J Crit Care 89 (2025) 155105

Contents lists available at ScienceDirect

Journal of Critical Care

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

Development and validation of a machine learning model for real-time prediction of invasive mechanical ventilation weaning readiness

Simone Zappalà [a,1], Vittorio Scaravilli [b,c,*,1], Lucrezia Rovati [d,e], Marco Bosone [f,g], Francesca Alfieri [a], Andrea Ancona [a], Giacomo Grasselli [b,h]

[a] U-Care Medical s.r.l., 10129 Turin, Italy
[b] Department of Anesthesia and Critical Care, Fondazione IRCCS Ca' Granda - Ospedale Maggiore Policlinico, Milan, Italy
[c] Department of Biomedical Surgical and Dental Sciences, University of Milan, Milan, Italy
[d] Department of Emergency Medicine, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy
[e] School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy
[f] School of Medicine and Surgery, University of Milan, Milan, Italy
[g] Interdepartmental Division of Critical Care Medicine, University of Toronto, Toronto, Canada
[h] Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy

ARTICLE INFO

Keywords:
Machine-learning
Respiratory failure
Weaning
Invasive mechanical ventilation

ABSTRACT

Purpose: To develop and validate a bedside machine learning (ML) decision support tool for prediction of invasive mechanical ventilation (IMV) weaning readiness.
Methods: Adults admitted after 2010 who underwent IMV (>24 h) were included from MIMIC-IV (development and internal validation) and AmsterdamUMCdb (external validation) databases. XGBoost boosted trees approach was used to develop three models predicting IMV weaning readiness within 24, 48, and 72 h by integrating electronic health record data. The areas under Receiver Operating Characteristic (auROC), the Precision-Recall curve (auPR) curves, and performance metrics were assessed. Sensitivity analyses evaluated the impact of gender, ethnicity, age and admission reason on model performance.
Results: 8565 patients from MIMIC-IV and 2626 from AmsterdamUMCdb were included. In the external validation cohort, the 24-, 48-, and 72-h models had auROCs of 0.847, 0.795 and 0.789, and auPR of 54.17, 54.56 and 59.4, respectively. Sensitivity was >0.75 for all models, but specificity decreased from 0.79 to 0.63 between the 24-h and 72-h models. Lower performances were observed for older (> 60 years) and neurosurgical patients.
Conclusions: This study presents three ML models for real-time prediction of IMV weaning readiness, offering a promising approach to enhance clinical decision-making and optimize patient care.

1. Introduction

Approximately 50 % of patients admitted to the Intensive Care Unit (ICU) during non-pandemic periods require Invasive Mechanical Ventilation (IMV) for more than 48 h [1]. During the recent SARS-CoV-2 pandemic, the global health systems collapsed due to the consumption of finite IMV accessibility [2], and the next pandemic may lead to similar scenarios [3]. In this context, timely and successful weaning from IMV is paramount. Moreover, many complications of IMV, including ventilator-associated pneumonia, critical illness polyneuropathy, and post-ICU

Abbreviations: AI, Artificial Intelligence; AmsterdamUMCdb, Amsterdam University Medical Center database; auPR, area under the Precision-Recall Curve; auROC, area under the Receiving Operating Curve; CI, confidence interval; ECMO, extracorporeal membrane oxygenation; EHR, electronic health record; FiO2, Fraction of inspired oxygen; ICU, Intensive Care Unit; IMV, Invasive Mechanical Ventilation; LR+, positive likelihood ratio; LR-, negative likelihood ratio; ML, Machine Learning; PaO2, Partial pressure of oxygen in arterial blood; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; SBT, spontaneous breathing trial; SD, standard deviation; SHAP, Shapley Additive exPlanation; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence; XGBoost, eXtreme Gradient Boosting.

* Corresponding author at: Department of Anesthesia and Critical Care, Fondazione IRCCS Ca' Granda - Ospedale Maggiore Policlinico, Via della Commenda 9, 20122 Milan, Italy.
E-mail address: vittorio.scaravilli@unimi.it (V. Scaravilli).
[1] contributed equally to the study.

https://doi.org/10.1016/j.jcrc.2025.155105
Received 14 August 2024; Accepted 30 April 2025
Available online 27 May 2025

frailty, correlate directly with the duration of IMV [4–6] rather than to its application per se.

Machine learning (ML) algorithms are designed to learn from existing data and generate predictions on specific outcomes. In critical care, ML algorithms can analyze vast amounts of real-time electronic health records (EHRs) data, identifying multi-parametric time-varying patterns that may elude clinicians' observation. Therefore, ML can have significant clinical, economic, and societal importance, particularly in resource-constrained settings or future pandemic contexts [7].

Several publications have explored the use of ML to optimize IMV [8], including models to predict weaning readiness [9,10]. However, few studies have validated these models and assessed their practical application in real-world clinical scenarios [11]. In addition, many existing prediction models rely on clinical variables that are not readily available at the bedside.

With this retrospective analysis of open-access European and American ICU databases, we aimed to 1) develop an ML model for prediction of IMV weaning readiness; 2) validate this ML model internally and externally; 3) assess the predictive ability of the model in diverse ICU sub-cohorts. We hypothesize that by integrating clinically relevant data captured automatically from EHRs, we could create an innovative tool capable of continuously suggesting weaning readiness throughout the IMV period.

2. Material and methods

This retrospective observational study utilizes freely accessible datasets with fully anonymized data; thus, no informed consent or institutional review board approval was required. The study complies with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD+AI) Statement [12] (see Online Supplements, TRIPOD+AI statement). While the full code cannot be made publicly available due to proprietary constraints and a pending patent (European Patent Application No. 24202572.4), interested researchers may contact the corresponding author to request a demo for external validation. Furthermore, to enhance transparency and reproducibility, a detailed step-by-step description of data processing pipeline is provided (see Online Supplement, Model Architecture Reproducibility).

