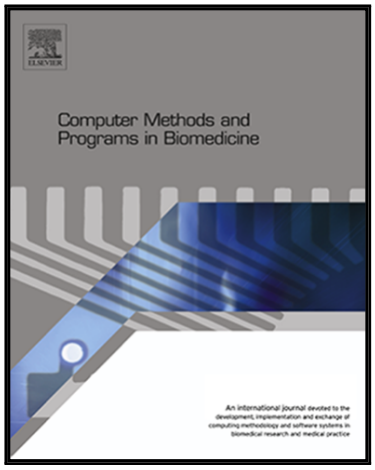
Journal Pre-proof



A Machine Learning decision-making tool for extubation in Intensive Care Unit patients

Alexandre Fabregat, Mónica Magret, Josep Anton Ferré, Anton Vernet, Neus Guasch, Alejandro Rodríguez, Josep Gómez, María Bodí

PII: S0169-2607(20)31702-8
DOI: https://doi.org/10.1016/j.cmpb.2020.105869
Reference: COMM 105869

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 5 August 2020
Accepted date: 13 November 2020

Please cite this article as: Alexandre Fabregat, Mónica Magret, Josep Anton Ferré, Anton Vernet, Neus Guasch, Alejandro Rodríguez, Josep Gómez, María Bodí, A Machine Learning decision-making tool for extubation in Intensive Care Unit patients, *Computer Methods and Programs in Biomedicine* (2020), doi: https://doi.org/10.1016/j.cmpb.2020.105869

Highlights

1. Machine Learning models are shown to potentially reduce unsuccessful extubation rate.
2. Monitor signals, patient admission data and medical records are used as predictors.
3. Support Vector Machines exhibit 92% accuracy in predicting extubation outcome.
4. Risks associated to prolonged invasive mechanical ventilation can be minimized.
5. Pre-processing challenges stress need for better data quality and curation protocols.

A Machine Learning decision-making tool for extubation in Intensive Care Unit patients

Alexandre Fabregat[a], Mónica Magret[b], Josep Anton Ferré[a], Anton Vernet[a], Neus Guasch[b], Alejandro Rodríguez[b], Josep Gómez[b,*], María Bodí[b]

[a] *Department of Mechanical Engineering, Universitat Rovira i Virgili. Av. Països Catalans, 26 (43007) Tarragona, Spain.*
[b] *Hospital Universitari de Tarragona Joan XXIII. Institut d'Investigació Sanitària Pere Virgili, Universitat Rovira i Virgili. C/. Dr. Mallafrè Guasch, 4 (43005) Tarragona, Spain*

Abstract

Background and Objective: To increase the success rate of invasive mechanical ventilation weaning in critically ill patients using Machine Learning models capable of accurately predicting the outcome of programmed extubations.
Methods: The study population was adult patients admitted to the Intensive Care Unit. Target events were programmed extubations, both successful and failed. The working dataset is assembled by combining heterogeneous data including time series from Clinical Information Systems, patient demographics, medical records and respiratory event logs. Three classification learners have been compared: Logistic Discriminant Analysis, Gradient Boosting Method and Support Vector Machines. Standard methodologies have been used for preprocessing, hyperparameter tuning and resampling.
Results: The Support Vector Machine classifier is found to correctly predict the outcome of an extubation with a 94.6% accuracy. Contrary to current decision-making criteria for extubation based on Spontaneous Breathing Trials, the classifier predictors only require monitor data, medical entry records and patient demographics.
Conclusions: Machine Learning-based tools have been found to accurately predict the extubation outcome in critical patients with invasive mechanical ventilation. The use of this important predictive capability to assess the extubation decision could potentially reduce the rate of extubation failure, currently at 9%. With about 40% of critically ill patients eventually receiving invasive mechanical ventilation during their stay and given the serious potential complications associated to reintubation, the excellent predictive ability of the model presented here suggests that Machine Learning techniques could significantly improve the clinical outcomes of critical patients.

