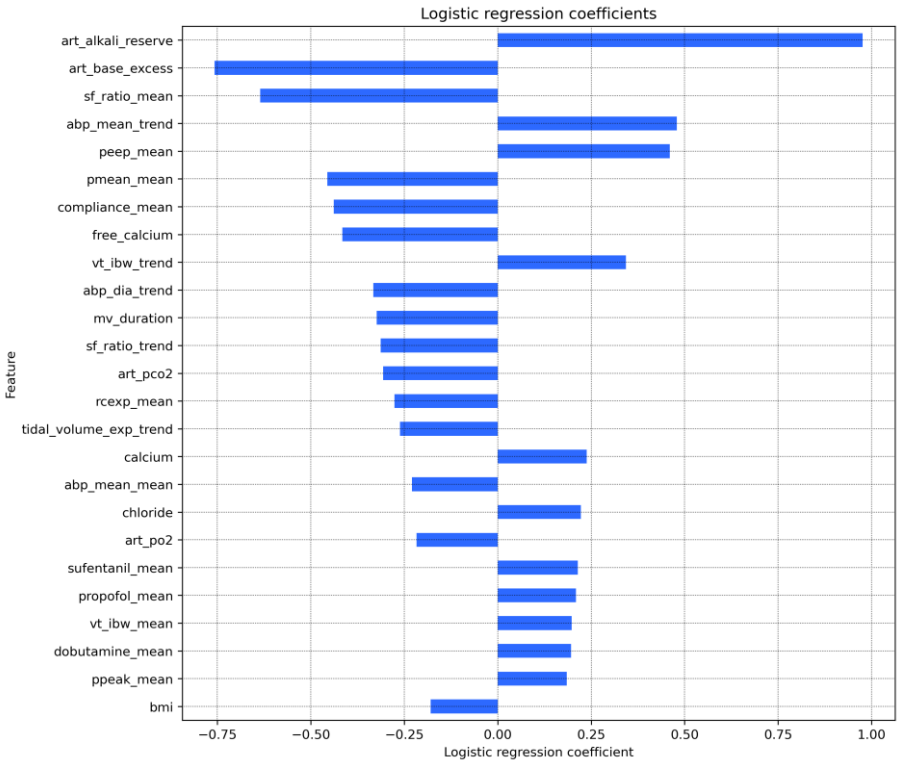


Figure B1. The 25 features with the largest absolute regression coefficients resulting from the logistic regression model fitted in Section 5.4.2.

Correlation between logistic regression coefficients and XGBoost feature importance

The correlation between the logistic regression coefficients and the XGBoost gain metric for each feature was explored using a scatter plot (Figure B2). Some degree of correlation is expected, as both metrics reflect a feature's influence on model output. Notable discrepancies were observed for, among others, FiO_2 (low regression coefficient, high gain) and arterial alkali reserve (high regression coefficient, low gain), whereas the SpO_2/FiO_2 ratio and PEEP showed relatively high values for both the regression coefficient and gain.

The inconsistencies can partially be explained by the substantial overlap in feature value distributions between event and control samples (Figure B3). Where XGBoost may still exploit these features effectively by performing multiple splits and combining them with other features, logistic regression is limited by its assumption of a linear relationship with the log-odds of the event probability. In addition, the relatively large number of input features compared with the sample size likely resulted in overfitting in both models. This may have further contributed to discrepancies between the observed regression coefficients and XGBoost gain and the true influence of these features on event probability.

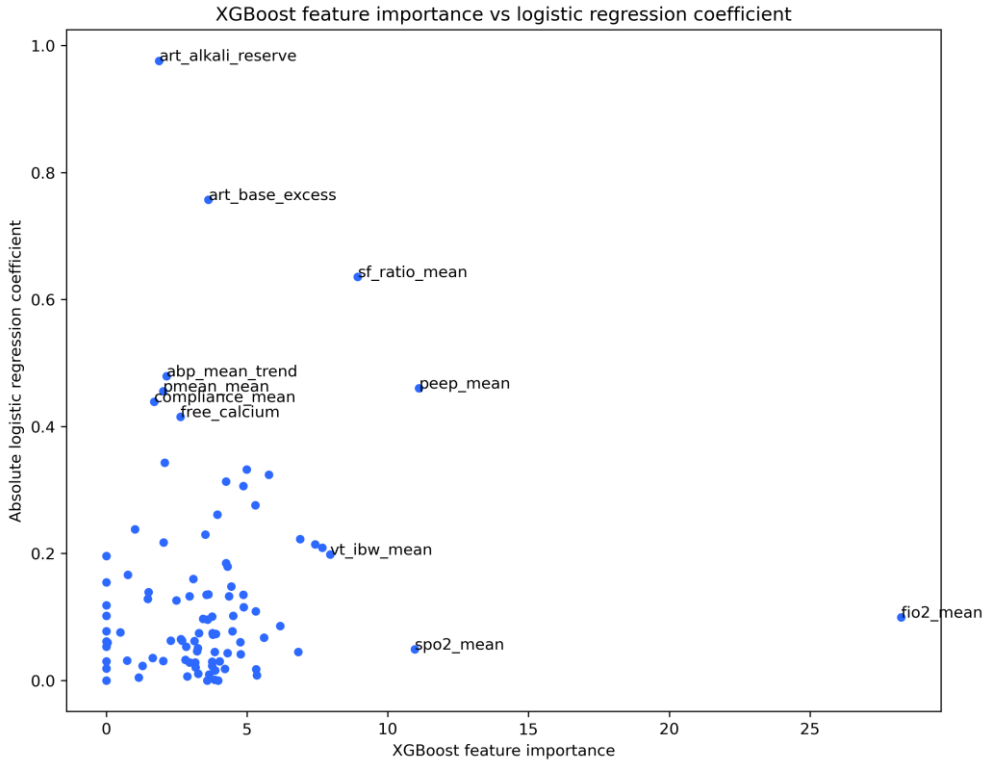


Figure B2. Scatterplot of the absolute logistic regression coefficient and the XGBoost feature importance (gain) for each feature, obtained by the fitted models in Section 5.4.2.

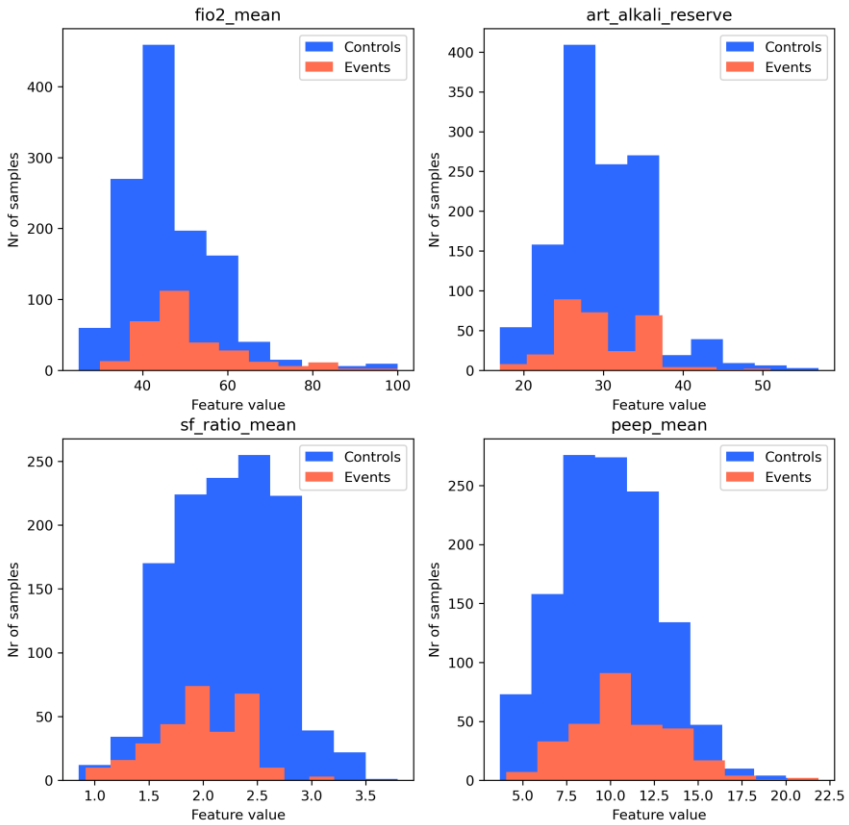


Figure B4. Feature value distributions for control and event samples of mean FiO_2, arterial alkali reserve, mean SpO_2/FiO_2 ratio and mean PEEP.

2 Model 2

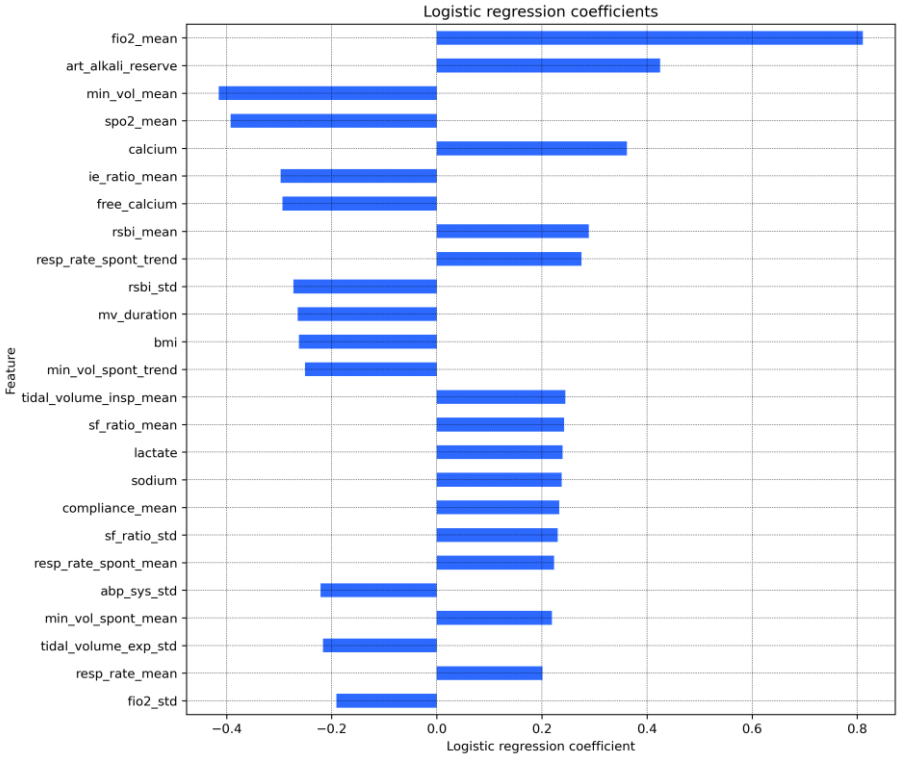


Figure B5. The 25 features with the largest absolute regression coefficients resulting from the logistic regression model fitted in Section 5.5.3.

C Feature selection

1 Model 1

1.1 Feature set 1

Mean, std, and trend over 1 hour, with a 1-hour observation window.

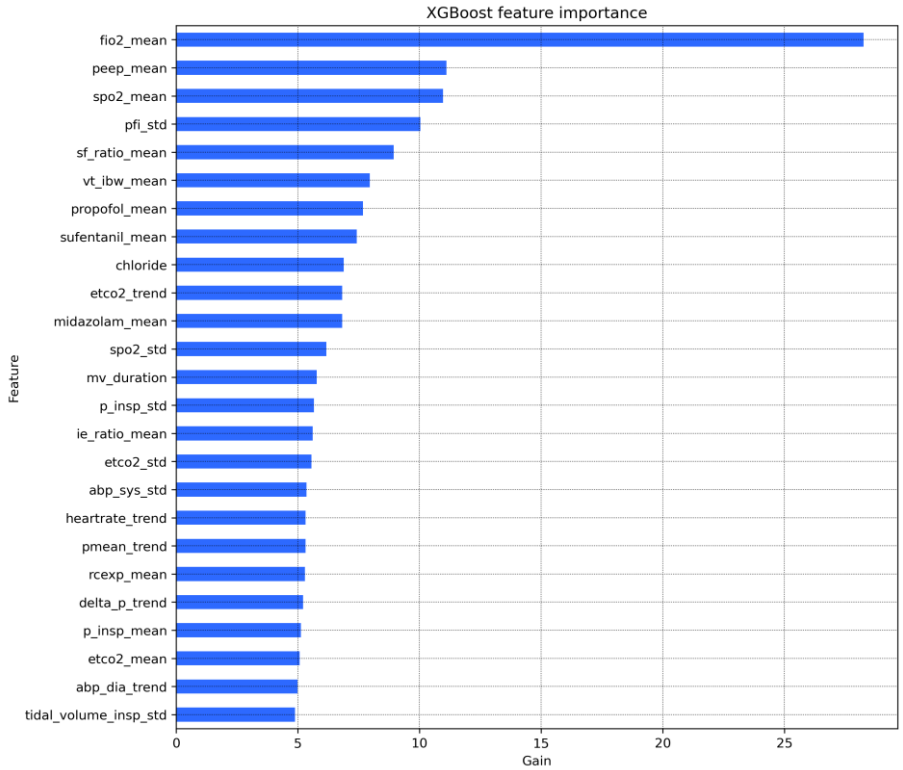


Figure C1. Top 25 feature importance scores, calculated as the gain, from XGBoost Model 1 with feature set 1.

1.2 Feature set 2

Mean over each 20-minute interval, with a 1-hour observation window.

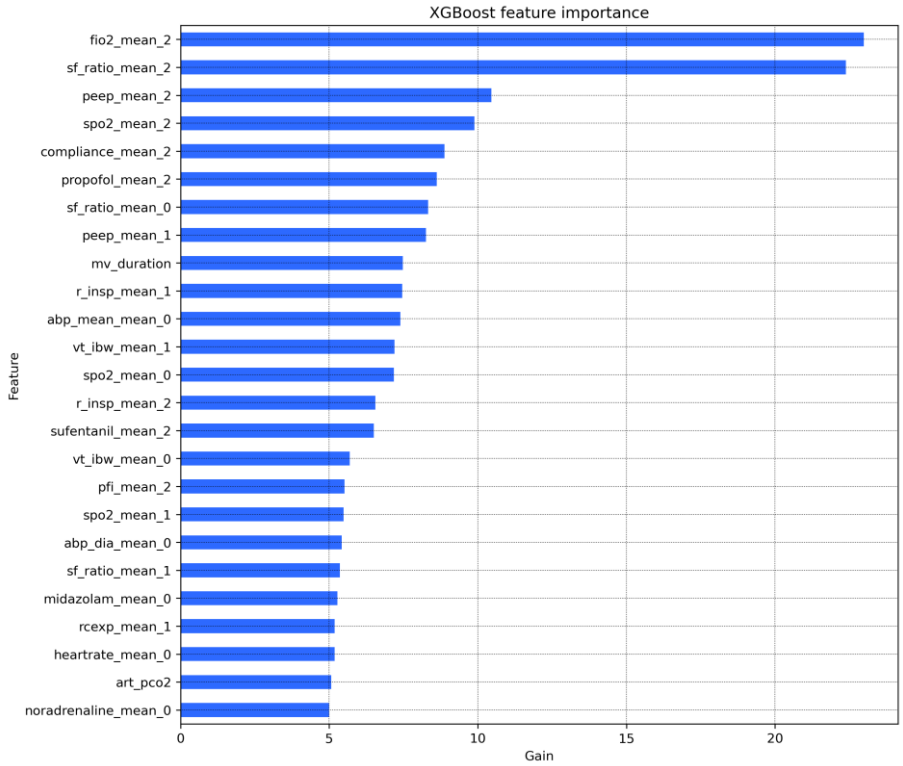


Figure C2. Top 25 feature importance scores, calculated as gain, from XGBoost Model 1 with feature set 2. 0 denotes minutes 0-20, 1 denotes minutes 20-40, and 2 denotes minutes 40-60.