2.1. Data source and population

The data used in this study was derived from the following sources:

- Amsterdam University Medical Center database (AmsterdamUMCdb) [13], a database of patients admitted at the University Hospital ICU of Amsterdam, in the Netherlands, between 2003 and 2016, for a total of 23,106 ICU admissions related to 20,109 distinct patients.
- Medical Information Mart for Intensive Care IV (MIMIC-IV) v2.2 [14], a database of patients admitted at the Beth Israel Deaconess Medical Center in Boston (MA) between 2008 and 2022, for a total of 73,181 ICU admissions related to 50,920 distinct patients.

Both databases include only adult patients (age $\geq$ 18).

For the current study, patient inclusion criteria were at least one episode of IMV during ICU stay and ICU admission since 2010. Exclusion criteria were duration of IMV $<$ 24 h and extracorporeal membrane oxygenation (ECMO) use. Only the first period of IMV for each ICU stay was examined. The development cohort comprised 70 % of randomly drawn patients from the MIMIC-IV database. The remaining 30 % was used as internal validation, while the AmsterdamUMCdb cohort was used for the model's external validation.

2.2. Outcomes

IMV was defined as ventilator-support periods documented by a corresponding IMV log generated by the mechanical ventilator. The primary outcome was successful weaning from IMV, defined as either 1) absence of IMV logs for 48 consecutive hours or 2) discharge alive from the study ICU within 48 h from IMV discontinuation, whichever came first. Notably, AmsterdamUMCdb and MIMIC-IV do not account directly or explicitly for re-intubation and extubation failure. To overcome this limitation, in our analyses, consecutive logs closer than 48 h apart were considered part of the same ventilation event. Patients who died within 48 h of IMV discontinuation were not considered successfully weaned, and all their prediction time points were marked as negative.

2.3. Predictor variables

Model input variables were selected a priori based on clinical relevance and accessibility through the EHR of both AmsterdamUMCdb and MIMIC-IV datasets (see Online Supplement Table S1). Before being analyzed, identification codes and units of measure were made uniform. These variables included patient demographics (e.g., age, sex, weight), clinical observations (e.g., Sedation Score), vital signs (e.g., heart rate, arterial blood pressure), mechanical ventilator settings (e.g., FiO_2, ventilation mode), blood gas analyses, and laboratory studies. To minimize imprecision and errors in retrospective datasets, we excluded clinical observations like the Glasgow Coma Scale (GCS) due to the lack of standardized sedation data. The only judgment-based measure included was the sedation score, applied using the unified approach proposed by Pham et al. [1] to ensure consistency and accuracy. Validity ranges were applied to filter input variables. Missing data imputation employed a forward-filling approach: vital signs and ventilatory settings were forwarded for 12 h, laboratory variables for 4 days, and sedation scores were considered valid for 24 h. No further data imputation was performed.

ICU admission-to-discharge data was employed to construct our model for the feature engineering process, including temporal fluctuations and trends. For instance, features such as the time series "respiratory_rate_std_12h_mean" were created by analyzing respiratory rate measurements, calculating standard deviations hourly, and then utilizing a 12-h rolling window to determine the mean. After completing the feature engineering process, a total of 2992 features per each hour of IMV were obtained. Using the internal validation cohort, the models were refined with recursive feature elimination, selecting 252 features for the 24-h model, 1232 for the 48-h model, and 1576 for the 72-h model (see Online Supplement, Additional Methods, Table S2).

2.4. Model definition

The eXtreme Gradient Boosting (XGBoost) [15] implementation of boosted trees was used to train 3 distinct classifier models, named the 24-h, 48-h, and 72-h models. In the models' classification scheme, each patient's IMV period is a series of hourly spaced prediction time points. At each full hour of ventilation, the models generate a new prediction, analyzing the patient's entire history up to that point, performing feature engineering, and classifying the hour as likely (positive class) or unlikely (negative class) to be within the 24, 48, or 72-h window for successful IMV discontinuation. The models operate in an almost real-time manner, providing ongoing predictions of liberation from IMV

Each XGBoost model was optimized by adjusting key parameters, including learning rate, number of trees, maximum depth, regularization (reg_alpha), and subsample (data fraction), as well as colsample_bytree (features per tree).

2.5. Statistical analyses

Data were summarized with mean $\pm$ standard deviation (SD) or

median [interquartile range], as appropriate. Based on previous literature [16], a sample size >2000 subjects was considered adequate. Patient cohorts were compared with the chi-square test, the Fisher's test, or the Wilcoxon rank-sum test, as appropriate. The predictive performance of the models is assessed at each hour of ventilation using the area under the Receiving Operating Curve (auROC) and the area under the Precision-Recall Curve (auPR). Internal validation from the MIMIC-IV cohort was used for threshold selection, aiming for sensitivity >0.8. Thus, *for each hourly prediction*, the sensitivity, specificity, precision, positive likelihood ratio (LR+), negative likelihood ratio (LR-), and F1 score and associated bootstrapped confidence intervals (CI) were computed. Moreover, the *per-patient* "Lead Time" was calculated, indicating the time between model prediction and actual event occurrence. Shapley Additive exPlanation (SHAP) [17] values were utilized to quantify each feature's impact on prediction. Sensitivity analyses were performed to assess the predictive capabilities of the models within different patients' sub-cohorts and the impact of gender, ethnicity, age, and reason for admission on the models' performances.

3. Results

3.1. Cohort and data preprocessing

Fig. 1 summarizes the patient population inclusion flowchart. AmsterdamUMCdb and MIMIC-IV datasets comprised 23,106 and 73,181 ICU admissions, respectively. Of those, 17,428 and 30,590 cases were managed with IMV, respectively. After applying the exclusion criteria, 2626 and 8565 ICU admissions were included from AmsterdamUMCdb and MIMIC-IV, respectively, for a total of 11,191 ICU admissions. The MIMIC-IV cohort was then randomly split into a development cohort (5995 ICU stays) and an internal validation cohort (2570 ICU stays). This resulted in 151,140 prediction points for development, 66,324 for internal testing, and 396,871 for external testing.

