Contents lists available at ScienceDirect

Journal of Critical Care

journal homepage: www.journals.elsevier.com/journal-of-critical-care

Development of a machine learning model for prediction of the duration of unassisted spontaneous breathing in patients during prolonged weaning from mechanical ventilation

Sebastian Johannes Fritsch [a,b,c,*], Morris Riedel [a,c,d], Gernot Marx [b], Johannes Bickenbach [b,1], Andreas Schuppert [e,1]

[a] Jülich Supercomputing Centre, Forschungszentrum Jülich, 52428 Jülich, Germany
[b] Department of Intensive Care Medicine, University Hospital RWTH Aachen, 52074 Aachen, Germany
[c] Center for Advanced Simulation and Analytics (CASA), Forschungszentrum Jülich, 52428 Jülich, Germany
[d] School of Engineering and Natural Science, University of Iceland, 107 Reykjavik, Iceland
[e] Joint Research Centre for Computational Biomedicine, University Hospital RWTH Aachen, 52074 Aachen, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Histogram-based gradient boosting
Machine learning
Predictive modelling
Prolonged weaning
Ventilator weaning

ABSTRACT

Purpose: Treatment of patients undergoing prolonged weaning from mechanical ventilation includes repeated spontaneous breathing trials (SBTs) without respiratory support, whose duration must be balanced critically to prevent over- and underload of respiratory musculature. This study aimed to develop a machine learning model to predict the duration of unassisted spontaneous breathing.

Materials and methods: Structured clinical data of patients from a specialized weaning unit were used to develop (1) a classifier model to qualitatively predict an increase of duration, (2) a regressor model to quantitatively predict the precise duration of SBTs on the next day, and (3) the duration difference between the current and following day. 61 features, known to influence weaning, were included into a Histogram-based gradient boosting model. The models were trained and evaluated using separated data sets.

Results: 18.948 patient-days from 1018 individual patients were included. The classifier model yielded an ROC-AUC of 0.713. The regressor models displayed a mean absolute error of 2:50 h for prediction of absolute durations and 2:47 h for day-to-day difference.

Conclusions: The developed machine learning model showed informed results when predicting the spontaneous breathing capacity of a patient in prolonged weaning, however lacking prognostic quality required for direct translation to clinical use.

1. Introduction

The ventilator weaning for patients who have undergone extended periods of mechanical ventilation (MV) is a complex and frequently challenging process. It can encompass a significant portion, approximately 40–50%, of the total MV duration [1], Therefore, it's crucial to focus on liberating patients from MV at the earliest time point when the underlying causes of respiratory failure are resolved. Patients, in whom weaning is protracted, usually have had a longer stay on an Intensive Care Unit (ICU) and show typical risk factors like advanced age, complex surgical procedures leading to complications, or severe comorbidities, especially cardiac and pulmonary conditions resulting in chronic ventilatory insufficiency [2]. The widely adopted definition of the International Consensus Conference in Intensive Care Medicine classifies patients to be in "prolonged weaning", if they fail more than three spontaneous breathing trials (SBTs) or stay under MV for >7 days after the first SBT [3]. Patients experiencing prolonged weaning typically suffer from an insufficiency of the respiratory musculature which is

Abbreviations: AI, Artificial Intelligence; BGA, Blood gas analysis; ICU, Intensive Care Unit; LOS, Length of stay; MAE, Mean absolute error; ML, Machine learning; MSE, Mean squared error; MV, Mechanical ventilation; NaN, Not a number = missing value; ROC-AUC, Receiver operating curve – Area under the curve; SBT, Spontaneous breathing trial.

* Corresponding author at: Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.
E-mail address: s.fritsch@fz-juelich.de (S.J. Fritsch).

[1] Both senior authors contributed equally to this work.

https://doi.org/10.1016/j.jcrc.2024.154795
Received 30 November 2023; Received in revised form 3 January 2024; Accepted 8 January 2024
Available online 26 March 2024

caused both by a pronounced contraction weakness of the diaphragm in combination with a deconditioning of the peripheral respiratory musculature [4-6]. Additionally, various factors contribute to diminished respiratory capacity during prolonged weaning, like mental impairments (e.g. delirium), recurrent infection, cardiac or renal failure, anaemia, malnutrition, imbalances of electrolytes and acid-base (e.g. hypomagnesaemia, hypochloremic alkalosis) or excessive positive fluid balance [7]. Retrospective studies indicate that biometric factors like age and sex also influence weaning outcomes [8-12]. Unsurprisingly, also the length of stay (LOS) on ICU and preceding duration of MV play significant roles in the course of weaning [8].

Given the high complexity of patients in prolonged weaning due to the indicated limiting organ dysfunctions and comorbidities, their treatment demands a systematic overall strategy in specialized units or centres with the corresponding interdisciplinary expertise of all professional groups involved and the necessary structural prerequisites [13]. The current weaning practice involves intermittent, structured and protocolized SBTs, where a patient is periodically disconnected from the respirator and breathes without respiratory support for a predefined period of time. After completion of the specified time or if the patient shows clinical signs of exhaustion, the SBT is terminated and ventilatory support is resumed [14]. Especially, in the beginning of a structured weaning approach, SBT durations of only a few minutes are not uncommon. In case of a clinical improvement of the patient, the duration of these SBTs can be extended from day to day until a full liberation from MV is achieved. Through continued repetition of this process, the respiratory musculature can be reconditioned to take over again the task of pulmonary gas exchange.

However, determining the optimal duration of SBTs has to be chosen wisely to balance the respiratory muscle load. Both ventilatory overloading, as well as underloading are unfavourable and could prolong weaning unnecessarily. Severe exhaustion of the respiratory muscles in particular should be avoided at all costs, since it may lead to diaphragmatic injury [15-17]. Thus, setting the target duration for spontaneous breathing per day is crucial and requires extensive clinical expertise and the ability to monitor and appraise a patient's condition well. Making this task even more challenging, there are neither clear criteria nor clear protocols which would support clinicians with this task. Clinicians typically not only use their impression on a patient but also structured data, like vital signs or laboratory parameters for assessment of a patient's condition. Thus, it can be hypothesized that the capacity of spontaneous breathing can be derived at least in part from these structured data. However, given the complexity and multitude of influencing factors, human caregivers, might miss crucial information. Due to its inherent ability to effortlessly analyse high-dimensional data and detect patterns in it, techniques of Artificial Intelligence (AI) could display their strengths in this area and support physicians and nurses with this task.