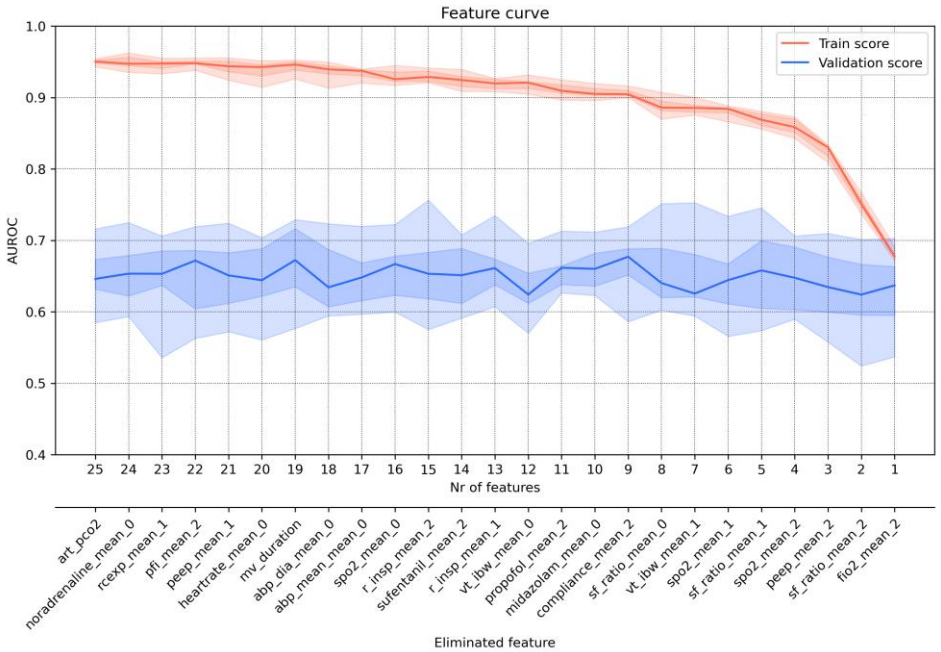


Figure C3. Feature curve illustrating the backward feature elimination process for Model 1 with feature set 2. 0 denotes minutes 0-20, 1 denotes minutes 20-40, and 2 denotes minutes 40-60. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

1.3 Feature set 3

Mean, std, and trend over 6 hours, with a 6-hour observation window.

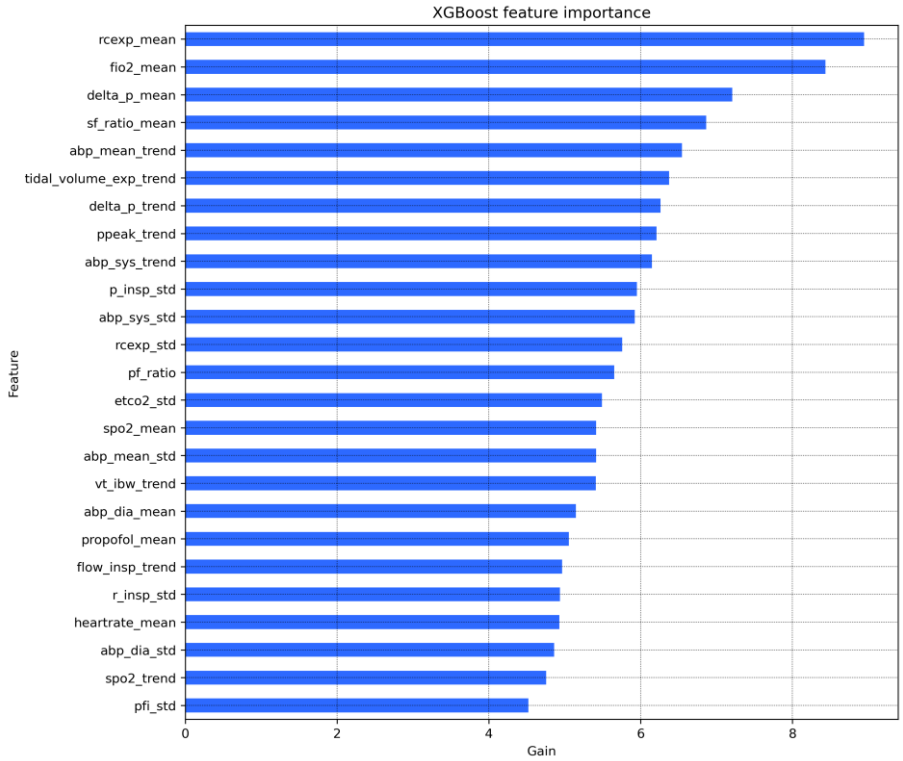


Figure C4. Top 25 feature importance scores, calculated as gain, from the XGBoost Model 1 with feature set 3.

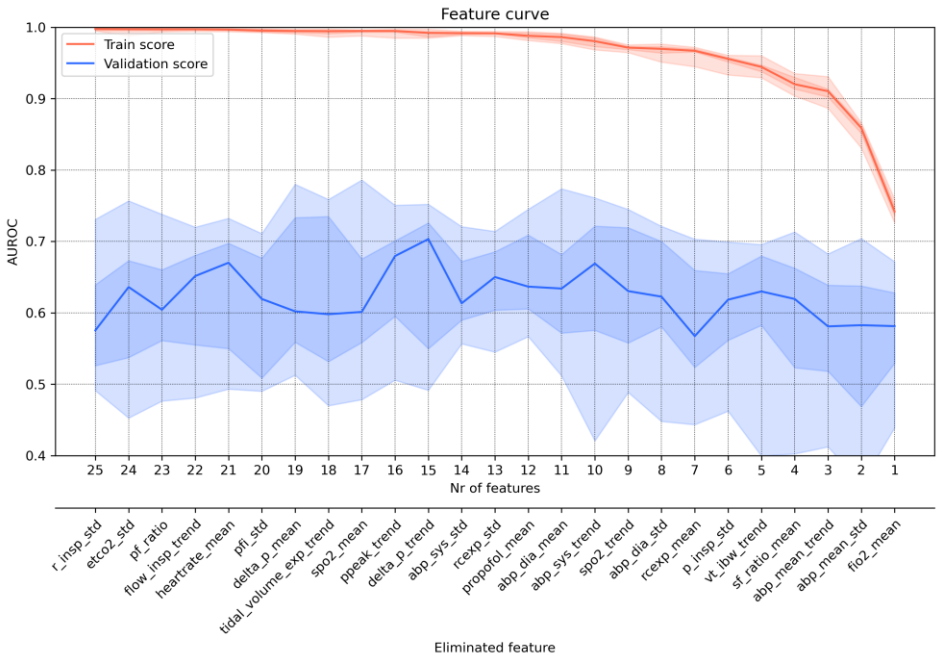


Figure C5. Feature curve illustrating the forward feature selection process for Model 1 with feature set 3. The 5th, 25th, 50th, 75th and 95th percentiles, obtained using 10-fold cross-validation, are indicated.

1.4 Feature set 4

Mean over each 2-hour interval, with 6-hour observation window.

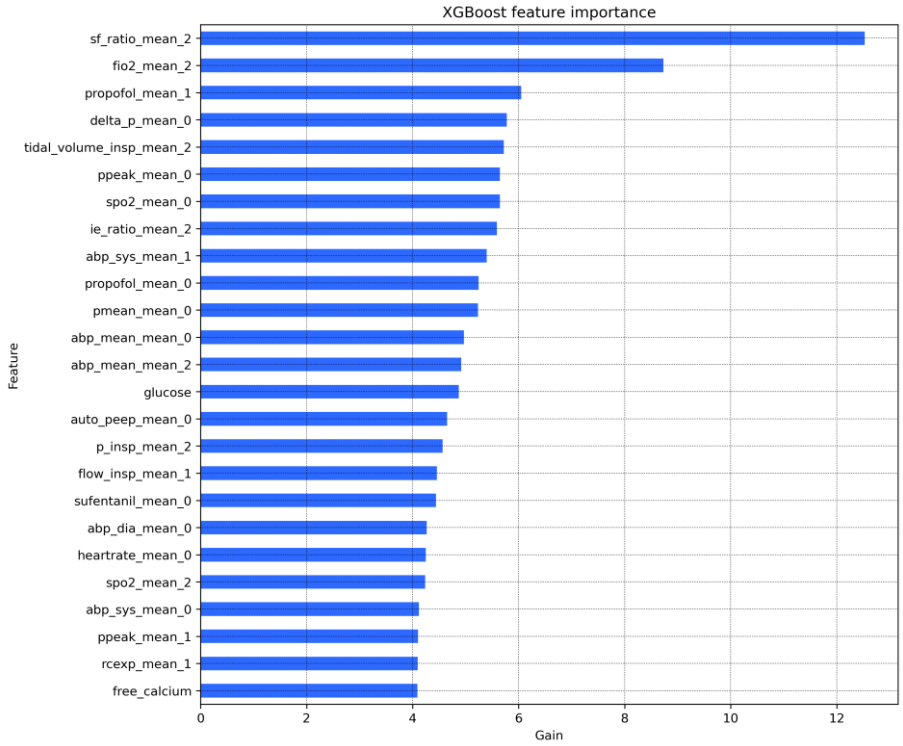


Figure C6. Top 25 feature importance scores, calculated as gain, from XGBoost Model 1 with feature set 4. 0 denotes hours 0-2, 1 denotes hours 2-4, and 2 denotes hours 4-6.

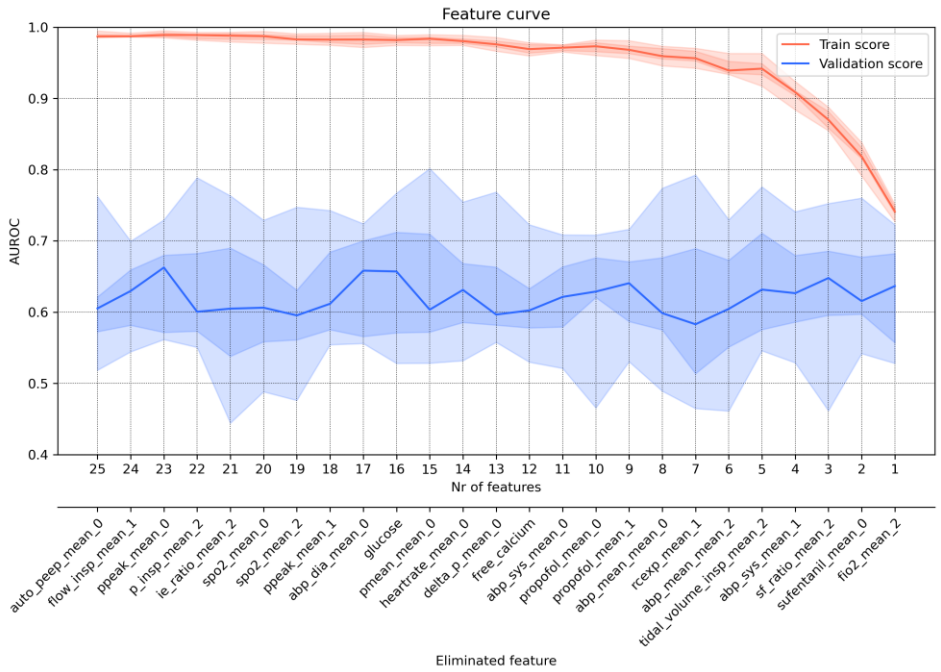


Figure C7. Feature curve illustrating the forward feature selection process for Model 1 with feature set 4. 0 denotes hours 0-2, 1 denotes hours 2-4, and 2 denotes hours 4-6. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

2 Model 2

2.1 Feature set 1

Mean, std, and trend over 2 hours, with a 2-hour observation window.

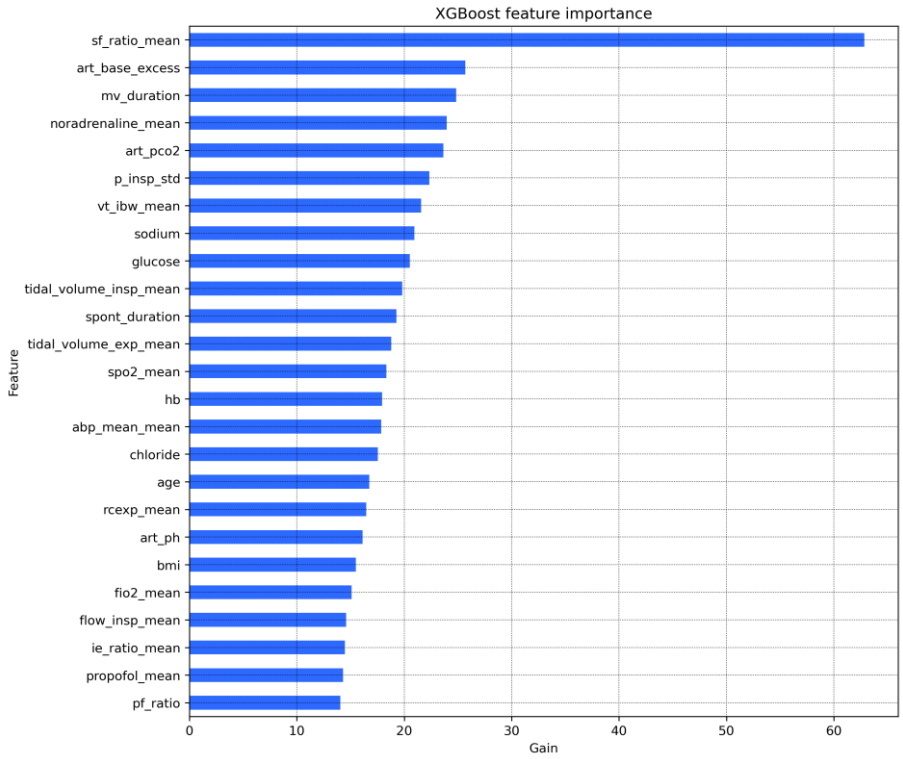


Figure C8. Top 25 feature importance scores, calculated as gain, from XGBoost Model 2 with feature set 1.

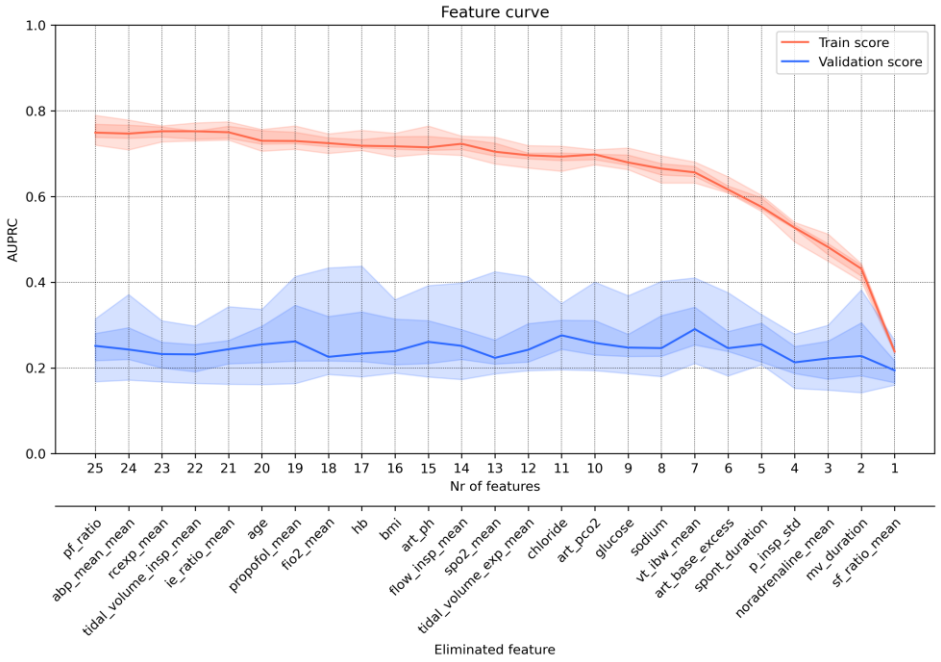


Figure C9. Feature performance curve illustrating the backward feature elimination process for Model 2 with feature set 1. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

2.2 Feature set 2

Mean over each 30-minute interval, with a 2-hour observation window.