Keywords: Invasive mechanical ventilation, Extubation, Reintubation, Machine Learning, Gradient Boosting, Support Vector Machine, Clinical decision support tool

1. Introduction

Invasive Mechanical Ventilation (IMV) is used to support gas exchange to preserve life when lungs and respiratory muscles are unable to maintain normal pulmonary ventilation and oxigenation, generally as a bridge to recovery. While IMV is not a treatment per se, it allows patients time to recover from the underlying cause of acute respiratory failure and is applied daily for a diverse spectrum of indications ranging from scheduled surgical procedures to acute organ failure.

*Corresponding author: josep.goal@gmail.com (Josep Gómez)
Email addresses: alexandre.fabregat@urv.cat (Alexandre Fabregat), mmagret@gmail.com (Mónica Magret), josep.a.ferre@urv.cat (Josep Anton Ferré), anton.vernet@urv.cat (Anton Vernet), guaschboque@gmail.com (Neus Guasch), arodri.hj23.ics@gencat.cat (Alejandro Rodríguez), mbodi.hj23.ics@gencat.cat (María Bodí)

Although IMV has demonstrated to be a life-saving procedure, it has several significant risks and potential complications usually identified under the term *ventilator-induced lung injury* [1]. The use of IMV is also associated to serious nosocomial infections such as *ventilator-associated pneumonia* and *ventilator-associated tracheobronchitis*. Moreover, prolonged IMV has been linked to long-term physical, cognitive, and mental health problems in Intensive Care Units (ICU) survivors [2].

Weaning is an essential procedure in the care of critically ill intubated patients receiving IMV. This procedure includes liberating the patient from both the IMV and endotracheal tube. Weaning begins once there has been an improvement of the cause that motivated the initiation of IMV. Medical protocols used by intensivists to decide on weaning usually include Spontaneous Breathing Trials (SBT) [3]. Despite having satisfactorily passed the SBT decision-making protocol, a fraction of the extubated patients will eventually fail to breath spontaneously and require reintubation [4].

The literature defines extubation failure as the need for reintubation within 2 days after a planned extubation [5]. Reintubation is associated to increased mortality (25–50%), prolonged IMV, increased frequency of ventilator-associated pneumonia, and longer ICU and hospital stays [6, 7, 8].

In recent years, Machine Learning (ML) approaches have been used to derive predictive tools in a vast range of applications [9, 10, 11]. However, few have focused on predicting the outcome of a weaning procedure or the optimal time for extubation. In 2013, Mueller *et al.* [12] applied different ML algorithms to predict extubation failure in a set of 486 premature infants, obtaining a poor model performance that did not outperform expert clinicians. In 2017, Shalish *et al.* [13] published a study protocol which aims to develop an Automated system for Prediction of EXtubation (APEX) for premature infants based on ML. This study protocol describes a promising methodology but no results are provided since it is still in a data collection step. Recently, in 2019, Tsai *et al.* [14] published how data science can be applied to predict extubation outcomes in surgical critical patients and evaluates the information value of the predictors in a set of patients discharged in a 1-year period.

The goal of this work is to create a ML model able to increase the current successful extubation rate in adult ICU patients under IMV. The model will be nourished with heterogeneous data routinelly collected into the Electronic Health Records (EHR) by the Clinical Information System (CIS) during a 5-year period.

The ML approach presented in this study can be used to estimate the probability of a weaning procedure failing, thus identifying situations in which it is advisable to re-evaluate the decision to extubate a patient. Because of the significant impact that reintubations have on patient outcomes, any reduction in the rate of failed extubations is beneficial to the patient, medical staff, and the rest of the healthcare system.