The patient characteristics and their comparison between development, internal and external validation cohorts are depicted in Table 1. Of note, in the external validation cohort, patients were more frequently male and older. Moreover, the external validation cohort showed longer IMV and less successful weaning.

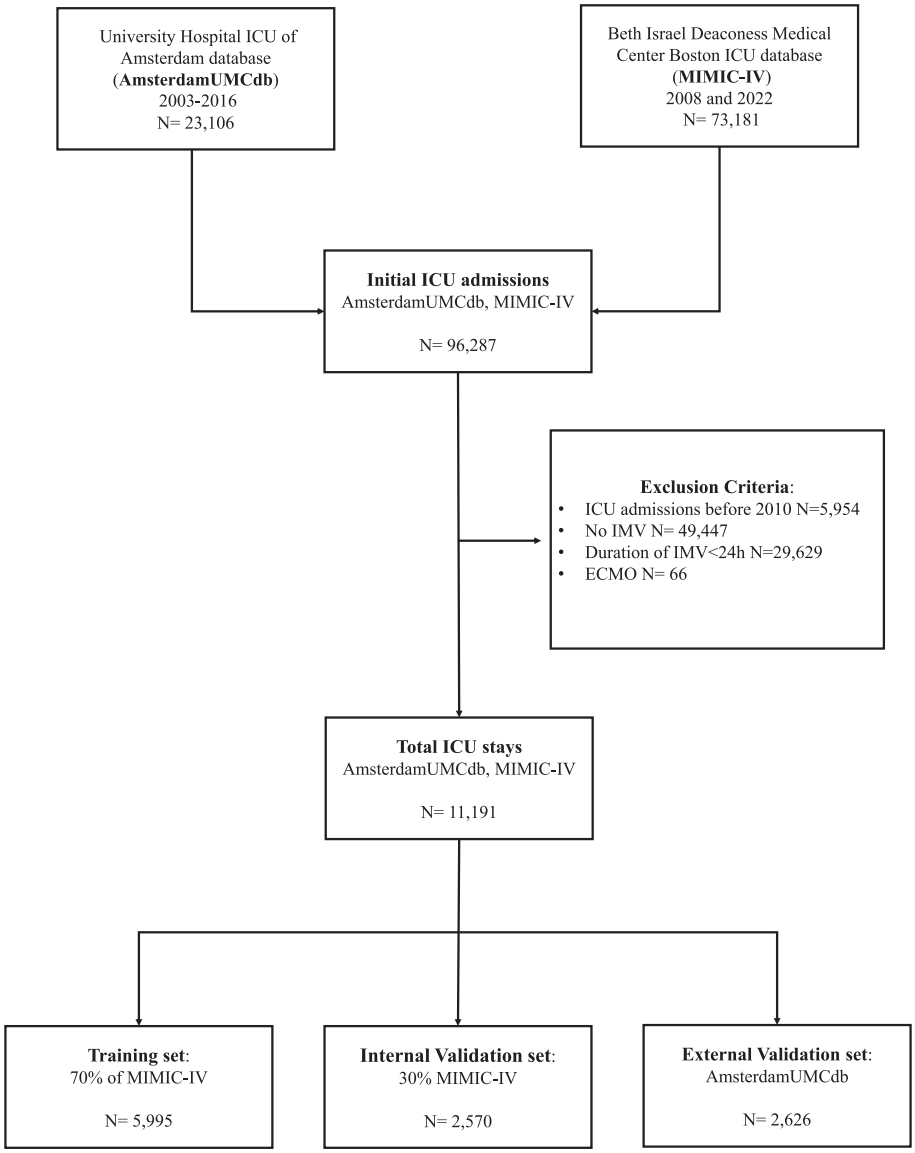

Fig. 1. Patients cohort flowchart.
ICU, Intensive Care Unit; MIMIC-IV, Medical Information Mart for Intensive Care IV; AmsterdamUMCdb: Amsterdam University Medical Center database; ICU, Intensive Care Unit; IMV, Invasive Mechanical Ventilation; ECMO, Extracorporeal Membrane Oxygenation.

Table 1
Patients demographics.

Parameter		MIMIC-IV Development (n = 5995)	MIMIC-IV Internal (n = 2570)	p MIMIC-IV Development vs. Internal	AUMCdb External (n = 2626)
Sex	Male	3473 (57.9 %)	1485 (57.78 %)	0.904	1724 (65.65 %)*
	Female	2522 (42.07 %)	1085 (42.22 %)		902 (34.35 %)*
Ethnicity	Caucasian	3130 (52.21 %)	1344 (52.30 %)	0.573	
	African-American	630 (10.51 %)	264 (10.27 %)		
	Hispanic	175 (2.92 %)	89 (3.46 %)		
	Asian	141 (2.35 %)	51 (1.98 %)		
	Other/Unknown	1858 (30.99 %)	798 (31.05 %)		
Age	Mean (std)	64.56 (16)	64.46 (15.61)	0.329	
	18–39	492 (8.21 %)	200 (7.78 %)		295 (11.23 %)*
	40–49	506 (8.44 %)	206 (8.02 %)		268 (10.21 %)*
	50–59	1061 (17.70 %)	480 (18.68 %)		470 (17.90 %)*
	60–69	1455 (24.27 %)	659 (25.64 %)		682 (25.97 %)*
	70–79	1359 (22.67 %)	586 (22.80 %)		644 (24.52 %)*
	80+	1122 (18.72 %)	439 (17.08 %)		267 (10.17 %)*
ICU Type	ICU	1560 (26.02 %)	685 (26.65 %)	0.408	1961 (74.68 %)*
	Cardiovascular	790 (13.18 %)	297 (11.56 %)		
	Coronary Care and Cardiothoracic	484 (8.07 %)	220 (8.56 %)		
	Neuro	240 (4.00 %)	97 (3.77 %)		
	Surgical	1989 (33.18 %)	859 (33.42 %)		
	Trauma	932 (15.55 %)	412 (16.03 %)		
Reason for Admission	ICU and MCU				640 (24.37 %)
	MCU				25 (0.95 %)
	Cardiovascular				289 (11.01 %)
	Neurological				118 (4.49 %)
	Respiratory				502 (19.12 %)
	Sepsis				510 (19.42 %)
	Surgical				723 (27.53 %)
	Trauma				271 (10.32 %)
	Other/Unknown				635 (24.18 %)
	ICU LOS	6.4 [3.8–11.2]	6.38 [3.76, 11.2]	0.959	7.8 [3.92, 16.54] *
	IMV Duration	3.0 [1.6–6.2]	3.04 [1.62, 6.03]	0.799	4.33 [2.09, 9.28] *
	Successful Weaning	4856 (81.0 %)	2081 (80.9 %)	0.976	2005 (76.3 %) *

