UNIVERSITÀ DEGLI STUDI DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Corso di Laurea Magistrale a ciclo unico
in Medicina e Chirurgia

DIPARTIMENTO DI SCIENZE CARDIO-TORACO-VASCOLARI E SANITÀ PUBBLICA
Direttore: Prof. Federico Rea

UNITÀ DI BIOSTATISTICA, EPIDEMIOLOGIA E SALUTE PUBBLICA
Direttore: Prof. Dario Gregori

TESI DI LAUREA

Artificial Intelligence for the prediction of weaning readiness outcome in a multi-central clinical cohort of mechanically ventilated patients

Relatore:

Corrado Lanera, Ph.D.

Laureando/a:

Andrea Pedot

ANNO ACCADEMICO 2022/2023

Università degli Studi di Padova

Università
degli Studi
di Padova

Università degli Studi di Padova

Corso di Laurea Magistrale a ciclo unico
in Medicina e Chirurgia

Dipartimento di Scienze Cardio-Toraco-Vascolari e Sanità pubblica
Direttore: Prof. Federico Rea

Unità di Biostatistica, Epidemiologia e Salute Pubblica
Direttore: Prof. Dario Gregori

TESI DI LAUREA

Artificial Intelligence for the prediction of weaning readiness outcome in a multi-central clinical cohort of mechanically ventilated patients

Relatore:

Corrado Lanera, Ph.D.

Laureando/a:

Andrea Pedot

Università
degli Studi
di Padova

TABLE OF CONTENTS

ABSTRACT [STRUCTURED SUMMARY]

When someone suffers from acute respiratory failure, mechanical ventilation (MV) is performed until they can breathe on their own again. The doctor checks every day whether the MV can be stopped. This screening consists of a first phase, the Readiness Testing (RT), which includes various clinical parameters. If this test is successful, 30 minutes of spontaneous breathing (SBT) is attempted. If also the SBT is passed successfully, the VM is stopped. On the contrary, if RT or SBT fails, the patient will be re-evaluated the next day. So, every day three mutually exclusive scenarios may happen: SBT will not be attempted, SBT will fail, or SBT will succeed.

Our artificial intelligence model is designed to infer early in the morning which of the three scenarios will probably occur during the day, starting from the patient's clinical data, from the information collected in the previous day’s clinical diary, and from whole minute-by-minute recording history of the various parameters of the mechanical ventilator, coming from a retrospective observational multi-central study, conducted in Italy over a course of 27 months.

Those data are processed with a deep learning approach, through a multi-source neural network topology, powered by multiple recurrent architectures. Hyper-parameters are optimized to select the purposed model through cross-validation, setting aside 36 out of 182 patients for testing final model performance over a variety of metrics, including a custom score designed to highlight clinical impact.

The final AI model had an accuracy of 79% [74, 83%], a custom score of 0.01 [-0.04, 0.05], a MCC of 0.28 [0.17, 0.39], scoring better than the other comparison models, including XG Boost that was trained on daily and baseline clinical data of the previous day only, which had an accuracy of 61% [56%, 66%], a MCC of 0.14 [0.06, 0.2] and a custom score of -0.05 [-0.08, -0.01].

Overall, AI model could approximate well what is the current clinical management throughout day-by-day providing suggestions early in the morning. Moreover, there are still space to improve the model clinical utility considering additional tailored training data.

RIASSUNTO

Quando un paziente soffre di insufficienza respiratoria acuta, viene praticata la ventilazione meccanica (VM) finché questa non riesce a respirare di nuovo in autonomia. Il medico di Terapia Intensiva verifica ogni giorno se la VM può essere interrotta. Questo screening consiste in una prima fase, il Readiness Test (RT), che è composta da vari parametri clinici. Se questo test ha esito positivo, si sottopone il paziente a 30 minuti di respirazione spontanea (SBT). Se anche l'SBT viene superato con successo, la VM viene interrotta. Al contrario, se l’RT o l’SBT falliscono, il paziente rimane in VM e verrà rivalutato il giorno successivo. Quindi ogni giorno possono verificarsi tre scenari mutuamente esclusivi: l’SBT non verrà tentato, l’SBT fallirà o l’SBT avrà successo (portando quindi all’estubazione del paziente).

Il modello di intelligenza artificiale sviluppato, è progettato per dedurre fin dalle prime ore del mattino quale dei tre scenari si verificherà probabilmente nel corso della giornata, partendo dai dati clinici del paziente, dalle informazioni raccolte nel diario clinico dei giorni precedenti e dall'intera storia di registrazione minuto-per-minuto dei vari parametri del ventilatore meccanico, provenienti da uno studio osservazionale retrospettivo multicentrico, condotto in Italia nel corso di 27 mesi.

Questi dati vengono elaborati con un approccio di Deep Learning, attraverso una topologia di rete neurale multi-sorgente, alimentata da architetture ricorrenti multiple. Gli iper-parametri sono ottimizzati per selezionare il modello desiderato attraverso la convalida incrociata, riservando 36 pazienti su 182 per testare le prestazioni finali del modello su una serie di metriche, tra cui uno score personalizzato progettato per evidenziare l'impatto clinico.

Il modello di intelligenza artificiale finale mostra un'accuratezza del 79% [74, 83%], uno score personalizzato di 0,01 [-0,04, 0,05], un MCC di 0,28 [0,17, 0,39], ottenendo un punteggio migliore rispetto agli altri modelli di confronto, tra cui XG Boost, addestrato solo sui dati clinici giornalieri del giorno precedente, che ha avuto un'accuratezza del 61% [56%, 66%], un MCC di 0,14 [0,06, 0,2] e uno score personalizzato di -0,05 [-0,08, -0,01].

Complessivamente, il modello di intelligenza artificiale è in grado di approssimare bene l'attuale gestione clinica giorno per giorno, fornendo suggerimenti al mattino presto. Inoltre, c'è ancora spazio per migliorare l'utilità clinica del modello considerando ulteriori dati di addestramento personalizzati.

INTRODUCTION

Rationale

[Clinical goal] Invasive Mechanical Ventilation (MV) is a life-saving medical procedure that supports a patient with Acute Respiratory Failure (ARF)(1), which can result from a wide variety of underlying diseases (2), as detailed in Table I. Once it is determined that a patient needs MV to support lung function while the primary cause of the disease is addressed, Intensive Care Unit (ICU) personnel's multiple choices should address. The topic is well covered in guidelines, both for ventilation mode setting (3), day-by-day management (4), and weaning, which address the safe withdrawal from MV (5–7). Weaning refers to the process of progressively reducing support provided by MV. Both prolonging MV and premature withdrawal could impact patient outcomes for the worst(8). This is why optimal weaning strategies have been proposed and validated (9–13). Current practice is a multi-step approach, which is composed of a daily screening liberation assessment composed of Readiness testing (RT) from clinical and laboratory criteria, which triggers a 30-minute Spontaneous Breathing Trial (SBT) (5,14,15). Only when the SBT succeeds the MV can be withdrawn.

Figure 1 - Stages occurring in a mechanically ventilated patient. ARF: acute respiratory failure; SBT: spontaneous breathing Trial. (16)

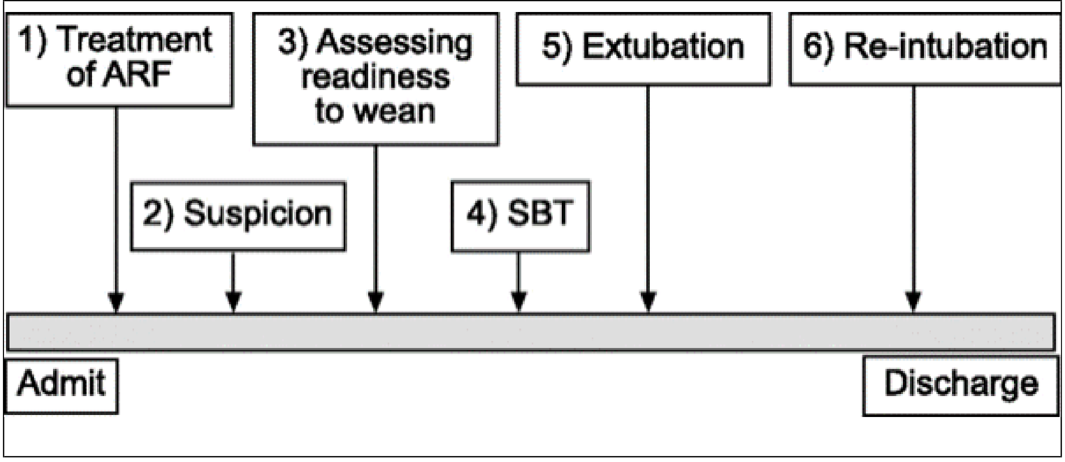
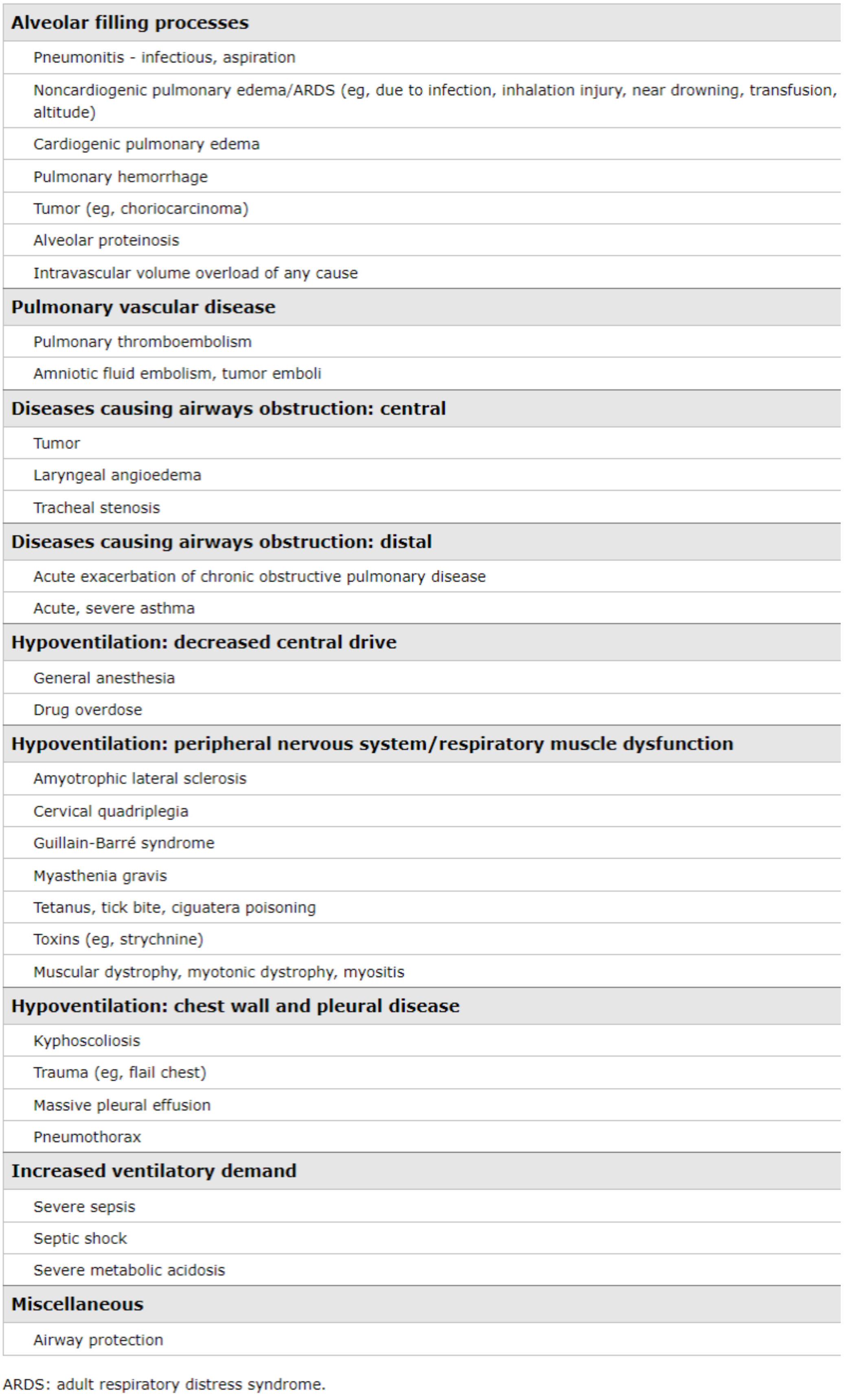


Table I - Examples of conditions requiring Mechanical Ventilation. (17)



Alveolar filling processes
Pneumonitis - infectious, aspiration
Noncardiogenic pulmonary edema/ARDS (eg, due to infection, inhalation injury, near drowning, transfusion, altitude)
Cardiogenic pulmonary edema
Pulmonary hemorrhage
Tumor (eg, choriocarcinoma)
Alveolar proteinosis
Intravascular volume overload of any cause
Pulmonary vascular disease
Pulmonary thromboembolism
Amniotic fluid embolism, tumor emboli
Diseases causing airways obstruction: central
Tumor
Laryngeal angioedema
Tracheal stenosis
Diseases causing airways obstruction: distal
Acute exacerbation of chronic obstructive pulmonary disease
Acute, severe asthma
Hypoventilation: decreased central drive
General anesthesia
Drug overdose
Hypoventilation: peripheral nervous system/respiratory muscle dysfunction
Amyotrophic lateral sclerosis
Cervical quadriplegia
Guillain-Barré syndrome
Myasthenia gravis
Tetanus, tick bite, ciguatera poisoning
Toxins (eg, strychnine)
Muscular dystrophy, myotonic dystrophy, myositis
Hypoventilation: chest wall and pleural disease
Kyphoscoliosis
Trauma (eg, flail chest)
Massive pleural effusion
Pneumothorax
Increased ventilatory demand
Severe sepsis
Septic shock
Severe metabolic acidosis
Miscellaneous
Airway protection

ARDS: adult respiratory distress syndrome.

[Current practice] Over the past 50 years, the criteria for deciding on a patient's weaning have gradually decreased and clinical judgment has given way to increasingly standardized protocols (5,9). This is because the mortality of extubation performed at the right time is 12%, if it is delayed it rises to 27% (12,18,19).

The individual predictors currently used in Readiness Testing practice have not proven particularly useful individually and are not standardized across all hospitals (except for one, the Rapid Shallow Breathing Index, already included in our study, and which has Sens 97% - Spec 65% when used alone (20)), so a protocolized weaning assessment involving multiple predictors reduce weaning time and hospital stay (11,21,22).

Patients who take a Daily Screening Liberation Assessment pass it in 25% of cases, of these 75% manage to pass the SBT(range from literature (23,24), in the primary study is 60%), and in 26-42% manage to breathe independently for at least 2 days(range from literature (ibidem), in ours are 25%).

Coming to the counterfactuals, 50% of patients who extubate by mistake (either by themselves or by accident) manage to be weaned permanently (25,26), while 35% of patients who do not pass RT could well be weaned(27). These are percentages from individual studies, but they give an idea of the size of the phenomenon. On the other hand, evidence is strong on the usefulness of SBT with a standardized protocol (16).

European and American guidelines (the latest European, from 2007) (5,16) point out that research is required on certain issues, which are investigated in the current study are:

(1) defining the minimum criteria required for assessment of readiness for weaning to allow earlier weaning)
(2) the need for a screening test before the SBT

We hypothesize that "minimal criteria" could be clinical variables collected in the preceding days along with a minute-by-minute recording of MV. This could answer point 2 about the usefulness of a RT as a screening for an SBT, although definitive confirmation would need an experimental setup since counterfactuals (e.g. SBT after a failed RT) in the current setting cannot be observed.

As highlighted in a recent review of AI on MV, concerns have been raised about the lack of reproducibility and generalisability due to methodological limitations (28,29). RT and SBT criteria are not chosen, their scoring system has been validated through Machine Learning or AI tools, and no widespread alternative has been validated in the clinical setting (28,30). Besides, clinicians adjust ventilation settings according to patients' status from time to time during ICU admission, but most of those data are not systematically recorded and assessed for potential to change clinical practice (31).

Objectives

[Target of prediction] the model aims to predict the outcome of RT and SBT for each MV patient early in the morning, ideally as part of morning rounds where all patient's cases are reviewed by the clinical team. The study aims to assess the potential for clinical implementation of the AI model, thus encompassing both generalisability and reproducibility through the adoption of the last international standards on model building, training, validation, and testing to minimize the risk of bias (32–37).

[How prediction may benefit clinical goal] RT and SBT are time-consuming processes, which require prolonged attention to be performed (38), while the ICU is an environment where timely decisions by the clinician can make the difference between life and death (39). A successful model will predict whether an SBT will probably never be attempted (because of RT failure), will fail, or will probably succeed, thus leading to MV weaning. Knowing this early in the morning could help ICU staff better allocate their resources throughout the day. Thus, enabling the proper selection of the subpopulation of patients that will benefit from it.

Moreover, the AI model in clinical settings needs to be trusted by physicians. When AI models are properly trained and scored fairly and transparently, systematically highlighting limitations and biases, their clinical value can be assessed by clinicians and subsequently integrated into clinical practice (37,40,41).

METHODS

Setting

[Clinical setting] To train, validate and test our model we used data from a multi-center randomized controlled trial that happened in Italy between 2013 and 2015 (42,43), where continuous data from MV were collected along with clinical data and laboratory findings for a cohort of 182 mechanically ventilated patients admitted to ICUs.

The study was conducted among 13 ICUs in Italy, equally distributed between universities (6 of 13) and community hospitals (7 of 13). For each of the admitted patients in the study, baseline clinical data were obtained, a daily clinical registry excerpt, a minute-by-minute summary statistic of all the ventilation variables recorded during MV (average, total sum, or proportion, according to the variable being recorded), and a log recording alarms and settings inputted by clinicians. All predictors can be seen in supplementary material, used predictors in Table II. All predictors were considered potentially relevant at the beginning, progressively reducing their number (as detailed in additional materials, "reason for exclusion").

Table II – Model Predictors

Source	*Predictor*	*Description*
Baseline	Ventilation type	Categorical variable, it defines ventilation type: • Pressure Support Ventilation (PSV) • Neurally Adjusted Ventilatory Assist (NAVA)
	Gender	
	Age	In years
	Body Mass Index	a measure of body fat based on height and weight
	Ideal Body Weight	It is based on height, gender, and age, and represents appropriate body weight
	Reason for Mechanical Ventilation	A categorical variable, it may be one of • Sepsis • Pneumonia • COPD exacerbation • Trauma/polytrauma • Post-surgical complication • Heart Failure • Acute Respiratory Distress Syndrome • Other (…)
	Simplified Acute Physiology Score	Score that predicts hospital mortality upon ICU admission

Daily registry	Sequential Organ Failure Assessment	Score that predicts ICU mortality based on lab results and clinical data
	Arterial pH	Represent acid-base equilibrium in arterial blood, taken after the SBT test (if performed)
	Arterial PaO2	The partial pressure of O2 in arterial blood, taken after the SBT test (if performed) in mmHg
	Arterial PaCO2	The partial pressure of CO2 in arterial blood, taken after the SBT test (if performed) in mmHg
	Readiness testing criteria	Set of clinical criteria, each a boolean variable 1. <2 aspirations/h 2. audible cough on aspiration 3. no distress (diaphoresis, accessory muscles, paradoxical rhinitis) 4. GCS ≥ 8 5. [RASS between -1 and +1] 6. Heart Rate (HR) ≤ 120bpm and Systolic Blood Pressure (SBP) between 90 and 180mmHg 7. Dopamine or dobutamine ≤ 5 and NorAdr ≤ 0.1 mcg/kg/min 8. Pao2/FiO2 ≥ 150mmHg 9. PEEP ≤ 8 cmH2O 10. Respiratory Rate (RR) ≤ 40/min 11. Tidal volume (Vt) ≥ 5m/kg (Ideal Body Weight) 12. pH ≥7.35
	Study days	Calculated from the date of study enrollment
TRD track data	Dynamic characteristics [ml/cmH2O]	Compliance
	End-expiratory Flow [L/min]	
	Positive end-expiratory Pressure [cmH2O]	PEEP is the pressure in the lungs (alveolar pressure) above atmospheric pressure that exists at the end of expiration. There are two types of PEEP: intrinsic and extrinsic. - Intrinsic PEEP depends on the progressive

		air trapping after incomplete expiration - extrinsic PEEP is set directly on the ventilator.
	Minute-expired Volume [L/min]	the volume of air that moves out of the lungs during a minute
	Current expired volume [ml]	the volume of air that moves out of the lungs during a single breath
	O2 concentration (%)	Percentage of oxygen in a specific volume of air
	Minute-inspired Volume [L/min]	the volume of air that moves into the lungs during a minute
	Current inspired volume [ml]	the volume of air that moves into the lungs during a single breath
	Mean Airway Pressure [cmH2O]	The average pressure in the airways during the inspiratory phase
	Measured Respiratory Rate [/min]	number of breaths per minute recorded
	Spontaneous respiratory Rate [/min]	number of spontaneous breaths per minute
	Edi peak [µV]	Diaphragm electromyography peak (NAVA specific)
	Edi min [µV]	Diaphragm electromyography minimum level (NAVA specific)
	Plateau Pressure [cmH2O]	Plateau Pressure: is the pressure applied to small airways and alveoli at the end of inspiration during positive-pressure mechanical ventilation
	Peak Pressure [cmH2O]	It is the highest pressure level applied to the respiratory system during inspiration. It depends on any airways resistance.
	Backup switches [/min]	Number of backup switches in a minute
	Backup percentage [%/min]	The percentage of breath switched to the backup mode in a minute
	P0.1 [cmH2O]	The negative pressure generated at 0.1 sec. from the beginning of the inspiratory phase
	Mechanical ventilator respiratory work [Joule/L]	Ventilator work of breathing
	Patient-ventilator respiratory work [Joule/L]	Patient work of breathing

	Spontaneous Breathing Index	Defined as the ratio of respiratory frequency to tidal volume (RR/Vt). People on a ventilator who cannot tolerate independent breathing tend to breathe rapidly (high frequency) and shallowly (low tidal volume), and will therefore have a high RSBI
	Spontaneous minute expired volume [L/min]	the volume of air that moves out of the lungs during a minute due to spontaneous breath

Prediction problem definition

[Nature of the study] Study is a retrospective analysis of data collected for a previous work, where a prognostic model is built to predict weaning outcomes.