2. Materials and methods

2.1. Study design

This is a single-center study that uses patient data from a 30-bed polyvalent ICU located in Spain. This ICU uses a commercial CIS provided by Centricity^{TM} Critical Care Suite — GE Healthcare that has been integrated with the EHR, the bedside equipment and several auxiliary information systems including admissions, laboratory and radiology. Demographics, procedures and lab measurement data as well as monitor records are all retrieved from the CIS database through an Extraction, Transform and Load (ETL) process and then stored into comma-separated values (CSV) files for posterior analysis.

All patients or their legal representatives provided written informed consent, and our center's research ethics committee approved the study protocol (CEIC Institut d'Investigació Sanitària Pere Virgili. Reference: 41/2016).

2.2. Study population and target event

2.2.1. Population

The population in this study is composed of all adult patients admitted to the ICU between January 2015 and June 2019 (both included) who received IMV for more than 12 consecutive hours. The target events are programmed extubations performed on ICU patients under IMV. Therefore, in this study we

discarded the non-programmed (accidental extubations or self-extubations) and patients who died or suffered a traqueostomy during their ICU stay.

2.2.2. Weaning process

All weaning procedures included in the study were decided according to our weaning protocol. This protocol was written based on the recommendations of the literature and medical guidelines ([3],[15]) and includes clinical stability criteria, predictors of SBT tolerance and performance. The clinical stability criteria included Arterial Oxygen Tension / Inspiratory Oxygen Fraction (PaFiO2) or Arterial Oxygen Saturation (SpO2) or Arterial Oxygen Tension, Arterial Carbon Dioxide Tension (PaCO2), Blood Pressure (BP) and Dose of Sedative and Analgesic Drugs (SAD), Temperature and Glasgow Coma Scale (GCS). The predictors of SBT were: Maximal Inspiratory Pressure (MIP), Airway Occlusion Tension at 0.1 second (P0.1), Rapid Shallow Breathing Index (RSBI), Vital Capacity (VC), Tidal Volume (V_T) and Minute Volume (MV). Additional tests include the SBT through a T-piece, Pressure Support (PS) of 7 cm H_2O or Continuous Positive Airway Pressure (CPAP) of 5 cmH_2O between 30 minutes and 2 hours.

2.3. Response and predictor features

In Supervised ML applications [16], the goal is to determine the relation between a *response* Y and a set of m *predictors* X_i, i.e.,

$$Y = f(X_i) + \varepsilon = \widehat{Y} + \varepsilon, i = \{1, \ldots, m\}. \tag{1}$$

where $\widehat{Y}$ represents the predicted response value and ε is the total error between the actual and the predicted responses. As ε decreases, the functionality f in Eq. (1) better captures the dependence between the response and the predictors and, therefore, the model capacity to accurately predict the variability in Y due to changes in the predictors increases. In the present application, knowing f may help to identify those extubations with large probabilities of failure which, in turn, can be used by the ICU intensivists to assess the extubation decision. If the 'model' predicts a failed extubation, the medical professional team may decide to keep the patient under mechanical ventilation until favorable conditions are eventually reached.

Here, the *response* is a binary categorical variable that represents each of the possible outcomes (or classes) of an extubation, namely *successful* and *failed*. The model is, therefore, a *classifier* or a *classification learner* which prediction must take one of two potential response classes. The list of variables used as predictors in this study is shown in Tab. 1. Each variable can be classified in one of the following types:

1. Type I: Time series (data streams) obtained from monitoring medical equipment connected to the ICU patient, e.g. the heart rate.
2. Type II: derived variables obtained from other Type I data, e.g. RSBI, defined as the ratio of Respiratory Rate (RR) to V_T.
3. Type III: discrete event information collected by the medical staff during the patient ICU stay, e.g. GCS.
4. Type IV: patient demographics and ICU admission information that are assumed not to change during the patient ICU stay, e.g, patient gender.

The Type I and II data (see Tab. 1) are obtained from the main dataset containing the records of 1570 patients under IMV over their ICU stay. Each record entry contains a unique Patient Identification Number, a time stamp and the values of the 8 Type I variables, namely, the time span under IMV (Δt), the ventilation mode (V-Mode), V_T, the heart rate (HR), RR, the peak inspiratory pressure (P_{IP}), the plateau pressure (P_{PLAT}) and the oxygen saturation to inspired fraction ratio (SpFiO2).