MIMIC-IV: Medical Information Mart for Intensive Care IV; AmsterdamUMCdb: Amsterdam University Medical Center database; ICU: intensive care unit; IC: intensive care unit; MCU: medium care unit; LOS: length of stay; IMV: invasive mechanical care. *$p < 0.001$ vs. MIMIC-IV development and validation cohorts.

3.2. Model performance

The MIMIC-IV development cohort included 151,140 prediction time points. Table S3 (see Online Supplements, Additional Results) describes the performance metrics of the models in the development cohort. Prevalence rates for positive prediction points (i.e., a time sample associated with possible successful weaning) at the 24-h, 48-h, and 72-h frames were 16.42 %, 27.07 %, and 35.30 %, respectively. In the MIMIC-IV validation cohort, among the 2081 successfully weaned patients, 2052 (98.61 %), 2059 (98.94 %), and 2061 (99.04 %) received at least one coherent prediction from the 24-, 48-, and 72-h models. In the AmsterdamUMCdb external validation cohort, among 2005 successfully weaned patients, 1984 (98.95 %), 1987 (99.10 %), and 1986 (99.05 %) received at least one coherent prediction from the 24, 48, and 72-h models. Table 2 and Fig. 2 depict the predictive performance of the three models for the validation cohorts. Overall, the auROC and the auPR were higher for the MIMIC-IV internal validation cohort vs. the Amsterdam UMCdb external validation cohort while maintaining high diagnostic accuracy (> 0.75 AUROC and > 0.5 auPR) in predicting successful weaning. Table 3 outlines model performance at an 80 % sensitivity threshold on the internal validation cohort. As expected, the models performed better in the internal validation cohort than in the external validation cohorts. Still, the models maintained a sensitivity >0.75 for each time endpoint in the external validation cohort.

Table 2
Performance metrics for the models measured on time samples of the validation cohorts.

Endpoint	Cohort	Prediction Time Points	Prevalence of Positive Prediction Time Points	auROC IQR	auPR IQR
24 h	MIMIC-IV Internal	66,324	16.30 %	91.70 [91.41, 91.99]	76.16 [75.43, 76.90]
	AmsterdamUMCdb External	396,871	9.48 %	84.69 [84.47, 84.90]	54.17 [53.65, 54.68]
48 h	MIMIC-IV Internal	66,324	27.03 %	88.84 [88.57, 89.11]	78.00 [77.43, 78.54]
	AmsterdamUMCdb External	396,871	16.92 %	79.52 [79.32, 79.71]	54.56 [54.14, 54.93]
72 h	MIMIC-IV Internal	66,324	35.23 %	86.93 [86.65, 87.21]	80.33 [79.85, 80.79]
	AmsterdamUMCdb External	396,871	23.25 %	78.93 [78.76, 79.09]	59.4 [59.07, 59.70]

MIMIC-IV: Medical Information Mart for Intensive Care IV; AmsterdamUMCdb: Amsterdam University Medical Center database; auROC: Area under the Receiver Operating Characteristic; auPR: Area under the precision recall; IQR: interquartile range.