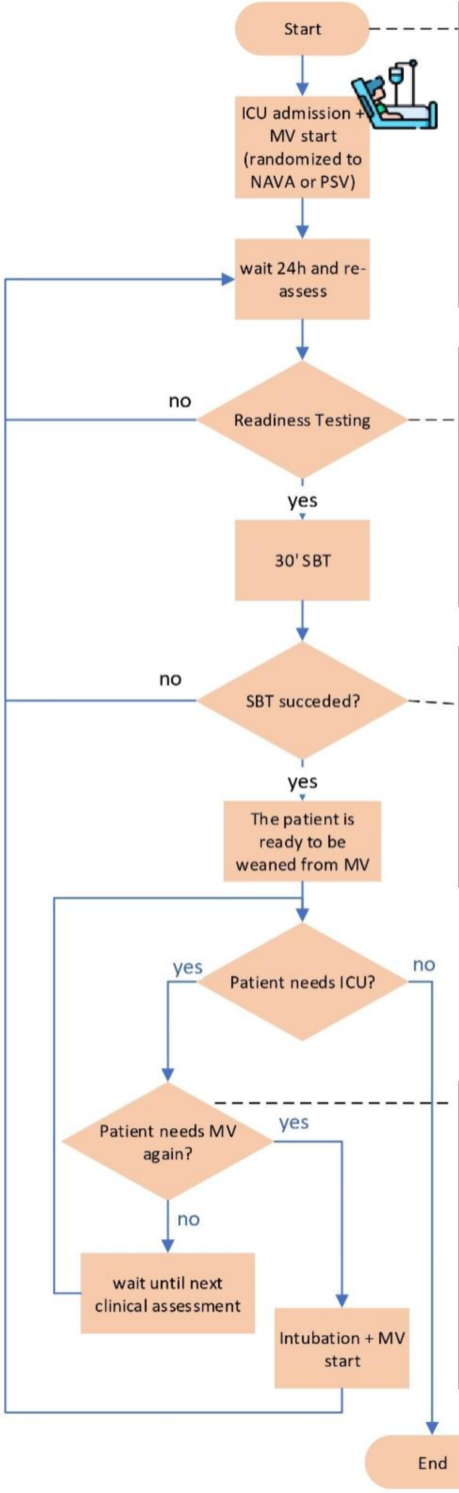
[Prediction goal measurement] The model is built to predict the outcome of RT and SBT that will be performed (or not) each day once for each patient while mechanically ventilated, starting from continuous MV data and clinical variables available at a given morning clinical round. RT success is necessary for performing SBT, so there is no data about SBTs after an RT failure. RT success is defined as respecting all 12 parameters of Tab.2 of Flowchart 1, while SBT success is defined by clinical stability during the 30-minute trial (defined by events in Tab.3 of Flowchart 1) and at the end of it (as in Tab.4 of Flowchart 1).

[Prediction model] The prediction target variable (weaning outcome) is a categorical variable that can have each different mutually exclusive values:

1. SBT not attempted (because of an RT failure)
2. SBT failure (attempted after a successful RT)
3. SBT success (attempted after a successful RT), that results in patient extubation on that day)

For each possible class, a probability is predicted, and the predicted class is the one with the highest predicted probability. In the dataset, there was a fourth possible value, which represents a day where a patient was under follow-up after a successful weaning in the previous days. Since no weaning was attempted on those days, they were considered outside the scope of the current model.

Flowchart 1 - Flowchart of the primary study. MV patients admitted to the primary study under criteria of Tab.0. RT criteria are shown in Tab.1, while SBT criteria are shown in Tab.2 (early stopping) and Tab.3 (success criteria). Weaning occurs if both RT and SBT succeed, while criteria for late weaning failure are detailed in Tab.4



Tab. 0 - Inclusion and exclusion criteria of the study

Inclusion:
- invasive MV patient (MV start >24h) capable of triggering the ventilator in an assisted mode
- no neuromuscular or neurological pathologies or nasogastric tube contraindicated
- PaO2/FiO2 >150 with PEEP >5cmH2O

Exclusion:
- <18 years, short-term poor prognosis, >3 organ failure, patient refusal, inclusion in other studies
- planned hospitalization for surgery

Tab. 2 - RT success criteria (SBT eligibility)
- <2 aspirations/h with audible cough on aspiration and no distress (diaphoresis, accessory muscles, paradoxical rhinitis)
- GCS ≥ 8 and RASS between -1 and +1
- HR ≤ 120bpm and SBP between 90 and 180mmHg
- Dopa or dobu ≤ 5 and NorAdr ≤ 0.1 mcg/kg/min
- Pao2/FiO2 ≥ 150mmHg
- PEEP ≤ 8 cmH2O
- RR ≤ 40/min and Vt ≥ 5m/kg (Ideal Body Weight)
- pH ≥7.35

Tab. 3 - SBT success criteria (MV stop eligibility)
- no distress (e.g. increased accessory muscle activity) and no dyspnoea with VAS ≤ 7
- RASS between -4 and +3
- SBP ≥ 90 mmHg
- PaO2 > 60mmHg with FiO2 ≤ 50%
- PaCo2 ≤ 50mmHg and delta ≤ 8mmHg
- RR/Vt ≥ 105 bpm/L
- pH > 7.32 and delta < 0.07
- no appearance of new cardiac arrhythmias

Tab. 4 - Weaning failure criteria
- emergencies (respiratory, cardiac arrest or gasping)
- worsening GCS or RASS (coma or agitation requiring continuous IV sedation)
- need for medications to maintain SBP ≥ 90 mmHg (despite adequate fluid infusion) such as Adr or Vasopressin (any dose), Dopa/Dobu >5 or NorAdr > 0.01 mcg/kg/min
- obstruction of the airways (stridor ot tirage) or tracheo-bronchial secretions that are difficult to manage
- Respiratory distress with SpO2 <90% and RR >35 bpm and clinical signs, despite oxygen administration

ICU = Intensive Care Unit; FiO2 = Oxyen inspired fraction; GCS = Glasgow Coma Scale; HR = Heart Rate; MV = Mechanical Ventilation; NAVA = Neurally Adjusted Ventilatory Assist; PaO2/PaCO2 = Oxygen/Carbon Dioxide Partial Pressure; PEEP = Positive End Expyratory Pressure; PSV = Pressure Support Ventilation; RASS = Richmond Agitatio-Sedation Scale; RR = Respiratory Rate; SBP = Systemic Blood Presure; SBT = Spontaneous Breathing Trial; VAS = Visual Analog Scale for pain; Vt = Tidal Volume

[Practical costs of Prediction errors] The model aims to guide clinicians to optimize the time spent on RT and SBT by reducing the number of patients to the most probable to be successfully weaned. Choosing to focus only on a reduced number of patients will unnecessarily prolong MV for weanable patients left behind. At the same time, including too many patients will prolong the time dedicated to each patient in a fast-paced environment like the ICU. Table III contains a summary of the consequences of each prediction compared to the observed current clinical management.

Table III – Confusion matrix of outcome consequences of predictions. On each given day, the observed classes (actual clinical management) are compared to model prediction, which is provided at the start of the day. The confusion matrix highlight for each possible combination of the clinical outcome of favoring SBT suggestion over standard-of-care

		Observed (actual clinical management)		
		0 – SBT not attempted	1 – SBT failed	2 – SBT succeded
Model prediction (suggested clinical management)	0: SBT not attempted	Unchanged	Improved (correct prediction of a future SBT failure, while it would be observed only after it happened)	Worsened (unnecessary extension in MV)
	1: SBT failure (thus weaning is unlikely)	Unchanged (SBT is not suggested both in actual and predicted)	Improved (correct prediction of a future SBT failure, while it would be observed only after it happened)	Worsened (unnecessary extension in MV)
	2: SBT success	30% worsened (Study protocol provides no data about not attempted SBT outcome, so we considered 35% Improved and 65% worsened as evidence (16,27))	Unchanged (SBT is wrongly suggested both in actual and predicted)	Unchanged

[Quality metrics] The model will be trained and validated to minimize cross-entropy as suggested in Introduction to Statistical Learning (44), as the main metric to boost model confidence in predicted classes. The model will be evaluated on balanced accuracy, MCC-score, Precision, and Recall to properly evaluate model performance (45). One-vs-all generalization for multiclass problems was implemented. Since not all classification errors have the same effect on clinical outcomes, predictions were also evaluated averaging a custom metric of clinical impacts. This metric assigns a score of 1 for "improved" predictions, -1 for "worsened" predictions, and 0 for "unchanged" ones. The final score, i.e. the average of the test predictions, can range between -1 (worse than RT), and 1 (better than RT).

Formula 1 - Categorical cross-entropy loss function, where **t_i** (ground truth, i.e. 0/1) and **s_i** (predicted probability) for each class **i** in **C**

$$CE = -\sum_{i}^{C} t_i \log{(s_i)}$$

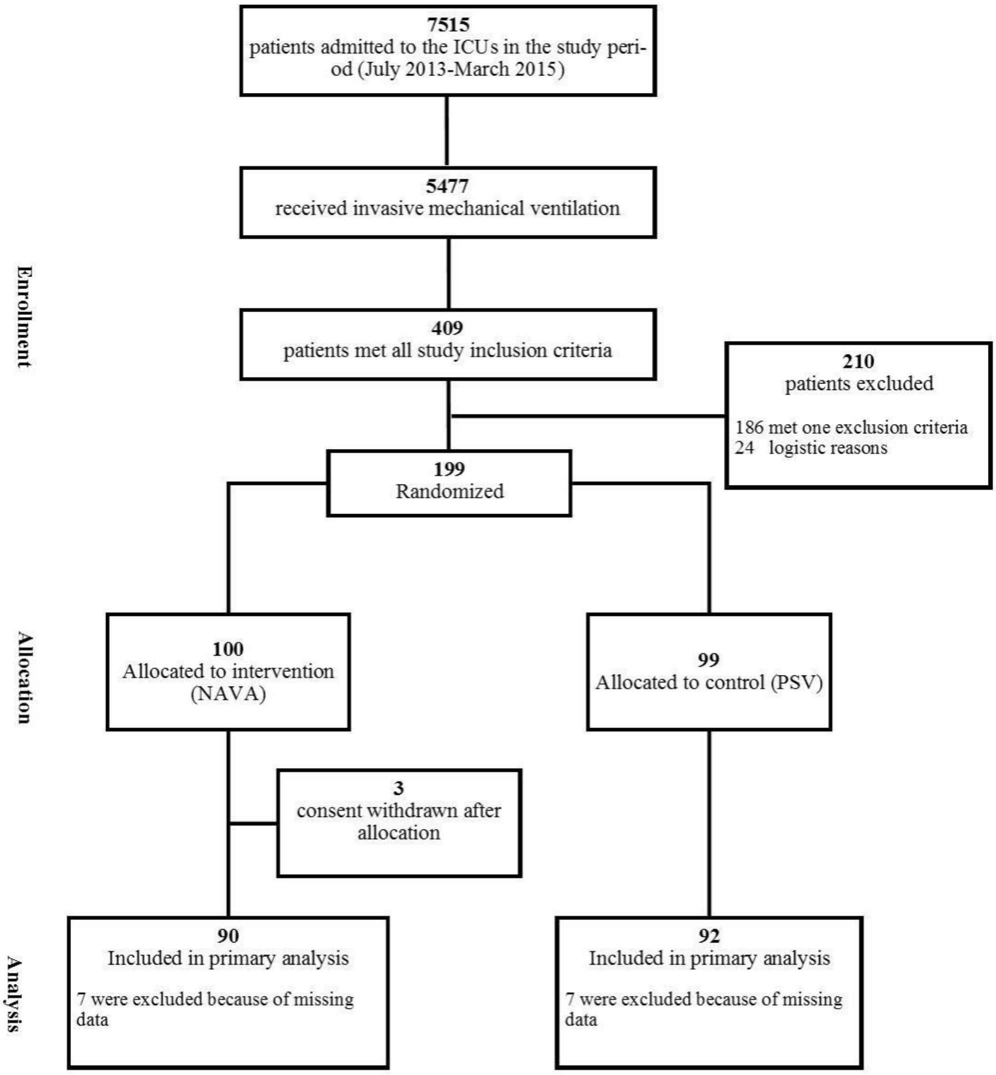
[Success criteria] To evaluate the model on the RT prediction, it will be tested against the current standard of care (which is shown in Flowchart 1), since a successful RT is not a perfect predictor of SBT success. A statistically significant difference in clinical outcome was assessed through bootstrapped confidence interval at a 95% level (41), against the hypothesis that the model is as good as the Readiness Testing, which has by construction a custom score of 0.

Data preparation

[Data sources] For all the centers the relevant ethics committee at that center approved the study protocol. Written informed consent was obtained from all the patients, their next of kin, or another surrogate decision maker as appropriate for the primary study (43). The trial was registered at the Australian New Zealand Clinical Trial Registry (ANZCTR) under the number ACTRN12612000815864. The trial was overseen by a steering committee and research assistants regularly monitored all the centers on site to check adherence to the protocol and the accuracy of the data recorded. An investigator at each center was responsible for enrolling patients in the study, ensuring adherence to the protocol, and completing the electronic case report form.

[Inclusion/exclusion criteria] In the primary study, patients were included if able to trigger the ventilator in partial support mode, did undergo invasive MV for at least 24h before study enrolment, and expected to need it for 48 hours or more. Additional criteria were age equal or greater than 18, absence of scheduled surgery, no inclusion in other research protocol, no neuromuscular or neurologic disease, no contraindications of nasogastric tube positioning, and PaO2/FiO2> 150 con PEEP≥ 5 cmH2O. Exclusion criteria were short-term poor prognosis, three or more organ failures, or refusal to participate, as detailed by Box0 of Flowchart 1, and resulted in a total of 182 patients, as detailed in Flowchart 2. In the current study, we were able to retrieve data on 180 patients, on which we conducted our analysis.

Flowchart 2 - CONSORT-style flowchart shows study admission and randomization criteria for the primary study. Data acquired during the primary study (RCT) are used as part of the current analysis (observational)


7515
patients admitted to the ICUs in the study period (July 2013-March 2015)
5477
received invasive mechanical ventilation
Enrollment
409
patients met all study inclusion criteria
210
patients excluded
186 met one exclusion criteria
24 logistic reasons
199
Randomized
Allocation
100
Allocated to intervention (NAVA)
99
Allocated to control (PSV)
3
consent withdrawn after allocation
Analysis
90
Included in primary analysis
7 were excluded because of missing data
92
Included in primary analysis
7 were excluded because of missing data

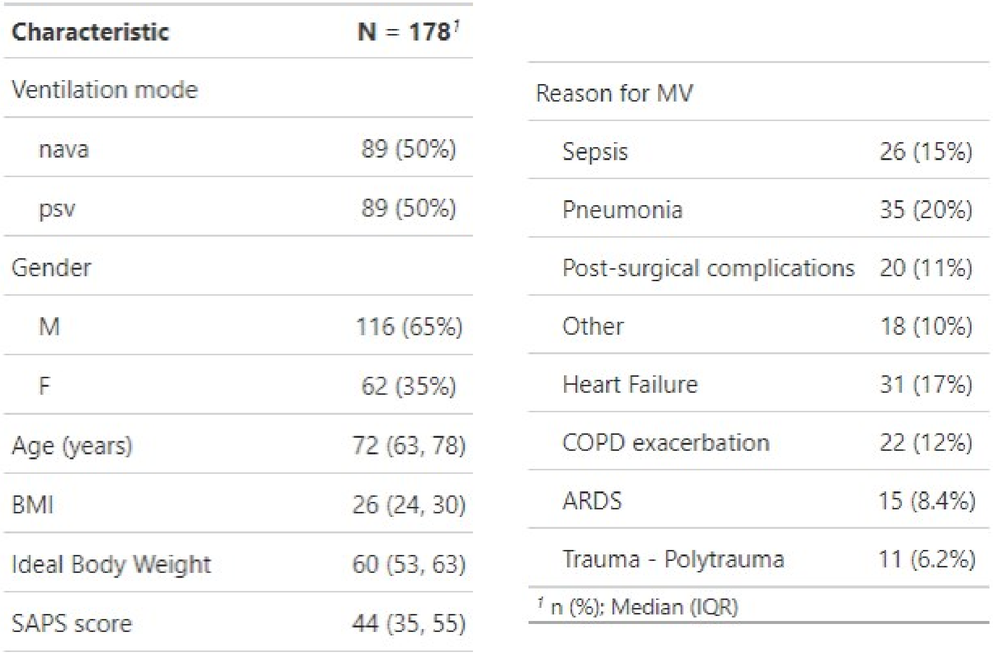
[Time span and cohort size] Data were collected between March 2013 and May 2015, from 13 centers across Italy. 180 patients were included in our study, with a total of 1929 days recorded in the daily registry. The basic demographics of patients can be found in Table IVa, while a summary of observations on daily records can be found in Table IVb. For each patient in the study, once per day a track record and log record were obtained by direct download from Servo-I Mechanical Ventilator (Maquet, Sölna, Sweden), resulting in 2,073,253 minutes observed across the cohort and 424,772 LOG events recorded, for which summary variables can be observed in Table IVc.

[Observational unit] As in Table II, each patient had recorded baseline information. Daily data about each patient were recorded in a clinical diary for each day of MV and at least two days after successful weaning, along with criteria relevant to clinical management (Box2-3 of Flowchart1). While the patient was ventilated, 20 variables were averaged or summed to the minute and recorded, 2 more variables were specific for the MV mode chosen and 12 other variables were almost or completely missing. Since the MV memory was limited to the equivalent of 24h of recording, data have been overwritten (thus lost) if the download didn't happen for more than 24h. LOG contained detailed information about the category of the message recorded associated with the specific time and message.

[Information leakage prevention] Information leakage can happen both in training examples and in features composing the model (41,46), inappropriately boosting confidence in model predictions. Leakage itself is defined as the introduction of information about the data mining target that should not be legitimately available to mine from (46). In our scenario, it may happen if the model trains with information that shouldn't be available during training time (e.g., information on a day D to predict the outcome of that day D, because prediction should happen early in the morning while the information would be collected during the day, so, after the prediction would be happened). To prevent information leakage, a test set composed of a subset of 36 out of 178 patients which comprise nearly 20% of overall days was set aside for final evaluation since it should not overlap with the training or validation set and be representative of a wider population to measure the model's generality (47). In addition to that, input data were screened for duplicates before the split, and eliminated by keeping only the first occurrence since they arose from the memory buffer problem outlined above. The split was performed at the patient level (e.g patient AB001, regardless of the total number of weaning attempts, which are the events on which the prediction is made), keeping the same metrics defined for the test set.

Table IV – Summary characteristics of the input datasets, highlighting predictors used in the model-building phase. IVa (top) shows baseline characteristics, IVb (middle) shows daily data, IVc (bottom) shows Mechanical Ventilation track data

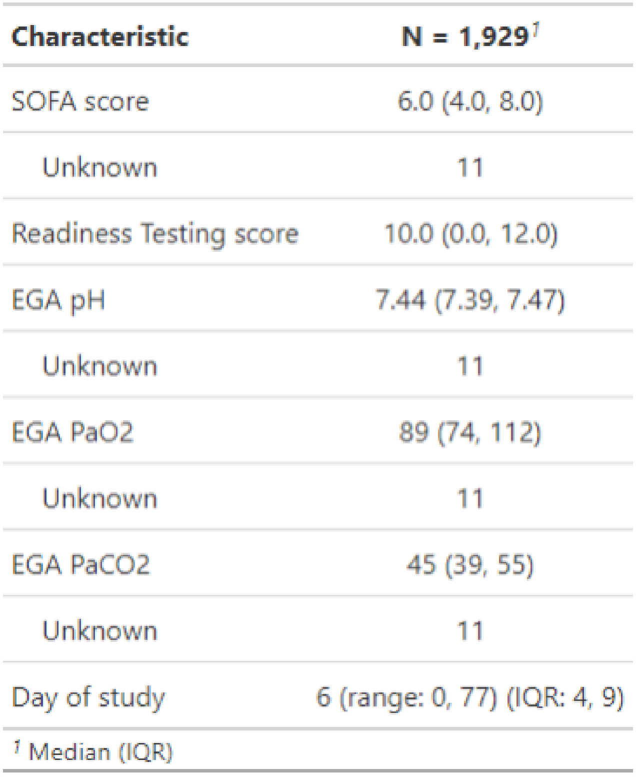
Tab. IVa



Characteristic	**N = 178**[1]
Ventilation mode	
nava	89 (50%)
psv	89 (50%)
Gender	
M	116 (65%)
F	62 (35%)
Age (years)	72 (63, 78)
BMI	26 (24, 30)
Ideal Body Weight	60 (53, 63)
SAPS score	44 (35, 55)
Reason for MV	
Sepsis	26 (15%)
Pneumonia	35 (20%)
Post-surgical complications	20 (11%)
Other	18 (10%)
Heart Failure	31 (17%)
COPD exacerbation	22 (12%)
ARDS	15 (8.4%)
Trauma - Polytrauma	11 (6.2%)

[1] n (%); Median (IQR)