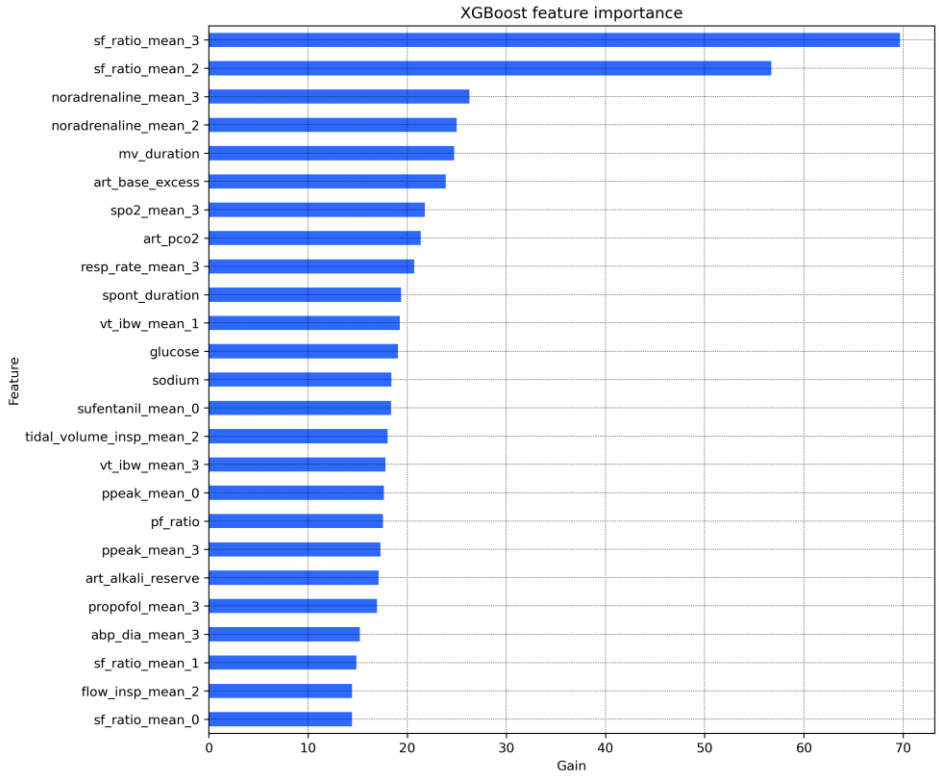


Figure C10. Top 25 feature importance scores, calculated as gain, from XGBoost Model 2 with feature set 2. 0 denotes minutes 0-30, denotes minutes 30-60, 2 denotes minutes 60-90, and 3 denotes minutes 90-120.

2.3 Feature set 3

Mean, std, and trend over 4 hours, with a 4-hour observation window.

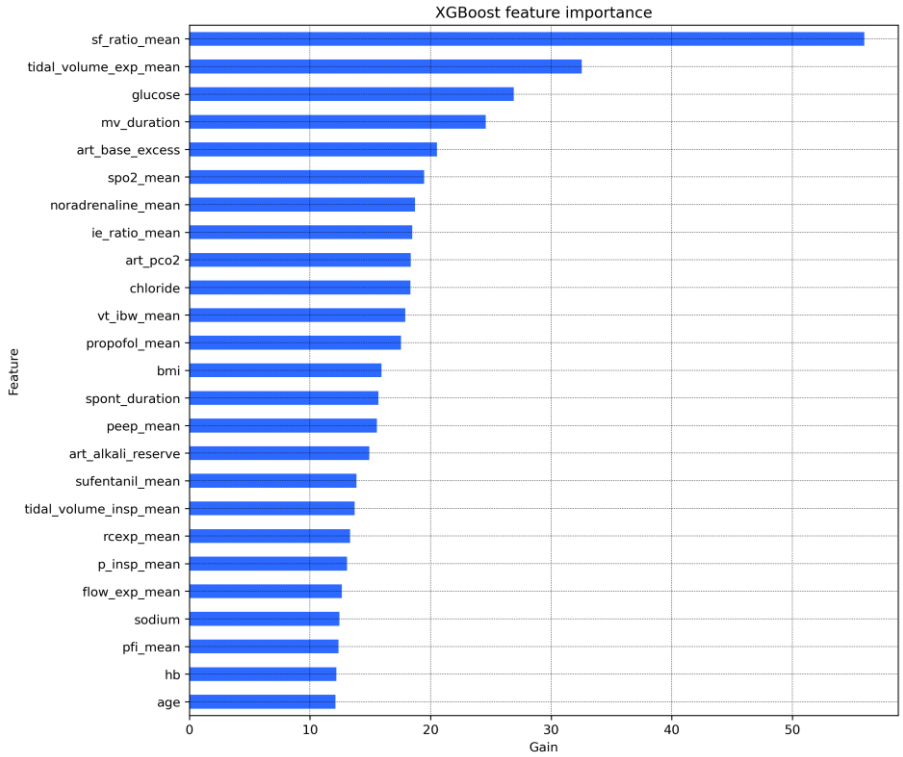


Figure C11. Top 25 feature importance scores, calculated as gain, from XGBoost Model 2 with feature set 3.

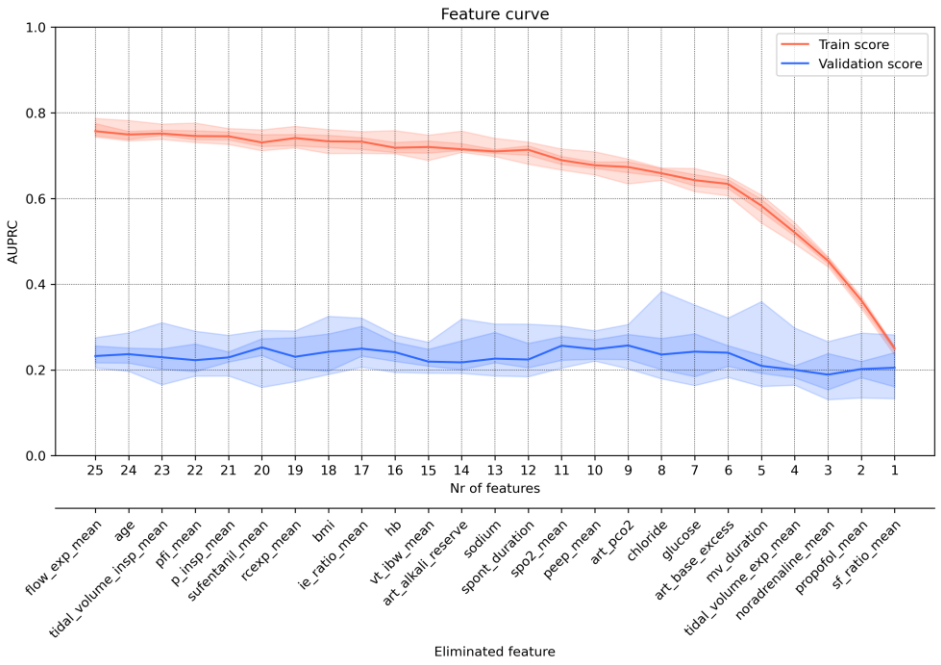


Figure C12. Feature performance curve illustrating the backward process for Model 2 with feature set 3. The 5th 25th, 50th, 75th and 95th percentiles are indicated.

2.4 Feature set 4

Mean over each 1-hour interval, with a 4-hour observation window.

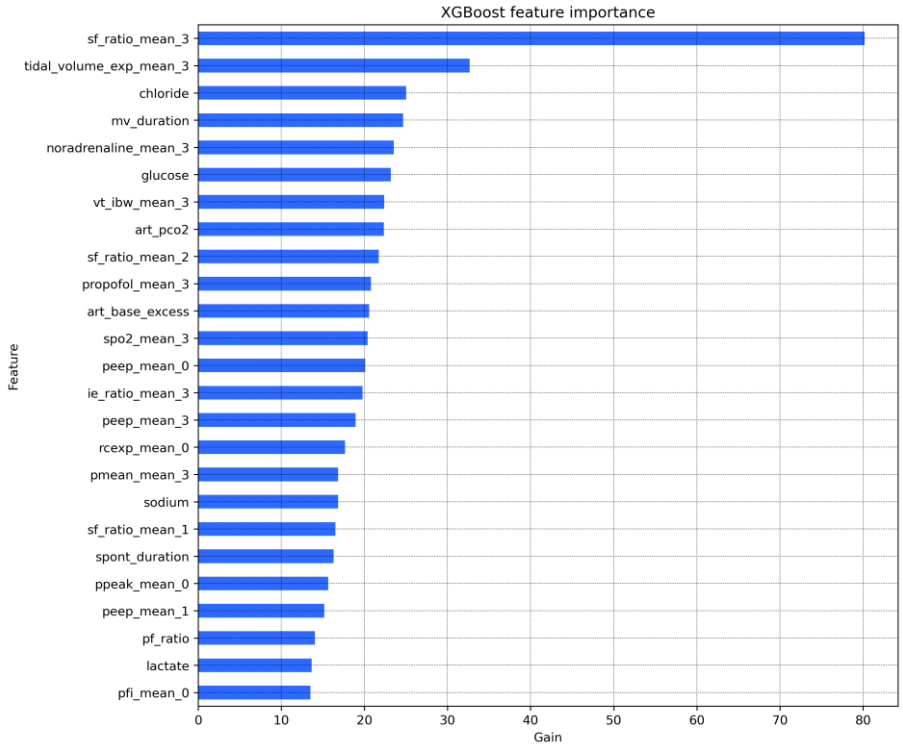


Figure C13. Top 25 feature importance scores, calculated as gain, from XGBoost Model 2 with feature set 4. 0 denotes hour 0-1, 1 denotes hour 1-2, 2 denotes hour 2-3, and 3 denotes hour 3-4.

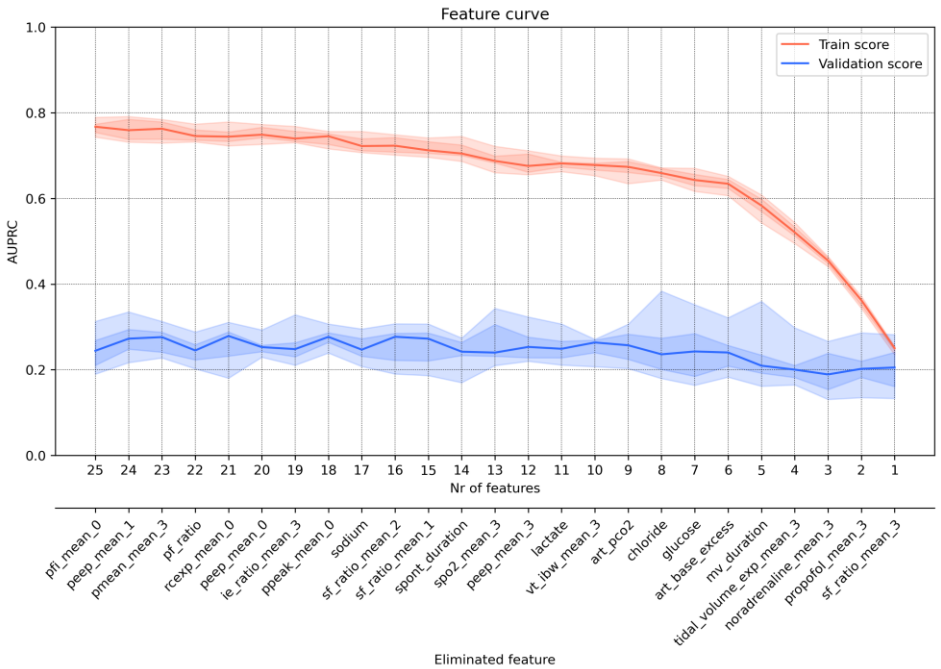


Figure C12. Feature performance curve illustrating the backward feature elimination process for feature set 2 (mean over each 1 hour). 0 denotes hour 0-1, 1 denotes hour 1-2, 2 denotes hour 2-3, and 3 denotes hour 3-4. The 5th 25th, 50th, 75th and 95th percentiles are indicated.

D Hyperparameter optimisation

1 Definitions of XGBoost hyperparameters

Table D1. Overview of the major hyperparameters for the XGBoost model.

Hyperparameter	Definition	Class
Learning rate	Scaling factor for newly added leaf weights after each boosting iteration. Increasing this value reduces the influence of individual trees.	Learning process
Number of estimators	Number boosting iterations, in each iteration one tree is added to the ensemble. The optimal number of estimators depends on the learning rate.	Learning process
Max depth	Maximum depth of a tree. Increasing this value increases model complexity and the risk of overfitting.	Complexity
Minimum child weight	Minimum required sum of instance weights (amount of information) in a leaf node to perform a split. Increasing this value, reduces tree complexity and thereby overfitting.	Complexity
Subsample	Fraction of random samples used for each boosting iteration. Using subsampling prevents overfitting.	Subsampling
Column sample by tree	Fraction of random features used for each boosting iteration. Using feature subsampling prevents overfitting.	Subsampling
Alpha (L1)	Regularisation term which sets leaf weights to zero, resulting in a simpler model. Reduces overfitting.	Regularisation
Lamda (L2)	Regularisation term which decreases leaf weights, resulting in a more stable model. Reduces overfitting.	Regularisation

2 Grid search values

Hyperparameter	Values
max_depth	3, 5, 7
min_child_weight	1, 3, 5
subsample	0.6, 0.8, 1.0
colsample_bytree	0.6, 0.8, 1.0
alpha	0, 0.01, 0.1, 1, 100
lambda	0, 0.01, 0.1, 1, 100
Learning rate	0.3, 0.1, 0.05, 0.01

3 Optimised settings Model 1

Hyperparameter	Value
n_estimators	20
max_depth	3
min_child_weight	1
subsample	1
colsample_bytree	0.6
alpha	0
lambda	1
learning_rate	0.05

4 Optimised settings Model 2

Hyperparameter	Value
n_estimators	20
max_depth	3
min_child_weight	5
subsample	0.6
colsample_bytree	0.6
alpha	0.01
lambda	1
learning_rate	0.1

4 Iterative learning curves Model 1

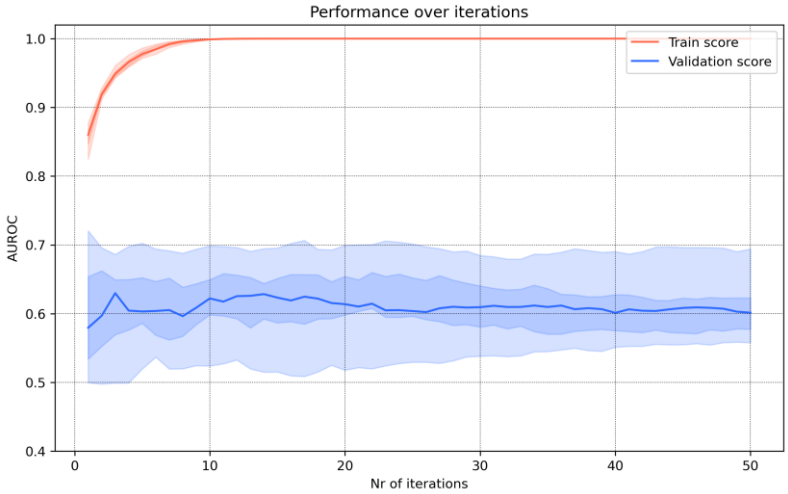


Figure D1. Performance over iterations for Model 1 with prior to feature selection with default hyperparameter settings.

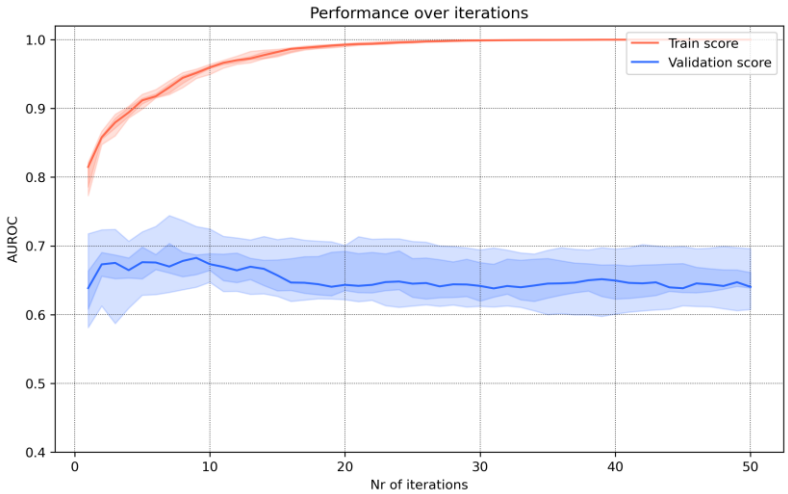


Figure D2. Performance over iteration for Model 1 after feature selection with default hyperparameter settings.