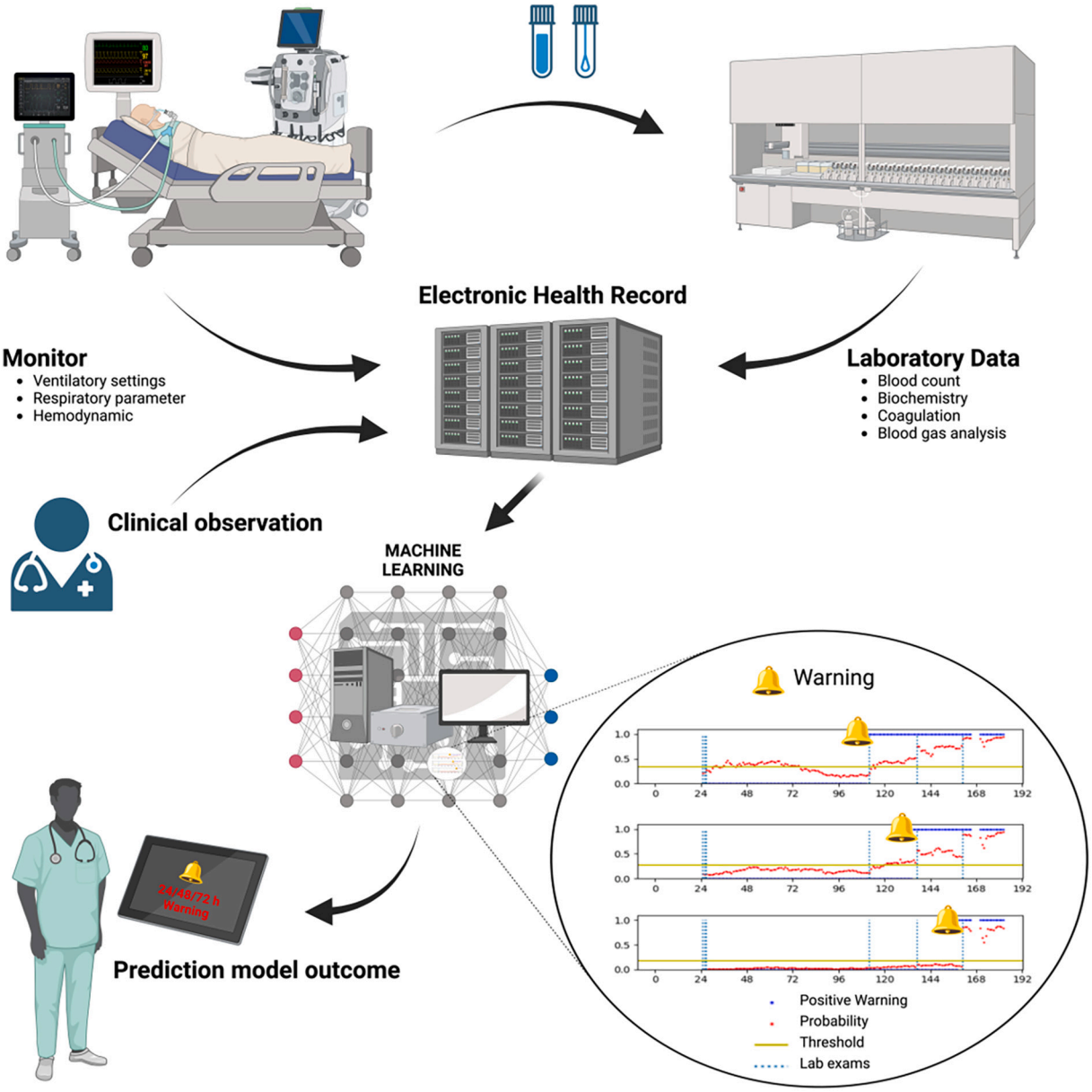

Fig. 2. Example of the model prediction framework.
Electronic Health Records are integrated, and the machine learning algorithm calculates the real-time probability of successful liberation from IMV within the 24, 48, or 72-h timeframes without any input from the physician. The likelihood of successful weaning is not displayed. Instead, a negative or positive "warning" is provided directly into the electronic medical chart and/or clinical monitor if the probability exceeds predefined sensitivity thresholds. In the *example derived from actual clinical data*, at hour 35, the 72-h model predicts liberation within the next 72 h, resulting in a false positive warning. By hour 120, the 72-h model surpasses the liberation threshold, the 48-h model also crosses its threshold, and the 24-h model remains below. This suggests liberation is likely within 48–72 h but not within the next day. At hour 140, new lab results prompt an update, increasing the liberation probability for both the 72-h and 48-h models. By hour 160, all models exceed their thresholds, indicating imminent liberation.

Specificity decreased from 0.79 to 0.63 from the 24-h to the 72-h model. Interestingly, the F1-score increased from the 24-h to the 72-h model, between 0.39 and 0.52. The models had overall high positive likelihood ratios.

3.3. Feature selection and importance

Table S2 (see Online Supplement, Additional Results) describes the features generated as input for the models. The distribution of SHAP values of the 20 most predictive features is given in Fig. S1. A significant effect on the predicted probability of weaning readiness was documented for sedation score, lactate, and several respiratory parameters (i.e., PaO_2, Plateau Pressure, respiratory rate). Additionally, the collective influence of remaining features still significantly shapes the models outcomes.

3.4. Lead-time

In the internal validation cohort, the median lead time was 19 [6–45] hours, while in the external validation cohort, the median lead time was 29 [6–80] hours. Fig. S2 represents the lead time analysis for the 24-h model stratified by ventilation duration. The lead time was consistently stable even for very long ventilation periods.

Table 3
Performance metrics for the 80 % sensitivity selected threshold.

Endpoint	Validation Cohorts	F1 score	Sensitivity	Specificity	Precision	Positive LR	Negative LR	Median Lead Time (hours)	Lead Time IQR (hours)
24 h	MIMIC-IV	0.63 [0.63–0.64]	0.8 [0.79–0.81]	0.86 [0.86–0.86]	0.52 [0.52–0.53]	5.66 [5.54–5.79]	0.23 [0.22–0.24]	19	[6.4]
	AmsterdamUMCdb	0.39 [0.39–0.39]	0.73 [0.73–0.74]	0.79 [0.79–0.79]	0.27 [0.26–0.27]	3.47 [3.44–3.5]	0.34 [0.33–0.34]	29	[8.9]
48 h	MIMIC-IV	0.69 [0.68–0.69]	0.8 [0.79–0.81]	0.8 [0.8–0.81]	0.6 [0.6–0.61]	4.09 [4.01–4.17]	0.25 [0.24–0.26]	30	[9.6]
	AmsterdamUMCdb	0.44 [0.44–0.45]	0.75 [0.74–0.75]	0.67 [0.67–0.67]	0.32 [0.31–0.32]	2.27 [2.25–2.28]	0.38 [0.37–0.38]	51	[15.1]
72 h	MIMIC-IV	0.72 [0.71–0.72]	0.8 [0.79–0.81]	0.77 [0.76–0.77]	0.65 [0.64–0.66]	3.41 [3.34–3.47]	0.26 [0.25–0.27]	35	[11.8]
	AmsterdamUMCdb	0.52 [0.51–0.52]	0.77 [0.77–0.78]	0.63 [0.63–0.63]	0.39 [0.39–0.39]	2.09 [2.08–2.1]	0.36 [0.35–0.36]	66	[20.1]

MIMIC-IV: Medical Information Mart for Intensive Care IV; AmsterdamUMCdb: Amsterdam University Medical Center database; LR: likelihood ratio; IQR: interquartile range.