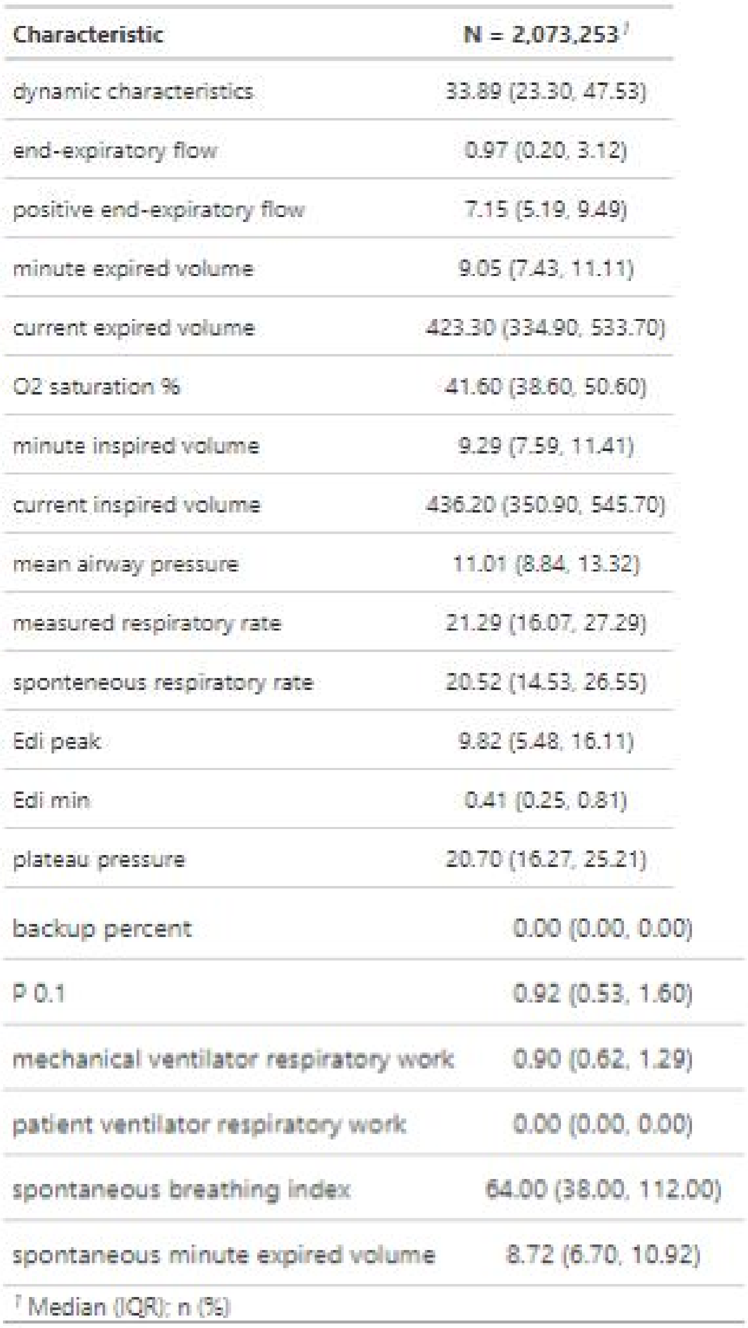
Tab IVb



Characteristic	**N = 1,929**[1]
SOFA score	6.0 (4.0, 8.0)
Unknown	11
Readiness Testing score	10.0 (0.0, 12.0)
EGA pH	7.44 (7.39, 7.47)
Unknown	11
EGA PaO2	89 (74, 112)
Unknown	11
EGA PaCO2	45 (39, 55)
Unknown	11
Day of study	6 (range: 0, 77) (IQR: 4, 9)

[1] Median (IQR)

Tab IVc



Characteristic	N = 2,073,253 [1]
dynamic characteristics	33.89 (23.30, 47.53)
end-expiratory flow	0.97 (0.20, 3.12)
positive end-expiratory flow	7.15 (5.19, 9.49)
minute expired volume	9.05 (7.43, 11.11)
current expired volume	423.30 (334.90, 533.70)
O2 saturation %	41.60 (38.60, 50.60)
minute inspired volume	9.29 (7.59, 11.41)
current inspired volume	436.20 (350.90, 545.70)
mean airway pressure	11.01 (8.84, 13.32)
measured respiratory rate	21.29 (16.07, 27.29)
sponteneous respiratory rate	20.52 (14.53, 26.55)
Edi peak	9.82 (5.48, 16.11)
Edi min	0.41 (0.25, 0.81)
plateau pressure	20.70 (16.27, 25.21)
backup percent	0.00 (0.00, 0.00)
P 0.1	0.92 (0.53, 1.60)
mechanical ventilator respiratory work	0.90 (0.62, 1.29)
patient ventilator respiratory work	0.00 (0.00, 0.00)
spontaneous breathing index	64.00 (38.00, 112.00)
spontaneous minute expired volume	8.72 (6.70, 10.92)

[1] Median (IQR); n (%)

At the same time, the model employs data only available at 7 am, referring to the previous day of MV, to prevent the use of features that would not be available at that time.

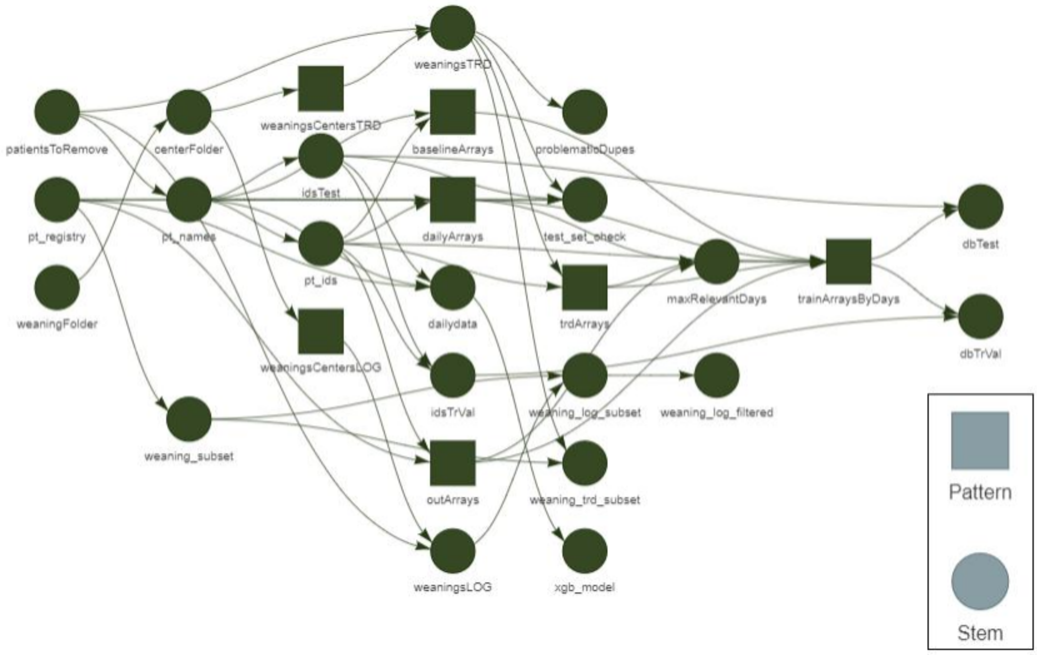
But leakage is not limited to the explicit use of illegitimate examples in the training process, but also through design decisions (46), and it wouldn’t be detectable. This is why each decision about model, split, or feature selection reported in the present thesis was set before seeing results.

[Data preprocessing and outlier removal] Patient baseline data and daily registry used in the primary study were available as .xlsx documents, while MV tracked data in a series of separate CSV files, one for each different download from the mechanical ventilation, spanning most 24h per patient. All analyses were performed using R Statistical Software (v4.2.2) (48), using data pre-processing and cleaning functions from the "tidyverse" meta-package (49,50). To ensure reproducibility, original data were not modified by hand, and software was developed as a stand-alone package developed under git version control and provided (https://github.com/UBESP-DCTV/weaning) through GitHub (51), with "targets" package to keep track, manage, and automate execution of workflow in a reproducible way (52), which can be seen in Additional Materials. "Quarto" was the main reporting tool (53), with "ggplot2" for charts (54) and "gtsummary" for tables (55).

Custom import functions were designed to check for inconsistencies and time discrepancies between different data sources referring to the same events and cross-check metadata. Patient NO021 was removed for clinical implausibility (extubated without being ready), while patient FE017 was removed because of a mismatch with MV metadata. The 12 criteria composing RT could not be analyzed as independent clinical predictors since they were not evaluated in a standardized way (e.g. it was common in some centers to record all criteria rows as negative, even if only one of them was not met). MV track data had missing parts, due to the limited size of buffer memory, and duplicate values for the same time with different recorded variables. This happened because of a known issue with the ventilator, wrong duplicates were identified and filtered. Moreover, no spectral analysis (e.g. on Respiratory Rate) could be performed since data were summarized to the minute.

After obtaining the dataset from each source (baseline, daily, MV track, and MV log), Exploratory Data Analysis was performed (56–58) in an iterative process to clean the dataset. A set of functions was designed to transform bidimensional tables into tensors (i.e., multidimensional arrays) ready to be used in model building.

Figure 2 - Network visualization of the analysis flow from the raw input (on the left) to a range of datasets that can be used in the analysis process, linked by arrows representing dependencies. The 4 sources of input data (baseline data, daily data, MV track data, and LOG) are imported, cleaned, and pre-processed in several different dataset to obtain the analysis-ready ones. Circles reports single objects, squares collections of homogeneous ones (e.g. baselineArrays is *the* object composed by all the arrays created separately for each patient).

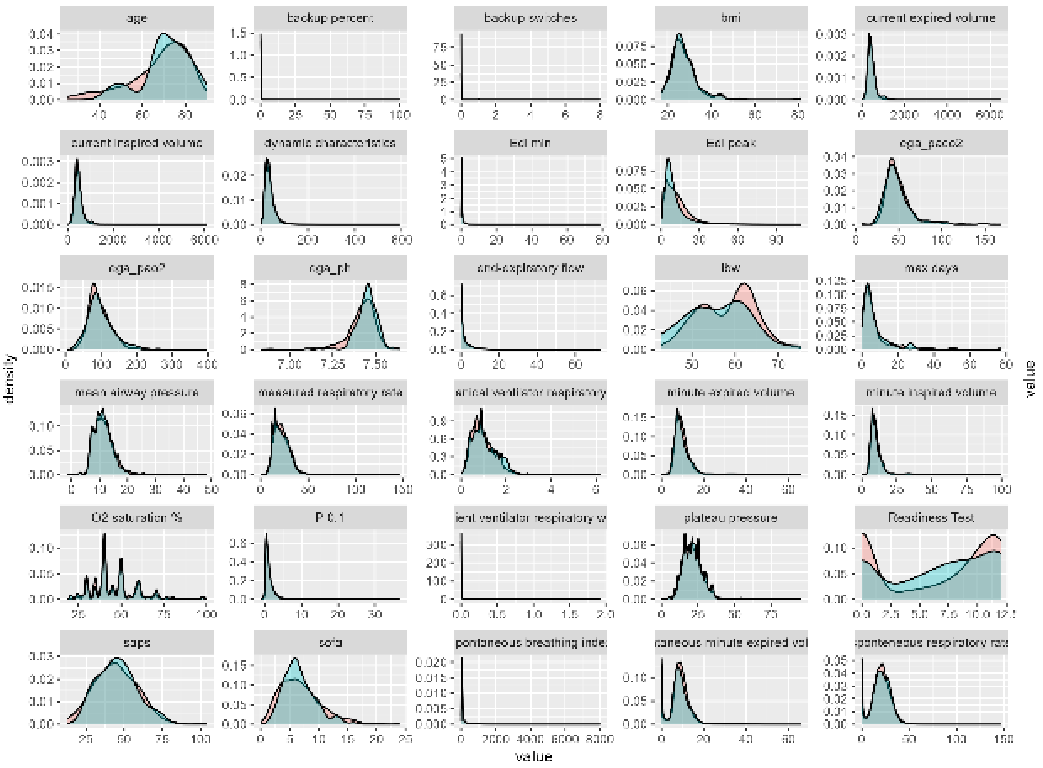


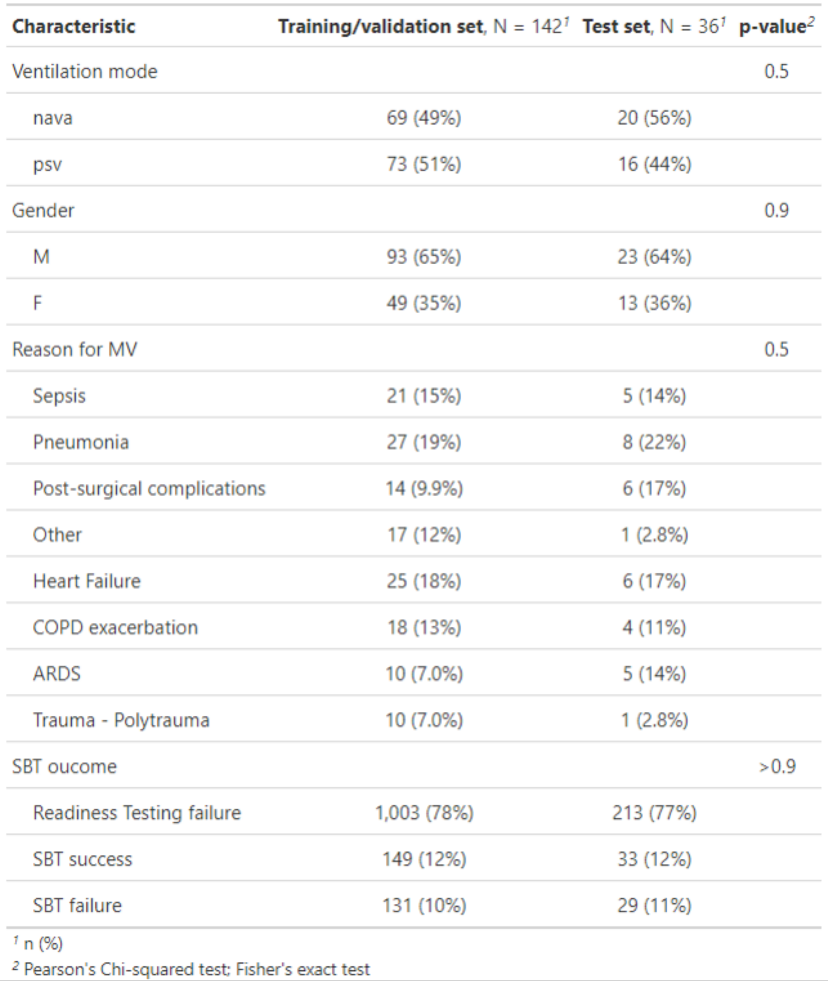
[Missing values] Missing values are common in the ICU setting, but they are not frequently addressed in the literature (28). Values missing more than 2/3 of the time in the training and validation subset were dropped from the analysis, while other missing data (NAs) were replaced by a fixed placeholder, a unique value that could be processed and learned to be "missing" by the neural network. Model predictors from baseline and daily were complete.

[Dataset's basic statistics] Each day weaning could potentially be attempted or not, usually because the patient died or was already extubated. Class distribution of remaining outcomes is unbalanced, with the majority class (RT failure) representing 70% of outcomes, as can be seen in Table V and Picture 2.

[Model validation] Data was split into a train (64%), validation (16%), and test set (20%), with 5-fold cross-validation.

Figure 3 - Distribution of all predictors among train and validation (red) and test set (blue) to check that the random split produced representative sets with similar distributions. Continuous variables are shown with a density plot to check for similar distributions, while categorical have a bar plot to make sure the split produced at least one training example and a test example for each class.





Characteristic	**Training/validation set**, N = 142[1]	**Test set**, N = 36[1]	**p-value**[2]
Ventilation mode			0.5
nava	69 (49%)	20 (56%)	
psv	73 (51%)	16 (44%)	
Gender			0.9
M	93 (65%)	23 (64%)	
F	49 (35%)	13 (36%)	
Reason for MV			0.5
Sepsis	21 (15%)	5 (14%)	
Pneumonia	27 (19%)	8 (22%)	
Post-surgical complications	14 (9.9%)	6 (17%)	
Other	17 (12%)	1 (2.8%)	
Heart Failure	25 (18%)	6 (17%)	
COPD exacerbation	18 (13%)	4 (11%)	
ARDS	10 (7.0%)	5 (14%)	
Trauma - Polytrauma	10 (7.0%)	1 (2.8%)	
SBT oucome			>0.9
Readiness Testing failure	1,003 (78%)	213 (77%)	
SBT success	149 (12%)	33 (12%)	
SBT failure	131 (10%)	29 (11%)	

[1] n (%)

[2] Pearson's Chi-squared test; Fisher's exact test

Table V – Outcome distribution of target variable across all data

Characteristic	N = 1,929
Daily attempt	
Already extubated	371 (19%)
Readiness Testing failure	1,216 (63%)
SBT success	182 (9.4%)
SBT failure	160 (8.3%)

A broader issue has been hyper-parameter tuning as part of model selection. In fact. low variance is at least as important as unbiasedness in model selection criteria, as the degradation in performance due to overfitting arising from selection bias can be surprisingly large (59). A model is defined by parameters (which weights are learned through the training process) and hyper-parameters (which are set by the user). A different range of hyper-parameters was tried, with pair of nested loops, with the hyper-parameters adjusted to optimize a model selection criterion in the outer loop (model selection) and the parameters set to optimize a training criterion in the inner loop (model fitting/training with better estimates). In a previous study (60), it was noted that validation set error is strongly biased since it was directly minimized during model selection, and thus should not be used for performance estimation. To mitigate this problem, and to improve the generality of the resulting model, a 5-fold Cross-validation was performed(61).

Ideally, since all our training and test data come from the same kind of equipment, its ability to obtain the same performance on data collected with other MV could not be tested. Moreover, no community benchmark was available for our data to test our model on an external dataset, so this part of the analysis could not be performed, relying only on an internal split to evaluate our model. Nevertheless, given the experimental protocol adopted and the multicentral nature of the study (including both community and university hospitals), we consider the study population drawn from the population of interest.

Prediction model building

[Redundant variables removal] In additional material, the list of all variables can be found, detailed with a reason for their exclusion. The most frequent reasons were clinical or missing values >67%. To assess for perfect separation, after randomly selecting the test set from the patient pool, each predictor was tested for statistical significance both for perfect separation screening and to assess the test set completeness of the wide range of clinical scenarios (e.g. at least one patient with each class of admission reason should be present both in testing and in training-validation subsets). No variable was removed after this step.

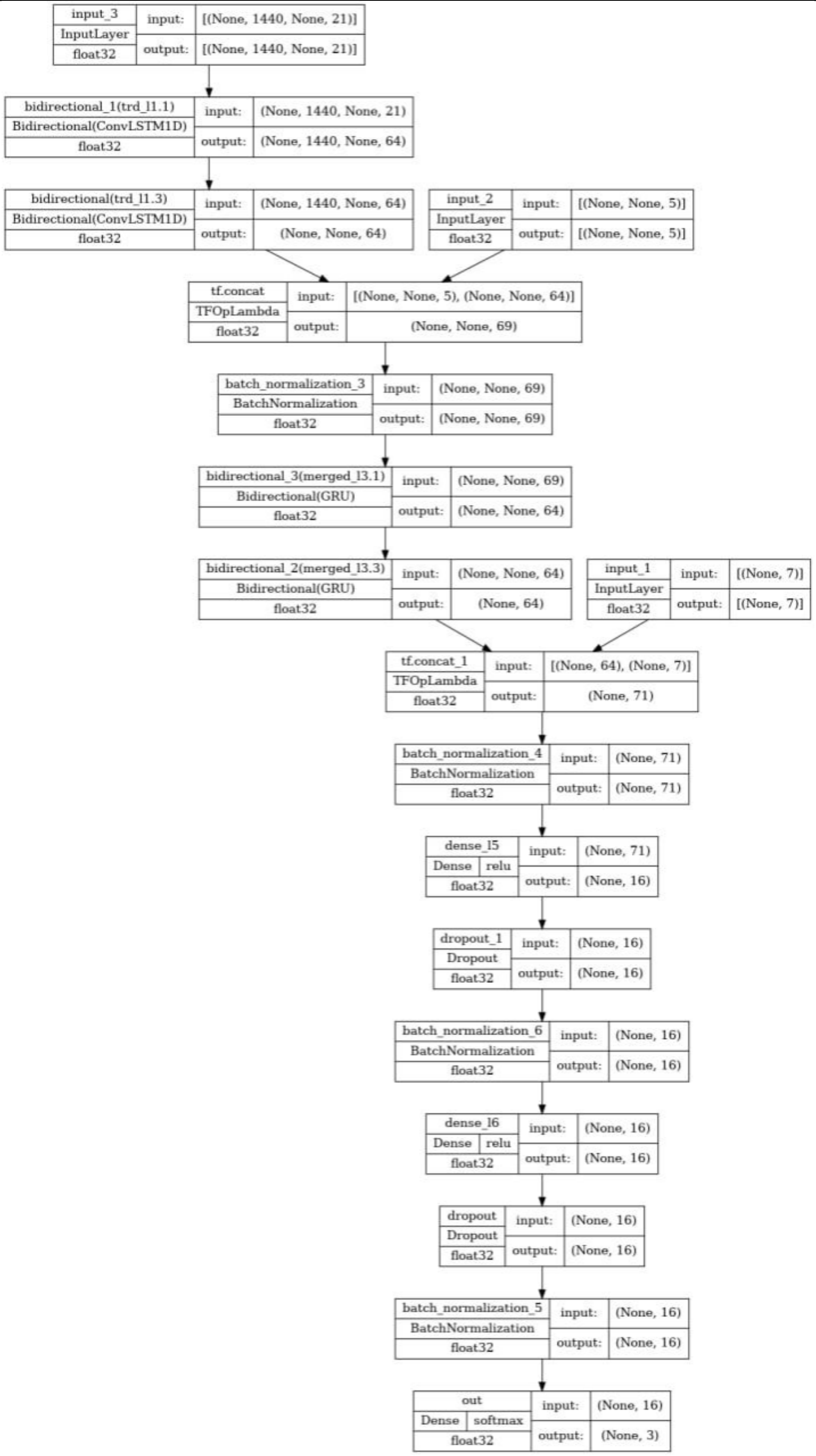
[Predictors and response association] Independent variables are shown in Table II, with a complete description of their fundamental characteristics in Table IV.

[Assess if sufficient data for good fit] To assess whether sufficient data were available for a good fit, goodness of fit was defined as the model's ability to outperform a set of extremely simple models as a baseline (three that predicted always the same classes, two more than predicted random classes), outperform an extreme gradient boosted tree model (trained and optimized thanks to the "familiar" package (62,63)) with only baseline and daily data as an intermediate point and to progressively improve its performance to reach perfect classification, eventually improving as it's shown in the composite metric derived from Table III. The flexibility of our RNN model was progressively increased recursively during model development.