The Type II variables, Respiratory rate-oxygenation index (ROX) and RSBI, are derived directly from Type I quantities as:

$$\mathrm{ROX} = \frac{\mathrm{SpFiO2}}{\mathrm{RR}} \tag{2}$$

$$\mathrm{RSBI} = \frac{\mathrm{RR}}{V_T} \tag{3}$$

Variable	Units	Symbol	Type	Comments
Time under IMV	h	Δt	I	
Ventilation mode	-	V-Mode	I	
Tidal Volume	L	V_T	I	
Heart Rate	min^{-1}	HR	I	
Respiratory rate	min^{-1}	RR	I	
Peak inspiratory pressure	cmH_2O	P_{IP}	I	
Plateau Pressure	cmH_2O	P_{PLAT}	I	
O_2 saturation to inspired fraction ratio	-	SpFiO2	I	
Respiratory rate-oxygenation index	min	ROX	II	SpFiO2/RR
Rapid Shallow Breathing Index	$L^{-1}\,s$	RSBI	II	RR/V_T
Number of previous MV events	-	NPE	III	
Total Cumulative Dose (sedatives and analgesics)	mg	TCD	III	
Total Given Dose (sedatives and analgesics)	mg	TGD	III	
Glasgow Coma Scale	-	GCS	III	
Richmond Agitation-Sedation Scale	-	RASS	III	
Age at admission to ICU	yr	AGE	IV	
APACHE II score	-	APACHEII	IV	
Body Mass Index at admission to ICU	$kg\,m^{-2}$	BMI	IV	
Gender	-	GENDER	IV	Categorical
SEMICYUC code	-	ICUAR	IV	Categorical

Table 1: List and characteristics of the variables used as model predictors.

Although the original RSBI definition is based on direct ventilometry data, it has been showed that the results are highly correlated with those obtained from mechanical ventilator records [17, 18].

The rest of the predictors in Tab. 1 (with each variable name in parenthesis) are obtained from other associated databases that contain ICU admission information including patient age (AGE), Acute Physiology Age Chronic Health Evaluation (APACHEII), Body Mass Index (BMI), patient gender (GENDER), and the *Spanish Society of Intensive, Critical Medicine and Coronary Units* (SEMICYUC in Spanish) classification code for ICU admission reason (ICUAR). Additional predictors include the number of respiratory events from medical log records (NPE), SAD including the Total Cumulative Dose (TCD) and Total Given Dose (TGD), and agitation/awareness state including GCS and Richmond Agitation-Sedation Scale (RASS).

The variable *ICUAR* has been one-hot encoded as 1 for a value of '09', corresponding to respiratory insufficiency of any aetiology, and 0 otherwise. Also, the NPE for a given programmed extubation, successful or failed, has been defined as the number of previous IMV events performed on a patient during her/his ICU stay, whether programmed/unplanned extubations or reintubations. This quantity have been determined after identifying each IMV event for each patient as described later in Sec. 2.5. The significant values of skewness of the TCD and TGD value distributions have been reduced using a Yeo-Johnson method [9].

2.4. Preprocessing

The data stream records collected by the different ICU monitors have different nominal sampling rates ranging from 2 to 15 minutes on average with non-synchronised time stamps. To map all variables records into a common time coordinate, Type I data was averaged over 20-minute bins with a common origin based on each patient admission time. Thus, for the variables with the largest sampling frequency (120 seconds), each bin value resulted from averaging 10 consecutive data points at most. Once the data is expressed over a unique time coordinate, it is possible to compute the combined variables of Type II including the ROX and the RSBI.

The continuous predictors (all except BMI and GENDER) have been normalized to have zero-mean and unity variance. Data with a normal z-score with $p \geq 0.985$ have been considered outliers and removed.