3.5. Sensitivity analyses

Several sensitivity analyses were conducted (detailed in Tables S4-S9, Online Supplement, Additional Results). In the internal validation cohort, the models showed auROC>0.85 and auPR >0.70 for all the timeframes across all the patients' sub-cohorts (i.e., gender, ICU type, comorbidities, and ethnicities). In the external validation cohort, the model had auROC >0.8 and auPR >0.5 for all the timeframes and patients' sub-cohorts. Notably, the models yielded better results for patients younger than 60, although auROC remained above 0.8 even for older patients. Conversely, patients admitted for neurological diseases consistently exhibited lower model performance across all models and validation cohorts.

4. Discussion

In this paper, we describe and validate a predictive model that provides clinical decision support for optimizing the timing of weaning attempts in patients undergoing IMV. By integrating clinical data from EHRs (i.e., monitors, mechanical ventilator, blood gas analyses, and laboratory data), the model provides a continuous, physician-unbiased assessment of the probability of successful weaning from IMV within the 24, 48, and 72-h timeframes. As shown in Fig. 3, these probabilities would be automatically generated in real time relying on data from the EHR. Data transmission would be carried out using the HL7 standard, ensuring secure and standardized communication.

The model was trained on data from over 5000 patients sourced from an open-access database and demonstrated robust predictive performance on an equally large cohort of patients from a separate dataset. Moreover, the model provided consistent predicting accuracy across diverse patients' sub-cohorts.

Timely weaning from IMV is of utmost importance in managing critically ill patients. While life-saving, IMV is associated with several complications causally linked to its duration [18] rather than its application per se. Recent findings [1] indicate that unnecessarily prolonged ventilation increases (rather than reduces) the risk of unsuccessful weaning. Several indexes, scores, and predictive tools have been tested to assess IMV weaning readiness [19,20], but no gold standard exists. Overall, these approaches are constrained by the limited information utilized for their calculation, which hampered their performances. In fact, the current best practice consists of simply testing whether spontaneous ventilation is possible by applying a spontaneous breathing trial (SBT) [21]. Consistent and timely application of a SBT has several limitations. Specifically, the patient and time selection for SBT may be biased by prior physician knowledge and expertise. Moreover, SBT might be particularly demanding in resource-constrained scenarios like pandemics. Thus, AI-based models for predicting IMV weaning readiness have been developed to overcome those limitations, supported by recent improvements in computing power, increased digitalization, and clinical data integration [22–25]. While earlier models used simpler logistic

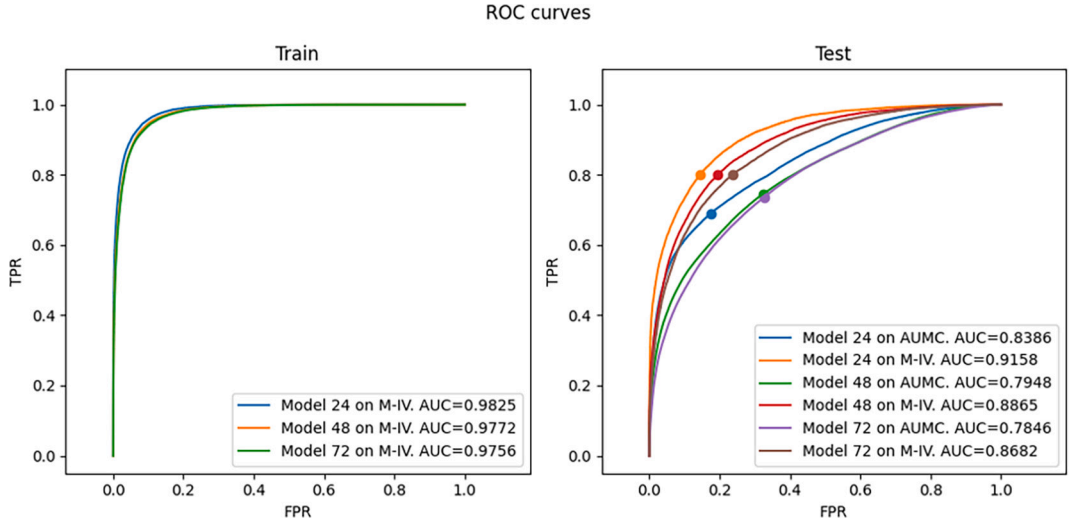

Fig. 3. Area under the Receiver Operating Characteristic of the models for the internal and external validation cohorts.
M-IV: Medical Information Mart for Intensive Care IV; AUMC: Amsterdam University Medical Center database; AUC: area under the curve.

regression for prediction [26], recent works developed models using Artificial Neural Networks, with performances constrained by small patient cohorts [25]. Comparing these models' performance is challenging due to different prediction endpoints, timing, and patient inclusion criteria. The key limitation of previous works is the absence of proper external validation [8]. Among the few externally validated models, Zhao et al. [27] created a model for anticipating extubation failure, which relies on static clinical variables. Such a model may be intrinsically biased since it is utilized only if and when the physician attempts weaning.

In contrast, our model possesses several strengths that enhance its potential for immediate clinical applicability. It was trained on a large, well-validated general ICU cohort and externally validated on a separate cohort, with no need for recalibration. The XGBoost [15] implementation of boosted trees is particularly robust to missing data, enabling the model to function effectively without the need for new data collection efforts. This allows each installation to seamlessly utilize routinely collected data, facilitating its integration into existing clinical workflows. To comply with the most recent developments in ventilation management, we included only patients admitted to ICU after 2010. However, we did not restrict the model development to specific patient sub-cohorts. A wide array of clinical variables was selected, including parameters directly obtainable from EHRs without additional interventions or maneuvers in the patient's clinical management. The variables were then subjected to extensive feature engineering, incorporating clinical expert knowledge before applying machine learning techniques through close collaboration between data scientists and experienced intensivists in our research team. This allowed us to leverage clinical expert knowledge to focus the ML algorithms on clinically meaningful relationships, which may also help foster the acceptance of ML in clinical practice.