Hence, this study aimed to develop a data-driven machine learning (ML) model which is able to predict the duration of unassisted spontaneous breathing in patients undergoing prolonged weaning from MV using one day's data to predict the SBT duration of the next day. Initially, the analysis sought to determine the model's ability to generally predict an increase in SBT duration per se. Subsequently, two other models were created to predict both the exact duration of spontaneous breathing at the next day and the achieved difference between the current and the following day.

2. Materials and methods

2.1. Ethical approval

This study was approved by the local ethical review board (EK 122/13, Ethics Committee, Faculty of Medicine, RWTH Aachen, Germany). Due to the retrospective character of the study, the Ethics Committee waived the need for an informed consent.

2.2. Data sources and patient data set

Data were retrieved from online patient data management system (IntelliSpace Critical Care and Anesthesia, ICCA Rev. F.01.01.001, Philips Electronics, Netherlands). The raw data were extracted as comma-separated value files.

The data set included data of all patients, who were admitted to the Interdisciplinary Weaning Unit at the University hospital RWTH Aachen over a five year period from 2017 to 2022 as a convenience sample. Admission criteria to the weaning unit include a previously performed tracheostomy not shorter than 24 h and a documented SBT (including reasons for withdrawal) as well as an assessment of the patient's readiness to wean [3], respectively, during the preceding ICU stay. Additionally, the absence of acute illness and multi-organ failure, absence of high-dosage or multiple vasopressor therapy, and the absence of continuous renal replacement therapy (RRT) is required. A surgical treatment, especially multistage procedures that require repeated performance of general anesthesia, should be preferably completed prior to admission. All extracted parameters were acquired and used in daily clinical routine and no additional data were collected.

2.3. Software and computational resources

For data pre-processing and development of the predictive model, the source-code editor Visual Studio Code (Version 1.71.2, Microsoft Corporation, Redmond, USA) and the programming language Python 3.10.5 (Python Software Foundation, Delaware, USA) in combination with its libraries numpy 1.23.3 [18], pandas 1.5.0 [19] and scikit-learn 1.1.2 [20] was used. Data visualisation was carried out using matplotlib 3.6.0 [21] and single components of seaborn 0.12.0 [22] and scipy 1.9.1 [23]. Computationally intensive calculations were carried out on the high-performance computing resources at the Juelich Supercomputing Centre of the Forschungszentrum Jülich, namely on the Data Analytics Module (DAM) of the Dynamical Exascale Entry Platform (DEEP) [24].

2.4. Choice of features

With respect to the evidence-based or perceived importance for the weaning outcome and the availability of data, a list of 61 clinical features was selected by group consent of two physicians who were experienced in the treatment of patients undergoing prolonged weaning. For all selected characteristics, there is evidence that they have either influence on the course of weaning themselves or that they represented a clinical condition that in turn had an impact. The included features comprised biometric information, vital signs, respirator settings, laboratory tests indicative for infectious processes or electrolyte imbalances, information on the administration of certain drugs, like vasopressors, sedatives, opioids, antipsychotic or anti-infective medication and transfusion of blood products as well as diuresis and the fluid balance over 24 h. For a full list of features please refer to the supplemental material (see Table S1).

2.5. Data preprocessing

For preparation of the analysis, extracted data were first revised for data-related inconsistencies, like diverging units, diverging decimal separators or values containing unreadable signs, as well as medical inconsistencies, like laboratory test results taken from wrong samples. For patients, who were transferred from an external hospital and thus lacking the information on the preceding ICU stay, length of stay on ICU was set to missing value. Dynamic parameters were transformed into one single value per day as mean, minimum or maximum values, depending on the pathophysiological medical background. These data were used as features for the predictive models. Finally, one data point represented one single day of one single patient. Obvious artifacts, e.g. due to a disconnected or misplaced sensor, were eliminated. Due to poor data quality, information on transfusions and anti-infective medication had to be converted to a Boolean parameter indicating that a patient received the respective therapeutic agent, but not containing information on a specific compound or dosing.

The data of respiratory parameters were calculated from manually entered respirator settings or from measurements which were automatically generated by the respirator depending on their availability. Wherever available, automatically measured values were preferred. A data point was labelled as "respirator-free", if no MV data was documented on the respective time point. Patient days, which showed no MV for 24 h on the respective as well as on the preceding and following days, were excluded from analysis, as the patient can be considered as successfully weaned. Finally, the data of the discharge day were removed, as logically it was not possible to predict a subsequent day.

2.6. Predicted outcome

Three predictive models were created to predict a patient's ability to breathe spontaneously on the next day in a qualitative and quantitative way. The predictive targets were (1) a patient's qualitative ability to increase their temporal proportion of spontaneous breathing, (2) the exact duration of spontaneous breathing on the following day, and (3) the exact difference in spontaneous breathing between the current and the following day.

2.7. Algorithm

For the generation of the predictive models, a Histogram-based Gradient Boosting Classifier Tree and a Histogram-based Gradient Boosting Regressor Tree from the scikit-learn library were used, respectively (Pedregosa et al. 2011). Histogram-based Gradient Boosting represents a further improvement of the popular Gradient Boosting algorithms LightGBM [25]. It takes advantage of the fact that the speed of the construction of decision trees is significantly higher if the number of values for continuous input features is reduced, what can be achieved by "binning" them into a fixed number of bins [26]. The number of unique values of a continuous feature thus can be reduced from tens of thousands down to a few hundred or even less. This binning procedure usually does not affect the model performance, but remarkably reduces training time [25]. For the binning of the input data, efficient data structures, like histograms, are used making it more efficient in both memory consumption and training speed than an algorithm dealing with continuous values. Compared to other algorithms, it is thus especially faster in big datasets with a sample size above 10,000 data points. Another characteristic, which is much more relevant in the context of the present work, is its native support for missing values (NaNs).

2.8. Validation of the models and hyperparameter tuning

For the validation of the models, the available dataset was split into a train data set of 80% and a test data set of 20% of the full data set, both selected using a random permutation process. To prevent a potential information leakage through the appearance of data points of one individual patient in both the train as well as in the test data set, the algorithm for data splitting "sklearn.model_selection.GroupKFold" was used. The classifier model was evaluated using the sensitivity (also known as true positive rate or recall), specificity (also known as true negative rate), positive predictive value (also known as precision) and negative predictive value, accuracy score and the ROC-AUC. For the evaluation of the regressor models, the adjusted R^2 score, the Mean squared error score (MSE) and the Mean absolute error score (MAE) were used. Both classifier and regressor models were analyzed for the most relevant features using the permutation feature importance technique.