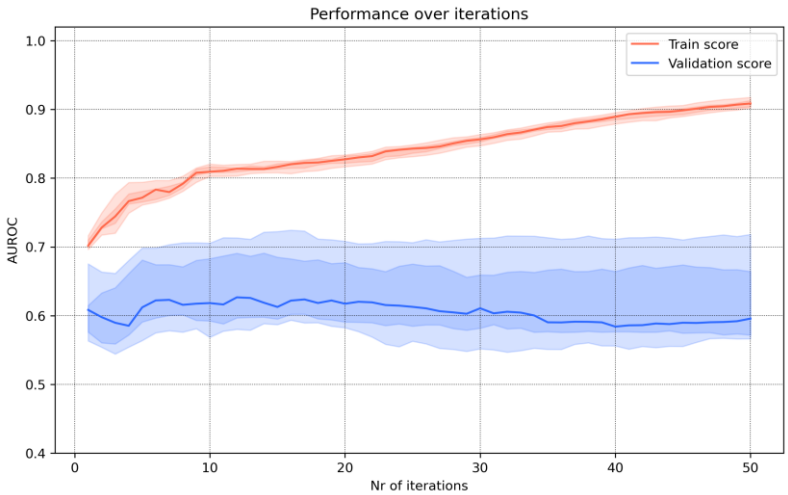


Figure D3. Performance over iterations for Model 1 after after feature selection and hyperparameter optimisation.

5 Iterative learning curves Model 2

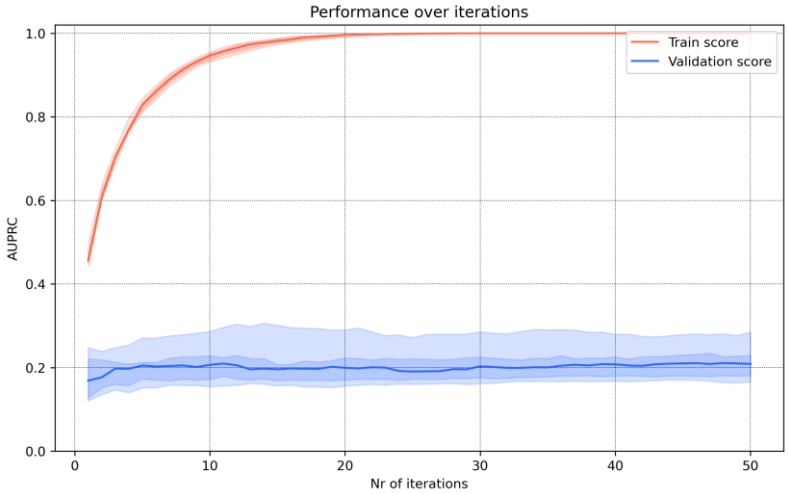


Figure D4. Performance over iterations for Model 2 with prior to feature selection with default hyperparameter settings.

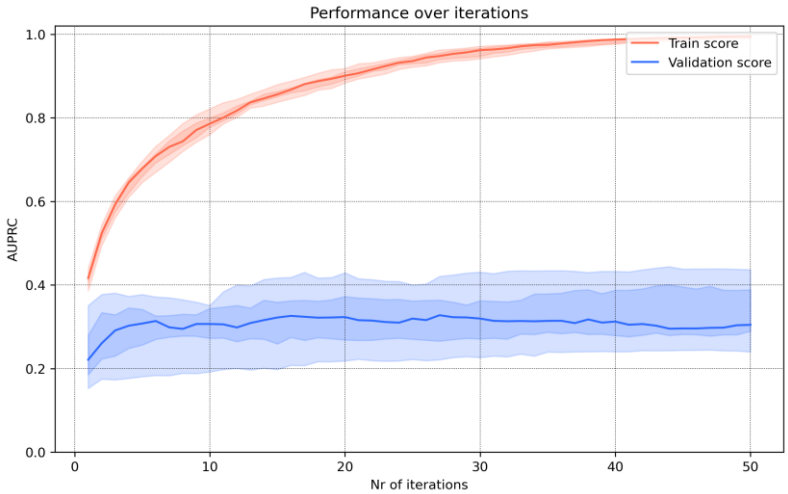


Figure D5. Performance over iteration for Model 2 after feature selection with default hyperparameter settings.

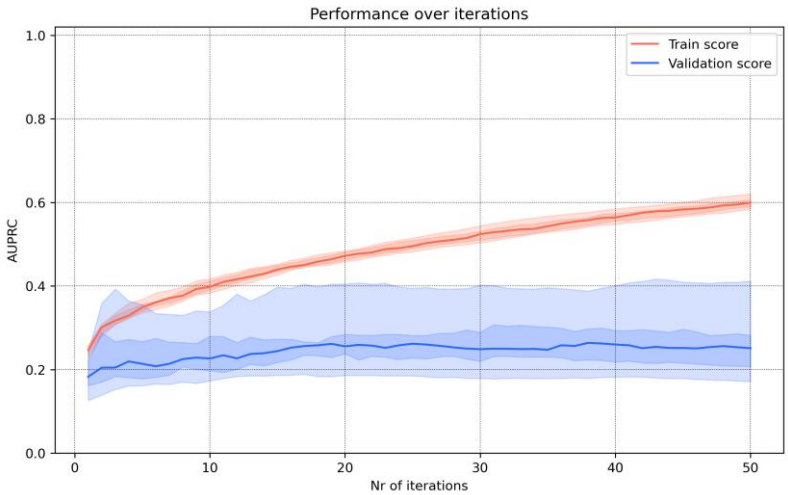


Figure D3. Performance over iterations for Model 2 after after feature selection and hyperparameter optimisation.

E Model calibration

1 Model 1

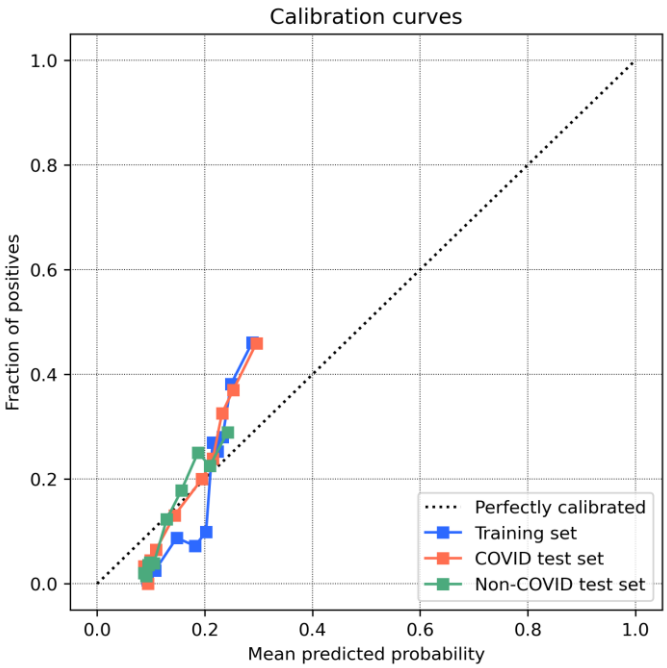


Figure E1. Calibration curves for Model 1 for the COVID training set, COVID test set, and non-COVID test set before calibration.

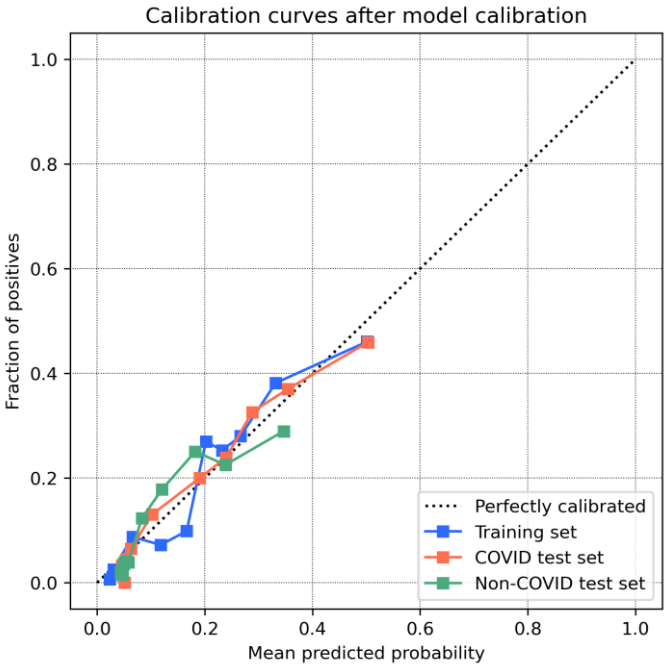


Figure E2. Calibration curves for Model 1 for the COVID training set, COVID test set, and non-COVID test set, after model calibration on each dataset.

2 Model 2

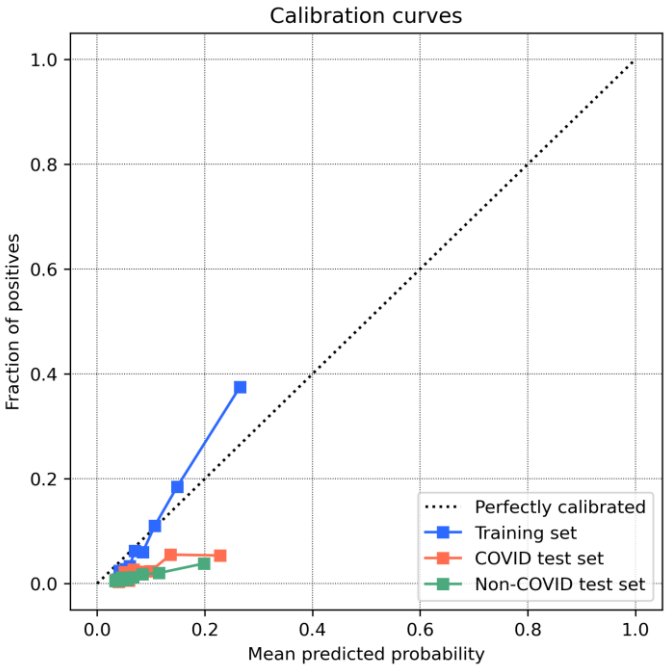


Figure E3. Calibration curves for Model 2 for the COVID training set, COVID test set, and non-COVID test set before calibration.

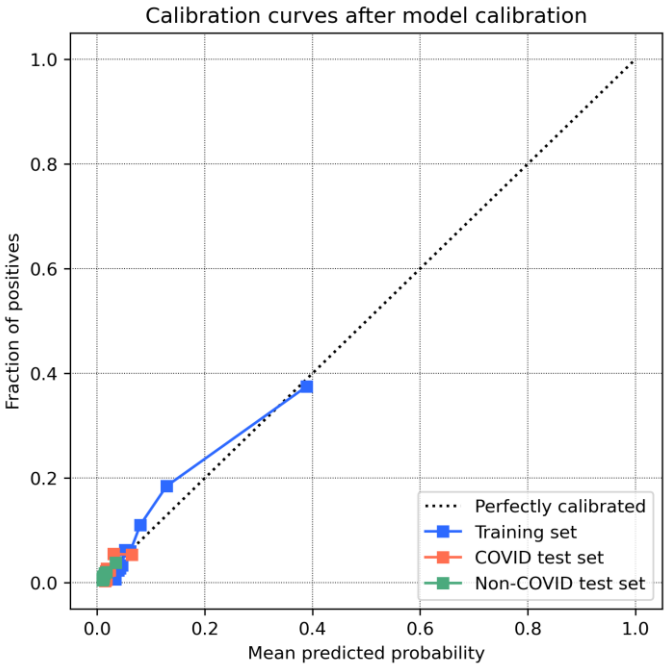


Figure E4. Calibration curves for Model 2 for the COVID training set, COVID test set, and non-COVID test set, after model calibration on each dataset.

F Case descriptions

1 Model 1

1.1 Event sample with a high predicted probability

A 61-year-old female patient with COVID-19 pneumonia was intubated due to hypoxaemia and respiratory exhaustion. After 2.5 hours of controlled mechanical ventilation, she was transitioned to assisted ventilation. Controlled ventilation was reinitiated three hours later (Figure F2). During assisted ventilation preceding the event, the respiratory rate ranged from 20 to 30 breaths per minute, with an rapid shallow breating index (RSBI) of 40–60. The PaO_2/FiO_2 ratio during assisted ventilation was 14.7 kPa and decreased to 13.0 kPa after the event. The FiO_2 around the event was 60 to 70%. The patient was sedated with propofol and sufentanil, with no dose adjustments made around the event.

The combination of a high FiO_2, a low SpO_2/FiO_2 ratio, a high propofol infusion rate, and a low SpO_2 resulted in a relatively high predicted event probability of 0.41 (uncalibrated) (Figure F1).

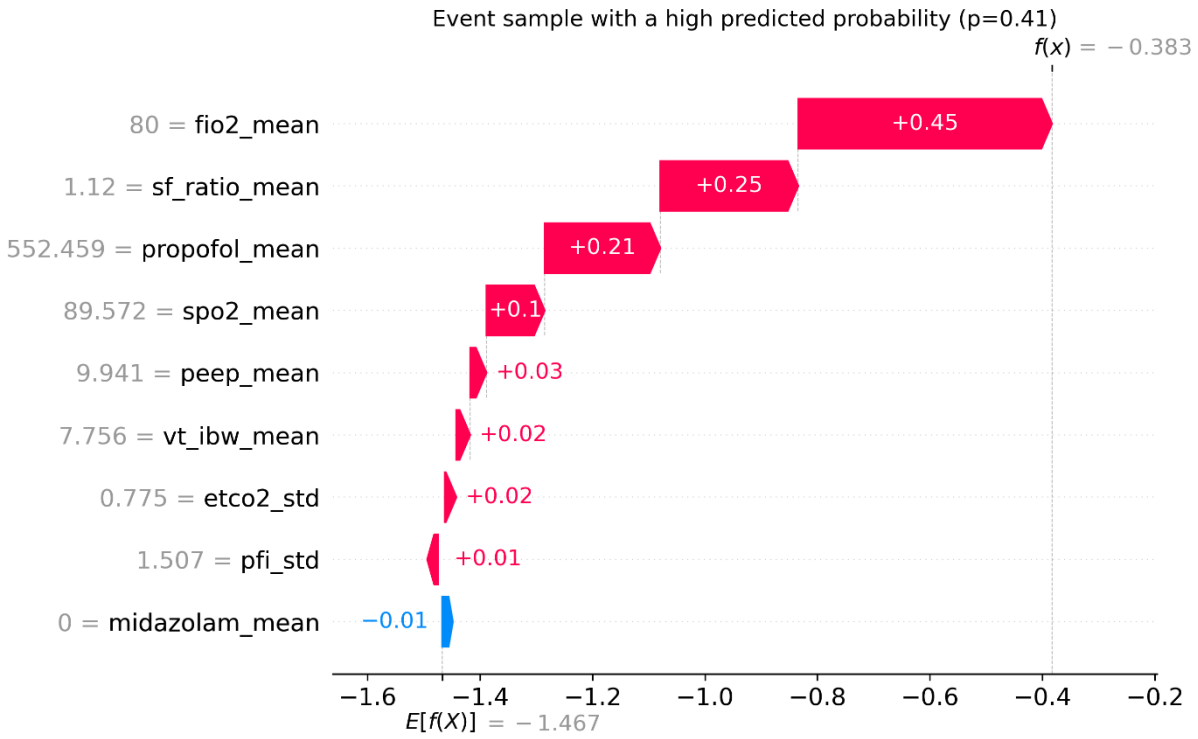


Figure F1. Local SHAP explanation for an event sample with a high predicted probability (p=0.41), showing the feature values and their additive contributions to the model output (expressed as log odds).