The performance of the models in the internal and external validation cohorts appeared robust, with high auROC and auPR, despite some statistically significant differences between the cohorts, indicating model accuracy and generalizability across new patient populations. The threshold for binary classification was selected using the internal validation cohort, and no calibration was performed. An important aspect of the model's real-world applicability is lead time, the interval between the first positive prediction and actual IMV discontinuation. Our analysis shows that lead time remains stable across different ventilation durations, indicating that the model consistently provides early predictions. This could help clinicians anticipate weaning readiness, enabling better preparation for extubation and resource allocation. Sensitivity analyses showed that the model maintained high accuracy and precision across diverse patient sub-cohorts, with comparable metrics among different sexes, ethnicities, ages, and reasons for admission. Notable exceptions were the slightly reduced performance in older patients and the limited applicability of the model in neurosurgical/neurological patients. Interpreting those results deserves further consideration. It is known that age, in and of itself, is not a predictor of unsuccessful weaning from mechanical ventilation [28]. Instead, higher age is commonly associated with higher burdens of comorbidities [29]. Since comorbidities were not consistently reported in the development and validation databases, they were not introduced in variable engineering. For the ML algorithm, age is the model's only variable accounting for comorbidities, and thus, higher age might introduce variability in outcomes that cannot be accurately predicted. Further improvement in the predictive performances of the model might, therefore, be obtained by introducing comorbidities as inputs in the model. The limited efficacy of the model in predicting outcomes of neurological/neurosurgical patients is because weaning feasibility in this cohort is primarily affected by the severity of the primary brain injury, whose severity was not documented and not considered in model engineering. Such limitations in the model may be improved by inputting further brain injury scores other than the Glasgow Coma Scale.

Our study has several limitations inherent to data sources and big data analysis. Firstly, its retrospective nature necessitates prospective validation to ensure robustness. Additionally, a prospective interventional study is needed to assess how the model's availability impacts clinical outcomes. We envision the possibility of conducting a prospective interventional study wherein patients undergoing IMV for >48 h are randomized to ML-assisted weaning vs. standard of care and their IMV length (and associated complications) compared. Still, for the moment being, the model is preclinical and exploratory. In this report, we did not aim at assessing whether the model might perform better than existing models or comparably to clinical assessment. Further prospective observational studies are necessary to assess these hypotheses. Secondly, to ensure our model's accuracy, data should be provided with hourly frequency, and imputation of missing data may introduce errors; however, we mitigated this issue using a forward-filling methodology consistent with previous investigations. Moreover, tracheostomy status was not explicitly considered, as we aimed to minimize reliance on clinical notes and instead focused on IMV duration rather than specific intubation or extubation events. Additionally, certain variables exhibit high importance scores, their clinical interpretation and causality warrant further investigation. The model's outcomes suggest that feature importances are not concentrated on specific medical variables or a small subgroup. Finally, further efforts will be required to develop a user-friendly graphical user interface.

5. Conclusions

We developed and validated three ML-based models with high accuracy in predicting the timing of IMV successful weaning. The models provide a continuous, hour-by-hour, physician-unbiased decision support tool readily implementable in current clinical practice workflow. Further, prospective observational studies are necessary to assess the models' performances in real-life scenarios.

CRediT authorship contribution statement

Simone Zappalà: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Vittorio Scaravilli:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation. **Lucrezia Rovati:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Marco Bosone:** Writing – review & editing, Visualization, Investigation. **Francesca Alfieri:** Writing – original draft, Methodology, Conceptualization. **Andrea Ancona:** Writing – review & editing, Writing – original draft, Resources, Project administration, Funding acquisition, Conceptualization. **Giacomo Grasselli:** Writing – review & editing, Writing – original draft, Supervision.

Financial support

This study was (partially) funded by Italian Ministry of Health—Current Research IRCCS.

Declaration of competing interest

VS and GG have no competing interest regarding the submitted paper. LR has received consulting fees from U-Care Medical for this research work. AA is CEO of U-Care Medical. FA and AA are shareholder of U-Care Medical. FA and SZ are employees of U-Care Medical. AA, FA and SZ filed for the European Patent Application No. 24202572.4 "System and method for management and prediction of invasive mechanical ventilation necessity".

Acknowledgments

We thank dr. Sebastiano Colombo for his invaluable graphic support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcrc.2025.155105.

References

[1] Pham T, Heunks L, Bellani G, Madotto F, Aragao I, Beduneau G, et al. Weaning from mechanical ventilation in intensive care units across 50 countries (WEAN SAFE): a multicentre, prospective, observational cohort study. Lancet Respir Med 2023;11:465–76. https://doi.org/10.1016/S2213-2600(22)00449-0.

[2] Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. JAMA 2020;323:1545–6. https://doi.org/10.1001/jama.2020.4031.

[3] WHO. WHO Respiratory pathogens: pandemic preparedness guidance n.d. https://cdn.who.int/media/docs/default-source/global-influenza-programme/who_respiratory_pathogens_011222.pdf?sfvrsn=44032ec7_3 (accessed June 26, 2024).

[4] Papazian L, Klompas M, Luyt C-E. Ventilator-associated pneumonia in adults: a narrative review. Intensive Care Med 2020;46:888–906. https://doi.org/10.1007/s00134-020-05980-0.

[5] Zink W, Kollmar R, Schwab S. Critical illness polyneuropathy and myopathy in the intensive care unit. Nat Rev Neurol 2009;5:372–9. https://doi.org/10.1038/nrneurol.2009.75.