[Modeling technique] Standard Machine Learning approaches to time-series analysis usually split data in various windows, trying to analyze them at once (44). To properly model a RNN was used. A complete sketch of the model architecture can be found in Figure 4, which shows data flow through network layers. ADAM was adopted as an optimizer (64,65).

Since the model is evaluated at each iteration against the validation set during the training process, it can quickly overfit. To prevent that, early stopping was implemented, combined with a learning rate scheduler and an optimizator that stops when validation metrics reach a plateau for a sufficient time. Keras API for R using TensorFlow as a deep learning backend to Python modules was used to design, compile, fit, and evaluate all the neural network models (66–68), run on an Intel(R) Xeon(R) E-2286M 16 Core CPU @ 2.40GHz equipped with 128 GB RAM - Ubuntu 22.04 LTS.

Figure 4 - Model architecture is represented by a series of neural network layers (boxes), linked together by arrows that represent information flow from input (labeled as "InputLayer") to output (labeled as "out"). Each box contains a brief description of the layer itself. The first column contains a unique name identifier (upper cell), a description of the layer (either in a single cell as in "BidirectionalGRU" or in two cells as "Dense | softmax"), and a type of input data (lower cell). The rest of the cell describes the input that is expected and the output that is produced, with the indication of the tensor's expected dimensions between brackets. Dimensions reported as "None" identify variable-length dimensions like number of patients processed (the first or the single ones) and the number of days processed (the second ones). 1440 is the number of minutes in a day.


input_3 | InputLayer | float32 | input: | [(None, 1440, None, 21)] | output: | [(None, 1440, None, 21)]
bidirectional_1(trd_l1.1) | Bidirectional(ConvLSTM1D) | float32 | input: | (None, 1440, None, 21) | output: | (None, 1440, None, 64)
bidirectional(trd_l1.3) | Bidirectional(ConvLSTM1D) | float32 | input: | (None, 1440, None, 64) | output: | (None, None, 64)
input_2 | InputLayer | float32 | input: | [(None, None, 5)] | output: | [(None, None, 5)]
tf.concat | TFOpLambda | float32 | input: | [(None, None, 5), (None, None, 64)] | output: | (None, None, 69)
batch_normalization_3 | BatchNormalization | float32 | input: | (None, None, 69) | output: | (None, None, 69)
bidirectional_3(merged_l3.1) | Bidirectional(GRU) | float32 | input: | (None, None, 69) | output: | (None, None, 64)
bidirectional_2(merged_l3.3) | Bidirectional(GRU) | float32 | input: | (None, None, 64) | output: | (None, 64)
input_1 | InputLayer | float32 | input: | [(None, 7)] | output: | [(None, 7)]
tf.concat_1 | TFOpLambda | float32 | input: | [(None, 64), (None, 7)] | output: | (None, 71)
batch_normalization_4 | BatchNormalization | float32 | input: | (None, 71) | output: | (None, 71)
dense_l5 | Dense | relu | float32 | input: | (None, 71) | output: | (None, 16)
dropout_1 | Dropout | float32 | input: | (None, 16) | output: | (None, 16)
batch_normalization_6 | BatchNormalization | float32 | input: | (None, 16) | output: | (None, 16)
dense_l6 | Dense | relu | float32 | input: | (None, 16) | output: | (None, 16)
dropout | Dropout | float32 | input: | (None, 16) | output: | (None, 16)
batch_normalization_5 | BatchNormalization | float32 | input: | (None, 16) | output: | (None, 16)
out | Dense | softmax | float32 | input: | (None, 16) | output: | (None, 3)

Model Selection

[Robust Model Evaluation] Evaluation of the test set is the final step of model building, which comes after proper validation, model refinement, and a fair comparison among all candidate models. It is necessary to have valid results from which reliable conclusions can be drawn (47). Evaluation is a multi-step process, which we document, share, and report to improve the reproducibility of our workflow and build confidence in our results (36). It is composed of an appropriate validation set and model comparison and selection, appropriate test set use, appropriate set of metrics selection, and statistical test for model comparison. The term “robust” is used to imply insensitivity to irrelevant experimental factors, such as sampling and partitioning of the data in training, validation, and test sets (59). Thus, the model on the test set is run only once on data that weren’t seen before, to prevent selecting a model that performs well on the test set for some specific reason, but not necessarily in real-world scenarios.

[Performance metrics] Since data are unbalanced, accuracy could be misleading. As a comparison, a model always predicting the majority class would result in a 75% accuracy (since it’s the frequency of the most common class). Also, the standard of care is suboptimal: we don’t have the counterfactual example where an SBT is tried after a negative RT, but a positive RT heightens the probability of a successful SBT to slightly more than 50% in our dataset. Moreover, successful SBT results in successful weaning in only 80% of cases (43).

Not only accuracy would be misleading in an unbalanced scenario, but it would also be incomplete, as we have seen. Thus, performance was reported with a multiplicity of metrics since each of them gives different information about the model. In the field, balanced accuracy and precision are commonly used metrics, which can be implemented with a generalization for multi-class classification tasks (45,69,70). Since F1-score may yield misleading results for classifiers biased towards predicting the majority class and it is susceptible to swapping of class labels (71) and it is unclear in the definition for multi-class problems (69), Matthews Correlation Coefficient (72–74).

Moreover, a custom function was built to translate predictions into a clinical impact metric. Starting from the confusion table with suggested vs observed patient management, we summarized clinical impact (either on the patient or on resources spent by the ICU staff). Whenever the algorithm suggested not to wean the patient (class 0 or 1), while the observed event was a successful weaning (class 2), the score was penalized (-1). On the contrary, if it didn’t impact clinical management (class 0 both predicted and observed, class 2 both predicted and observed, class 2 predicted but class 1 was observed, or class 1 predicted and class 0 observed), it was scored as 0. If it improved clinical management (anticipating a failure to wean with a class 0 or 1 prediction, while class 1 was observed on that day), it was positively rewarded (+1). Unfortunately, given the limitations of an observational study, since SBT was not attempted after a failed RT, predicted classes 2 on observed class 0 were evaluated considering literature knowledge proportions on unattempt SBT, i.e. 35% success and 65% of failure, leading to an overall mean negative 30% of impact caused but this kind of systematic error (i.e., -0.3 score). The total score is the mean of the individual scores, thus ranging between -1 (model is harmful respect to RT) to +1 (model improves clinical management respect RT). Readiness testing, by construction (and computation, i.e., if use as model prediction early in the morning it will fill scored 0 cells only), has a custom score of 0, providing a meaningful comparison of the SBT success rate against the actual clinical management.

[Fairness] To identify relevant subpopulations where the model could have a differential impact (Fairness), baseline information about Sex and Age was collected and reported. Ethnicity and social status were not recorded at the time of the study and couldn't be inferred from available data, so that differential split was not performed (35,46).

[Model selection strategies] RNNs are the best option for complex time-series analysis, and a comparable explainable alternative is yet to be developed. Given the only partial explainability, given the limited number of patients, individual plots of the test set have been developed. A discussion of more general range of techniques on AI model explainability can be found in the discussion.

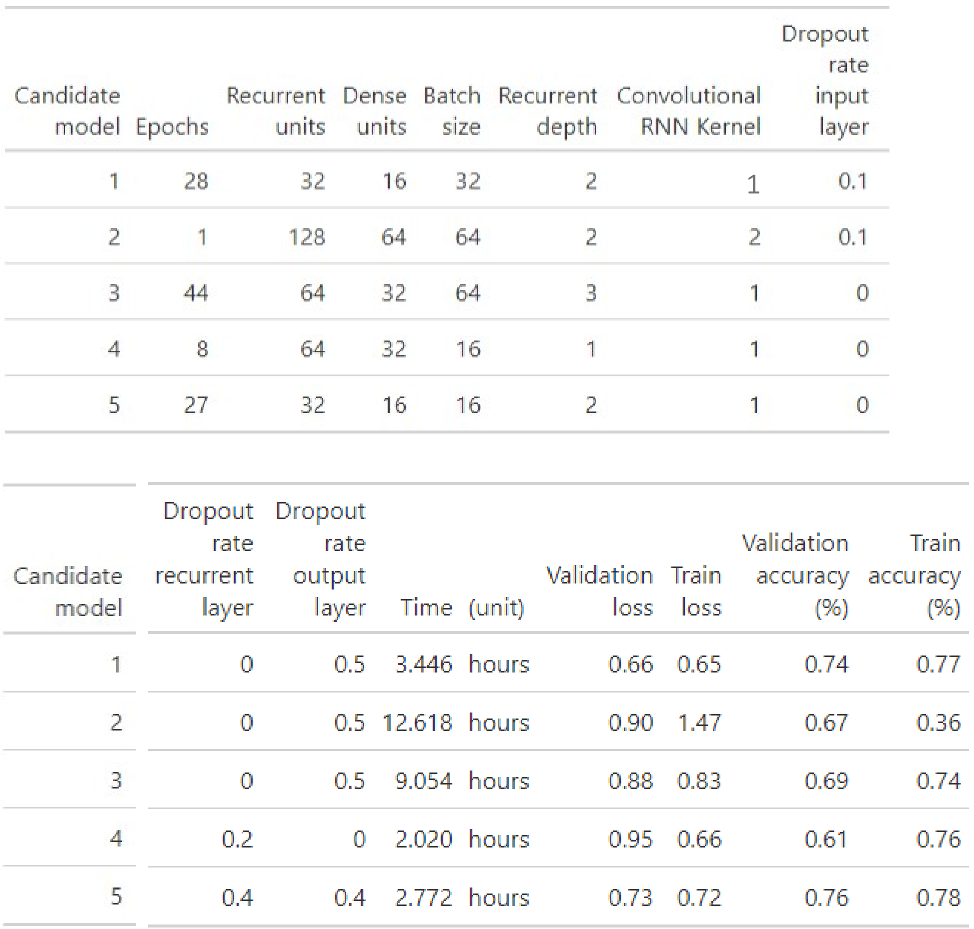
To select the final model, the one that represents the actual RNN model which takes as input baseline and daily clinical information as well as the MV data recorded up to the moment when the prediction is generated for that patient, many candidates have been developed and assessed. After completing the 5-fold CV, the hyper-parameter tuning and visually inspecting the losses, and balanced accuracy curves to identify a point where metrics were satisfactory enough without starting to worsen, the final model was selected.

RESULTS

Preliminary models performance

Out of the 5 different candidates models developed, the last candidate was selected. Performance on validation seemed to reach a plateau around epoch 15 (as shown in Figure 6), so it was selected. The final model was used as training data for all the previous datasets (training and validation) while being evaluated on the test set.

Table VII - Comparison of hyper-parameters and results of tried model architectures, showing the 5 candidates. Epochs represent the number of epoch leading to the lower cross-validate validation average loss, recurrent and dense units are the number of neurons in each type of layer, batch size are the number of sample the network explore in a single step (one epoch reached when all sample where elaborated by the network, batch by batch); drop-out rate are the proportion of random neurons ignored at each step for the corresponding layer. Time are reported for a single CV run. Losses are reported as average of the CV runs. Accuracies are balanced.

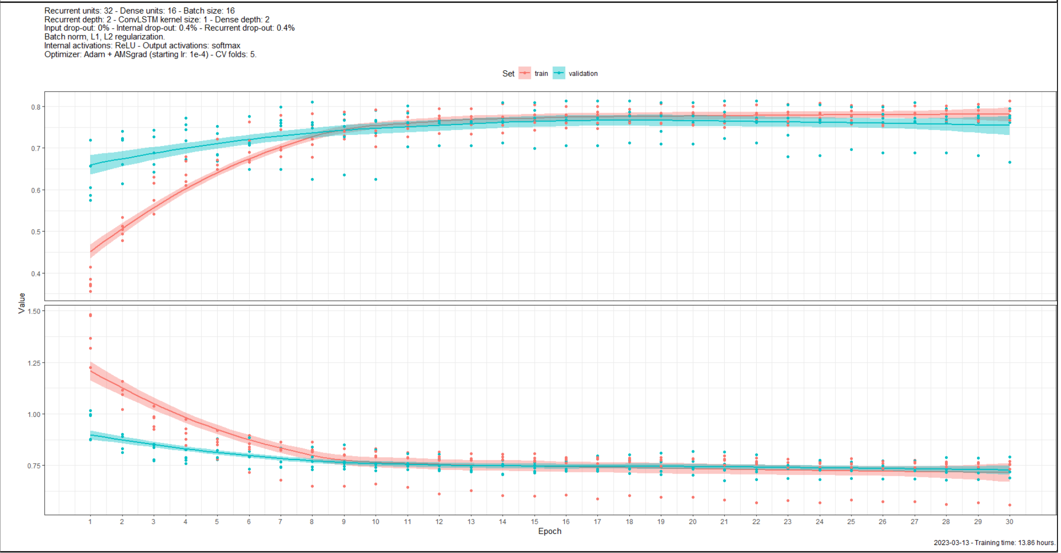


Candidate model	Epochs	Recurrent units	Dense units	Batch size	Recurrent depth	Convolutional RNN Kernel	Dropout rate input layer
1	28	32	16	32	2	1	0.1
2	1	128	64	64	2	2	0.1
3	44	64	32	64	3	1	0
4	8	64	32	16	1	1	0
5	27	32	16	16	2	1	0

Candidate model	Dropout rate recurrent layer	Dropout rate output layer	Time	(unit)	Validation loss	Train loss	Validation accuracy (%)	Train accuracy (%)
1	0	0.5	3.446	hours	0.66	0.65	0.74	0.77
2	0	0.5	12.618	hours	0.90	1.47	0.67	0.36
3	0	0.5	9.054	hours	0.88	0.83	0.69	0.74
4	0.2	0	2.020	hours	0.95	0.66	0.61	0.76
5	0.4	0.4	2.772	hours	0.73	0.72	0.76	0.78

The same normalization metrics used in the training of the selected model were used also to pre-process test data, a set consisting of 36 patients for a total of 273 days of weaning, which were not seen by the model until the testing phase. Figure 3 shows how predictors don't have significant differences from the training and validation data so that the test set is representative of the assumed clinical scenario. It is composed of 213 (77%) days when SBT was not attempted, 29 (11%) days when it failed, and 33 (12%) SBT successes that led to weaning (% on the total is reported in parentheses).

Figure 5 - Training and Validation learning curves of the selected architecture training (Candidate model 5), showing a plateau in loss and accuracy around epoch 15, selected to train the final model.



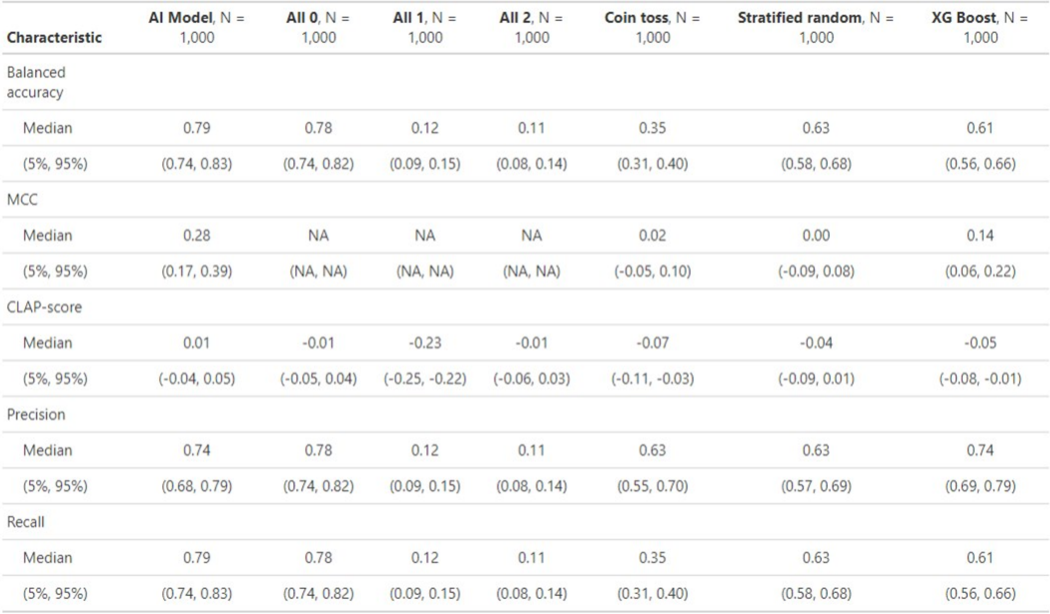
Recurrent units: 32 - Dense units: 16 - Batch size: 16
Recurrent depth: 2 - ConvLSTM kernel size: 1 - Dense depth: 2
Input drop-out: 0% - Internal drop-out: 0.4% - Recurrent drop-out: 0.4%
Batch norm, L1, L2 regularization.
Internal activations: ReLU - Output activations: softmax
Optimizer: Adam + AMSgrad (starting lr: 1e-4) - CV folds: 5.

2023-03-13 - Training time: 13.86 hours.

Final model performance

Table VI shows the model comparison. Models predicting a single class, randomly, and XGB served as benchmarks. The metrics chosen are Balanced Accuracy, Matthews Correlation Coefficient (in its R3 multiclass implementation), the proposed custom CLAP score, Precision, and Recall. Comparison have been estimated through 1000 bootstrap intervals to allow comparison among the models.

Table VI - Comparison of each model against the observed results in the test set, showing balanced accuracy, MCC, CLAP custom score, precision, and recall. For metrics defined for binary classifiers, a weighted average on the class proportion of one-vs-all is implemented.



Characteristic	**AI Model**, N = 1,000	**All 0**, N = 1,000	**All 1**, N = 1,000	**All 2**, N = 1,000	**Coin toss**, N = 1,000	**Stratified random**, N = 1,000	**XG Boost**, N = 1,000
Balanced accuracy							
Median	0.79	0.78	0.12	0.11	0.35	0.63	0.61
(5%, 95%)	(0.74, 0.83)	(0.74, 0.82)	(0.09, 0.15)	(0.08, 0.14)	(0.31, 0.40)	(0.58, 0.68)	(0.56, 0.66)
MCC							
Median	0.28	NA	NA	NA	0.02	0.00	0.14
(5%, 95%)	(0.17, 0.39)	(NA, NA)	(NA, NA)	(NA, NA)	(-0.05, 0.10)	(-0.09, 0.08)	(0.06, 0.22)
CLAP-score							
Median	0.01	-0.01	-0.23	-0.01	-0.07	-0.04	-0.05
(5%, 95%)	(-0.04, 0.05)	(-0.05, 0.04)	(-0.25, -0.22)	(-0.06, 0.03)	(-0.11, -0.03)	(-0.09, 0.01)	(-0.08, -0.01)
Precision							
Median	0.74	0.78	0.12	0.11	0.63	0.63	0.74
(5%, 95%)	(0.68, 0.79)	(0.74, 0.82)	(0.09, 0.15)	(0.08, 0.14)	(0.55, 0.70)	(0.57, 0.69)	(0.69, 0.79)
Recall							
Median	0.79	0.78	0.12	0.11	0.35	0.63	0.61
(5%, 95%)	(0.74, 0.83)	(0.74, 0.82)	(0.09, 0.15)	(0.08, 0.14)	(0.31, 0.40)	(0.58, 0.68)	(0.56, 0.66)

Considering the bootstrapped 90% confidence intervals, AI model was significantly more accurate than the XG model, which has as predictors only the clinical information of the previous day. It also scores significantly better than the random ones in balanced accuracy, MCC, and recall .

Similarity in most of the metrics with the model that suggested every day not to attempt SBT ("All 0" in Table VI), as well as a CLAP-score around zero, means that the model would have suggested similar course of action as the observed clinical management (which assumes that on most of the days, RT will fail until the respiratory failure cause is addressed).

To better understand predictions, the confusion matrix for the AI model is provided in Table VIII, which predicted early in the morning a similar behaviour to the clinical one that later happened during the day.

Table VIII – Confusion matrix of predicted vs observed classes by the final model on the Test set.