2.9. Learning curves

Learning curves for all models were created to gain insight into the variance and bias of the trained models. For this purpose, the data set was split into training and test data set using a 5-fold cross-validation for every step. The number of training data points was increased in ten steps up to the full training data set. For every randomly selected number of training data points, a new model was trained. The accuracy score for the classifier and the MSE for the regressor as quality metrics were averaged over all 5 runs and plotted against the size of the training data set.

For all models, the respective hyperparameters were optimized using grid search with stratified 5-fold cross-validation on the training set. The optimized hyperparameters comprised the maximum number of iterations of the boosting process ("max iter"), the learning rate, the minimum number of samples per leaf, the L2 regularization, the maximum number of bins to use for non-missing values ("max bins") and the maximum depth of each tree ("max depth"). In case of the classifier, a repeated stratified k-fold cross validation with 10 folds and 3 iterations was carried out and the best hyperparameters were chosen with respect to the accuracy metric. Contrasting, for the regressor, a repeated k-fold cross validation with likewise 10 folds and 3 iterations was carried out and the selected metric was a negative MAE. Subsequently, the generated hyperparameters were adjusted in the predictive models. The maximum number of iterations of the boosting process resulted in a high tendency to create overfitting without relevantly contributing to improvement of the validation score. Therefore, boosting was kept on default value ('max_iter' = 100). The optimized hyperparameters are given in the supplementary material (see Table S2).

3. Results

3.1. Clinical characteristics

The complete dataset encompassed 1018 individual patients who were admitted to the Interdisciplinary Weaning Unit after November 25, 2016, and discharges before January 22, 2022. The average age of the patients was 65.5 years, with males comprising nearly two-thirds of the group. Prior to being transferred, patients spent an average of approximately 20 days receiving treatment in an ICU. Information about the previous ICU stay was unavailable for 10.4% of patients, as they were transferred from an external hospital to the weaning unit. The average length of stay on the weaning unit was nearly 26 days. Since the data extracted from the PDMS did not include diagnoses, it was not possible to definitively identify patients who were discharged as fully weaned. Around 17.1% of patients were respirator-free on the last full day before discharge, since there was no MV documented on that day. Furthermore, at least 68.9% of patients required ventilation for <6 h on this day. Detailed clinical characteristics of the included patients can be found in Table 1.

Table 1
Clinical characteristics of the included patients. IQR Interquartile ratio, ICU intensive care unit, WEA weaning unit.

Parameter	Median (IQR) or n (%)
Included patients	1018 (100)
Age [years]	66.97 (18.29)
Length of stay on ICU [days]	15.83 (14.15)
MV duration on ICU [days]	14.93 (12.97)
Transfer from external ICU (*i.e. ICU data missing)*	106 (10.41)
Length of stay on WEA [days]	19.79 (19.81)
Male sex	693 (68.08)
MV free time at last full day on the weaning ward	
- 24 h ("respirator-free")	174 (17.09)
- > 18 h	701 (68.86)

3.2. Data set

The complete data set encompassed 22,887 data points, each representing one day of treatment for a specific patient in the specialized weaning unit. From these, 3.939 “respirator-free” datapoints without documented MV over a period of 72 h were excluded, resulting in a final data set of 18.948 data points. Within this data set, there were 9037 data points indicating an increase in SBT duration, while 9911 data points showed no increase, making the distribution of classes nearly balanced. Due to varying lengths of stay among patients, each patient contributed differently to the respective data set. Some features exhibited notably low data density, particularly certain laboratory parameters such as BNP, IL-6, chloride, magnesium, and phosphate, missing in >80% of data points. A full list of missing data proportions for each feature can be found in the supplementary material (refer to Table S3).

3.3. Predictive accuracy of the trained models

3.3.1. Classification models

When assessing a patient’s ability to extend SBT duration the next day, the classifier achieved an overall accuracy score of 0.65 on the test data set, indicating that around two-thirds of the predictions were correct. Sensitivity and negative predictive value slightly surpassed specificity and positive predictive value. The ROC-AUC for this model was 0.713 (refer to Fig. 1). In comparison to the metrics from the training data set, the model exhibits an absolute decrease in accuracy of approximately 15% when applied to the test data. Detailed metrics for the classifier model concerning both the training and test data sets can be found in Table 2.

The analysis of feature importance using permutation revealed that only seven features showed a notable influence, represented by a decrease of accuracy of >0.005 when values of the respective parameter were scrambled. These crucial parameters included the MV-free time on the current day, which emerged as the most pivotal feature. Similarly, the differences between the SBT duration of the current day and the day 1, day 2 and day 3 before had relevant influence. Additionally, the urine output, the FiO$_2$ and the age showed relevant influence. A visualisation of the full set of features and their respective importance can be found in the supplementary material (see Fig. S1).

The learning curve for the classifier model reveals that the validation error shows practically constant behaviour and reaches its plateau already with a model trained on 10% of the final training data. Even with a higher amount of training samples, the model is unable to increase the predictive accuracy on test data. In contrast, the accuracy of the prediction on training data starts at a high point and decreases only slowly with increase of data points. After reaching the full size of the training dataset, there is still a relevant gap between the predictive performance achieved, indicating a decrease in accuracy when moving from predicting training data to predicting test data. The learning curve for the classifier is given in Fig. 2.

Table 2
Evaluation metrics of the classifier models applied on the test and training data set. Classifier target: Increase of duration of MV-free time at the following day. PPV: positive predictive value, NPV: negative predictive value, ROC: receiver operating characteristic.