[6] Herridge MS, Chu LM, Matte A, Tomlinson G, Chan L, Thomas C, et al. The RECOVER program: disability risk groups and 1-year outcome after 7 or more days of mechanical ventilation. Am J Respir Crit Care Med 2016;194:831–44. https://doi.org/10.1164/rccm.201512-2343OC.

[7] Zwerwer LR, van der Pol S, Zacharowski K, Postma MJ, Kloka J, Friedrichson B, et al. The value of artificial intelligence for the treatment of mechanically ventilated intensive care unit patients: an early health technology assessment. J Crit Care 2024;82:154802. https://doi.org/10.1016/j.jcrc.2024.154802.

[8] Stivi T, Padawer D, Dirini N, Nachshon A, Batzofin BM, Ledot S. Using artificial intelligence to predict mechanical ventilation weaning success in patients with respiratory failure, including those with acute respiratory distress syndrome. J Clin Med 2024;13:1505. https://doi.org/10.3390/jcm13051505.

[9] Liu C-F, Hung C-M, Ko S-C, Cheng K-C, Chao C-M, Sung M-I, 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. https://doi.org/10.3389/fmed.2022.935366.

[10] Hsu J-C, Chen Y-F, Chung W-S, Tan T-H, Chen T, Chiang JY. Clinical verification of a clinical decision support system for ventilator weaning. Biomed Eng Online 2013;12:S4. https://doi.org/10.1186/1475-925X-12-S1-S4.

[11] Gallifant J, Zhang J, del Pilar Arias Lopez M, Zhu T, Camporota L, Celi LA, et al. Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias. Br J Anaesth 2022;128:343–51. https://doi.org/10.1016/j.bja.2021.09.025.

[12] 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:e078378. https://doi.org/10.1136/bmj-2023-078378.

[13] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam university medical centers database (AmsterdamUMCdb) example. Crit Care Med 2021;49:e563–77. https://doi.org/10.1097/CCM.0000000000004916.

[14] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x.

[15] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. vol. 19; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[16] Rajput D, Wang W-J, Chen C-C. Evaluation of a decided sample size in machine learning applications. BMC Bioinf. 2023;24:48. https://doi.org/10.1186/s12859-023-05156-9.

[17] Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Prog Biomed 2022;214:106584. https://doi.org/10.1016/j.cmpb.2021.106584.

[18] Schmidt GA, Girard TD, Kress JP, Morris PE, Ouellette DR, Alhazzani W, et al. Official executive summary of an American Thoracic Society/American College of Chest Physicians Clinical Practice Guideline: liberation from mechanical ventilation in critically ill adults. Am J Respir Crit Care Med 2017;195:115–9. https://doi.org/10.1164/rccm.201610-2076ST.

[19] Lewis KA, Chaudhuri D, Guyatt G, Burns KEA, Bosma K, Ge L, et al. Comparison of ventilatory modes to facilitate liberation from mechanical ventilation: protocol for a systematic review and network meta-analysis. BMJ Open 2019;9:e030407. https://doi.org/10.1136/bmjopen-2019-030407.

[20] Blackwood B, Alderdice F, Burns K, Cardwell C, Lavery G, O'Halloran P. Use of weaning protocols for reducing duration of mechanical ventilation in critically ill adult patients: Cochrane systematic review and meta-analysis. BMJ 2011;342:c7237. https://doi.org/10.1136/bmj.c7237.

[21] Burns KEA, Khan J, Phoophiboon V, Trivedi V, Gomez-Builes JC, Giammarioli B, et al. Spontaneous breathing trial techniques for extubating adults and children who are critically ill. JAMA Netw Open 2024;7:e2356794. https://doi.org/10.1001/jamanetworkopen.2023.56794.

[22] Fabregat A, Magret M, Ferré JA, Vernet A, Guasch N, Rodríguez A, et al. A machine learning decision-making tool for extubation in intensive care unit patients. Comput Methods Prog Biomed 2021;200:105869. https://doi.org/10.1016/j.cmpb.2020.105869.

[23] Chen T, Xu J, Ying H, Chen X, Feng R, Fang X, et al. Prediction of Extubation failure for intensive care unit patients using light gradient boosting machine. IEEE Access 2019;7:150960–8. https://doi.org/10.1109/ACCESS.2019.2946980.

[24] Hsieh M-H, Hsieh M-J, Chen C-M, Hsieh C-C, Chao C-M, Lai C-C. An artificial neural network model for predicting successful Extubation in intensive care units. J Clin Med 2018;7:240. https://doi.org/10.3390/jcm7090240.

[25] Kuo H-J, Chiu H-W, Lee C-N, Chen T-T, Chang C-C, Bien M-Y. Improvement in the prediction of ventilator weaning outcomes by an artificial neural network in a medical ICU. Respir Care 2015;60:1560–9. https://doi.org/10.4187/respcare.03648.

[26] Zeggwagh AA, Abouqal R, Madani N, Zekraoui A, Kerkeb O. Weaning from mechanical ventilation: a model for extubation. Intensive Care Med 1999;25:1077–83. https://doi.org/10.1007/s001340051015.

[27] Zhao Q-Y, Wang H, Luo J-C, Luo M-H, Liu L-P, Yu S-J, et al. Development and validation of a machine-learning model for prediction of Extubation failure in intensive care units. Front Med 2021:8. https://doi.org/10.3389/fmed.2021.676343.

[28] Stieff KV, Lim F, Chen L. Factors influencing weaning older adults from mechanical ventilation: an integrative review. Crit Care Nurs Q 2017;40:165–77. https://doi.org/10.1097/CNQ.0000000000000154.

[29] de Hesselle ML, Borgmann S, Rieg S, Vehreschild JJ, Rasch S, Koll CEM, et al. Age and comorbidity burden of patients critically ill with COVID-19 affect both access to and outcome of ventilation therapy in intensive care units. J Clin Med 2023:12. https://doi.org/10.3390/jcm12072469.