The dataset exhibits significant frequency imbalance due to a larger number of successful extubations in comparison to failed events. In order to ameliorate this disparity and ensure that both extubation outcome

classes have similar frequency in the train set, randomly selected data points of the most frequent class (successful extubation) were removed from the working dataset. In the case of Support Vector Machines (SVM) and Gradient Boosting Machines (GBM), any small remaining difference in the frequency of each class was *corrected* by assigning to each data point a weight equal to each class respective frequency in the training dataset.

The resampling strategy aimed to measure the model dependence on the training set was carried out using z-fold cross-validation (CV). The A-test [19] found that a 7-fold CV leads to appropiate values of the structural risk of a classification method $\Gamma_{\zeta,z}$ as shown in Appendix B.1.

The hyperparameters tuned to optimal values in this work include the cost C and γ value of radial kernel in SVM and the Number of Trees N_t and the Interaction Depth i_d in GBM. Details on implementation and specific optimal values can be found in Appendix B.2.

2.5. Event time stamp identification

Since in predicting the outcome of an extubation we use data prior to the IMV event that spans backwards in time to include any SBT procedure impact on the patient state, a crucial step in this study is to identify the time stamps of extubations for each patient. Each extubation may then be classified as *successful* or *failed* depending on whether the patient required or not a reintubation within the 48-hour time span following the extubation.

To correctly identify extubation time stamps, two sources of information were combined:

1. IMV data streams obtained from the monitors of $\mathrm{P_{IP}}$, RR, $\mathrm{V_T}$ and $\mathrm{P_{PLAT}}$.
2. Extubation and reintubation event entries recorded by the medical staff.

On the one hand, extubations leave a footprint in the IMV monitors signal in the form of data stream interruptions. Analogously, reintubations are associated to IMV signal resumptions. Thus, an extubation/reintubation pair is associated to a 'gap' in the data stream. If the gap is equal or shorter than 48 hours, the programmed extubation is classified as 'failed'. Otherwise it is considered 'successful'. Of course, for each patient there is always a terminal extubation at the end of data stream that indicates the removal of IMV support preceding the ICU discharge. These extubations are also considered 'successful' since patients who died during their ICU stay have not been considered.

Unfortunately, other non-relevant events may also generate 'gaps' in the IMV data streams. Therefore, this approach is not sufficient to guarantee the correct identification of time stamps for extubations (and reintubations, if any). To discriminate legit extubations, the time stamps obtained from IMV data stream interruptions are compared with the time stamps in the IMV log records. These records are also found to contain incorrect time stamps due to errors during time stamp data entering in the computerized system. For these reasons, a legit extubation will be considered only when there is a coincidence in the time stamps between the IMV data stream interruptions (and resumptions) and the medical record entries.

A formal description of this extubation detection algorithm is detailed in the Appendix A. The result of this procedure is a set $\{t^k_{0,i}\}$ that contains the $i = \{1 \dots M^k\}$ time stamps of actual extubations for each patient k. Note that a patient might have more than one IMV event, i.e., $M^k > 1$. For instance, the database contains patients that have suffered a failed extubation followed by a a successful one some time later.

This extubation detection procedure is illustrated in Fig. 1 that shows the temporal evolution of 6 predictors of the Type I for a selected patient (only 1 out of every 5 points in the signals is shown for the sake of clarity). HR and SpFiO2 are shown in the top panel. RR, $\mathrm{P_{PLAT}}$ and $\mathrm{P_{IP}}$ are shown in the central panel and $\mathrm{V_T}$ is shown in the bottom panel.