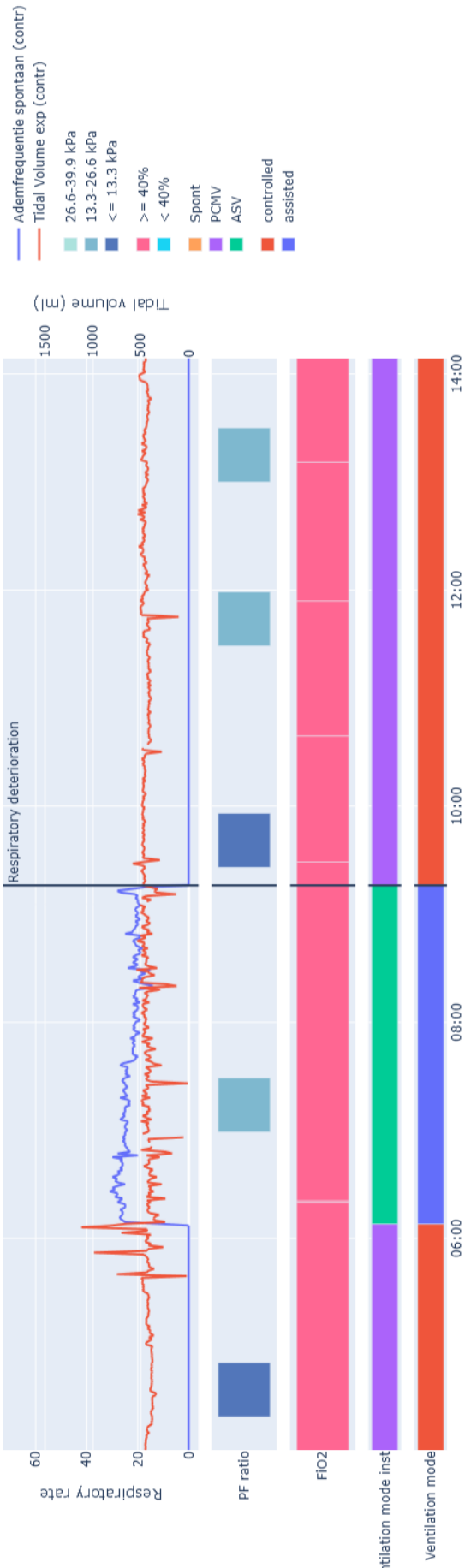


Figure F2. Timeline of respiratory rate, tidal volume, PaO_2/FiO_2 ratio, FiO_2, and ventilation mode surrounding the event, for an event sample with a high predicted probability (p=0.41).

1.2 Event sample with a low predicted probability

A 67-year-old female patient with COVID-19 pneumonia was intubated due to respiratory insufficiency. On day 28 of mechanical ventilation, after four days of continuous assisted ventilation, she was transitioned to controlled ventilation, based on spontaneous respiratory rate, for four hours. She was subsequently switched back to assisted ventilation after which an event occurred five hours later (Figure F4).

During assisted ventilation preceding the event, the respiratory rate ranged from 20 to 30 breaths per minute, with an RSBI of 45-75. The PaO_2/FiO_2 ratio shortly after the event was 22.5 kPa, and the FiO_2 around the event ranged from 40 to 45%. One hour before the event, the clonidine infusion rate was increased. The patient was hypotensive, and the noradrenaline infusion rate was increased one hour after the event. Following this episode, the patient recovered and was successfully weaned from mechanical ventilation within 12 days.

The combination of a relatively low FiO_2, a high SpO_2/FiO_2 ratio, and the absence of propofol infusion resulted in a relatively low predicted event probability of 0.10 (uncalibrated) (Figure F3).

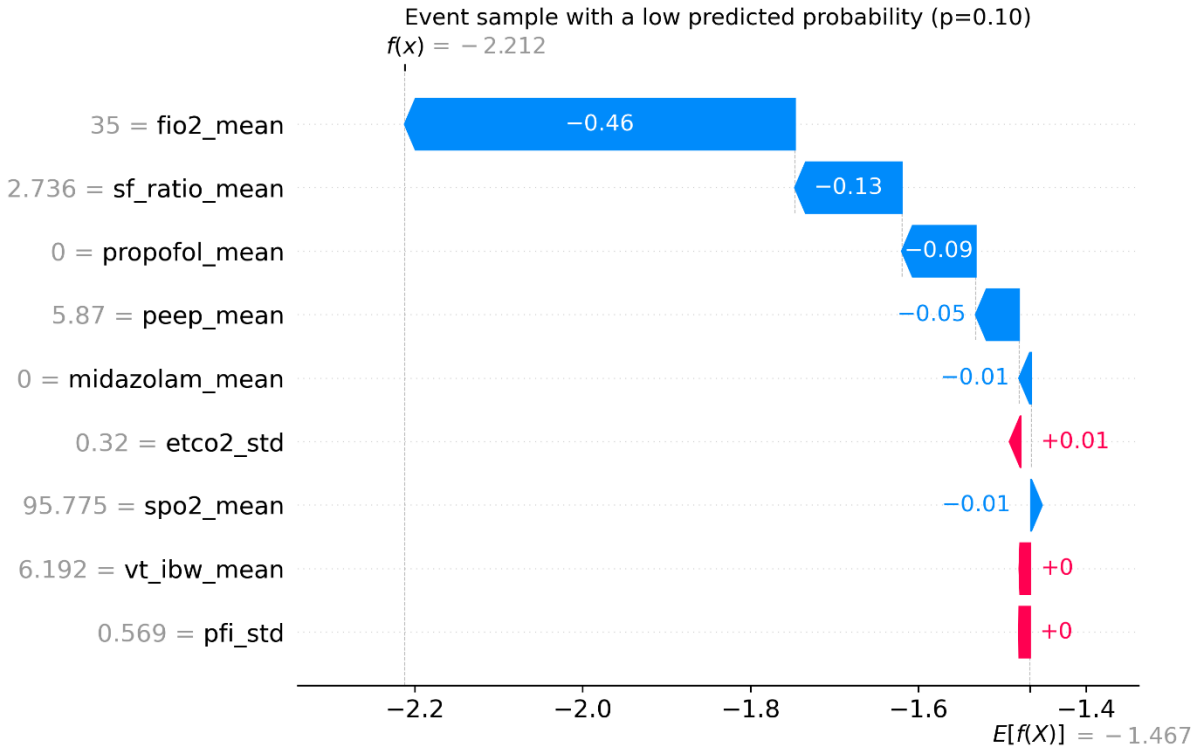


Figure F3. Local SHAP explanation for an event sample with a low predicted probability (p=0.10), showing the feature values and their additive contributions to the model output (expressed as log odds).

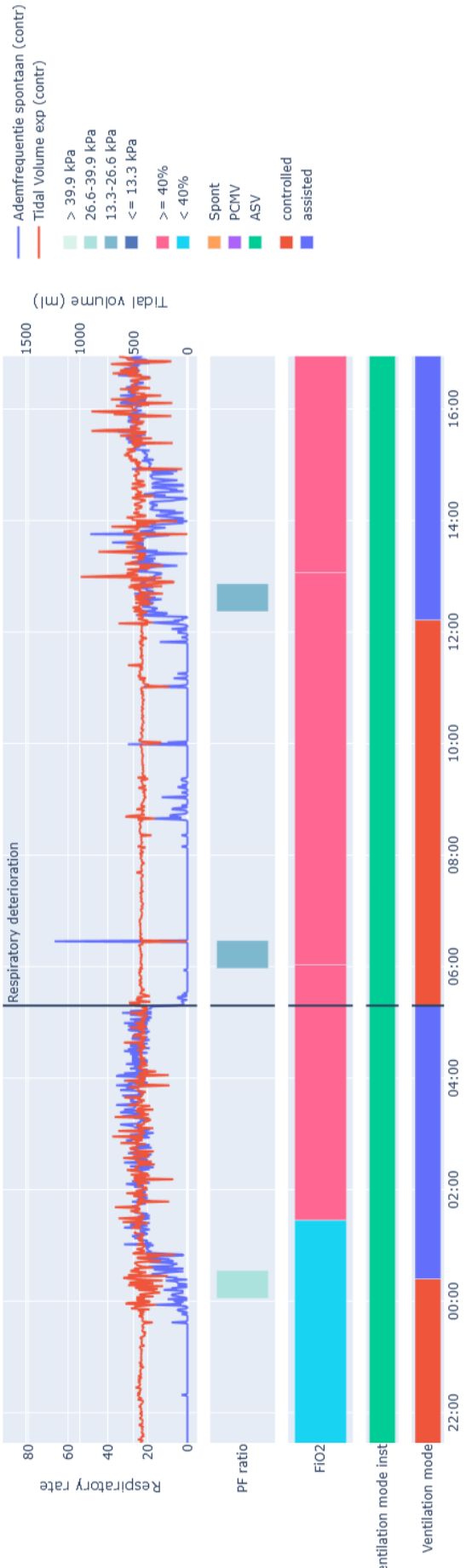


Figure F4. Timeline of respiratory rate, tidal volume, PaO_2/FiO_2 ratio, FiO_2, and ventilation mode surrounding the event, for an event sample with a low predicted probability (p=0.10).

1.3 Control sample with a low predicted probability

A 63-year-old female patient with COVID-19 pneumonia was intubated due to respiratory insufficiency. This control sample was drawn from week 7 of mechanical ventilation during a 2.5-hour period of controlled ventilation, based on spontaneous respiratory rate (Figure F6). At that time, the patient had been on assisted ventilation since one week, alternated with multiple short periods of controlled ventilation similar to this control sample.

PaO_2/FiO_2 ratio was approximately 30 kPa, and the FiO2 remained stable at 35%. Following this time point, periods of controlled ventilation became less frequent, and the patient was successfully weaned from mechanical ventilation within 12 days.

The combination of a relatively low FiO_2, a high SpO_2/FiO_2 ratio, and a low tidal volume per kg ideal body weight in a relatively low predicted event probability of 0.08 (uncalibrated) (Figure F5).

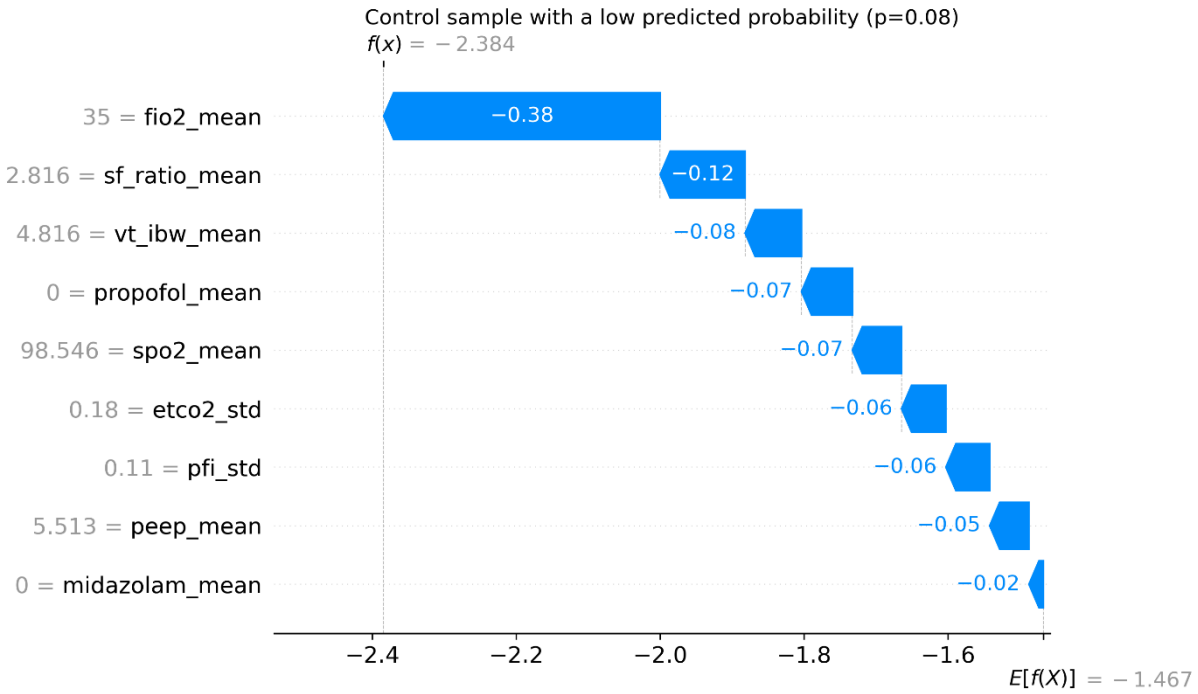


Figure F5. Local SHAP explanation for a control sample with a low predicted probability (p=0.08), showing the feature values and their additive contributions to the model output (expressed as log odds).

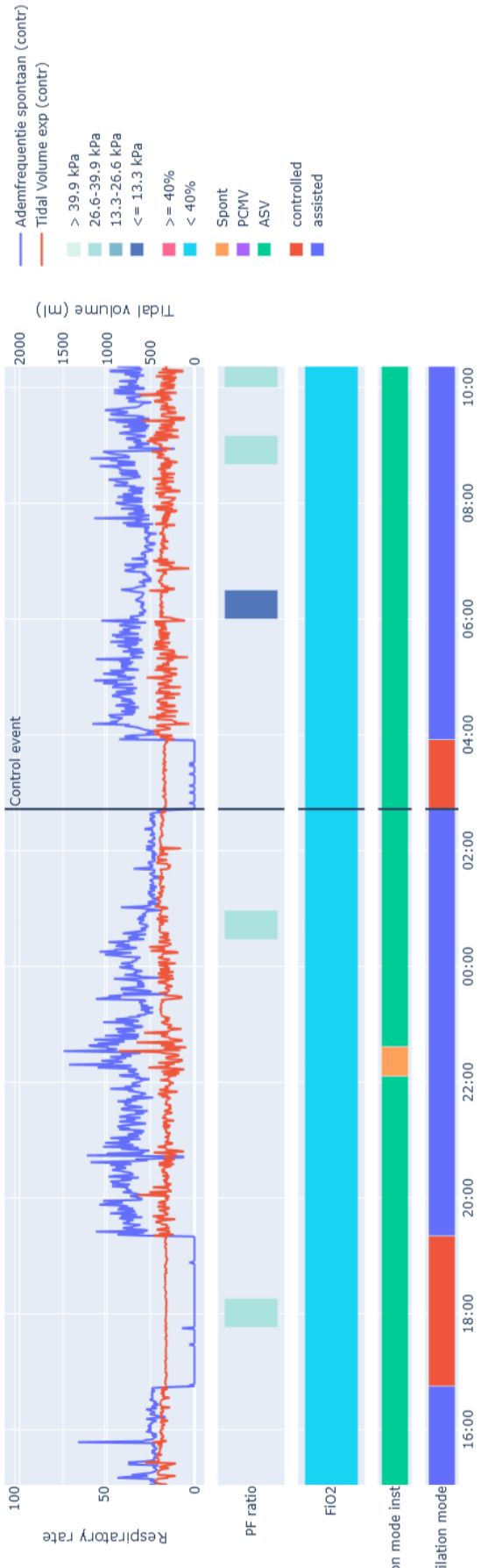


Figure F6. Timeline of respiratory rate, tidal volume, PaO_2/FiO_2 ratio, FiO_2, and ventilation mode surrounding the event, for a control sample with a low predicted probability (p=0.08).

1.4 Control sample with a high predicted probability

A 70-year-old female patient with COVID-19 pneumonia was intubated due to respiratory insufficiency. She was switched to assisted ventilation on day 10 of mechanical ventilation and did not return to controlled ventilation for longer than three hours thereafter, but died 30 hours later. The PaO_2/FiO_2 ratio remained consistently around 10 kPa, with the FiO_2 at 80% or higher throughout this period (Figure F8). The patient was ventilated in prone position; however, ventilation and oxygenation failed to improve, and treatment was subsequently withdrawn.

The combination of a high FiO_2, a low SpO_2/FiO_2 ratio, a high propofol infusion rate, and low SpO_2 resulted in a relatively high predicted event probability of 0.40 (uncalibrated) (Figure F7).