	Accuracy	Sensitivity	Specificity	PPV	NPV	ROC score
Test data set	0.65	0.70	0.61	0.62	0.68	0.713
Training data set	0.80	0.85	0.76	0.76	0.85	0.891

3.3.2. Regressor models

The regression model predicting the absolute duration of SBT on the following day displayed a MAE of 2.84 h (i.e. 2 h and 50 min) for the test data. Regarding the prediction of the difference between the current and

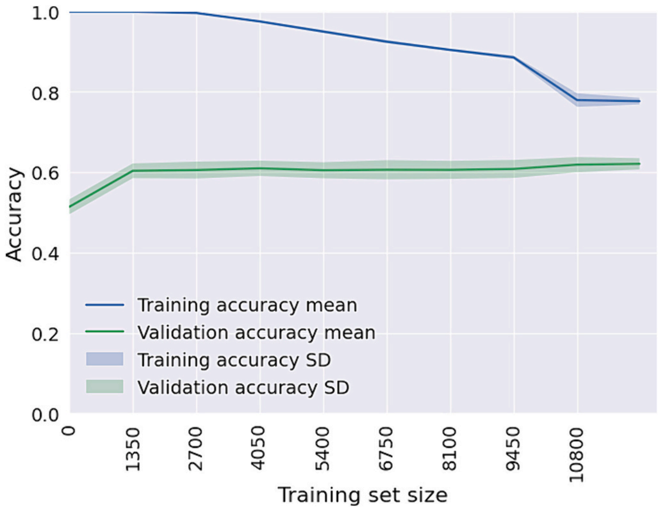

Fig. 2. Learning curve for the classification model. The models were trained using a 5-fold cross validation. The accuracy of training and validation is plotted against the size of available training data set of the respective model. The lines indicate the mean value, while the standard deviation is indicated by the shaded area.

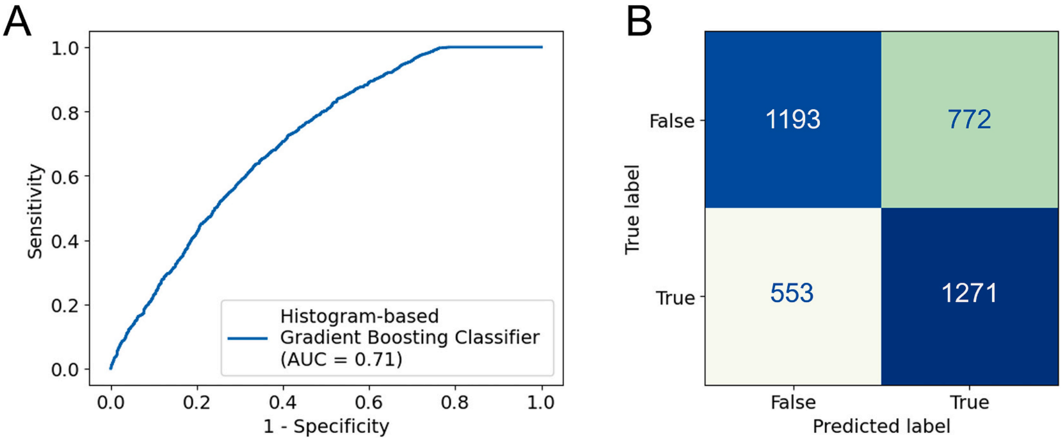

Fig. 1. ROC curve and confusion matrix of the classifier model. A. ROC curve of the classifier model applied on the test data set. B. Confusion matrix of the classifier model applied on test data set showing the combination of true and predicted labels.

the next day, the MAE was marginally improved, measuring at 2.79 h (i.e. 2 h and 47 min). The full metrics of the regressor models can be found in Table 3.

Although both MAE values were nearly identical, a plotting of predicted against the actual durations shows that the distribution of predictions is relevantly diverging between the two models. In the case of predicting the absolute SBT time, the data points representing true vs. predicted values predominantly cluster around the identity line (see Fig. 3). Conversely, when predicting the difference between consecutive days, the plot of predicted versus true values displays two clusters centred around predicted values of 0 and 10 h (see Fig. 3B). However, these clusters are elliptical in shape, indicating that the dispersion of the true values is significantly greater than that of the predicted ones, so that the pattern clearly deviates from a linear distribution.

The permutation analysis emphasized the significance of features previously identified as crucial for the classifier algorithm. Key features included MV-free time, 24-h urine volume, the difference in SBT duration between the current day and one, two, and three days before, as well as respiratory parameters, like FiO_2 and inspiratory pressure support (ASB). Interestingly, urine volume was more important in predicting differences, whereas MV-free time turned out to be more influential in predicting the absolute value of SBT. Detailed plots of the permutation analysis can be found in the supplementary material (see Figs. S2 and S3).

In contrast to the classifier's learning curves, the learning curves of the regressor models showed a gradual, albeit marginal decrease of MSE with growing size of the training data set when predicting the test data (see Fig. 4). Correspondingly, the MSE of a prediction of training data increases slowly and steadily with growing data set. While the slope becomes less pronounced with more training data, it still remains possible that the curves converge with additional data. However, even at the maximal sample size of the training data, a notable performance gaps persists between training and test data within the models.

4. Discussion

Assessing a patient's ability to breathe without ventilatory support during prolonged weaning from MV poses a significant challenge for clinicians. These patients form a diverse group with unique characteristics that may exceed a physician's full comprehension due to their complexity. While the analysis of complex high-dimensional data can be a problem for clinicians, ML algorithms theoretically offer a potential solution. However, the models generated in this study produced only moderate results, with approximately two-thirds of qualitative predictions being correct regarding increased duration. On average, there was nearly a 3-h difference between predicted and achieved durations.

The complexity of this field might explain why, to the authors' knowledge, publications with a similar focus are not available. Most literature focuses on successful extubation as the primary outcome, usually after short-term MV durations, often less than a week before inclusion, which doesn't represent patients undergoing prolonged weaning [27]. However, in patients in prolonged weaning, the road to extubation or decannulation is much more complex than the act of extubation itself. Or in other words: if a physician can already think about extubating a patient, the most difficult part of the job is already done. Models explicitly focussing on patients in prolonged weaning were, for instance, developed by Yang et al. and Lin et al. both using data sets of patients under MV for >21 days and focussing on successful extubation as primary outcome parameter [28,29]. Another model explicitly developed for the use in long-term MV patients by Hadjitodorov and Todorova included patients to the training dataset with a mean MV duration of >26 days aiming to determine the time point when a patient is ready to start the weaning procedure, i.e. the transfer from controlled ventilation to assisted spontaneous breathing. Notably, their model was finally applied prospectively in a real-world setting [30].

Table 3
Evaluation metrics of the regressor models applied on the test and training data sets. Target: Prediction of absolute duration of MV-free time at the following day and of difference of duration of MV-free time from current to the following day. MAE: mean absolute error, MSE: mean squared error.