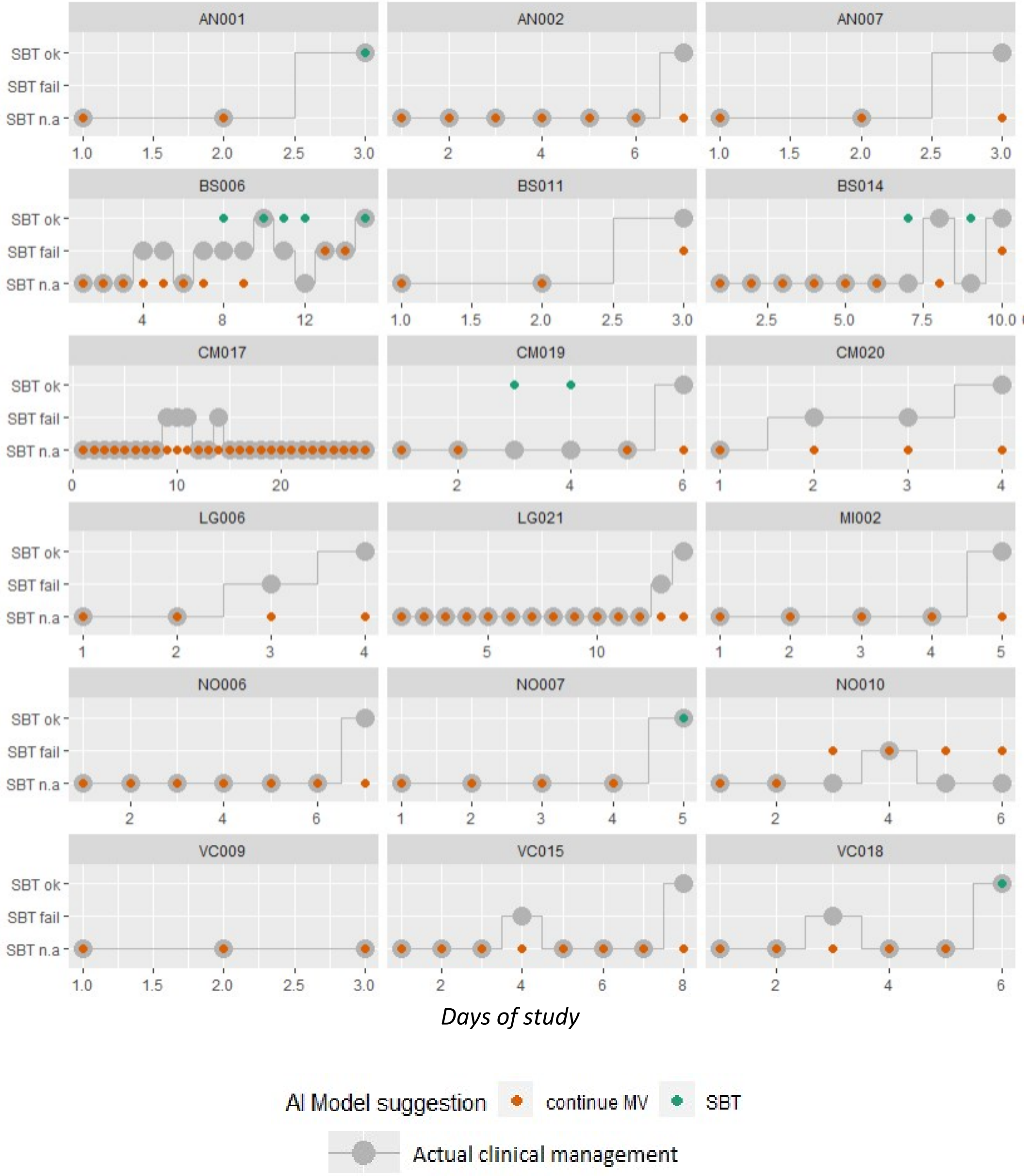
		Observed (actual clinical management)		
		0 – SBT not attempted	1 – SBT was attempted but failed	2 – SBT succeded
Model prediction (suggested clinical management)	0: SBT not attempted (because of RT failure)	204	24	21
	1: SBT failure* *thus weaning is unlikely	3	3	3
	2: SBT success (leading to weaning)	5	2	8

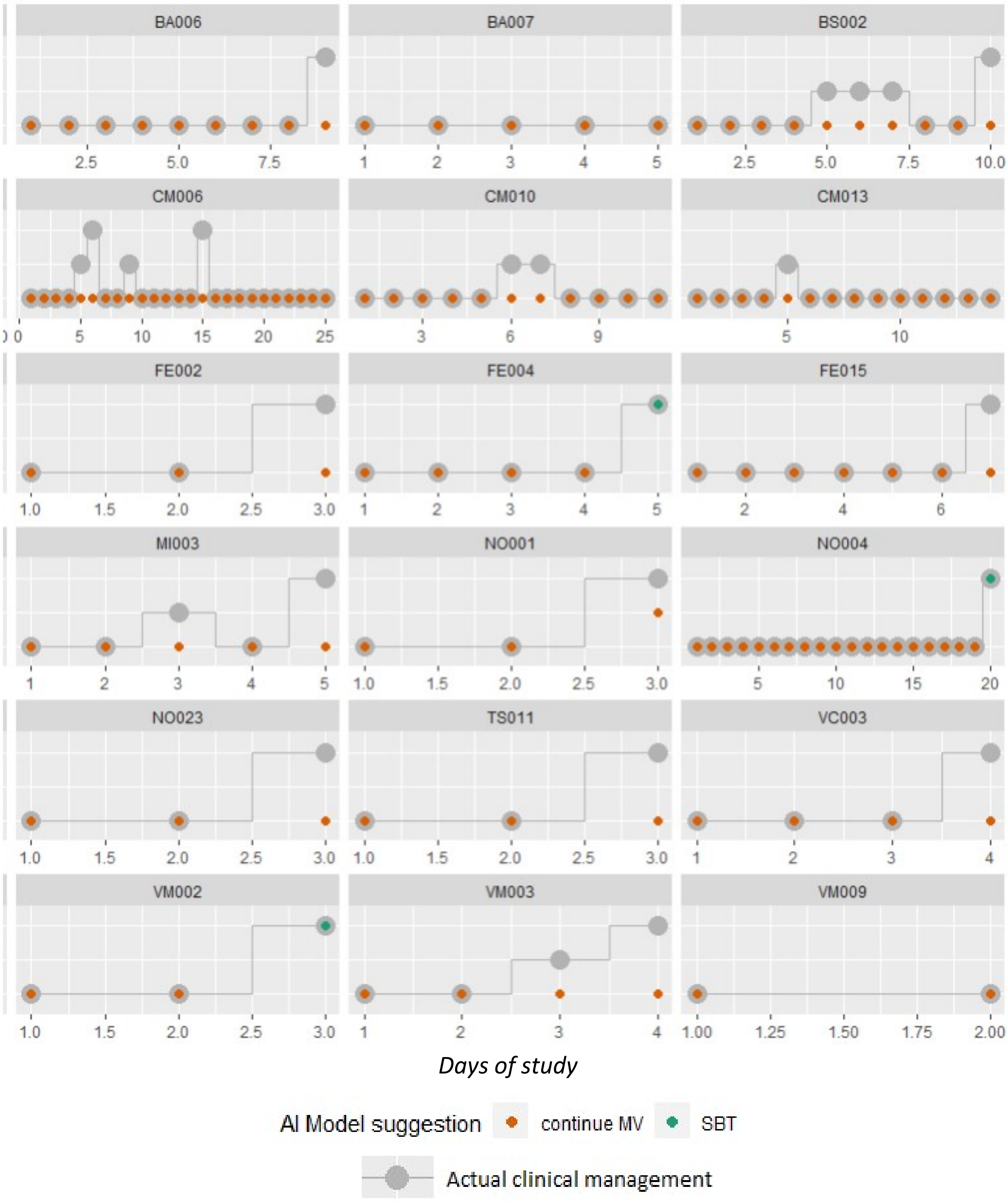
Even if each error is scored according to the custom metric we proposed, wrong predictions may mean different things, according to the individual patient history, which is plot alongside the AI model prediction that happened early in the morning in Figure 6.

Patients BS006 and BS014 both have a "SBT suggested" prediction on a day where RT was negative, and both were successfully weaned the following day. Given the limitations of the observational study, there is no way to know if SBT would have succeded on the day RT failed, thus leading to our model to anticipate it.

Similarly, CM006 suffers from weaning failure, requiring reintubation in 48h after their successful SBT, while the model always advised to continue MV. It was evaluated as an error in the performance evaluation, but implementing AI model's suggested course of action may have prevented the patient from weaning failure.

Figure 6 - AI Model evaluation is displayed by showing a different subplot for each patient in the test set. Clinical history is displayed with a series of grey dots connected by a line, showing what happened (Y axis, with the 3 possible outcomes of Table III) during each day (on the X axis) from study admission to weaning. Predictions of the model are displayed as a colored dot for each day. Whenever a colored dot (predicted) is displayed in the same location as a grey dot (observed), it means the model predicted early in the morning what was observed on that day of the study. Color represents our AI model suggestion: either it is to continue mechanical ventilation (in red, valid both for "SBT not attempted" and "SBT fail") or to suggest a Spontaneous Breathing Trial (blue)

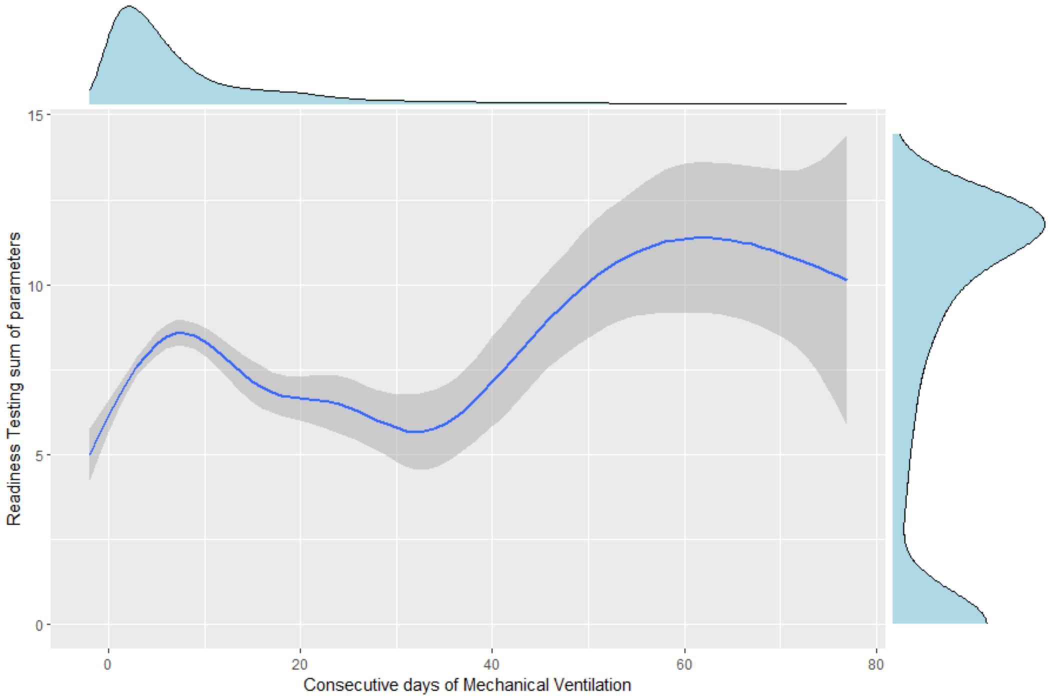

AN001
AN002
AN007
BS006
BS011
BS014
CM017
CM019
CM020
LG006
LG021
MI002
NO006
NO007
NO010
VC009
VC015
VC018
SBT ok
SBT fail
SBT n.a
1.0
1.5
2.0
2.5
3.0
2
4
6
8
12
5.0
7.5
10.0
0
10
20
1
3
5
Days of study
AI Model suggestion
● continue MV
● SBT
Actual clinical management


BA006
BA007
BS002
CM006
CM010
CM013
FE002
FE004
FE015
MI003
NO001
NO004
NO023
TS011
VC003
VM002
VM003
VM009
Days of study
AI Model suggestion
continue MV
SBT
Actual clinical management

DISCUSSION

None of the predictors alone is consistently able to identify whether a patient could be weaned or not, despite most of them carrying information, even in a non-linear way, such as in Figure 5. This is known from previous studies and it has been confirmed in our preliminary analysis.

Figure 7 - Readiness Test (RT) is a score that is used to identify the patients that would likely succeed in a Spontaneous Breathing Trial (SBT). It is comprised of 12 different parameters which are measured once per day, so the sum of them ranges between 0 and 12, where 12 represents the Readiness for SBT. The score among all patients shows a complex non-linear behavior, which is represented by a smooth line fitted to the individual data points (not shown), which represents a generalized additive model. The marginal distribution of days in the study shows that most of the data are collected in the first 10 days of Mechanical Ventilation, while the marginal distribution of RT shows that most of the values are either around minimum or maximum.



Moreover, both RT and SBT are not perfect tests, and further research, as stated in the introduction section, is needed to investigate a combination of predictors (clinical or ventilatory variables) to optimize the weaning process. Data-intensive settings like this one have the potential to be investigated with Machine Learning or Artificial Intelligence approaches.

To provide a meaningful comparison a series of models have been built, each of them evaluated with multiple metrics. From those comparisons, we can affirm that there is sufficient information in MV track data to improve on the current state-of-the-art. A model that always predicted the same class or random provided reasonable estimates for classification metrics, that acted as the lower threshold. On the contrary, the Extreme Gradient Boosted Trees model was trained as the industry standard for tabular data (only baseline and daily clinical data), so it represented a scenario mostly similar to the clinician who had no access to all the data about MV in the previous days. This model, as expected, did not significantly improve clinical management, since the latter has improved throughout the years to use the best available information.

To improve further, an architecture that could make full use of available time series was necessary. Recurrent Neural Networks are the state-of-the-art option to do that, and they haven't been applied to weaning, yet. Their ability to outperform other models can be seen by the difference in the metrics displayed.

A particular mention is to the custom score. It has been developed to highlight improvements in the current clinical management. As a design choice, it was extremely conservative. Given the restrictions of the observational study, we didn't have access to SBT results after RT failure or weaning outcome after a failed SBT, even if the literature suggests that a non-negligible fraction of patients can be weaned in either case. We chose to evaluate as -30% those cases, with respect of the difference in average proportions between improved and worsened cases. At the same time, we evaluated errors in each prediction that prolonged the MV while the observed result was weaning, without regard to a late failure of the weaning attempt. Lastly, we counted as errors all the cases where SBT was suggested on the day before it succeeded, even if it is plausible that part of those happen on a ready-to-wean patient. In conclusion, the score was built to highlight "testable improvement", disregarding due to the impossibility to test it in an observational experiment the real "potential improvement".

Limitations of the model

Data from the ICU setting are usually noisy and/or missing, and our setting is no different. Some predictors had to be excluded due to too many missing values, but a consistent fraction of them remains. The limited size of the memory of the mechanical ventilator, which started to overwrite after 24h had passed since the last download, prevented a truly complete series of continuous MV track data. But even if we had access to all available data recorded from the ventilator, there are still well-known challenges to developing predictive models using vital sign data, which include the presence of recording errors, omissions, and outliers in measurements during the ICU stay. The papers of Arcentales and Precup implement spectral analysis of heart and respiratory variables. Since we did not have such high-frequency data, we could not include those predictors in our work. Table 2 of the article details the extracted features from the data available in the DB. Also, Power Index analyses (Chaparro & Giraldo, 2014) and expiratory peak characterization are not possible on data that are not high-frequency (Correa et al., 2010) In a similar way, Trapero's analyses on the same database are also not replicable, given that it is not possible to visualize in sufficient detail the variability of heart rate and respiratory rate behavior, which generates a synchronization rhythm in the high-frequency interval and a secondary rhythm in the low-frequency range, as can be seen in the works of Arcentales and Orini (Arcentales et al., 2015; Orini et al., 2008).

On the contrary, other authors have preferred different approaches to the task, all stemming from similar datasets. Castiñeira predicts the number of days spent in the ICU starting from monitoring the data collected in the first 24 hours. The weaning identification problem is thus transformed into a classification (duration of ICU hospitalization >4 days) instead of the success or failure of weaning. Hsu's work (Hsu et al., 2013) is structured as an RCT. It starts by highlighting how the precision of successful weaning of a doctor is around 40% and then proposes a Clinical Decision Support System to increase this predictive capacity. In general, the judgments of doctors are tending to be unreliable. Therefore, decreasing dependence on their knowledge, experience, and skills is promising for increasing the rate of predictability if objective data and effective variables can be identified to determine the success of weaning from mechanical ventilation. The CDSS development is tested in a real clinical setting comparing it with the medical decision that involves only the evaluation of the clinician. The developed CDSS is effective in identifying the earliest time for weaning from the ventilator for a patient to resume and sustain breathing spontaneously. Yu's paper (Yu et al., 2019, 2020) proposes an Inverse Reinforcement Learning approach to understand how to learn the latent reward function of the clinician and reproduce decision-making capacity through an algorithm. Using real treatment trajectories, they infer the latent reward functions of physicians during their decisions on mechanical ventilation and sedative dosing in intensive care units. Finally, Hagan (Hagan et al., 2020) discuss a way to validate mechanical protective ventilation of the lungs via an early clinical alert system.

[Generalizability of the model] Model generalizability refers to the possibility to apply it to different clinical settings from the ones in which it was developed. Since it was trained and tested on the same set of equipment, external validation is advised before wider use. Not all mechanical ventilators acquire the same variables, with the same patterns of missing data and with the same artifacts or errors during reading. This applies also to the clinical workflow: it was common that some centers didn’t report each of the 12 predictors composing the RT. Since all of them were needed to proceed with the SBT, it was a sound clinical decision, but it prevented our model from learning the differential impact of the predictors on the SBT outcome.

Clinical implications of model adoption

As stated before, the AI model has never been meant to replace clinical judgment, but to enhance it with customized suggestions. Further studies need to be conducted, preferably in an experimental setting, to properly evaluate AI improvement over current care.

Yet, there are some barriers to clinical adoption. The first one is calibration, since “SBT suggested” doesn’t need to be the most plausible prediction, for it to be part of a decision support system. The second one is the explainability of the model, which has been widely discussed. The third one, probably the most important, is the structure of the model itself. At present, the model is trained to best separate the three classes, but a wrong (or correct) prediction doesn’t have the same clinical consequences for all of them. Moreover, data are incomplete since the counterfactual can’t be observed and consequently act as an input to the model. To obtain a better model, new data collected in a way that prevents those limitations, are required.

Interpretability and Explainability of the model

[Interpretability with global methods] Along with prediction performance, a barrier to widespread clinical adoption is the interpretability of models. Especially neural networks could be seen as black boxes. Interpretability can be defined as the degree to which a human can understand the cause of a decision. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. It helps the developer debug and improve the model, build trust, justify model predictions, and gain insights. The increased need for machine learning interpretability is a natural consequence of the increased use of machine learning (83).

Moreover, the European Union General Data Protection Regulation requires a right to an explanation, stating as “[the data subject should have] the right … to obtain an explanation of the decision reached” (84).

Interpretability is built of many layers: the first one is algorithm transparency, which is intrinsic to the chosen approach. Another one helps to understand how different parts of the model affect predictions, while the last one could explain why a model makes a particular prediction for a specific instance. While interpretability and explainability can both be used, there is a difference with the term explanation, which will be used for individual predictions (85)

From all the published methods, the global model-agnostic method of choice for this task was Feature Importance through Permutation(86,87), since it has a nice interpretation that provides a highly compressed global insight into the model's behavior. For interpretation purposes, feature importance does not add up, since the interaction between features is represented in the relative importance of both of them (88,89).

[Explanation of a single prediction] SHAP (SHapley Additive exPlanations) (90) is a method to explain individual predictions. SHAP is based on the game's theoretically optimal Shapley values, which suffer from extreme computational burden to be calculated and are prone to be misinterpreted. The Shapley value of a feature value is not the difference of the predicted value after removing the feature from the model training. Instead, they are the difference between the actual prediction and the mean prediction. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction (91,92).

Clinical implications of broader AI tools adoption

The artificial Intelligence model's adoption of clinical practice has already shown great potential to improve human health and life quality. Even if it can't be completely understood, an analytical approach should favor AI model adoption in some specific clinical settings. AI is not necessarily useful in all contexts and not for all problems, but further research is required to identify specific clinical areas where an improvement can be brought.

Whenever there is a decision to be made, the potential for AI adoption can be assessed. In our paper, we proposed an advisory role (in the form of a non-compulsory decision support system), but the AI model's adoption could be imagined with a wide variety of degrees of autonomy. Trust is not given, but it is built progressively by final users (usually, clinicians) in a multi-step process, starting from the data on which the model is built, passing onto algorithms and performance, and ending on implementation research considerations, including the liability profile of a doctor using it in the day-by-day practice.

Dwelling deeper on the first (DATA), good AI models are built from a good dataset. Unfortunately, since datasets are usually built by humans, they incorporate the same biases and judgment errors, from which the AI model learns how to best approximate them. This exposes the risk of gold-plating expert AI prediction, overinflating its expected performance, especially on poorly represented groups like minorities or rare conditions. Moreover, since what that data to collect, how to preprocess them (e.g. which categories to use), and the setting in which are collected, are all decided by humans, they are prone both to incorporate additional biases or to represent, as in a photograph, only one side of reality. All of this without ever considering that the “wider” (more predictors screened), the easier is to overfit and learn patterns that are only by chance present in the data, and the issue of missing data, which in some settings (e.g. ICU) are common and may potentially be informative, since they represent a result of the daily ward activity.

Moreover, some considerations on the algorithm itself and how the model is trained and built should be made. Not all techniques are created equal, and different problems require different approaches. Nevertheless, they all train to optimize a single measure (either single itself or composed of various weighted elements) on a population level, while humans can take into account different competing metrics and optimize them on a single clinical case. Also, the metric to optimize is defined by humans, thus it represents implicitly or explicitly judgments and ethical considerations on which is the optimal reality to reach through progressive increments in the model’s performance. A model that is used in a clinical setting requires a robust scientific design to be trusted, as it enables to exposure and audit of the entire process from its ideation to its adoption.

In addition to that, since at the moment CLAP-score is only used in the evaluation phase, the training and validation processes do not try to optimize it. Instead of a loss function that aims to perfectly separate the prediction classes (as the cross-entropy, used in the current study to train the RNN), a custom one may be developed to truly capture the clinical impact of the predicted versus observed clinical management. Discussions on how to implement this type of measure will have to be taken in thorough concert with specialists in the field.

A special note should be made on errors since they also embody ethical and technical considerations in their scoring. But they have a further level of analysis, which is they substantially differ from the one committed by humans, even if we refer to them with the same word. From the point where the model is trained, an error is not made by distraction/chance/tiredness, but it is already embodied in the model’s parameter, and it arises from a specific combination of input in a straightforward fashion (even if it can’t be known in advance, and thus prevented). Guilt, as it is defined in humans, is a concept that can be translated only by distributing it to all the actors involved in model design (who decided which data to collect and how, who decided which metric and how to optimize, who chose the algorithm and wrote the experimental protocol, who audited it before its adoption, who decided to use it on that specific case,…). Furthermore, it is composed of a reducible (e.g. by improving the model’s parameters or data) and an irreducible portion, but both of them are implicitly accepted when the model is deployed. The choice of whether that is acceptable or not is, once again, made by humans. The reducible one (at least, theoretically) is the bias, but unbiasedness has a context-specific meaning.

Stemming from the fact that biases in humans or models are potentially similar to each other (since the latter is optimized to mimic the behavior of the former), a critical aspect required to adopt a model is its explainability, both on the global level and on the single prediction. Interpretability (which in this context we use as a synonym) is a human right, especially for decisions that can impact one's own life directly, as it is common in the medical field. This is no different from asking an expert opinion (e.g. on a fellow senior consultant) and then asking to explain it, being subjected to a similar problem: is it a post hoc explanation or does it represent the true process of thought? Or also, do their biases and limits to his/her knowledge translate well to my patient or my reference population?

In conclusion, building a case for AI adoption is a challenging task, with different intertwined layers of complexity both on the technical and ethical side, but given its potential use to advance people's health access and access, clinicians should take part in all the steps of its design, evaluation, and adoption process.