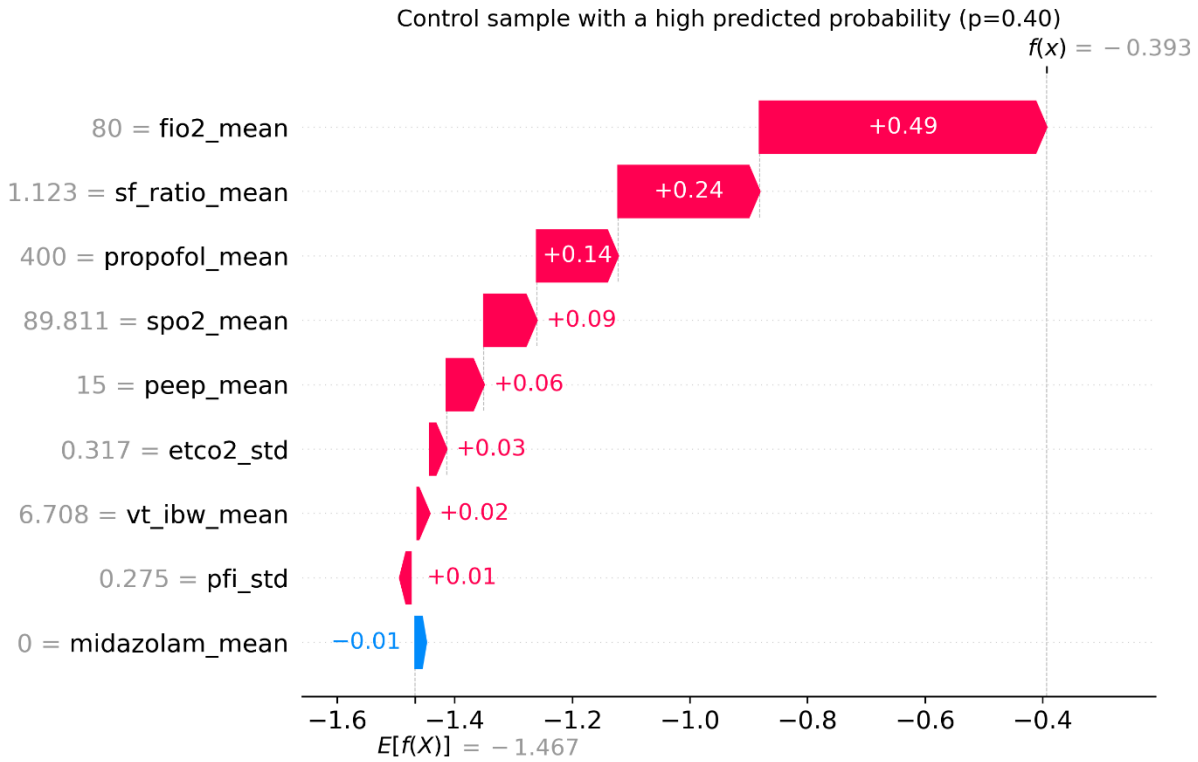


Figure F7. Local SHAP explanation for a control sample with a high predicted probability (p=0.40), showing the feature values and their additive contributions to the model output (expressed as log odds).

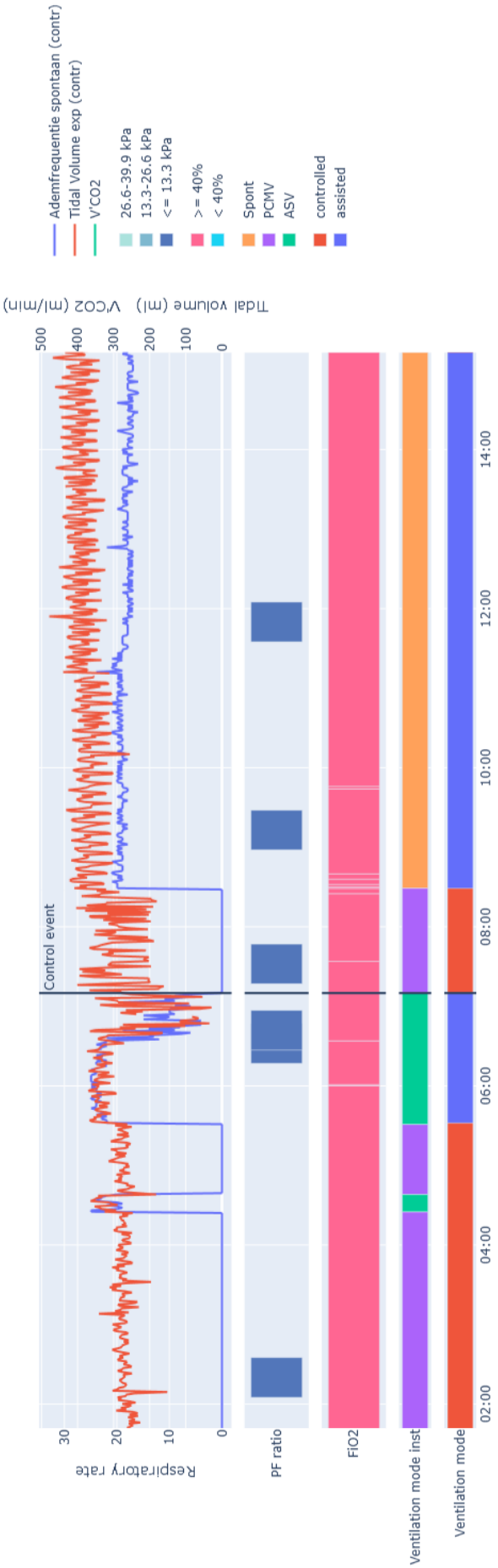


Figure F8. Timeline of respiratory rate, tidal volume, PaO_2/FiO_2 ratio, FiO_2, and ventilation mode surrounding the control-event, for a control sample with a high predicted probability (p=0.40).

2 Model 2

2.1 Event sample with a high predicted probability

Figure F9 shows an event sample characterised by relatively low SpO_2/FiO_2 ratios, a short total IMV duration, and a low tidal volume per kg ideal body weight, resulting in a relatively high predicted probability of 0.60.

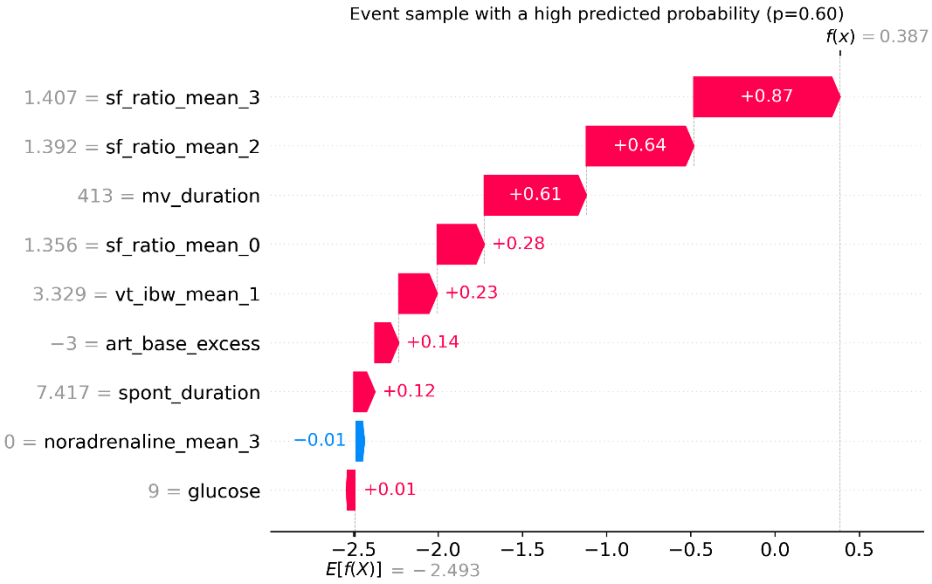


Figure F9. Local SHAP explanation for an event sample with a high predicted probability (p=0.60), showing the feature values and their additive contributions to the model output (expressed as log odds).

2.2 Event sample with a low predicted probability

Figure F10 shows an event sample with high SpO_2/FiO_2 ratios and a relatively long duration of assisted ventilation, which resulted in a relatively low predicted probability of 0.03.

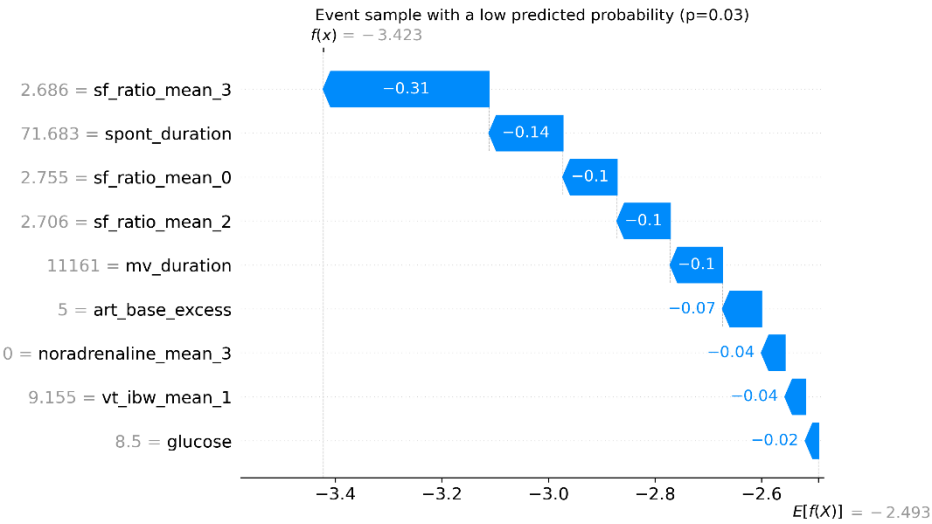


Figure F10. Local SHAP explanation for an event sample with a low predicted probability (p=0.03), showing the feature values and their additive contributions to the model output (expressed as log odds).

2.3 Control sample with a low predicted probability

Figure F11 shows a control sample with high SpO_2/FiO_2 ratios and a long duration of assisted ventilation, which resulted in a relatively low predicted probability of 0.03.

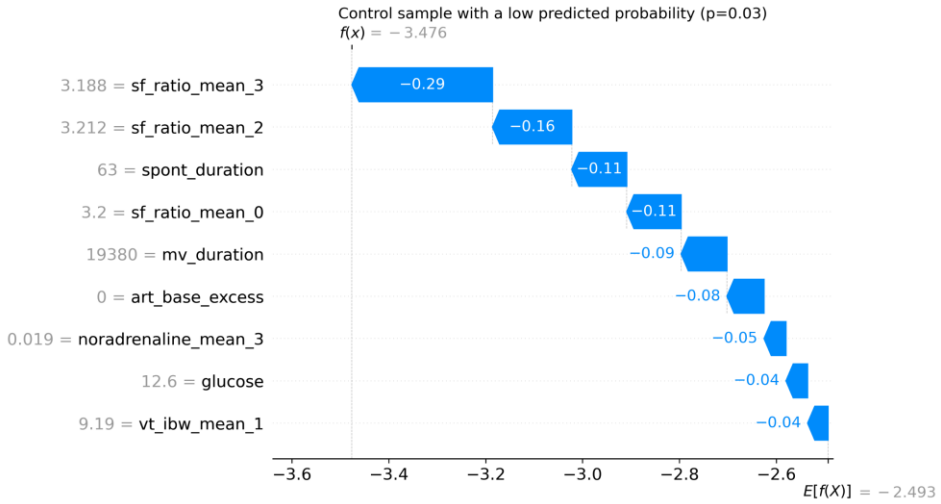


Figure F11. Local SHAP explanation for a control sample with a low predicted probability (p=0.03), showing the feature values and their additive contributions to the model output (expressed as log odds).

2.4 Control sample with a high predicted probability

Figure F12 shows a control sample characterised by relatively low SpO_2/FiO_2 ratios, a short total IMV duration, resulting in a relatively high predicted probability of 0.44, despite a relatively high tidal volume per kg ideal body weight.

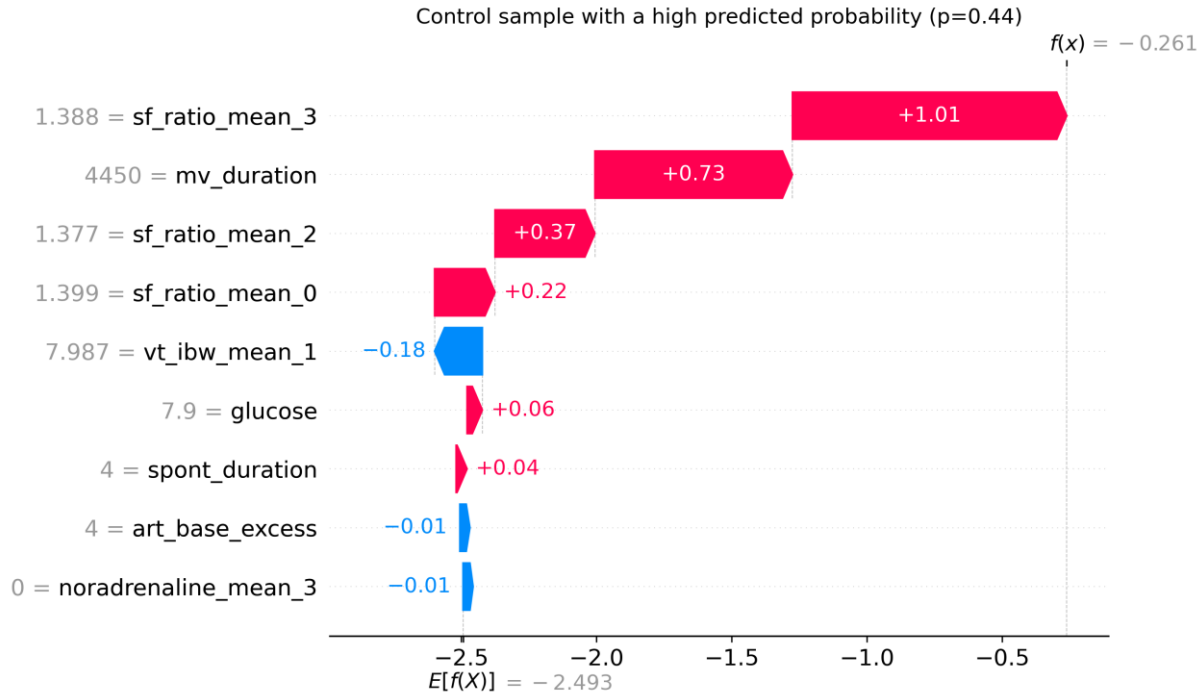


Figure F12. Local SHAP explanation for a control sample with a high predicted probability (p=0.44), showing the feature values and their additive contributions to the model output (expressed as log odds).

G Decision tree visualisations

1 Model 1

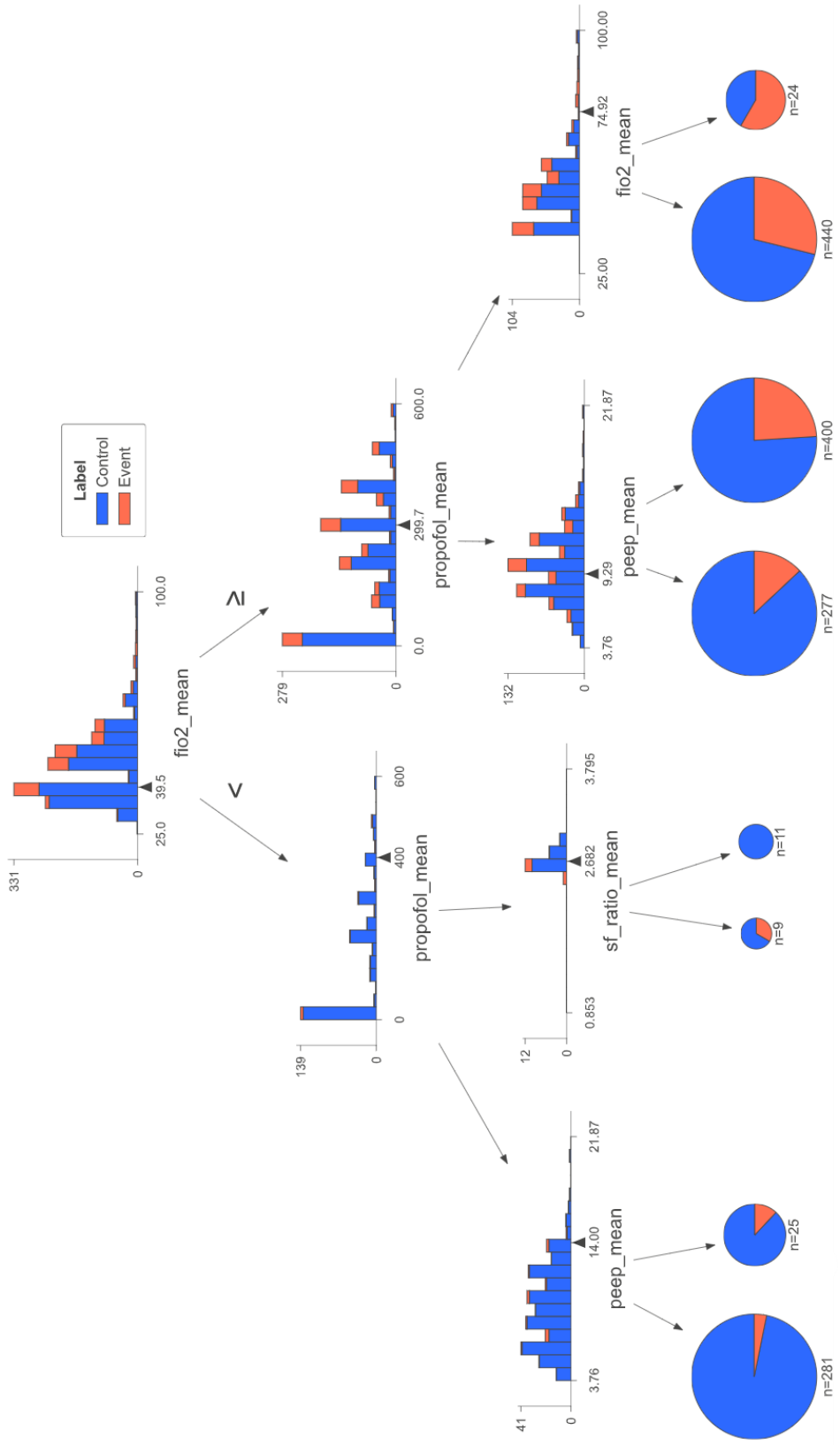


Figure G1. Visualisation of the first decision tree from the XGBoost ensemble Model 1.