	MAE [hour]	MSE [$hour^2$]	Adjusted R^2
Prediction of absolute MV-free time at the next day			
Test data set	2.84	15.20	0.754
Training data set	2.50	11.45	0.806
Prediction of the difference of MV-free time to the next day			
Test data set	2.79	14.99	0.324
Training data set	2.51	11.86	0.460

The present work included 1018 individual patients, forming a high number compared to other observational studies involving this patient cohort, usually involving fewer than 300 patients [2]. Despite some minor differences, like an about eight days shorter duration of MV in median before transfer to the weaning unit, the examined population appeared to be comparable with the patient population in a network of 70 German pneumological weaning centres „WeanNet“[8]. A clear deviation from the registry data occurred in the attempt to derive a successful weaning exclusively from the structured data. While the WeanNet registry reports 64.3% of successfully weaned patients, only 17.01% of the patients from the data set used, no MV is documented at the last full day before discharge from the weaning unit. This notably low percentages are also in contrast to prior evaluations in a subsample from the same weaning unit reporting 80% successfully weaned patients [31]. This apparent contradiction could be caused by the fact that the structured data could be biased due to a standard of care. In clinical routine on ICU and on the weaning ward as well, even spontaneously breathing patients, often receive positive pressure ventilation – usually using a non-invasive interface – for the purpose of pneumonia prophylaxis. Thus, durations of up to 4 × 1.5 h daily, which is a widely-used routine, are usually considered as a prophylactic measure and therefore not assessed as MV in the narrower sense.

The available data set encompassed a high proportion of missing values, especially in certain features. However, it must be considered that the absence of measured values can contain information in a medical context as well. For instance, a missing parameter can reflect the opinion of a physician on charge that a laboratory analysis will currently not reveal any relevant new information since a patient has stabilized [32]. Thus, an imputation of missing values seemed not justified. These considerations, in turn, significantly constrained the options when selecting the algorithm leading to the choice of a Histogram-based Gradient Boosting algorithm. Although this choice provided not only a native support for missing values and proves to be a very efficient when training on large datasets, it also has some drawbacks, such as a tendency to overfitting, which is common in decision tree-based methods [33].

This problem could also be clearly demonstrated using the learning curves generated during the model development. Small training data sets lead to a nearly full memorization of data when predicting training data. Also, with a growth of the training data set, the accuracy on training data stayed constantly better than on the test data indicating an overfitting of the models. The course of the performance metric in the test data, however, is rather surprising. It reaches its plateau already with a model trained on 10% of the final training data and from that point on a further increase of accuracy is not achievable or with respect to a decrease of MAE in the regressor only to minimum extent. Due to these results, it can be concluded that even a further increase of data points in the training data set will not lead to a better predictive performance.

The classifier model for the prediction of an increase in MV-free duration achieved a ROC-AUC score of 0.73 which is considered as "acceptable" [34]. From a clinical point of view this kind of model might

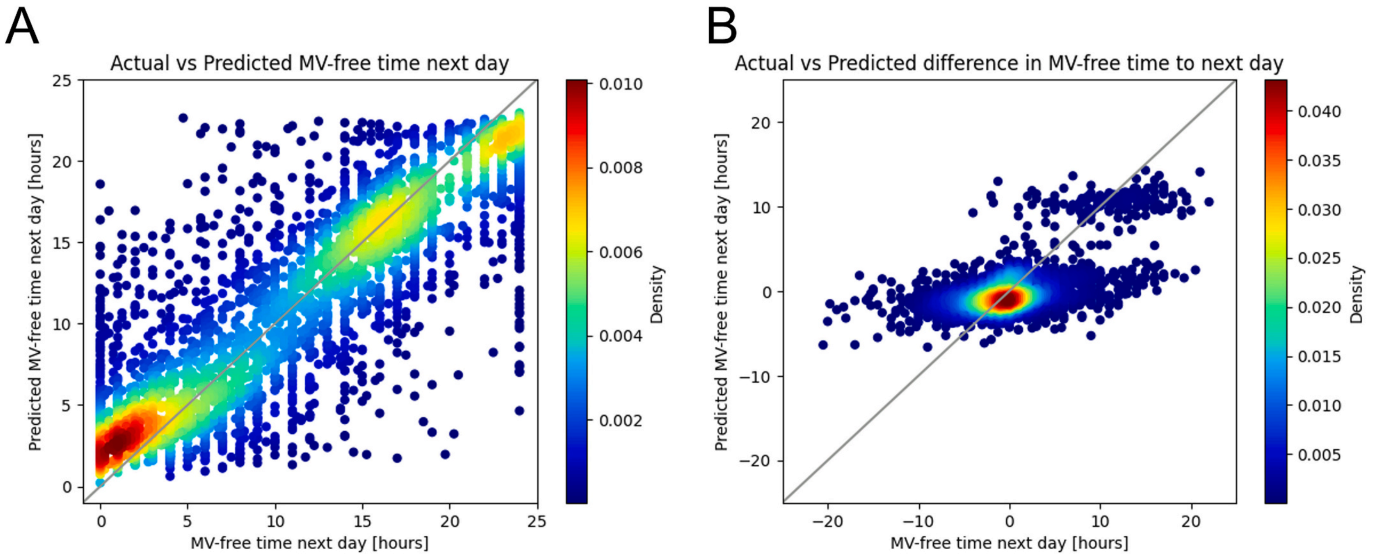

Fig. 3. Plots of the predicted vs. the true values of the regressor models. Evaluation was carried out on the test data set. The density of points is represented by a Gaussian kernel density estimation. A. Prediction for time without MV at the following day B. Prediction for the difference of MV-free time from one day to the next.