CONCLUSIONS

The model we developed shows ability to anticipate well and early in the morning what would be the clinical management that observed later in the day, without significant differences from actual clinical choices. Further studies overcoming analyzed challenges and limitations, as the experimental design and the availability of quality data, may improve its performances. Conservative choices on scoring may have hidden potential benefit of it. Overall, as the best of our knowledge, this is the first RNNs-based networks trained on the whole patient history of minute-by-minute MV data to provide day-by-day predictions of weaning outcome; the technology proved superior to the analysis of clinical data only, even using state-of-art non-deep machine learning models. Retaining a lower enough complexity to be used in the clinical practice (simultaneous predictions for 32 admitted patients in the ICU took around 15s on a standard laptop), the purposed approach provides both space for further investigation and improvement to support clinical decisions, in a data-intensive settings like ICUs.

BIBLIOGRAFIA

1. Fan E, Brodie D, Slutsky AS. Acute Respiratory Distress Syndrome: Advances in Diagnosis and Treatment. JAMA. 2018 Feb 20;319(7):698.

2. Wilson JG, Calfee CS. ARDS Subphenotypes: Understanding a Heterogeneous Syndrome. In: Vincent JL, editor. Annual Update in Intensive Care and Emergency Medicine 2020 [Internet]. Cham: Springer International Publishing; 2020 [cited 2023 Feb 14]. p. 67–79. (Annual Update in Intensive Care and Emergency Medicine). Available from: http://link.springer.com/10.1007/978-3-030-37323-8_5

3. MDShijing Jia, MD RCH. Modes of Mechanical Ventilation. Post TW Ed UpToDate [Internet]. [cited 2023 Feb 14];Waltham, MA: UpToDate Inc. Available from: http://www.uptodate.com

4. Mechanical ventilation. American College of Chest Physicians' Consensus Conference - PubMed [Internet]. [cited 2023 Feb 14]. Available from: https://pubmed.ncbi.nlm.nih.gov/8252973/

5. Girard TD, Alhazzani W, Kress JP, Ouellette DR, Schmidt GA, Truwit JD, et al. An Official American Thoracic Society/American College of Chest Physicians Clinical Practice Guideline: Liberation from Mechanical Ventilation in Critically Ill Adults. Rehabilitation Protocols, Ventilator Liberation Protocols, and Cuff Leak Tests. Am J Respir Crit Care Med. 2017 Jan 1;195(1):120–33.

6. Ouellette DR, Patel S, Girard TD, Morris PE, Schmidt GA, Truwit JD, et al. Liberation From Mechanical Ventilation in Critically Ill Adults: An Official American College of Chest Physicians/American Thoracic Society Clinical Practice Guideline. Chest. 2017 Jan;151(1):166–80.

7. 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 Jan 1;195(1):115–9.

8. International consensus conferences in intensive care medicine: Ventilator-associated Lung Injury in ARDS. This official conference report was cosponsored by the American Thoracic Society, The European Society of Intensive Care Medicine, and The Societé de Réanimation de Langue Française, and was approved by the ATS Board of Directors, July 1999. Am J Respir Crit Care Med. 1999 Dec;160(6):2118–24.

9. Blackwood B, Burns KEA, Cardwell CR, O'Halloran P. Protocolized versus non-protocolized weaning for reducing the duration of mechanical ventilation in critically ill adult patients. Cochrane Database Syst Rev. 2014 Nov 6;2014(11):CD006904.

10. Burns KEA, Meade MO, Lessard MR, Hand L, Zhou Q, Keenan SP, et al. Wean earlier and automatically with new technology (the WEAN study). A multicenter, pilot randomized controlled trial. Am J Respir Crit Care Med. 2013 Jun 1;187(11):1203–11.

11. Ely EW, Baker AM, Dunagan DP, Burke HL, Smith AC, Kelly PT, et al. Effect on the duration of mechanical ventilation of identifying patients capable of breathing spontaneously. N Engl J Med. 1996 Dec 19;335(25):1864–9.

12. Esteban A, Frutos F, Tobin MJ, Alía I, Solsona JF, Valverdú I, et al. A comparison of four methods of weaning patients from mechanical ventilation. Spanish Lung Failure Collaborative Group. N Engl J Med. 1995 Feb 9;332(6):345–50.

13. Kollef MH, Shapiro SD, Silver P, St John RE, Prentice D, Sauer S, et al. A randomized, controlled trial of protocol-directed versus physician-directed weaning from mechanical ventilation. Crit Care Med. 1997 Apr;25(4):567–74.

14. Burns KEA, Rizvi L, Cook DJ, Lebovic G, Dodek P, Villar J, et al. Ventilator Weaning and Discontinuation Practices for Critically Ill Patients. JAMA. 2021 Mar 23;325(12):1173–84.

15. Girard TD, Kress JP, Fuchs BD, Thomason JWW, Schweickert WD, Pun BT, et al. Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (Awakening and Breathing Controlled trial): a randomised controlled trial. Lancet Lond Engl. 2008 Jan 12;371(9607):126–34.

16. Boles JM, Bion J, Connors A, Herridge M, Marsh B, Melot C, et al. Weaning from mechanical ventilation. Eur Respir J. 2007 May 1;29(5):1033–56.

17. Conditions needing ventilation - UpToDate [Internet]. [cited 2023 Feb 14]. Available from: https://www.uptodate.com/contents/image?imageKey=PULM%2F67457&topicKey=PULM%2F1640&search=mechanical%20ventilation&rank=1~150&source=see_link

18. Coplin WM, Pierson DJ, Cooley KD, Newell DW, Rubenfeld GD. Implications of extubation delay in brain-injured patients meeting standard weaning criteria. Am J Respir Crit Care Med. 2000 May;161(5):1530–6.

19. Unroe M, Kahn JM, Carson SS, Govert JA, Martinu T, Sathy SJ, et al. One-year trajectories of care and resource utilization for recipients of prolonged mechanical ventilation: a cohort study. Ann Intern Med. 2010 Aug 3;153(3):167–75.

20. Trivedi V, Chaudhuri D, Jinah R, Piticaru J, Agarwal A, Liu K, et al. The Usefulness of the Rapid Shallow Breathing Index in Predicting Successful Extubation: A Systematic Review and Meta-analysis. Chest. 2022 Jan;161(1):97–111.

21. Ely EW, Baker AM, Evans GW, Haponik EF. The prognostic significance of passing a daily screen of weaning parameters. Intensive Care Med. 1999 Jun;25(6):581–7.

22. Hall JB, Wood LD. Liberation of the patient from mechanical ventilation. JAMA. 1987 Mar 27;257(12):1621–8.

23. Esteban A, Alía I, Tobin MJ, Gil A, Gordo F, Vallverdú I, et al. Effect of spontaneous breathing trial duration on outcome of attempts to discontinue mechanical

ventilation. Spanish Lung Failure Collaborative Group. Am J Respir Crit Care Med. 1999 Feb;159(2):512–8.

24. Vallverdú I, Calaf N, Subirana M, Net A, Benito S, Mancebo J. Clinical characteristics, respiratory functional parameters, and outcome of a two-hour T-piece trial in patients weaning from mechanical ventilation. Am J Respir Crit Care Med. 1998 Dec;158(6):1855–62.

25. Betbesé AJ, Pérez M, Bak E, Rialp G, Mancebo J. A prospective study of unplanned endotracheal extubation in intensive care unit patients. Crit Care Med. 1998 Jul;26(7):1180–6.

26. Epstein SK, Nevins ML, Chung J. Effect of unplanned extubation on outcome of mechanical ventilation. Am J Respir Crit Care Med. 2000 Jun;161(6):1912–6.

27. Vitacca M, Vianello A, Colombo D, Clini E, Porta R, Bianchi L, et al. Comparison of two methods for weaning patients with chronic obstructive pulmonary disease requiring mechanical ventilation for more than 15 days. Am J Respir Crit Care Med. 2001 Jul 15;164(2):225–30.

28. 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 Feb;128(2):343–51.

29. Marshall DC, Komorowski M. Is artificial intelligence ready to solve mechanical ventilation? Computer says blow. Br J Anaesth. 2022 Feb;128(2):231–3.

30. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical Verification of A Clinical Decision Support System for Ventilator Weaning. Biomed Eng OnLine. 2013;12(Suppl 1):S4.

31. Arizmendi C, Romero E, Alquezar R, Caminal P, Diaz I, Benito S, et al. Data mining of patients on weaning trials from mechanical ventilation using cluster analysis and neural networks. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. Minneapolis, MN: IEEE; 2009 [cited 2022 Jul 19]. p. 4343–6. Available from: http://ieeexplore.ieee.org/document/5332742/

32. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015 Jan 6;162(1):55–63.

33. G.M. Moons K, G. Altman D, B. Reitsma J, P.A. Ioannidis J, Macaskill P, W. Steyerberg E, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med [Internet]. 2015 Jan 6 [cited 2023 Feb 14]; Available from: https://www.acpjournals.org/doi/10.7326/M14-0698

34. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec 16;18(12):e323.

35. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2019. p. 220–9. (FAT* ’19). Available from: https://doi.org/10.1145/3287560.3287596

36. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d’Alché-Buc F, et al. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program) [Internet]. arXiv; 2020 [cited 2022 Sep 9]. Available from: http://arxiv.org/abs/2003.12206

37. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circ Cardiovasc Qual Outcomes [Internet]. 2020 Oct [cited 2022 Sep 9];13(10). Available from: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.120.006556

38. Vitacca M, Clini E, Porta R, Ambrosino N. Preliminary results on nursing workload in a dedicated weaning center. Intensive Care Med. 2000 Jun;26(6):796–9.

39. Tehrani FT, Roum JH. Intelligent decision support systems for mechanical ventilation. Artif Intell Med. 2008 Nov;44(3):171–82.

40. Clemen RT. Making hard decisions with decision tools. 3rd ed. Mason, OH: South-Wester/Cengage Learning; 2010.

41. Lee CY, Chien CF. Pitfalls and protocols of data science in manufacturing practice. J Intell Manuf. 2022 Jun;33(5):1189–207.

42. Crimaldi F, Della Corte F. Gestione della ventilazione durante la fase di weaning. Analisi dei dati campionati dal ventilatore durante uno studio multicentrico italiano. Università del Piemonte Orientale; 2015.

43. Carenzo L, Navalesi P. NAVA vs. PSV: randomized multicenter trial. [Novara]: Università del Piemonte Orientale; 2014.

44. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. Second edition. New York NY: Springer; 2021. 607 p. (Springer texts in statistics).

45. Blagec K, Dorffner G, Moradi M, Samwald M. A critical analysis of metrics used for measuring progress in artificial intelligence [Internet]. arXiv; 2021 [cited 2022 Sep 9]. Available from: http://arxiv.org/abs/2008.02577

46. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. ACM Trans Knowl Discov Data. 2012 Dec;6(4):1–21.

47. Lones MA. How to avoid machine learning pitfalls: a guide for academic researchers [Internet]. arXiv; 2022 [cited 2022 Sep 9]. Available from: http://arxiv.org/abs/2108.02497

48. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/

49. Grolemund HW and G. Welcome | R for Data Science [Internet]. [cited 2023 Feb 16]. Available from: https://r4ds.had.co.nz/

50. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. J Open Source Softw. 2019;4(43):1686.

51. Marwick B, Boettiger C, Mullen L. Packaging Data Analytical Work Reproducibly Using R (and Friends). Am Stat. 2018 Jan 2;72(1):80–8.

52. Landau WM. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. J Open Source Softw. 2021;6(57):2959.

53. Allaire JJ. quarto: R Interface to 'Quarto' Markdown Publishing System [Internet]. 2022. Available from: https://github.com/quarto-dev/quarto-r

54. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

55. Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible Summary Tables with the gtsummary Package. R J. 2021;13(1):570–80.

56. Cox V. Exploratory Data Analysis: What Data Do I Have? In: Translating Statistics to Make Decisions [Internet]. Berkeley, CA: Apress; 2017 [cited 2023 Feb 16]. p. 47–74. Available from: http://link.springer.com/10.1007/978-1-4842-2256-0_3

57. Grolemund HW and G. 7 Exploratory Data Analysis | R for Data Science [Internet]. [cited 2023 Feb 16]. Available from: https://r4ds.had.co.nz/exploratory-data-analysis.html

58. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns. 2021 Nov;2(11):100336.

59. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res. 2010;11(70):2079–107.

60. Cawley GC, Talbot NLC. Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters.

61. James G, Witten D, Hastie T, Tibshirani R. Resampling Methods. In: An Introduction to Statistical Learning [Internet]. New York, NY: Springer US; 2021 [cited 2022 Jul 19]. p. 197–223. (Springer Texts in Statistics). Available from: https://link.springer.com/10.1007/978-1-0716-1418-1_5

62. Zwanenburg A, Löck S. familiar: End-to-End Automated Machine Learning and Model Evaluation [Internet]. 2021. Available from: https://github.com/alexzwanenburg/familiar

63. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 785–94. (KDD '16). Available from: http://doi.acm.org/10.1145/2939672.2939785

64. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 [cited 2022 Jul 31]; Available from: https://arxiv.org/abs/1412.6980

65. The Berkeley Review. ADAM: A Method for Stochastic Optimization [Internet]. theberkeleyview. 2015 [cited 2023 Mar 2]. Available from: https://theberkeleyview.wordpress.com/2015/11/19/berkeleyview-for-adam-a-method-for-stochastic-optimization/

66. Chollet F, Allaire J, others. R Interface to Keras [Internet]. GitHub; 2017. Available from: https://github.com/rstudio/keras

67. Gullì A, Pal S. Deep learning with Keras: implement neural networks with Keras on Theano and TensorFlow. Birmingham Mumbai: Packt Publishing; 2017. 303 p.

68. Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

69. Opitz J, Burst S. Macro F1 and Macro F1 [Internet]. arXiv; 2021 [cited 2023 Mar 2]. Available from: http://arxiv.org/abs/1911.03347

70. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009 Jul 1;45(4):427–37.

71. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation [Internet]. arXiv; 2020 [cited 2023 Mar 2]. Available from: http://arxiv.org/abs/2010.16061

72. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE. 2017 giu;12(6):e0177678.

73. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Jan 2;21(1):6.

74. Fowlkes EB, Mallows CL. A Method for Comparing Two Hierarchical Clusterings. J Am Stat Assoc. 1983 Sep 1;78(383):553–69.

75. Chaparro JA, Giraldo BF. Power index of the inspiratory flow signal as a predictor of weaning in intensive care units. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. Chicago, IL: IEEE; 2014 [cited 2022 Jul 19]. p. 78–81. Available from: http://ieeexplore.ieee.org/document/6943533/

76. Correa LS, Laciar E, Mut V, Giraldo BF, Torres A. Multi-parameter analysis of ECG and Respiratory Flow signals to identify success of patients on weaning trials. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology [Internet]. Buenos Aires: IEEE; 2010 [cited 2022 Jul 19]. p. 6070–3. Available from: http://ieeexplore.ieee.org/document/5627623/

77. Arcentales A, Caminal P, Diaz I, Benito S, Giraldo BF. Classification of patients undergoing weaning from mechanical ventilation using the coherence between heart rate variability and respiratory flow signal. Physiol Meas. 2015 Jul 1;36(7):1439–52.

78. Orini M, Giraldo BF, Bailon R, Vallverdu M, Mainardi L, Benito S, et al. Time-frequency analysis of cardiac and respiratory parameters for the prediction of ventilator weaning. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. Vancouver, BC: IEEE; 2008 [cited 2022 Jul 31]. p. 2793–6. Available from: https://ieeexplore.ieee.org/document/4649782/

79. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical Verification of A Clinical Decision Support System for Ventilator Weaning. Biomed Eng OnLine. 2013;12(Suppl 1):S4.

80. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak. 2019 Apr;19(S2):57.

81. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak. 2020 Jul;20(S3):124.

82. Hagan R, Gillan CJ, Spence I, McAuley D, Shyamsundar M. Comparing regression and neural network techniques for personalized predictive analytics to promote lung protective ventilation in Intensive Care Units. Comput Biol Med. 2020 Nov;126:104030.

83. Molnar C. Chapter 10 Neural Network Interpretation | Interpretable Machine Learning [Internet]. [cited 2023 Feb 27]. Available from: https://christophm.github.io/interpretable-ml-book/neural-networks.html

84. Vollmer N. Recital 71 EU General Data Protection Regulation (EU-GDPR) [Internet]. SecureDataService; 2022 [cited 2023 Feb 23]. Available from: https://www.privacy-regulation.eu/en/recital-71-GDPR.htm

85. Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences [Internet]. arXiv; 2018 [cited 2023 Feb 20]. Available from: http://arxiv.org/abs/1706.07269

86. LSTM Feature Importance [Internet]. [cited 2023 Feb 23]. Available from: https://kaggle.com/code/cdeotte/lstm-feature-importance

87. Molnar C. 8.5 Permutation Feature Importance | Interpretable Machine Learning [Internet]. [cited 2023 Feb 23]. Available from: https://christophm.github.io/interpretable-ml-book/feature-importance.html

88. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously [Internet]. arXiv; 2019 [cited 2023 Feb 23]. Available from: http://arxiv.org/abs/1801.01489

89. Wei P, Lu Z, Song J. Variable importance analysis: A comprehensive review. Reliab Eng Syst Saf. 2015 Oct 1;142:399–432.

90. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–77. (NIPS'17).

91. Molnar C. 9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning [Internet]. [cited 2023 Feb 23]. Available from: https://christophm.github.io/interpretable-ml-book/shap.html

92. Molnar C. 9.5 Shapley Values | Interpretable Machine Learning [Internet]. [cited 2023 Feb 23]. Available from: https://christophm.github.io/interpretable-ml-book/shapley.html#shapley

SUPPLEMENTARY MATERIALS

- Table 1 – Included and Excluded Predictors
- Preliminary Literature Review

Additional Table I – Candidate predictors, with inclusion or exclusion criteria

Source	***Feature***	***Description***	***Included as a predictor (if excluded, reason)***
Baseline	Patient ID		Already accounted for
	Informed consent	A categorical variable, it defines whether consent was obtained	Irrelevant
	Dropout	Categorical variable, the patient dropped out	Irrelevant
	Unique ID	5 letters identifying each patient in the study: • 2 letters for the hospital • 3 numbers for patient ID e.g. FE002 -> was the patient n.2 of Ferrara hospital	Yes
	Hospital name	A categorical variable (13 distinct values), defines the hospital location: • Vercelli • Novara • Ancona • Brescia • Busto • Castellammare • Ferrara • Forlì-Cesena • Legnano • Milano Fatebenefratelli • Trieste • Vimercate • Bari	Already accounted for
	Hospital level	A categorical variable, it defines the hospital level (community or university)	Irrelevant
	Doctor (user) ID		Irrelevant
	Ventilation type	Categorical variable, defines ventilation type: • Pressure Support Ventilation (PSV) • Neurally Adjusted Ventilatory Assist (NAVA)	Yes
	Gender		Yes
	Age	In years	Yes

	Weight	In Kg	Irrelevant
	Height	In cm	Irrelevant
	Body Mass Index	measure of body fat based on height and weight	Yes
	Ideal Body Weight	It is based on height, gender, and age, and represents appropriate body weight	Yes
	Reason for Mechanical Ventilation	A categorical variable, it may be one of • Sepsis • Pneumonia • COPD exacerbation • Trauma/polytrauma • Post-surgical complication • Heart Failure • Acute Respiratory Distress Syndrome • Other (…)	Yes
	Simplified Acute Physiology Score	Score that predicts hospital mortality upon ICU admission	Yes
	Sequential Organ Failure Assessment	Score that predicts ICU mortality based on lab results and clinical data, on admission day	Daily measurements were preferred
	Arterial pH	Represent acid-base equilibrium in arterial blood, on admission day	Daily measurements were preferred
	Arterial PaO2	The partial pressure of O2 in arterial blood, on admission day	Daily measurements were preferred
	Arterial PaCO2	The partial pressure of CO2 in arterial blood, on admission day	Daily measurements were preferred
	FiO2	Fraction of inspired O2 at which Arterial Blood Gases were obtained, on admission day	Daily measurements were preferred
	Post-expiratory end pressure (PEEP)	↓	Continuous measurements were preferred
	PaO2/FiO2		Daily measurements were preferred
	Sedation type	A categorical variable, it may be oral or intravenous administration of anaesthetics	Irrelevant

	Sedation drug and dose	Categorical and numerical variable (the dose is expressed in dose/time or dose/weight/time, accordingly), one or more of • Thiopenthal • Propofol • Clonidine • Midazolam • Diazepam • Remifentanil • Fentanyl • Morphine • Hydroxyzine • Benzodiazepines • Chlorpromazine • Promazine • Haloperidol • Fentanyl transcutaneous	Temporal inconsequentiality, it was calculated only at the end of the ICU admission period
	Richmond Agitation Sedation Scale	The score used to measure the agitation or sedation level of a person, ranging from +4 (combative) to -5 (unarousable)	Daily measurements were preferred
	Date-time of enrolment		Irrelevant
	Date of hospital admission		Irrelevant to the model (but it was used to check data consistency among datasets during the cleaning phase)
	Date of ICU admission		Irrelevant to the model (but it was used to check data consistency among datasets during the cleaning phase)
	Date of mechanical ventilation start		Irrelevant to the model (but it was used to check data consistency among datasets during the cleaning phase)

	Date of mechanical ventilation end		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Date of ICU discharge		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Date of hospital discharge		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Total days of mechanical ventilation		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Total days of ICU stay		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Total days of Hospital stay		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Days of Mechanical ventilation before study enrolment		Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Patient outcome at discharge	Categorical, it may be: ● Discharged ● ICU death ● Hospital death	Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Patient outcome after 90 days	Categorical, it may be: ● Alive ● Dead ● Still admitted to hospital	Temporal inconsequentiality, it was inserted only at the end of the ICU admission period
	Complications	↓	Daily measurements were preferred