2 Model 2

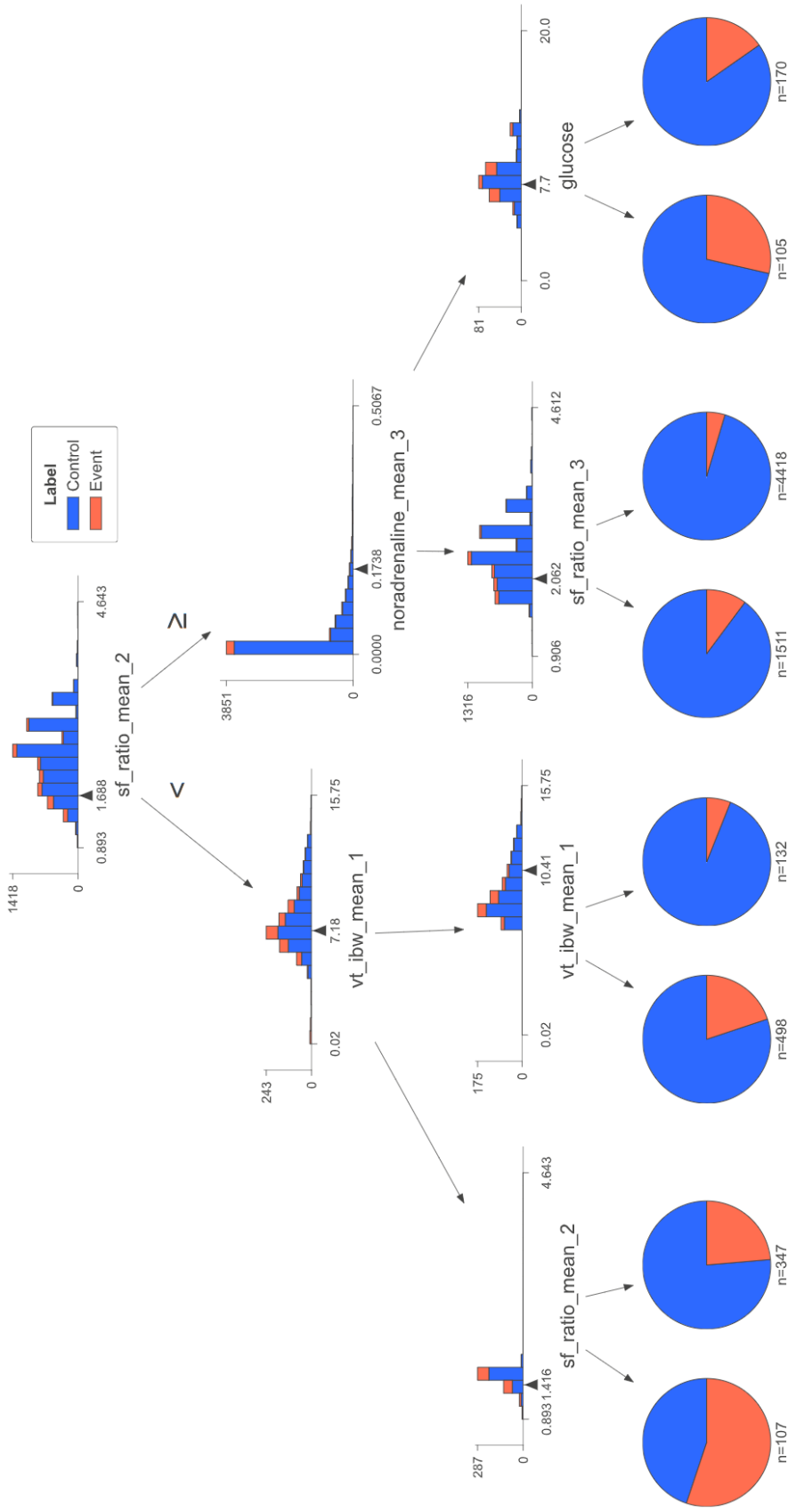


Figure G2. Visualisation of the first decision tree from the XGBoost ensemble Model 2.

H Model 1B

1 Model design

A variation on Model 1 was developed, with model input from one hour before and one hour after the switch from controlled to assisted ventilation (Figure G1).

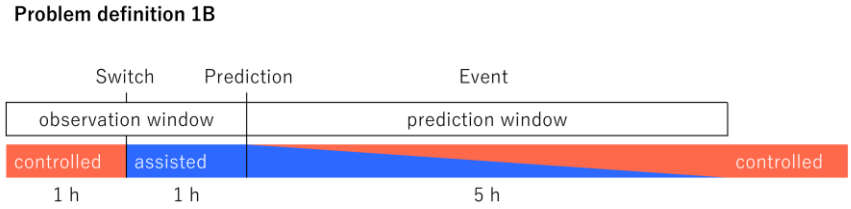


Figure G1. Overview of the observation window and prediction window aligned with ventilation modes and moment of switching from controlled to assisted ventilation.

2 Model development

Four different types of input features were used for ventilatory and haemodynamic parameters:

- The mean over the last hour before the switch X_0
- The mean over the first hour after the switch X_1
- The absolute change: X_1-X_0
- The relative change: $\frac{X_1-X_0}{X_0}$

Mean feature importance was determined via 10-fold cross-validation and forward feature selection in descending order of importance was employed (Figure 2). Eight features were selected, resulting in a median AUROC of 0.70 (IQR 0.63-0.75).

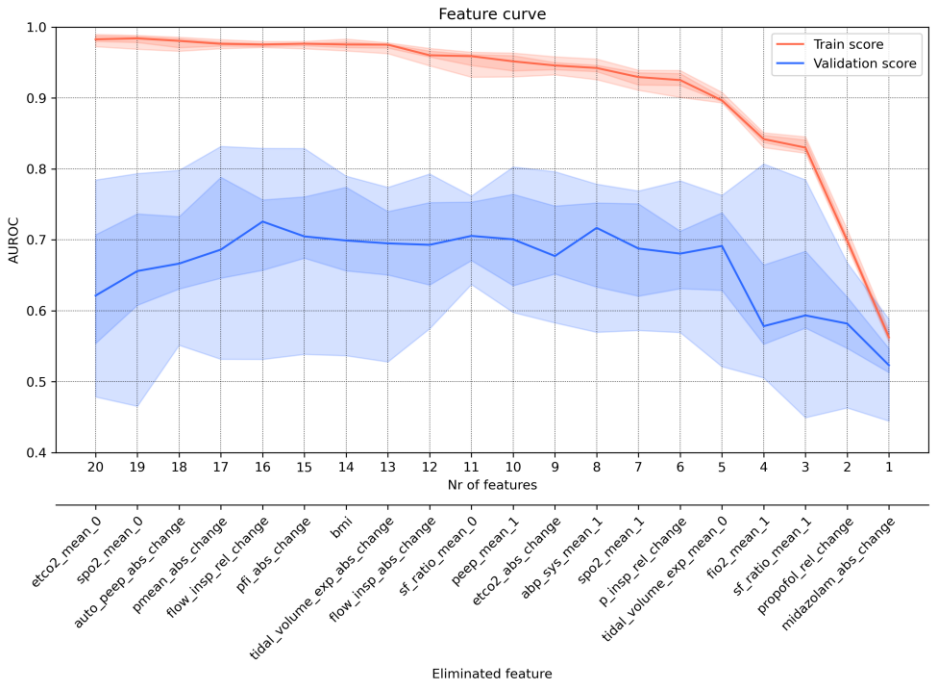


Figure G2. Feature performance curve illustrating the backward feature elimination process. 0 denotes the hour before the switch and 1 denotes the hour after the switch. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

After hyperparameter optimisation the AUROC yielded 0.72 (IQR 0.63-0.79). Learning curves obtained after feature selection and hyperparameter optimisation are presented in Figures G3 and G4.

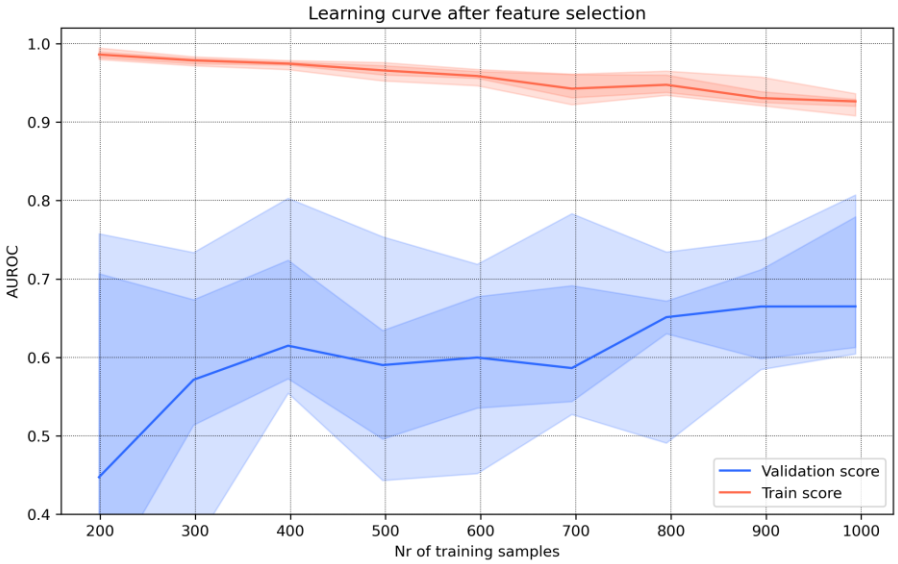


Figure G3. Learning curve, illustrating the performance over the number of training samples, obtained after feature selection. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

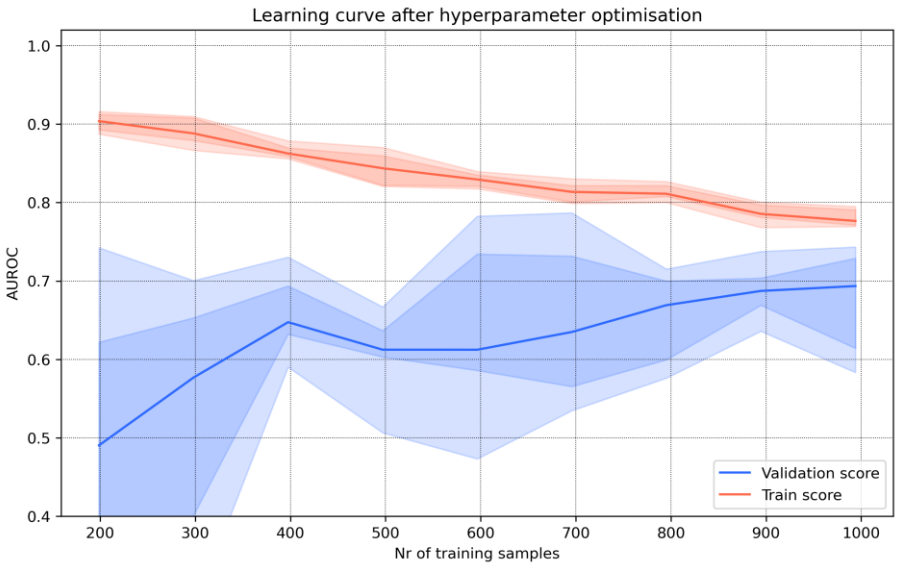


Figure G4. Learning curve, illustrating the performance over the number of training samples, obtained after feature selection and hyperparameter optimisation. The 5th, 25th, 50th, 75th and 95th percentiles are indicated.

3 Model validation

Validation on the COVID and non-COVID test sets yielded an AUROC of 0.70 and 0.58 respectively (Figure G5).

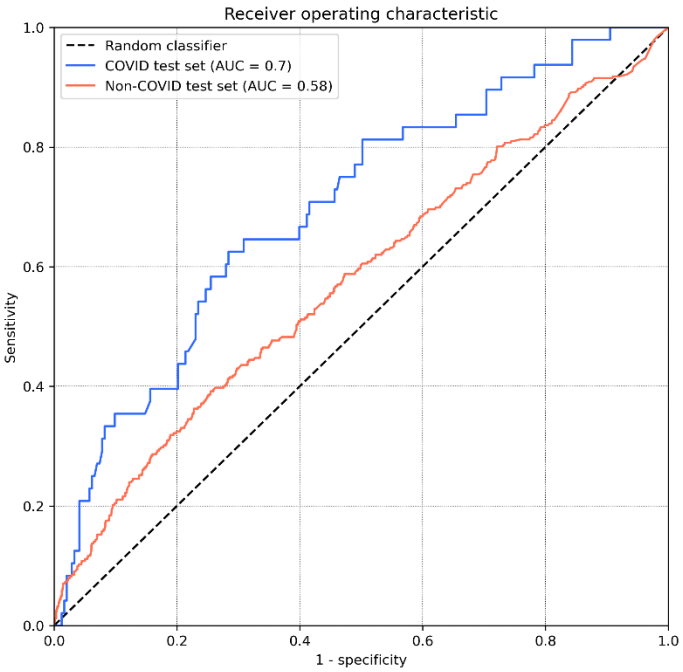


Figure G5. Receiver operating characteristic of validation on the COVID test dataset (AUC = 0.70) and non-COVID test dataset (AUC = 0.58).

4 Model interpretation

A global SHAP analysis is presented in Figure G6. This analysis indicates the relative change in propofol and the mean SpO_2 after the switch influence model output most strongly.

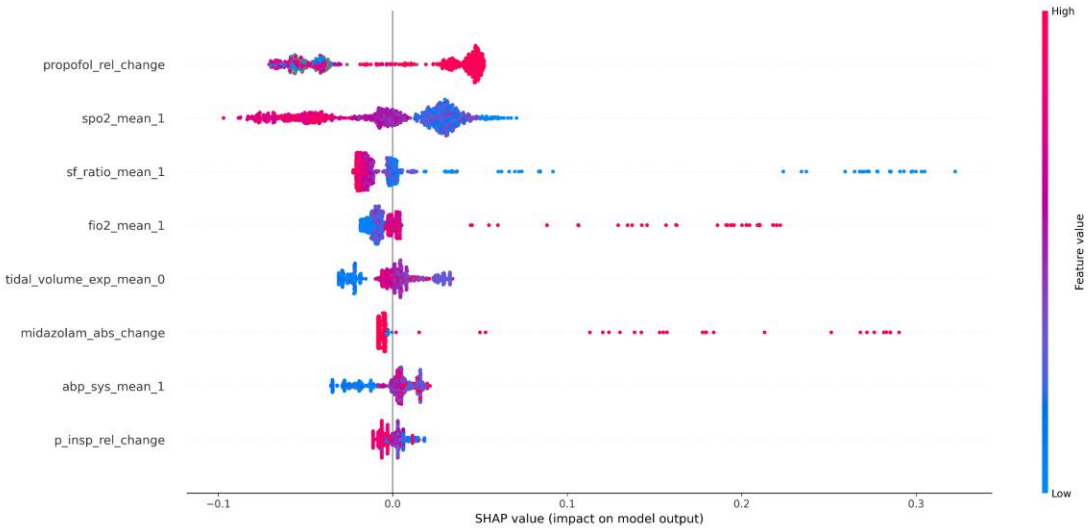


Figure G6. Global SHAP analysis showing the impact of feature values on the model output by aggregating local SHAP values across all training samples. 0 denotes the hour before the switch and 1 denotes the hour after the switch.