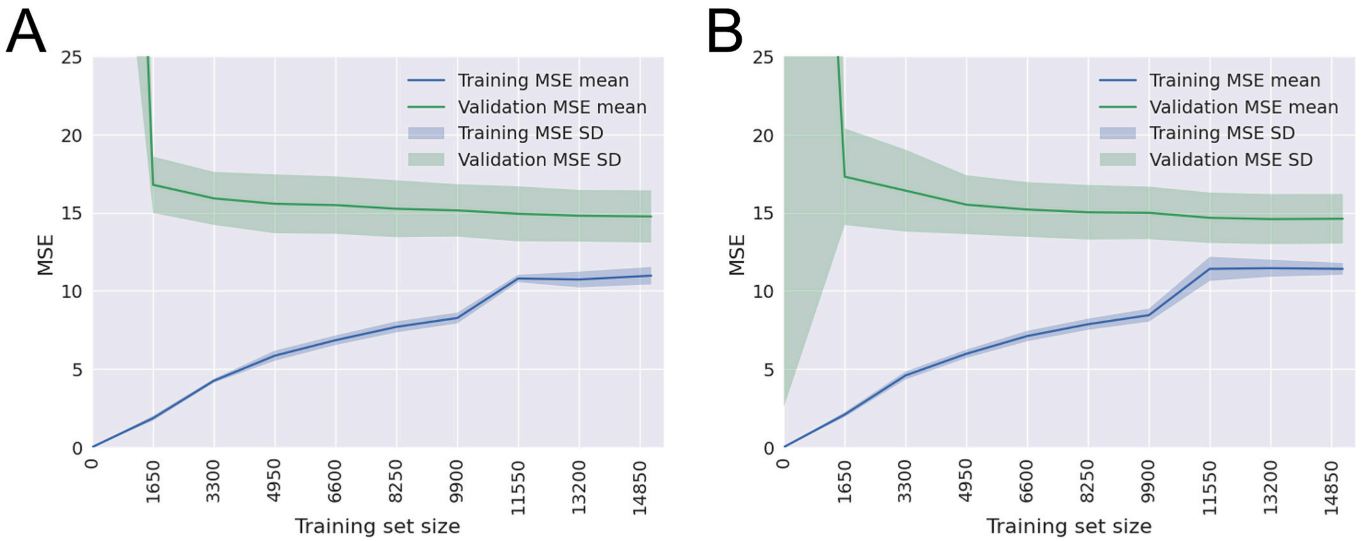

Fig. 4. Learning curves for the regression models. Models were trained using a 5-fold cross validation. The mean squared error of training and validation is plotted against the size of the available training data set of the respective model. The lines indicate the mean value, while the standard deviation is indicated by the shaded area. A. Regressor for prediction of the absolute duration of SBT on the next day. B. Regressor for prediction of the difference of the SBT duration from one to the next day. MSE: mean squared error.

be even more useful since it would be able to support physicians with the decision whether a patient can proceed in the weaning process or not. This prediction, if it is done accurately enough, is much more important and relevant than trying to predict the MV-free time of the next day to the minute and then, due to the high inaccuracy of the model, predicting a deterioration instead of an in fact possible increase in MV-free time. The prediction of precise time spans or their difference for spontaneous breathing on the following day showed just this behaviour with predicted values deviating by 2:47 h to 2:50 h from the real values on average. Regarding the MAE, the target of the model (absolute MV-free time vs. change between the days) was not relevant. But looking at the scatter plot of prediction vs. true values, salient distributions become evident. Although, the predictions of absolute MV-free times scattered symmetrically around the identity line, the Gaussian kernel density estimation revealed two clusters, namely one big cluster of patients, which still have a low spontaneous breathing capacity, i.e. under five hours, and another cluster with a centre at about 16 h gets visible. These two clusters represent typical clinical findings. The first observation indicates that especially the start of weaning is sometimes challenging, with several patients requiring several days of recurrent, very short SBTs adding up to one or two hours only before they start making bigger progress leading to a relevant increase of MV-free time. The following steps to increase MV-free time frequently are run through quite easily. The last step before complete liberation from the ventilator is again - also mentally - a big step and physicians, as well, are sometimes hesitant to take back the last hours of MV, which are applied, on one hand, as intermittent pneumonia prophylaxis of up to six hours, as already explained above, or on the other hand as a continuous MV of 6 h for night rest. Maintaining ventilation at night is an established treatment option for patients with persisting respiratory insufficiency, as well, e.g. in advanced cases of COPD [35]. While these findings for the model predicting the absolute MV-free time are reasonably explainable, this is much more difficult for the model predicting the difference of MV-free time between two days. Looking at the plot of predicted and true differences, there are two clusters with centres on the identity line at true values of 0 h and about 10 h. Looking at the plot of the test data, it appears that the predicted values are elliptically distributed around the centres, i.e. the model predicts either no difference or an increase of 10 h, while the true values deviate considerably from this value in both directions. However, a look at the plot of training data, which has a higher data density, reveals that the ellipses have indeed a positive slope, which is, however, much too small to fit the data correctly (see Fig. 5). Thus, it can be stated that the model is able to detect a trend but is much too conservative in its prediction. The question of how this pattern occurs ultimately remains unclear.

Although only features which were shown to have an influence on the weaning were included into the model, the vast majority of features influenced the metrics of the models just minimally. Beside the features, which represent the current SBT capacity of a patient on the neighbouring days and some features representing the respiratory situation, the most relevant features contained the daily urine volume and creatinine representing the absence of a renal failure. The relationship between renal failure requiring renal replacement therapy and a bad outcome in prolonged weaning is in concordance with several publications in this cohort [36,37]. Interestingly, the fluid balance had much lower influence, although its inclusion still improved the model. Moreover, the connection between length of stay and MV duration on ICU, which were important for the performance, was demonstrated repeatedly before, as well [8]. It must also be considered that the remaining features might be predictive for a weaning success in the long run but are not useful for the small-scale prediction of the next day. This would also agree with the finding that die adjacent MV-free time spans are obviously more important than other features, which, in contrast, might show bigger influence on the final outcome.

To the best of the author's knowledge, there are no examinations which assess the accuracy of physicians or similar predictive model when predicting a patient's respiratory capacity in the next 24 h. Thus, a comparison of the developed models against an existing standard is not possible. It would be of high interest, whether human physicians would achieve similar or better accuracy rates. Due to the lack of these examinations, it is also not possible to appraise the clinical usefulness of the predictive models. From a clinical perspective, a model that correctly classifies only 2 out of 3 cases or which deviates from the real duration on average >2 h might have difficulty being accepted in routine clinical practice. Considering the serious deviations of some outliers, which can reach even >20 h, it becomes clear that considerable improvements of the models would be necessary in order to consider further clinical testing.