Daily registry	Type	↑	A baseline measurement was preferred
	Patient ID	↑	Already accounted for
	Informed consent	↑	Irrelevant
	Dropout	↑	Irrelevant
	Unique ID	↑	Yes
	User (doctor) ID	↑	Irrelevant
	Hospital ID	↑	Irrelevant
	Date of insertion	Date of when the entry in the database was physically recorded	Irrelevant
	Date of clinical event	Date of when the entry is referring to	Already accounted for
	Sequential Organ Failure Assessment	Score that predicts ICU mortality based on lab results and clinical data	Yes
	Arterial pH	Represent acid-base equilibrium in arterial blood, taken after the SBT test (if performed)	Yes
	Arterial PaO2	The partial pressure of O2 in arterial blood, taken after the SBT test (if performed) in mmHg	Yes
	Arterial PaCO2	The partial pressure of CO2 in arterial blood, taken after the SBT test (if performed) in mmHg	Yes
	Clinical Pulmonary Infection Score	Score that assists in diagnosing ventilator-associated pneumonia by predicting the benefit of pulmonary cultures	Irrelevant
	Nasogastric tube substitution	Categorical variable (true/false) and if substituted, the reason was listed	Irrelevant
	Clinical notes	In free text	Most of the values were NAs, remaining values were not standardized
	LOG file	Link to the appropriate log file	Already accounted for
	REG file	Link to appropriate registry file	Already accounted for
	TRD file	Link to appropriate track file	Already accounted for
	Extubation	A categorical variable representing patient status: “was the patient in invasive MV on that day?”	Yes (part of the outcome)
	Tracheostomy	A categorical variable stating if the patient did undergo a tracheostomy	Excluded since it depended on study protocol in this setting

	Reintubation	A categorical variable representing if the patient did undergo a reintubation on that day	Yes (part of the outcome)
	Patient death		Yes (part of the outcome)
	Non-invasive ventilation (NIV)	A categorical variable representing if the patient was non-invasively ventilated after successful extubation	Irrelevant to our study
	NIV reason	Categorical variable, either Respiratory Failure prevention or treatment	Irrelevant to our study
	NIV interface	The device used for NIV is one of the following: • Helmet • Full-face • Nasal mask • Mouth-nose mask	Irrelevant to our study
	NIV type	Categorical NIV ventilation type, one of the: • NAVA-NIV • PSV • NIV-CPAP	Irrelevant to our study
	Readiness testing criteria	Set of clinical criteria, each a binary dichotomous variable 1. <2 aspirations/h 2. audible cough on aspiration 3. no distress (diaphoresis, accessory muscles, paradoxical rhinitis) 4. GCS ≥ 8 5. [RASS between -1 and +1] 6. Heart Rate (HR) ≤ 120bpm and Systolic Blood Pressure (SBP) between 90 and 180mmHg 7. Dopamine or dobutamine ≤ 5 and NorAdr ≤ 0.1 mcg/kg/min 8. Pao2/FiO2 ≥ 150mmHg 9. PEEP ≤ 8 cmH2O 10. Respiratory Rate (RR) ≤ 40/min 11. Tidal volume (Vt) ≥ 5m/kg (Ideal Body Weight) 12. pH ≥7.35	Yes (but only the total sum)
	SBT early stopping criteria	Set of clinical criteria, each a binary dichotomous variable 1. [RASS between -1 and +1] 2. RR >= 35	Excluded to limit predictors quantity in an over-fitting prone situation

		3. Respiratory Distress symptoms 4. SpO2 <90%, even though FiO2 increase 5. SBP outside the 90-180 range 6. HR > 140bpm	
	SBT failure criteria	Set of clinical criteria, each a binary dichotomous variable 1. [1 or more of SBT early stopping criteria] 2. RASS outside -4 / +3 range 3. no distress (e.g. increased accessory muscle activity) 4. no dyspnoea (VAS >= 7) 5. RR/Vt < 105 bpm/L 6. PaO2 < 60mmHg with FiO2 > 50% 7. [PaO2/FiO2 <200] – not used 8. pH < 7.32 or delta > 0.07 9. PaCo2 > 50mmHg or delta > 8mmHg 10. SBP < 90 mmHg 11. the appearance of new cardiac arrhythmias	Excluded to limit predictors quantity in an over-fitting prone situation
	Reintubation criteria	Set of clinical criteria, each a binary dichotomous variable 1. clinical emergencies (respiratory, cardiac arrest or gasping) 2. worsening GCS or RASS (coma or agitation requiring continuous IV sedation) 3. need for medications to maintain SBP ≥ 90 mmHg 4. airway obstruction (stridor or tirage) or tracheobronchial secretions that are difficult to manage 5. Respiratory distress with SpO2 <90% and RR >35 bpm and clinical signs, despite oxygen administration	Irrelevant for current research
	Daily complications	Categorical, one or more among • Myocardial infarction or cardiogenic shock • Sepsis • Renal failure • Pancreatitis • Critical Illness Polyneuropathy • Sinusitis • Pneumonia	Excluded to limit predictors quantity in an over-fitting prone situation

		• Massive Blood Loss • Pulmonary embolism • Infection at study entry • Pneumothorax • New intubation beyond 48h	
	Study days	Calculated from date of study enrollment	Yes
TRD track data	Dynamic characteristics [ml/cmH2O]	Compliance	Yes
	Elastic power [cmH2O/L]	A measure of the work that has to be exerted by the muscles of inspiration to expand the lungs	No, missing values >67%
	End-expiratory Flow [L/min]		Yes
	Positive end-expiratory Pressure [cmH2O]	PEEP is the pressure in the lungs (alveolar pressure) above atmospheric pressure that exists at the end of expiration. There are two types of PEEP: intrinsic and extrinsic. - Intrinsic PEEP depends on the progressive air trapping after incomplete expiration - extrinsic PEEP is set directly on the ventilator.	Yes
	Expiratory resistance [cmH2O/L/s]	Is the resistance of the respiratory tract to airflow during expiration This includes the resistance of the upper airways and oro-tracheal tube	No, missing values >67%
	Minute expired Volume [L/min]	the volume of air that moves out of the lungs during a minute	Yes
	Current expired volume [ml]	the volume of air that moves out of the lungs during a single breath	Yes
	O2 concentration (%)	Percentage of oxygen in a specific volume of air	Yes
	Inspiratory resistance [cmH2O/L/s]	Is the resistance of the respiratory tract to airflow during inhalation. This includes the resistance of the upper airways and oro-tracheal tube	No, missing values >67%
	Minute inspired Volume [L/min]	the volume of air that moves into the lungs during a minute	Yes
	Current inspired volume [ml]	the volume of air that moves into the lungs during a single breath	Yes

	Mean Airway Pressure [cmH2O]	The average pressure in the airways during the inspiratory phase	Yes
	Measured Respiratory Rate [/min]	number of breaths per minute recorded	Yes
	Spontaneous respiratory Rate [/min]	number of spontaneous breaths per minute	Yes
	Edi peak [µV]	Diaphragm electromyography peak (NAVA specific)	Yes
	Edi min [µV]	Diaphragm electromyography minimum level (NAVA specific)	Yes
	Plateau Pressure [cmH2O]	Plateau Pressure: is the pressure applied to small airways and alveoli at the end of inspiration during positive-pressure mechanical ventilation	Yes
	Peak Pressure [cmH2O]	It is the highest pressure level applied to the respiratory system during inspiration. It depends on any airways resistance.	Yes
	Backup switches [/min]	Number of backup switches in a minute	Yes
	Backup percentage [%/min]	The percentage of breath switched to the backup mode in a minute	Yes
	Static compliance [ml/cmH2O]	Statice Compliance. It is the reciprocal of elastance measured in a dynamic manner	No, missing values >67%
	P0.1 [cmH2O]	The negative pressure generated at 0.1 sec. from the beginning of the inspiratory phase	Yes
	Mechanical ventilator respiratory work [Joule/L]	Ventilator work of breathing	Yes
	Patient-ventilator respiratory work [Joule/L]	Patient work of breathing	Yes
	Spontaneous Breathing Index	Defined as the ratio of respiratory frequency to tidal volume (RR/Vt). People on a ventilator who cannot tolerate independent breathing tend to breathe rapidly (high frequency) and shallowly (low tidal volume), and will therefore have a high RSBI	Yes
	Spontaneous minute expired volume [L/min]	the volume of air that moves out of the lungs during a minute due to spontaneous breath	Yes
	Airleak percentage [%]	Percentage of air leaks during not invasive ventilation	No, missing values >67%
LOG	Date		Irrelevant
	Time		Irrelevant

	Message ID	Numeric code associated to the type of message	Irrelevant
	Type of message	A categorical variable representing an alarm, a piece of information to the clinician or a change of settings (including switch on/off the ventilator)	Too many parameters without a clear link to the clinical outcome
	Additional info	About message, stored in free text	Too much free text without a clear link to the clinical outcome

PRELIMINARY LITERATURE REVIEW

CONTENTS

To answer the questions outlined in the thesis's objectives, it was necessary to define an analysis plan as detailed as possible starting from the existing literature.

To do this, we started with the only existing systematic review on the topic of mechanical ventilation (the other systematic review on the subject (Kwong et al., 2019) is partial and collects far less data on studies and their quality). The work in question is interesting for two main reasons: it analyzes 1342 articles available on the various databases to identify the 95 articles available on the subject in February 2021 and analyzes the Risk of Bias according to the items of the TRIPOD Statement (Collins et al., 2015) allowing to quickly evaluate the quality of the design and analysis.

If, however, on the one hand, the systematic review lays solid foundations for the use of artificial intelligence algorithms during mechanical ventilation of patients, on the other hand, the extreme heterogeneity of the primary outcomes evaluated and the methodologies used did not allow to directly compare the studies with each other except for the overall risk of bias. In addition, none of the selected articles deals with data from ventilations with NAVA.

For these reasons, a deeper analysis was performed on the supplementary materials, firstly by skimming the articles analyzed in the systematic review to identify those that dealt only with predicting the success of the weaning attempts, then by collecting detailed information from each of them regarding the methods.

SELECTION OF RELEVANT ARTICLES

Of the 95 articles included in the systematic review, 28 were deemed directly relevant to our work. The summary table of evidence of the systematic review had as its column "Main Prediction Outcome", which allowed only articles containing some reference to "Weaning failure or success" or similar descriptions to be selected. To verify that none of the relevant works had been excluded, the titles were also screened. In cases where neither the title nor other information gathered by the authors of the systematic review was sufficient to determine its degree of relevance to the current study, the abstract was analyzed. Some of the doubtful or partially relevant works were however included in the subsequent analysis, to include them in the comparative analysis of materials and methods.

INFORMATION COLLECTED FOR EACH ITEM

The selected articles were read, getting together materials and methods from the section

- The type of "raw" data collected, grouping it into three categories: data from a database of patients undergoing weaning (where each patient represents a record), continuously recorded data collected by ventilators/other equipment, or only clinical data from patients
- The frequency of data collection at the source, also grouping here into various categories
    - High-frequency data (>=1 Hz)
    - Medium frequency data (between 1 Hz and a record every 30 minutes)
    - Low-frequency data (with the time between two consecutive records >30m)
    - no data was recorded "continuously", for those articles that collected a single measurement for a given ventilation parameter or directly a summary statistic
- data frequency at the end of the cleaning and pre-processing phase, to capture the type of data entering the model, with the same frequency categories seen in the previous point.

The categories were created with the high-frequency data in mind that they had the potential to allow frequency domain transformations, for example by allowing heart rate (HR) variability over time to be calculated as a predictive model parameter, as well as a simple average on the minute, hence the limit at 1Hz. The 30-minute threshold instead was chosen because it allows highlighting patterns that have longer-term variability.

Some peculiarities emerged from the analysis: the two most used databases were WEANDB (data collected continuously at high frequency but only during a Spontaneous Breathing Trial) and MIMIC (a complete clinical database with also clinical notes in free text and examinations, whose data collected for ventilation parameters are medium frequency) (.

In addition to this preliminary information, a short note of free text has also been recorded to capture specific peculiarities of the article with potential relevance.

The table obtained is as follows:

Table I – Literature Review, displaying the original bibliographic reference, the type of data that are collected, either from (W)eaning attempts, from (V) ventilation or (C)linical variables, and the Frequency (High, Medium, Low or No frequency if there was no real-time recording of variables). WeanDB and MIMIC refer to the most common benchmark databases available for studies.

Author	*Raw data type*	*Raw data frequency*	*Pre-analysis frequency*	*Notes*
(4)	WC	HF	HF	Spectral analyses
(5)	WEANDB	HF	MF	Cross-validated feature selection
(6)				
(7)	W	NF	NF	
(8)	VC	MF	MF	Early warning system + missing values handling
(9)	WEANDB	HF	MF	Inspiratory cycle power analysis
(10)	WEANDB	HF	LF	
(11)	WEANDB	HF	HF	Spectral analysis
(12)	VC	MF	LF	Normalization
(13)	WEANDB	HF	HF	Symbolic feature generation

(14)	WEANDB	HF	MF	Sliding window and clustering approach
(15)	W	NF	NF	Neural Network (not Deep Learning)
(16)	V	MF	MF	SVM Feature selection and automatic predictor forecast
(17)	WC	NF	NF	Neural Network
(18)	WC	NF	NF	
(19)	CW	LF	LF	SVM Feature Selection is a RCT with a Clinical Decision Support System
(20)	WC	NF	NF	Argues about selection bias
(21)	C	NF	NF	Resampling to correct the class imbalance
(22)	C	NF	NF	
(23)	?	?	?	Clustering (data mining approach)
(24)	V (W) C	HF	HF	Breath power estimate + RR and HR correlation analyses
(25)	WEANDB	HF	HF	Symbolic feature generation and Sliding Window Variance Analysis for dimensionality reduction

(26)	VC	?	NF	Variable Inflation Factor to address collinearity and Bayesian decision analysis
(27)	VC	HF	HF/LF	Temporal pattern classification
(28)				
(29)	MIMIC3	MF	MF	SVM for missing values
(30)	MIMIC3	MF	MF	Actor critic reinforcement learning
(31)	MIMIC2	NF	NF	On newborn

Starting from the preceding articles, a series of ideas have been collected on how to analyze the data collected for the current work, detailed in the following subsections.

Definition of the outcome

Most articles agree in defining a weaning attempt as a "success" if it allows the patient not to be re-intubated for the following 48 (or 72) hours, see as an example the work of Fabregat.

To discriminate true extubation from an artifact, Fabregat (Fabregat et al., 2021) and the logs contained in the medical record. Other studies, on the other hand, do not address the problem of determining when extubation occurred, often because they are based solely on Spontaneous Breathing Trial data.

Patient selection

A particular focus goes on the management of the patient who died during the study. Different studies had varying approaches in this regard. Some studies, such as all those based on WEANDB do not deal directly with the issue, while others such as Castiñeira and Fabregat (Castiñeira et al., 2020; Fabregat et al., 2021) exclude them. However, information about the patient's future death is never available at the time of prediction in a hypothetical real-world setting, so it cannot be used as a weaning predictor, as it is a post-intervention variable. The risk that follows is to create an optimized predictive model on a slice of patients not detectable a priori.

Hsieh (M.-H. Hsieh et al., 2018) on the contrary, considers failed attempts at extubation, including them in the analysis in cases where death occurs less than 72h after extubation, a preferable approach.

In studies with an approach like that of the NAVA vs PSV dataset, in which a Spontaneous Breathing Trial is carried out to determine subsequent extubation: the use of the model is placed in a phase after the evaluation of the clinician, without suggesting extubation from the data recorded continuously but limiting itself to giving a "second opinion" on the doctor's decision. In most of the studies, see as an example the work of Fabregat or Hsieh (Fabregat et al., (Fabregat et al., 2021; M. H. Hsieh et al., 2019; M.-H. Hsieh et al., 2018) the clinician begins with the analysis of a series of variables that allow determining the adequacy for the suspension of mechanical ventilation, called Readiness Testing. If the patient meets the criteria, the staff proceeds to perform the spontaneous breathing (SBT) trial, then evaluates a set of 60 different predictors (Fabregat) or 34 (Hsieh); and if SBT is successful, ventilatory support is removed. This corresponds to the experimental protocol applied in the study on whose data the thesis is based.

This limits the patient population to those assessed as weanable by the clinician, therefore training data contains only patients considered ready for extubation. The main consequence is that it is not possible to directly investigate what would have happened in cases where the model suggested a different course of action than that of the clinician, see for example the work of Hsieh or Kuo (M.-H. Hsieh et al., 2018; Kuo et al., 2015) in which the clinical database contains only those patients for whom the physician was confident enough to trust the model used, excluding possible extubations suggested by the model but not recommended. Paradoxically, as Yu suggests ((Yu et al., 2019) the more competent the clinician is (the more consistent his decisions are and the variability in treating similar cases is reduced) the more difficult it is for the algorithm to learn the consequences of behaviors that are not explored.

Feature Extraction I – summaries and transformations

The inputs for the predictive model don't use raw clinical data, but they find appropriate transformations and representations to summarize the most important characteristics. All articles contain a section dedicated to extracting features from the collected data series.

- Castiñeira (Castiñeira et al., 2020) extracts summary stats of time series such as correlation structure, distribution, entropy, and stationarity. In the article, the projection of the time series of vital signs in the "feature space" allows for capturing the underlying statistical and temporal behavior of the vital signs.
- Other relevant characteristics can be obtained from ventilator data, as in Chaparro's paper, where moving average models (ARMA) and autoregressive models with exogenous input (ARX) are extracted with autoregressive models (AR) and are then used to transform continuous data into input for models.
- These parameters can be combined to compose clinical scores (Fabregat et al., 2021) already present in the literature and guidelines. Please refer to the table below for a detailed list of these parameters, either recorded directly by the fan or calculated later.
- For continuous variables (all and except BMI and gender), see as an example Fabregat (Fabregat et al., 2021), where they were normalized to have zero mean and variance equal to 1. In the proposed analysis, data with a normal z-score with $p \geq 0.985$ were considered outliers and removed. The same normalization approach applied to all variables is also found in the work of Hsieh (M.-H. Hsieh et al., 2018) and others, except for the removal of outliers, which are retained instead.
- To the various approaches that consider only the analysis of Spontaneous Breathing Trials is added the work of Tsai (Tsai et al., 2019) that analyzes the 48 before extubation, more similar to the dataset analyzed for this thesis work

Feature extraction II—Moving window

To find the optimal size of the time window, several approaches are proposed in the papers analyzed:

- In Castineira’s paper (Castiñeira et al., 2020), patients for whom there is a signal interruption greater than 20 minutes were excluded, while inferior intervals of an absence of signal in the data have been filled with a Gaussian process, as defined by Rasmussen (Rasmussen & Williams, 2006) and O'Hagan (O’Hagan, 1978)
- In Fabregat's paper (Fabregat et al., 2021) the variables are subjected to a 2-hour moving average with 20-minute "bins" and common origin at the patient's entry into the ICU. From this, the derived variables (e.g. clinical scores) are then calculated. This is also done to reduce the number of missing data points due to errors during the vital signs or ventilation recording process.
- In Yu (Yu et al., 2019) as well as in other articles, it is emphasized that data in intensive care can be irregular, prone to errors and that some physiological parameters are detected several times an hour, while others are measured only once in several hours. As a solution, classical interpolation methods are insufficient and it is therefore suggested to use Support Vector Machines (SVM) to predict missing ones, obtaining complete data for each patient, with a temporal resolution of 10 minutes, from the time of admission to that of discharge.
- In Arizmendi ((Arizmendi et al., 2009) it is proposed to determine the optimal width of the rolling window in the range from 3 to 100 consecutive time points using two U-Mann-Whitney tests between groups S (successful weaning) and F (failure) and between groups S and R (late failure). From the total p-values of the comparison between these groups, minimum local and global values were obtained for a given window width. A similar approach is in Giraldo (Giraldo et al., 2006) where is the p-value used on a series of progressively longer movable windows that identifies the optimal length of respiratory cycles for analysis, then selecting the lowest value

Once the optimal width of the window has been determined and the completeness ensured, the classification of predictors according to Fabregat (Fabregat et al., 2021) is proposed for the types of variables recorded, completed also with an additional list of possible predictors from the various studies analyzed:

Table II – Candidate predictors. A list of candidate predictors has been extracted from the relevant studies, citing the original study whenever possible. Classification provides a distinction between categories of possible predictors.