I Switch conditions

1 Methods

To gain insight into current practices regarding the first switch from controlled to assisted ventilation at the ICU of the LUMC, conditions during the hour preceding the first switch were analysed. All patients included in this study (COVID and non-COVID patients with ≥ 48 h of invasive mechanical ventilation and a PaO_2/FiO_2 ratio < 40 kPa) were included if invasive mechanical ventilation was started in a controlled ventilation mode.

Time to the first switch, mean haemodynamic and ventilatory parameters over the final hour before the switch, and the most recent arterial blood gas values were reported as median (IQR). Comparison were made between patients who experienced an event of respiratory deterioration (as defined in *Section 4.4)* within six hours (failure) and those who did not (success), as well as for patients who experienced an events within 72 hours and those who did not. Differences between the success and failure groups were assessed using the Mann-Whitney U test with Holm-Bonferroni correction for multiple testing, applying a two-sided significance level of 0.05.

2 Results

In total, first switch attempts in 746 patients were analysed, of whom 121 (16.2%) experienced failure within six hours and 148 (19.8%) within 72 hours (Tables I1 and I2). The median time from initiation of mechanical ventilation to the first switch to assisted ventilation was 47.9 hours (IQR 19.0-87.9). Time to switch was longer in the success groups (48.2 and 48.5 hours) than in the failure groups (43.0 and 42.9 hours), although these differences were not statistical significant.

Small, but significant differences between the success and failure groups were observed for the FiO_2, P_{peak}, P_{mean}, PEEP, SpO_2/FiO_2 ratio, PaO_2/FiO_2 ratio, and arterial O_2 saturation, for both six-hour and 72-hour comparisons. In addition, minute volume and inspiratory flow were significantly lower for the success 72-h groups. A complete overview of all analysed parameters is provided Tables I1 and I2.

3 Interpretation

The timing of the first switch to assisted ventilation varied substantially between patients. The median time from the start of mechanical ventilation to the first switch was 47.9 hours, which is relatively late compared with the findings of Smit et al. (1.3-1.8 days) (6) and Haudebourg et al. (9 hours) (8), but earlier than reported by Pérez et al. (3 days) (5). In addition, Smit et al. observed that failed switch attempts occurred earlier than successful ones, which is consistent with the trend observed in our analysis.

A failure rate of 19.8% within 72 hours is remarkably low compared with previously reported rates of 30%-67% (5–8), three of which also included COVID-19 patients. A plausible explanation is the stricter event definition applied in the present analysis. Transitions back to controlled ventilation lasting less than three hours, or occurring with with an FiO_2 below 40% were not classified as failure, which likely resulted in lower failure rates compared with similar studies. Furthermore, the seemingly different timing of switches may contribute to this low failure rate.

Differences in pre-switch parameters between the success and failure groups were observed for FiO_2, P_{peak}, P_{mean}, PEEP, SpO_2/FiO_2 ratio, PaO_2/FiO_2 ratio, and arterial O_2 saturation. This is consistent with previous studies reporting associations between switch failure and FiO_2, PaO_2/FiO_2 ratio, and ventilatory pressures (6–8). Other studies have also reported small differences gas exchange parameters (pH, PaO_2, $PaCO_2$, base excess, lactic acid) which were not observed in the present analysis (6–8).

Table I1. Overview of clinical parameters in one hour before the first switch attempt to assisted ventilation. P values indicate the difference between the success group (no event within six hours) and failure group (event within six hours).

	Overall			Success (> 6h)			Failure (≤ 6h)			Succes vs. Failure	
n (%)	746			625 (83.8)			121 (16.2)				
	Median	**(IQR)**		**Median**	**(IQR)**		**Median**	**(IQR)**		**p**	**corrected p**
Time to switch (h)	47.9	(19.0	87.9)	48.2	(19.1	89.7)	43.0	(18.5	84.4)	0.575	1.000
Heart rate (/min)	82.5	(70.2	96.8)	83.4	(70.6	97.1)	76.9	(69.2	90.3)	0.018	0.366
MAP (mmHg)	77.0	(72.0	84.7)	76.9	(71.8	84.8)	77.8	(72.8	83.5)	0.556	1.000
Noradrenaline (µg/kg/min)	0.05	(0.00	0.14)	0.05	(0.00	0.15)	0.04	(0.01	0.10)	0.114	1.000
FiO_2 (%)	40.2	(34.7	49.9)	40.0	(32.5	49.3)	45.3	(40.0	51.0)	<0.001	<0.001*
SpO_2 (%)	95.6	(93.8	97.3)	95.6	(93.9	97.4)	94.8	(93.1	96.9)	0.017	0.366
Respiratory rate (/min)	17.6	(15.2	20.9)	17.5	(15.2	20.6)	18.8	(15.1	22.2)	0.049	0.923
P_{insp} (cmH_2O)	12.8	(11.0	15.0)	12.8	(11.1	14.9)	12.8	(10.1	15.1)	0.896	1.000
ΔP (cmH_2O)	10.2	(8.4	12.1)	10.1	(8.4	12.0)	10.3	(8.4	12.6)	0.553	1.000
P_{plat} (cmH_2O)	20.4	(17.0	24.0)	20.0	(16.9	24.0)	22.0	(18.8	24.7)	0.014	0.309
P_{peak} (cmH_2O)	22.2	(18.7	25.4)	21.9	(18.4	25.1)	23.5	(20.7	26.3)	<0.001	0.014*
P_{mean} (cmH_2O)	13.2	(10.7	15.8)	13.1	(10.3	15.5)	14.9	(12.0	17.0)	<0.001	0.001*
PEEP (cmH_2O)	8.7	(6.3	11.7)	8.3	(6.0	11.0)	10.0	(7.9	12.2)	<0.001	0.002*
Auto PEEP (cmH_2O)	0.6	(0.3	1.2)	0.6	(0.3	1.2)	0.7	(0.3	1.3)	0.202	1.000
V_T (mL)	486	(412	561)	487	(413	557)	473	(411	592)	0.797	1.000
V_T/IBW (mL/kg)	7.0	(6.2	7.7)	7.0	(6.2	7.7)	7.1	(6.3	7.9)	0.580	1.000
Minute ventilation (L/min)	8.5	(7.3	9.8)	8.5	(7.3	9.7)	8.9	(7.9	10.1)	0.013	0.297
Inspiratory flow (mL/s)	40.8	(35.5	46.0)	40.6	(35.2	45.5)	41.8	(37.7	48.4)	0.003	0.085
Expiratory flow (mL/s)	38.7	(34.0	43.4)	38.5	(33.8	43.2)	39.4	(35.4	43.9)	0.101	1.000
Compliance (mL/cmH_2O)	53.9	(40.1	71.8)	54.0	(40.3	72.0)	53.5	(38.1	69.8)	0.816	1.000
R_{insp} (cmH_2O)	11.6	(9.0	14.4)	11.8	(9.2	14.5)	10.6	(8.3	12.9)	0.005	0.132
$ETCO_2$ (kPa)	5.0	(4.5	5.7)	5.1	(4.5	5.7)	5.0	(4.4	5.6)	0.318	1.000
V'CO2 (mL/min)	182.0	(151.8	219.0)	180.4	(151.4	216.3)	186.7	(157.3	233.0)	0.316	1.000
SpO_2/FiO_2 ratio	2.4	(2.0	2.8)	2.4	(2.0	3.0)	2.1	(1.8	2.4)	<0.001	<0.001*
PaO_2/FiO_2 ratio	25.4	(19.8	33.7)	26.5	(20.8	35.1)	21.8	(17.6	26.8)	<0.001	<0.001*
Art. PH	7.40	(7.35	7.44)	7.40	(7.35	7.44)	7.39	(7.35	7.44)	0.804	1.000
$PaCO_2$ (kPa)	5.6	(5.0	6.3)	5.6	(5.0	6.2)	5.9	(5.2	6.8)	0.003	0.083
PaO_2 (kPa)	10.2	(9.1	11.9)	10.3	(9.2	12.0)	10.0	(8.8	11.3)	0.006	0.149
Art. O_2 saturation (%)	95.0	(93.0	96.0)	95.0	(93.8	96.0)	94.0	(93.0	96.0)	<0.001	0.010*
Art. Alkali Reserve (mmol/L)	25.0	(22.0	30.0)	25.0	(22.0	29.0)	27.0	(22.0	31.5)	0.055	0.997
Art. Base Excess (mmol/L)	0.0	(-3.0	5.0)	0.0	(-3.0	4.0)	2.0	(-3.0	6.0)	0.146	1.000
Lactate (mmol/L)	1.5	(1.2	2.0)	1.5	(1.2	2.0)	1.5	(1.2	1.8)	0.433	1.000

* statistically significant difference between groups ($p < 0.05$) after Holm-Bonferroni correction

Table I2. Overview of clinical parameters in one hour before the first switch attempt to assisted ventilation. P values indicate the difference between the success group (no event within 72 hours) and failure group (event within 72 hours).

	Overall			Success (> 72h)			Failure (≤ 72h)			Succes vs. Failure	
n (%)	746			598 (80.2)			148 (19.8)				
	Median	**(IQR)**		**Median**	**(IQR)**		**Median**	**(IQR)**		**p**	**corrected p**
Time to switch (h)	47.9	(19.0	87.9)	48.5	(19.3	89.5)	42.9	(17.6	84.5)	0.419	1.000
Heart rate (/min)	82.5	(70.2	96.8)	83.3	(70.2	97.2)	79.8	(70.4	91.1)	0.087	1.000
MAP (mmHg)	77.0	(72.0	84.7)	76.9	(71.9	84.7)	77.7	(72.3	83.6)	0.658	1.000
Noradrenaline (µg/kg/min)	0.05	(0.00	0.14)	0.05	(0.00	0.15)	0.04	(0.01	0.12)	0.471	1.000
FiO_2 (%)	40.2	(34.7	49.9)	40.0	(32.3	49.2)	45.0	(40.0	50.3)	<0.001	<0.001*
SpO_2 (%)	95.6	(93.8	97.3)	95.7	(93.9	97.4)	95.0	(93.1	96.9)	0.015	0.335
Respiratory rate (/min)	17.6	(15.2	20.9)	17.3	(15.1	20.4)	19.2	(15.4	22.2)	0.005	0.125
P_{insp} (cmH_2O)	12.8	(11.0	15.0)	12.8	(11.0	14.8)	13.0	(10.2	15.4)	0.553	1.000
ΔP (cmH_2O)	10.2	(8.4	12.1)	10.1	(8.3	12.0)	10.5	(8.4	12.6)	0.342	1.000
P_{plat} (cmH_2O)	20.4	(17.0	24.0)	20.0	(16.9	24.0)	21.9	(17.8	23.4)	0.093	1.000
P_{peak} (cmH_2O)	22.2	(18.7	25.4)	21.9	(18.3	25.0)	23.2	(20.2	26.2)	0.001	0.018*
P_{mean} (cmH_2O)	13.2	(10.7	15.8)	13.1	(10.4	15.5)	14.5	(11.7	16.6)	<0.001	0.010*
PEEP (cmH_2O)	8.7	(6.3	11.7)	8.3	(6.0	11.1)	9.8	(7.7	12.0)	0.001	0.024*
Auto PEEP (cmH_2O)	0.6	(0.3	1.2)	0.6	(0.3	1.2)	0.8	(0.3	1.3)	0.083	1.000
V_T (mL)	486	(412	561)	487	(412	558)	477	(413	575)	0.885	1.000
V_T/IBW (mL/kg)	7.0	(6.2	7.7)	7.0	(6.2	7.7)	7.1	(6.3	7.9)	0.705	1.000
Minute ventilation (L/min)	13.2	(10.7	15.8)	13.1	(10.4	15.5)	14.5	(11.7	16.6)	<0.001	0.010*
Inspiratory flow (mL/s)	40.8	(35.5	46.0)	40.5	(35.2	45.3)	41.8	(37.4	48.3)	0.001	0.033*
Expiratory flow (mL/s)	38.7	(34.0	43.4)	38.4	(33.7	42.9)	39.5	(35.2	44.6)	0.033	0.662
Compliance (mL/cmH_2O)	53.9	(40.1	71.8)	54.1	(40.3	73.0)	53.3	(38.7	67.9)	0.484	1.000
R_{insp} (cmH_2O)	11.6	(9.0	14.4)	11.8	(9.2	14.5)	10.9	(8.3	13.5)	0.020	0.428
$ETCO_2$ (kPa)	5.0	(4.5	5.7)	5.1	(4.5	5.7)	5.0	(4.3	5.6)	0.113	1.000
V'CO2 (mL/min)	182.0	(151.8	219.0)	180.8	(151.4	217.0)	185.9	(156.1	224.1)	0.480	1.000
SpO_2/FiO_2 ratio	2.4	(2.0	2.8)	2.4	(2.0	3.1)	2.1	(1.8	2.4)	<0.001	<0.001*
PaO_2/FiO_2 ratio	25.4	(19.8	33.7)	26.5	(21.0	35.2)	22.3	(17.6	27.7)	<0.001	<0.001*
Art. PH	7.40	(7.35	7.44)	7.40	(7.35	7.44)	7.39	(7.34	7.43)	0.406	1.000
$PaCO_2$ (kPa)	5.6	(5.0	6.3)	5.6	(5.0	6.2)	5.8	(5.2	6.6)	0.012	0.272
PaO_2 (kPa)	10.2	(9.1	11.9)	10.3	(9.2	12.0)	10.0	(8.8	11.5)	0.008	0.186
Art. O_2 saturation (%)	95.0	(93.0	96.0)	95.0	(94.0	96.0)	94.0	(93.0	96.0)	0.001	0.017*
Art. Alkali Reserve (mmol/L)	25.0	(22.0	30.0)	25.0	(22.0	30.0)	26.0	(22.0	30.8)	0.399	1.000
Art. Base Excess (mmol/L)	0.0	(-3.0	5.0)	0.0	(-3.0	4.0)	1.0	(-3.0	5.0)	0.633	1.000
Lactate (mmol/L)	1.5	(1.2	2.0)	1.5	(1.2	2.0)	1.5	(1.2	2.0)	0.889	1.000

* statistically significant difference between groups ($p < 0.05$) after Holm-Bonferroni correction

J TRIPOD+AI Checklist

TRIPOD+AI

Version: 11-January-2024

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	in section:
ABSTRACT				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	1
INTRODUCTION				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	2
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	2
	3c	D;E	Describe any known health inequalities between sociodemographic groups	-
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	2
METHODS				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	4.1
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	4.1
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	4.1
	6b	D;E	Describe the eligibility criteria for study participants	4.2
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	4.2
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	4.3, 4.5
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	4.4
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	4.4
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	4.4
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	4.5
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	4.5
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	-
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	4.1, 4.2, 4.7
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	4.5
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	4.2, 4.7, 4.8
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation)	4.5
	12c	D	Specify the type of model, rationale[2], all model-building steps, including any hyperparameter tuning, and method for internal validation	4.7, 4.8
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations[3]	-
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	4.8
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	-
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	4.10
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	4.7
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	-
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	4.4, 4.8

[1] D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

[2] Separately for all model building approaches.

[3] TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

TRIPOD+AI

Version: 11-January-2024

Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	4.7, 4.8
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	4.11
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	-
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	-
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	-
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	-
Data sharing	18e	D;E	Provide details of the availability of the study data	-
Code sharing	18f	D;E	Provide details of the availability of the analytical code[4]	4.10
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	-
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	5.1
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	5.1
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	5.1
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	5.4, 5.5
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)[5]	4.10, D
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	5.4, 5.5
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details[3].	-
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	-
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	6.1, 6.2, 6.3
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	6.4
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	-
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	6.5
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	6.6

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378