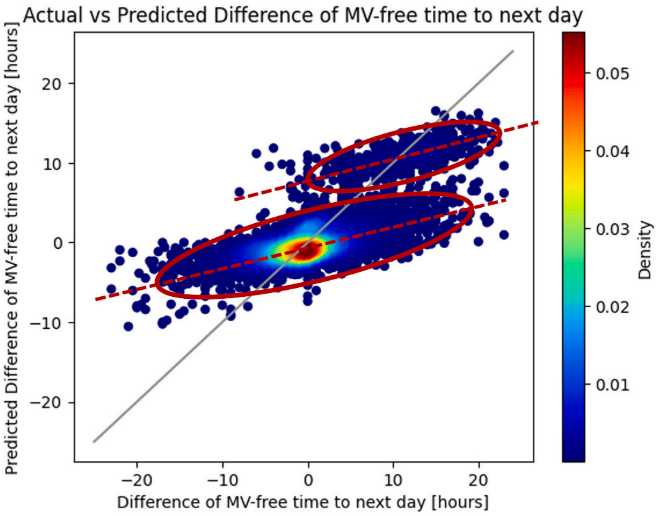

Fig. 5. Depiction of the conservative prediction behaviour of the model predicting the differences of MV-free time between two days. Plot of the predicted vs. the true values of the differences of the time without MV between the current and the following day for the training data set. The density of points is represented by a Gaussian kernel density estimation. The two indicated clusters (red ellipses) show a slight positive slope (dotted red line) which is however too small for a good fit of the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

If all these considerations are summarized, the question ultimately remains as to why the predictive accuracy of the models has fallen so far short of expectations. One possible interpretation of the results would be that the information from one single day represents a very limited section of data, which is not suitable for making respective predictions. The course that a patient has taken over the last several days may play a much greater role than the one-day-snapshot. Thus, it definitely would be worth examining if the predictive performance can be increased by extending the present models with algorithms and techniques from time series forecasting to integrate and process longer time periods, like e.g. long short-term memory (LSTM) models [38]. A much more fundamental question, however, is whether the information that would enable a correct prediction is contained in the structured clinical data at all. For example, the clinical impression that a patient makes on physicians and nurses, but which is not documented anywhere in this form, could have a significantly greater impact on the patient's ability of spontaneous breathing. For instance, it was shown that a clinical concern of an experienced healthcare worker was able to detect patient deterioration better than standardized scores [39]. Similar considerations could apply to the context of prolonged weaning. The same applies for the reason to terminate an SBT, which is documented just in an unreliable way. However, an SBT that was interrupted because the patient complained of severe respiratory exhaustion would of course have a different meaning for the next day than an SBT that was terminated as planned after the preset time had elapsed.

Another relevant limitation of this study was the missing information on the diagnoses of patients. This information would have included relevant information, which also might had been useful as potential features indicating the presence of certain conditions like Chronic obstructive pulmonary disease (COPD) or ICU-acquired weakness (ICUAW). Moreover, a list of diagnoses also had included information about the final weaning outcome. A common obstacle in medical data science is the heterogeneous and sometimes poor data quality of the raw data, which made an extensive data cleaning necessary. In the present work, for example, data on blood products or medication contained a lot of free text entries including incomplete and or inconsistent data, which prevented the extraction of the absolute number of administered red blood cell concentrates, although it was shown to have a clinical impact as well [31]. Similar problems arised during the work on data from blood gas analyses (BGA) giving valuable insights into oxygenation and removal of CO_2. Due to unclear time stamps attached to the BGA parameters, it was not possible to reliably determine whether a BGA sample was taken under MV or under an SBT.

The problems that became evident during pre-processing and data cleaning at least gave rise to the assumption that the design of either the patient data management system or at least the respective data extraction algorithms had not been considered in terms of a secondary use of the extracted data. It would therefore be desirable for subsequent versions to at least facilitate secondary data usage and possibly update systems that have already been implemented accordingly. Clinicians and researchers should actively address this aspect when communicating with the manufacturers of such systems.

This study focussed on an increase of spontaneous breathing time as topic of highest interest and thus created a binary classifier. However, it might also have been of interest to predict a deterioration with decrease of the SBT duration as a "pre-warning system" which had resulted in a multiclass classifier. Such an approach might be up to further examinations.

5. Conclusions

In the present work, an attempt was made to approach the challenging task of determining the spontaneous breathing capacity of a patient in prolonged weaning using a data-based ML model to predict an improvement of SBT duration as well as prediction of specific durations. Although a large number of 61 features was included in the model, for which an influence on the weaning process was demonstrated throughout, the results showed predictive qualities below clinical needs. In particular, the actual prediction of a duration showed such serious deviations between predicted and real value that implementation in a clinical context appears inappropriate.

Ethics approval details

This study was approved by the local ethical review board (EK 122/13, Ethics Committee, Faculty of Medicine, RWTH Aachen, Germany). Due to the retrospective character of the study, the Ethics Committee waived the need for an informed consent.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 3.5 in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Sebastian Johannes Fritsch: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Morris Riedel:** Writing – review & editing, Supervision, Resources. **Gernot Marx:** Writing – review & editing, Supervision, Resources. **Johannes Bickenbach:** Writing – review & editing, Supervision, Conceptualization. **Andreas Schuppert:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Data availability

The datasets of patients of University hospital RWTH Aachen, which were used and analyzed during the current study, are not publicly available due to medical confidentiality and German data protection legislation, but are available from the corresponding author on reasonable request. The developed software code is also available from the corresponding author on reasonable request.

Acknowledgements

The authors would like to thank Dr. Albert Esser from the IT department of the University hospital RWTH Aachen for his assistance with the data extraction. Special thanks are given to Prof. Rainer Röhrig from the Department of Medical Informatics of the University hospital RWTH Aachen for his valuable contributions and the fruitful discussions during this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcrc.2024.154795.