Typology	Description	List of variables (with the source if to be calculated)
1	Time series directly obtained from the patient data stream	• Inspiratory time (TI) • Expiratory time (TE) • Breathing cycle duration (TTot) • Tidal Volume (VT) in mL/kg of ideal body weight or liters • Ventilation Mode • Airway occlusion voltage at 0.1 seconds (P0.1) • Peak/Maximum inspiratory pressure (MIP) • Peak flow setting • Maximum expiratory pressure (MEP) • Plateau pressure • Pressure support level • PEEP • Negative Inspiratory Force (NIF) • Static resistance (/compliance) from the inspiratory pause • Dynamic compliance • FiO2 • Respiratory rate (RR) • Heart rate (HR) • O2 Saturation (SpO2)

		• Days of mechanical ventilation • Days of ICU stay (/hospital stay)
2	Variables derived from type 1	• Respiratory rate-oxygenation index (Fabregat et al., 2021) • Work of breathing index (Mikhno & Ennett, 2012) • Rapid Shallow Breathing Index (Fabregat et al., 2021) both in itself and how much has changed between the beginning and end of SBT • Inspiratory fraction (Chaparro et al., 2012; Chaparro & Giraldo, 2014) • Half-inspired flow (Chaparro et al., 2012; Chaparro & Giraldo, 2014) • O2 saturation to inspired fraction ratio [check if it is PaO2 / FiO2 (M.-H. Hsieh et al., 2018)] • Alveolar-arterial oxygen pressure difference (Hsu et al., 2013) • From RR, HR, Ttot, and VT (Correa et al., 2010) ○ Mean Value ○ Median Absolute Deviation ○ Variation Coefficient ○ Temporal Interquartile range ○ Standard Deviation frequency ○ Aymmetry coefficient ○ Kurtosis coefficient ○ 95% cumulated energy frequency ○ First- and third-quartile frequencies ○ Spectral interquartile range ○ Lempel-Ziv complexity

		• From any continuously recorded variable (Verduijn et al., 2007) ○ Mean value ○ Median value ○ Soft minimum (0.05 percentile) ○ Soft maximum (0.95 percentile) ○ Soft empirical range (difference between the soft minimum and maximum), ○ Change (difference between first and last value) ○ Variance around the mean ○ Slope coefficient of a linear model fitted to the data
3	Information on (discrete) events obtained from ICU personnel	• Number of previous MV events • Number of weaning attempts • Weaning Method ○ T-piece ○ Pressure support <8 cmH2O ○ Spontaneous Breathing Trial • Total Cumulative Dose (sedatives and analgesics) • Total Given Dose (sedatives and analgesics) • Glasgow Coma Scale • RASS - Richmond Agitation-Sedation Scale • APACHE II Score [MOF/Immuno-/AKI + Age + Temperature + MAP + pH + HR + RR + Na+/K+ + S-Cr + Hct% + WBC + GCS + FiO2] • TISS scale - simplified therapeutic intervention scoring system

		• Systolic/Mean/Diastolic Arterial Pressure • Cardiac Output • Blood chemistry tests ○ Arterial pH ○ PaCO2 ○ PaO2 ○ Hemoglobin ○ White Blood Cells ○ Hematocrit (%) ○ Blood Urea Nitrogen ○ Creatinine ○ Sodium ○ Potassium ○ Calcium ○ Phosphate ○ Albumin ○ Glucose ○ Base excess
4	Demographic variables obtained at entry (assumed not to change during hospitalization)	• Age at admission to ICU (both continuous and categorized at >=65 years) • Body Mass Index • Gender • Main diagnosis (etiology of Respiratory Failure) ○ Chronic Heart Failure (Congestive or MI) ○ Neurological Disease (Central or Neuromuscular) ○ Pulmonary Disease (Pneumonia or COPD)

		o Abdominal Disease (Renal or gastrointestinal) o Post-operative o Other (incl. Cancer) • Reason for intubation o Hypoventilation o Airway obstruction o Pneumonia o Cardiogenic pulmonary edema o Septic shock o COPD o Post-operative • Number of comorbidities • Comorbidities o Cardiovascular accident o Chronic heart failure o Chronic lung disease o Chronic hemodialysis (renal disease) o Chronic liver disease o Diabetes o Old stroke or neurological disorder o Active cancer o Immunocompromised • SEMICYUC Code [Spanish ICU only, not applicable] • APGAR1 and APGAR5 [only for studies containing infants, not applicable]

However, considering the features generated by data windowing as predictors of the model is a problem both from a practical point of view and also from a theoretical point of view (Verduijn et al., 2007) In practice, it is difficult to obtain "tabulable" data from the time series due to the difficulties in the different sampling times and time extensions of the recorded data. From a theoretical point of view, on the other hand, the reasons are three: the risk of overfitting that derives from an increase in the number of features; the risk that the analysis models are based on the assumption that two contiguous values are not correlated with each other (an assumption which is not valid in the case of time series) and the risk that, ignoring these correlations, you lose part of the information that was contained in these (imagine for example two slightly out of phase time series, same data but the correlation information is lost).

Feature selection

- Arcentales identifies the most relevant features through Sequential Floating Feature Selection (Pudil et al., 1994) which maximizes the proportion of patients belonging to a given class with optimal accuracy.
- 10-Fold Cross Validation is proposed in various studies. Note the case of Arizmendi (Arizmendi et al., 2009) in which it was then once the features were identified, another 10-fold CV was carried out to estimate the value of the "Final Average Test Classification result" using only the previously selected variables. The result is to conflate the results of the model, which is therefore not comparable with the others.
- Castiñeira (Castiñeira et al., 2020) proposes several "baseline" approaches, in particular removing features with low variance, using Principal Component Analysis, and removing highly correlated features to reduce the amount of redundant information. However, the approach used in the paper consists of clustering to identify a reduced number of characteristics, applying the affinity propagation method. The biggest advantage over PCA is that the resulting features can be interpreted.
- Correa (Correa et al., 2010) proposes a step-forward selection method in which variables are progressively eliminated if they are not statistically significant (with p-value <0.05). This is also not a method capable of producing reliable subsets of features, making the model incomparable with others.
- Mikhno (Mikhno & Ennett, 2012) takes a similar approach, where the Pearson correlation coefficient for each feature concerning each other feature is calculated and all features with R^2>0.7 and p-value>0 are deleted in bulk. 05. In addition to this, features recorded in <15 patients in the weaning failure group are also excluded to maintain a reasonable cohort for statistical analysis.
- Hsu (Hsu et al., 2012, 2013) implements backward feature selection.
- Tsai (Tsai et al., 2019) proposes to use the variance inflation factor (VIF) to solve the problem of collinearity, recursively eliminating one by one the most highly correlated variables, which otherwise would potentially have resulted in the

inclusion of non-relevant variables. This is not complemented by an assessment of the clinical usefulness of the collinear variables to decide which of the two to eliminate.

- Tsai (Tsai et al., 2019) suggests three possible different methods for feature selection: Multivariate adaptive regression splines (MARS), stepwise logistic regression (SLR), and random forest (RF) to rank relative importance among variables. It is not suggested a priori one over the other, but all three are implemented in parallel. Note how the relative frequency of one variable concerning the others is then used after a validation process on 100 randomly generated but stratified subsets (to reflect the proportions observed in the population of that department) to have a 1:1 balance between failed and successful weaning.
- Verdujin (Verduijn et al., 2007) discretizes all summary variables into five categories using quintile values of patient distribution; values missing in the calculated summaries were charged with the median value of that summary. Second, it selects the subset of features based on the cross-validated 10-fold information gain versus the outcome of the mechanical ventilation disruption on the univariate probability tree model.

Finally, Tsai (Tsai et al., 2019) recommends that variable variables obtained from the feature selection process (both retained and eliminated) undergo clinical validation.

Classification algorithm

Various techniques have been proposed, helping to broadly cover the entire spectrum presented in textbooks on the subject (James et al., 2021a) :

- Arcentales (Arcentales et al., 2015) proposes a "fuzzy K-NN" in which the degree of proximity is calculated with the Euclidean distance for the various classes to which they belong. However, with several predictors, Euclidean distance can’t be considered an appropriate measure.
- Mikhno (Mikhno & Ennett, 2012) proposes a simple logistic regression, trying as candidate models all combinations of 3 characteristics, with and without

interaction terms, then bootstrapped 100 times to obtain robust estimates of the area under the ROC curve.

- Castiñeira (Castiñeira et al., 2020) after an initial selection proposes a Gradient Boosting Tree due to its greater resistance to overfitting, which depends on the fact that observations that are difficult to classify have a progressively greater weight. It adds also a shrinking factor to favor smaller trees, which are also less prone to overfitting (see the relationship between Bias and Variance in the textbook (James et al., 2021b))
- Other papers try several algorithms, as in Chaparro's works (Chaparro et al., 2012; Chaparro & Giraldo, 2014) in which LDA, SVM, and CART are tested. Tsai (Tsai et al., 2019) proposes various alternatives, including Support Vector Machines, boosted logistic regression, and neural network backpropagation, then evaluating performance for each through a confusion matrix. The main issue of this approach is that, when they are evaluated on the same test data, it is not possible to determine the best, and the choice of the algorithm becomes another "parameter to be optimized", which should be optimized on the validation data and not on test one.
- Hsieh (M.-H. Hsieh et al., 2018) presents a paper based on Neural Network, then improved in the subsequent work of 2019 (M. H. Hsieh et al., 2019), in which a multilayer perceptron was used, (M. H. Hsieh et al., 2019) with a 10-fold CV to select hyperparameters, optimizers, and the most performing loss function. The model consists of an input layer (37 variables), a hidden layer of 19 dimensions, and an output layer of 2 dimensions. The network was optimized using Adam with predefined parameters as described by Kingma (Kingma & Ba, 2014) The SeLU (Scaled Exponential Linear Unit) activation function was used for each layer, and a Softmax for the output layer. A 20% dropout (to mitigate overfitting) was applied to the input layer and a 50% dropout rate to the output layer, according to suggestions in Srivastava (Srivastava et al., 2014). The categorical cross-entropy error function was used as the loss function for binary classifiers. The balanced accuracy was then calculated.

Many of these values have made it possible to obtain an area under the curve (AUC) as a performance evaluation tool. To check if this is significantly different between a simple use of the calculated variables (e.g. TISS score) and the models used, (M.-H. Hsieh et al., 2018) suggest using DeLong's AUC tests (DeLong et al., 1988) Moreover, to obtain credible estimates of confidence intervals, Mueller (Jonas S Almeida, 2013) proposes to create 100 datasets through random sampling before proceeding with cross-validation, then choosing the median performance of each algorithm and calculating its variability.

Validation (internal)

Of the various validation techniques, heterogeneous among the reference articles, there are considerations to be made:

- Leave-One-Out Cross-Validation as in (Arcentales et al., 2015; Gottschalk et al., 2000) has not been considered a viable solution. Although particularly useful since it does not tend to overestimate the Error Rate (almost all available data are used for model training), it has an important disadvantage, as described in ((James et al., 2021c). When performing LOOCV, models are trained on an almost identical set of observations. Because the average of highly correlated quantities has greater variance than the average of as many quantities that are not as highly correlated, the estimation of the test error resulting from LOOCV tends to have greater variance than the estimate of the test error resulting from other methods, such as a k-fold CV
- Castiñeira ((Castiñeira et al., 2020) validation) – 10% (test) repeated 300 times to obtain accuracy distributions and mitigate problems of specific subsets of the original data. On the choice of cut-offs for the various splits, however, no further details are provided.
- To mitigate the problem of imbalance between successful and failed attempts of weaning (class imbalance), Fabregat (Fabregat et al., 2021) proposes a validation approach "weighted" by weaning results so that the creates balanced test sets first by randomly removing part of the points of the ruling class and then

correcting with a weighted selection any remaining differences. A similar approach is that of Tsai (Tsai et al., 2019) which provides for the selection for each record of the minority class to select a random one of the majority class, thus composing the training set. This data selection process is then repeated 100 times to validate these generated random subsets.

- In which for each neural network examined, 25 independent runs of an 8-fold Cross Validation were performed, to obtain different data partitions (6 parts for training, 1 for validation, and 1 to test the model). This is followed by a further phase of feature selection (always the k-fold CV, this time with a sequential backward feature selection) repeated several times, and finally a last round of CV with only the features generated to obtain the average classification performance. However, this practice of using test data multiple times is not recommended in subsequent articles (Lee & Chien, 2022)

Calibration of the model on the population on which it could be used

It is one of the central points both for the current literature on the subject (Walsh et al., 2017) and concerning the (Walsh et al., 2017) bias (Collins et al., 2015) and in so that it was "calibrated" on the underlying epidemiological situation (in the first case) or at least that they were considered in the study of population determinants such as age or ethnicity (in the second).

Most articles are recorded as "does not mention ethnicity, generalizability or representativeness of the model" or "does not provide information on population variables" or "does not discuss the potential effects of adding variables such as ethnicity in the model" (Gallifant et al., 2022) Some positive exceptions are the work of Hsu (Hsu et al., 2013) matched cohorts" or Hagan.

The comparison, as Tsai recalls (Tsai et al., 2019) must be made between the tools used in clinical practice (eg APACHE score) and the predictive model, to compare the two real situations and be able to determine the advantage given by the use of a predictive model.

CHOICE OF THRESHOLD DECISION VALUE

As suggested in Fabregat's work incorrect prediction of a "failed" extubation would result in an unnecessary prolongation of the time on mechanical ventilation, while predicting a "successful" extubation destined to fail can potentially create complications for the patient. The risks of incorrect intervention are summarized by Garde and Boles (Boles et al., 2007; Garde et al., 2013) pointing out that the need for reintubation involves an 8-fold higher probability ratio of nosocomial pneumonia and a mortality risk 6 to 12 times higher.

Tsai's work (Tsai et al., 2019) fits into this discussion, proposing a method for evaluating the decision as a whole. In fact, in addition to the accuracy results of the predictive model, a real Bayesian Decision Analysis is carried out (better described in the work of Clemen (Clemen, 2010))). Starting from the Expected Monetary Value (EMV) and the Expected Value of Experimentation (EVE), it calculates the expected value of profit/loss of all possible actions.

Finally, a sensitivity analysis is then carried out to assess the value of the information provided by the forecasting model and identify the best weaning failure rate that maximizes the value of information to validate the forecasting model. This phase is fundamental to providing a link from predictive to prescriptive analysis, as also pointed out in the work of Lee and Chien.

BIBLIOGRAPHY

1. Kwong MT, Colopy GW, Weber AM, Ercole A, Bergmann JHM. The efficacy and effectiveness of machine learning for weaning in mechanically ventilated patients at the intensive care unit: a systematic review. Bio-Des Manuf. 2019 Mar;2(1):31–40.

2. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015 Jan 6;162(1):55–63.

3. Johnson, Alistair, Pollard, Tom, Mark, Roger. MIMIC-III Clinical Database [Internet]. PhysioNet; 2015 [cited 2022 Jul 19]. Available from: https://physionet.org/content/mimiciii/1.4/

4. Arcentales A, Caminal P, Diaz I, Benito S, Giraldo BF. Classification of patients undergoing weaning from mechanical ventilation using the coherence between heart rate variability and respiratory flow signal. Physiol Meas. 2015 Jul 1;36(7):1439–52.

5. Arizmendi C, Romero E, Alquezar R, Caminal P, Diaz I, Benito S, et al. Data mining of patients on weaning trials from mechanical ventilation using cluster analysis and neural networks. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. Minneapolis, MN: IEEE; 2009 [cited 2022 Jul 19]. p. 4343–6. Available from: http://ieeexplore.ieee.org/document/5332742/

6. Arizmendi C, Viviescas J, González H, Giraldo B. Patients classification on weaning trials using neural networks and wavelet transform. Stud Health Technol Inform. 2014;202:107–10.

7. Ashutosh K, Lee H, Mohan CK, Ranka S, Mehrotra K, Alexander C. Prediction criteria for successful weaning from respiratory support: statistical and connectionist analyses. Crit Care Med. 1992 Sep;20(9):1295–301.

8. Castiñeira D, Schlosser KR, Geva A, Rahmani AR, Fiore G, Walsh BK, et al. Adding Continuous Vital Sign Information to Static Clinical Data Improves the Prediction of Length of Stay After Intubation: A Data-Driven Machine Learning Approach. Respir Care. 2020 Sep;65(9):1367–77.

9. Chaparro JA, Giraldo BF. Power index of the inspiratory flow signal as a predictor of weaning in intensive care units. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. Chicago, IL: IEEE; 2014 [cited 2022 Jul 19]. p. 78–81. Available from: http://ieeexplore.ieee.org/document/6943533/

10. Chaparro JA, Giraldo BF, Caminal P, Benito S. Performance of respiratory pattern parameters in classifiers for predict weaning process. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. San Diego, CA: IEEE; 2012 [cited 2022 Jul 19]. p. 4349–52. Available from: https://ieeexplore.ieee.org/document/6346929/

11. Correa LS, Laciar E, Mut V, Giraldo BF, Torres A. Multi-parameter analysis of ECG and Respiratory Flow signals to identify the success of patients on weaning trials. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology [Internet]. Buenos Aires: IEEE; 2010 [cited 2022 Jul 19]. p. 6070–3. Available from: http://ieeexplore.ieee.org/document/5627623/

12. 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 Programs Biomed. 2021 Mar;200:105869.

13. Garde A, Voss A, Caminal P, Benito S, Giraldo BF. SVM-based feature selection to optimize sensitivity–specificity balance applied to weaning. Comput Biol Med. 2013 Jun;43(5):533–40.

14. Giraldo B, Arizmendi C, Romero E, Alquezar R, Caminal P, Benito S, et al. Patients on Weaning Trials from Mechanical Ventilation Classified with Neural Networks and Feature Selection. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. New York, NY: IEEE; 2006 [cited 2022 Jul 19]. p. 2195–8. Available from: https://ieeexplore.ieee.org/document/4462225/

15. Gottschalk A, Hyzer MC, Geer RT. A Comparison of Human and Machine-based Predictions of Successful Weaning from Mechanical Ventilation. Med Decis Making. 2000 Apr;20(2):160–9.

16. Hagan R, Gillan CJ, Spence I, McAuley D, Shyamsundar M. Comparing regression and neural network techniques for personalized predictive analytics to promote lung protective ventilation in Intensive Care Units. Comput Biol Med. 2020 Nov;126:104030.

17. Hsieh MH, Hsieh MJ, Chen CM, Hsieh CC, Chao CM, Lai CC. An Artificial Neural Network Model for Predicting Successful Extubation in Intensive Care Units. J Clin Med. 2018 Aug 25;7(9):240.

18. Hsieh MH, Hsieh MJ, Cheng AC, Chen CM, Hsieh CC, Chao CM, et al. Predicting weaning difficulty for planned extubation patients with an artificial neural network. Medicine (Baltimore). 2019 Oct;98(40):e17392.

19. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical Verification of A Clinical Decision Support System for Ventilator Weaning. Biomed Eng OnLine. 2013;12(Suppl 1):S4.

20. Kuo HJ, Chiu HW, Lee CN, Chen TT, Chang CC, Bien MY. Improvement in the Prediction of Ventilator Weaning Outcomes by an Artificial Neural Network in a Medical ICU. Respir Care. 2015 Nov 1;60(11):1560–9.

21. Jonas S Almeida MM. Can Machine Learning Methods Predict Extubation Outcome in Premature Infants as well as Clinicians? J Neonatal Biol [Internet]. 2013 [cited 2022 Jul 19];02(02). Available from: http://www.omicsgroup.org/journals/can-machine-learning-methods-predict-extubation-outcome-in-premature-infants-as-well-as-clinicians-2167-0897.1000118.php?aid=15779

22. Mueller M, Wagner CL, Annibale DJ, Hulsey TC, Knapp RG, Almeida JS. Predicting Extubation Outcome in Preterm Newborns: A Comparison of Neural Networks with Clinical Expertise and Statistical Modeling. Pediatr Res. 2004 Jul;56(1):11–8.

23. Alvaro Silva SO. Clustering Data Mining models to identify patterns in weaning patient failures. Int J Biol Biomed Eng [Internet]. 2016;10. Available from: http://ijdri.com/ijbbe/2016/a422010-068.pdf

24. Precup D, Robles-Rubio CA, Brown KA, Kanbar L, Kaczmarek J, Chawla S, et al. Prediction of extubation readiness in extreme preterm infants based on measures of cardiorespiratory variability. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. San Diego, CA: IEEE; 2012 [cited 2022 Jul 19]. p. 5630–3. Available from: http://ieeexplore.ieee.org/document/6347271/

25. Trapero JI, Arizmendi CJ, Gonzalez H, Forero C, Giraldo BF. Nonlinear dynamic analysis of the cardiorespiratory system in patients undergoing the weaning process. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. Seogwipo: IEEE; 2017 [cited 2022 Jul 19]. p. 3493–6. Available from: https://ieeexplore.ieee.org/document/8037609/

26. Tsai TL, Huang MH, Lee CY, Lai WW. Data Science for Extubation Prediction and Value of Information in Surgical Intensive Care Unit. J Clin Med. 2019 Oct 17;8(10):1709.

27. Verduijn M, Sacchi L, Peek N, Bellazzi R, de Jonge E, de Mol BAJM. Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. Artif Intell Med. 2007 Sep;41(1):1–12.

28. Wise ES, Stonko DP, Glaser ZA, Garcia KL, Huang JJ, Kim JS, et al. Prediction of Prolonged Ventilation after Coronary Artery Bypass Grafting: Data from an Artificial Neural Network. Heart Surg Forum. 2017 Feb 24;20(1):E007-E014.

29. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak. 2019 Apr;19(S2):57.

30. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak. 2020 Jul;20(S3):124.

31. Mikhno A, Ennett CM. Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. San Diego, CA: IEEE; 2012 [cited 2022 Jul 19]. p. 5094–7. Available from: http://ieeexplore.ieee.org/document/6347139/

32. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, Mass: MIT Press; 2006. 248 p. (Adaptive computation and machine learning).

33. O'Hagan A. Curve Fitting and Optimal Design for Prediction. J R Stat Soc Ser B Methodol. 1978 Sep;40(1):1–24.

34. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognit Lett. 1994 Nov;15(11):1119–25.

35. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007 Feb 16;315(5814):972–6.

36. Cover TM, Thomas JA. Elements of information theory [Internet]. New York: Wiley; 2001 [cited 2022 Jul 20]. Available from: http://www3.interscience.wiley.com/cgi-bin/booktoc?ID=86512731

37. Hsu JC, Chen YF, Du YC, Huang YF, Jiang X, Chen T, et al. Design of a clinical decision support for determining ventilator weaning using support vector machine. Int J Innov Comput Inf Control. 2012;8(1):933–52.

38. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. Second edition. New York NY: Springer; 2021. 607 p. (Springer texts in statistics).

39. James G, Witten D, Hastie T, Tibshirani R. Introduction. In: An Introduction to Statistical Learning [Internet]. New York, NY: Springer US; 2021 [cited 2022 Jul 19]. p. 1–14. (Springer Texts in Statistics). Available from: https://link.springer.com/10.1007/978-1-0716-1418-1_1

40. Lee CY, Chien CF. Pitfalls and protocols of data science in manufacturing practice. J Intell Manuf. 2022 Jun;33(5):1189–207.

41. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 [cited 2022 Jul 31]; Available from: https://arxiv.org/abs/1412.6980

42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15(56):1929–58.

43. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988 Sep;44(3):837.

44. James G, Witten D, Hastie T, Tibshirani R. Resampling Methods. In: An Introduction to Statistical Learning [Internet]. New York, NY: Springer US; 2021 [cited 2022 Jul 19]. p. 197–223. (Springer Texts in Statistics). Available from: https://link.springer.com/10.1007/978-1-0716-1418-1_5

45. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk. J Biomed Inform. 2017 Dec;76:9–18.

46. 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 Feb;128(2):343–51.

47. Boles JM, Bion J, Connors A, Herridge M, Marsh B, Melot C, et al. Weaning from mechanical ventilation. Eur Respir J. 2007 May 1;29(5):1033–56.

48. Clemen RT. Making hard decisions with decision tools. 3rd ed. Mason, OH: South-Wester/Cengage Learning; 2010.