References

[1] Esteban A, Alía I, Ibañez J, Benito S, Tobin MJ. Modes of mechanical ventilation and weaning. A national survey of Spanish hospitals. The Spanish lung failure collaborative group. Chest 1994;106(4):1188–93.
[2] Trudzinski FC, Neetz B, Bornitz F, Müller M, Weis A, Kronsteiner D, et al. Risk factors for prolonged mechanical ventilation and weaning failure: a systematic review. Respiration 2022;101(10):959–69.
[3] Boles JM, Bion J, Connors A, Herridge M, Marsh B, Melot C, et al. Weaning from mechanical ventilation. Eur Respir J 2007;29(5):1033–56.
[4] Barchuk A, Barchuk SA, Roebken CK, Ahn J. Prevalence of diaphragmatic dysfunction in the long-term acute care setting and its effects on ventilator weaning outcomes: a retrospective cohort study. Am J Phys Med Rehabil 2022;101(6):555–60.
[5] Dres M, Dubé BP, Mayaux J, Delemazure J, Reuter D, Brochard L, et al. Coexistence and impact of limb muscle and diaphragm weakness at time of liberation from mechanical ventilation in medical intensive care unit patients. Am J Respir Crit Care Med 2017;195(1):57–66.
[6] Pu L, Zhu B, Jiang L, Du B, Zhu X, Li A, et al. Weaning critically ill patients from mechanical ventilation: a prospective cohort study. J Crit Care 2015;30(4). 862.e7-13.
[7] Heunks LM, van der Hoeven JG. Clinical review: the ABC of weaning failure–a structured approach. Crit Care 2010;14(6):245.
[8] Windisch W, Dellweg D, Geiseler J, Westhoff M, Pfeifer M, Suchi S, et al. Prolonged weaning from mechanical ventilation: results from specialized weaning centers. Dtsch Arztebl Int 2020;117(12):197–204.
[9] Warnke C, Heine A, Müller-Heinrich A, Knaak C, Friesecke S, Obst A, et al. Predictors of survival after prolonged weaning from mechanical ventilation. J Crit Care 2020;60:212–7.
[10] Ghiani A, Paderewska J, Sainis A, Crispin A, Walcher S, Neurohr C. Variables predicting weaning outcome in prolonged mechanically ventilated tracheotomized patients: a retrospective study. J Intensive Care 2020;8(1):19.
[11] Huang C. Gender differences in prolonged mechanical ventilation patients - a retrospective observational study. Int J Gen Med 2022;15:5615–26.
[12] Ma J-G, Zhu B, Jiang L, Jiang Q, Xi X-M. Gender- and age-based differences in outcomes of mechanically ventilated ICU patients: a Chinese multicentre retrospective study. BMC Anesthesiol 2022;22(1):18.
[13] Bingold T, Bickenbach J, Coburn M, David M, Dembinski R, Kuhnle G, et al. DGAI-Zertifizierung anästhesiologische Intensivmedizin: Entwöhnung von der Beatmung Modul 1. Anästh Intensivmed 2013;54:212–6.
[14] Esteban A, Alía I, Gordo F, Fernández R, Solsona JF, Vallverdú I, et al. Extubation outcome after spontaneous breathing trials with T-tube or pressure support ventilation. Am J Respir Crit Care Med 1997;156(2):459–65.
[15] Orozco-Levi M, Lloreta J, Minguella J, Serrano S, Broquetas JM, Gea J. Injury of the human diaphragm associated with exertion and chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2001;164(9):1734–9.
[16] Jiang TX, Reid WD, Belcastro A, Road JD. Load dependence of secondary diaphragm inflammation and injury after acute inspiratory loading. Am J Respir Crit Care Med 1998;157(1):230–6.
[17] Jiang TX, Reid WD, Road JD. Delayed diaphragm injury and diaphragm force production. Am J Respir Crit Care Med 1998;157(3 Pt 1):736–42.
[18] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357–62.
[19] McKinney W. Data structures for statistical computing in Python. 2010.
[20] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[21] Caswell TA, Lee A, Droettboom M, Sales de Andrade E, Hoffmann T, Klymak J, et al. Matplotlib/matplotlib: REL: v3.6.0. Zenodo. 2022.
[22] Waskom ML. Seaborn: statistical data visualization. J Open Source Softw 2021;6(60):3021.
[23] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–72.
[24] Riedel M, Sedona R, Barakat C, Einarsson P, Hassanian R, Cavallaro G, et al. Practice and experience in using parallel and scalable machine learning with heterogenous modular supercomputing architectures. In: 2021 IEEE international parallel and distributed processing symposium workshops (IPDPSW); 2021. p. 76–85.
[25] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. LightGBM: a highly efficient gradient boosting decision tree; 2017. p. 30.
[26] Guryanov A. Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In: van der Aalst WMP, Batagelj V, Ignatov DI, Khachay M, Kuskova V, Kutuzov A, et al., editors. Analysis of images, social networks and texts. Cham: Springer International Publishing; 2019. p. 39–50.
[27] Ambrosino N, Gabbrielli L. The difficult-to-wean patient. Expert Rev Respir Med 2010;4(5):685–92.
[28] Yang H, Hsu J, Chen Y, Jiang X, Chen T. Using support vector machine to construct a predictive model for clinical decision-making of ventilation weaning. In: 2008

IEEE international joint conference on neural networks (IEEE world congress on computational intelligence); 2008. p. 3981–6.

[29] Lin M-Y, Li C-C, Lin P-H, Wang J-L, Chan M-C, Wu C-L, et al. Explainable machine learning to predict successful weaning among patients requiring prolonged mechanical ventilation: a retrospective cohort study in Central Taiwan. Front Med 2021;8:663739.

[30] Hadjitodorov S, Todorova L. Consultation system for determining the patients' readiness for weaning from long-term mechanical ventilation. Comput Methods Programs Biomed 2010;100(1):59–68.

[31] Fritsch SJ, Dreher M, Simon T-P, Marx G, Bickenbach J. Haemoglobin value and red blood cell transfusions in prolonged weaning from mechanical ventilation: a retrospective observational study. BMJ Open Respir Res 2022;9(1):e001228.

[32] Houben PH, van der Weijden T, Winkens B, Winkens RA, Grol RP. Pretest expectations strongly influence interpretation of abnormal laboratory results and further management. BMC Fam Pract 2010;11:13.

[33] Bramer M. Avoiding overfitting of decision trees. In: Bramer M, editor. Principles of data mining. London: Springer London; 2007. p. 119–34.

[34] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.

[35] Ergan B, Oczkowski S, Rochwerg B, Carlucci A, Chatwin M, Clini E, et al. European respiratory society guidelines on long-term home non-invasive ventilation for management of COPD. Eur Respir J 2019;54(3):1901003.

[36] Lai CC, Shieh JM, Chiang SR, Chiang KH, Weng SF, Ho CH, et al. The outcomes and prognostic factors of patients requiring prolonged mechanical ventilation. Sci Rep 2016;6:28034.

[37] Vemuri SV, Rolfsen ML, Sykes AV, Takiar PG, Leonard AJ, Malhotra A, et al. Association between acute kidney injury during invasive mechanical ventilation and ICU outcomes and respiratory system mechanics. Crit Care Explor 2022;4(7): e0720.

[38] Kaushik S, Choudhury A, Sheron PK, Dasgupta N, Natarajan S, Pickett LA, et al. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. Front Big Data 2020:3.

[39] Forcey DS, Dyer JC, Hopper IK. Clinical concern and the deteriorating patient: a review of rapid response 2018&#x2013;20. Aust Health Rev 2022;46(6):679–85.