The vertical green and red vertical solid lines show interruptions and resumptions in the signal for the IMV-associated variables (central and bottom panels). The black dashed vertical lines show the time stamps of IMV event entries in the medical records. Time axis origin is set to the time stamp of the first Type I variable data record. Two proper extubations, occurring when both the IMV signal interruption coincides

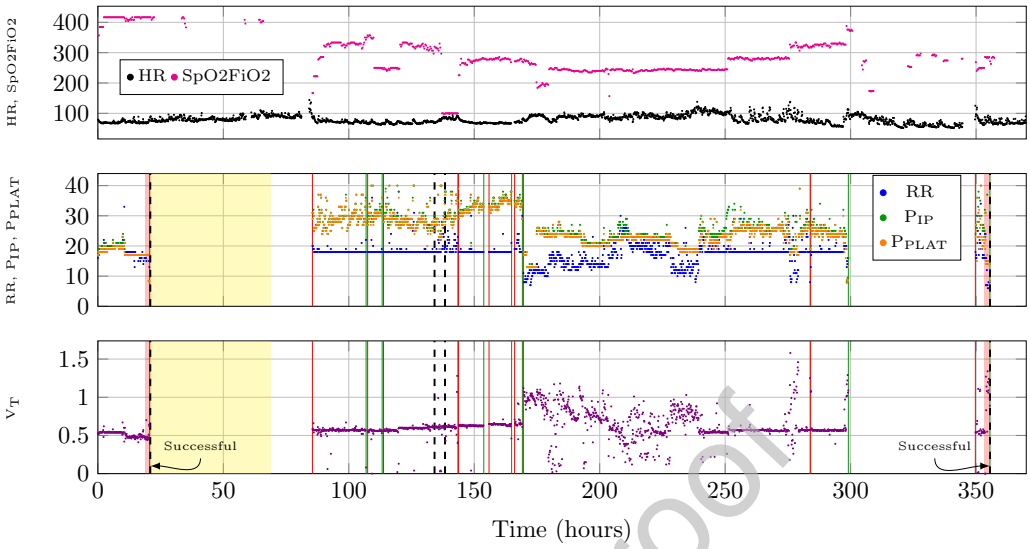


Figure 1: Signal of Heart Rate, O_2 Saturation to Inspired Fraction Ratio (top panel), Respiratory Rate, Plateau and Peak Inspiratory Pressures (central) and Tidal Volume (bottom) for a selected patient. Time stamps of interruptions and resumption of IMV-associated predictors (central and bottom panels) are shown as green and red vertical solid lines respectively. IMV events entry records are shown as dashed vertical black lines. Two proper extubations time stamps occur when both signal interruptions coincide with a IMV event record. Both are considered as 'successful' since there are not IMV signal resumptions over the 48-hour time span after the extubation time stamp (indicated as a yellow box). Working dataset used for model training uses the data over the ΔL =2-hour time stamp before the extubation (red box).

with the IMV event records, have been identified for this patient (see annotations in Fig. 1). The first one, approximately ∼21 hours after the beginning of the ICU stay, is a successful event since there has not been a reintubation over the following 48-hour (indicated by the yellow box). The second one, ∼355 hours since the ICU admission, is associated to the terminal extubation at the end of the ICU stay. The thin red box spanning backwards in time from both extubations shows the ΔL =2-hour time span that contain the data used in the training of the classifiers. As show, two of the IMV entries recorded around 135 hours into the ICU stay do not coincide with any IMV signal interruption (or resumption) and therefore they have been discarded as extubation (or reintubation) events.

3. Working dataset and statistics

The medical records for the period of time considered in this study contain 1108 programmed extubations. Of those, 100 failed. Therefore, the overall failed extubation rate in the ICU of the single center considered in this study is $100 \times \frac{100}{1108} = 9.0\%$. Due to the screening for properly identified extubations detailed in Sec. 2.5 and Appendix A the size of the dataset is restricted to 50 failed and 647 successful extubations. By averaging the signal over a $\Delta L = 2$ hour period prior to each extubation using 20 minute bins the 697 extubations should generate 4182 data points. This data span of 2 hours also ensures that the working dataset contains the values of the predictors collected during the SBT performed by ICU physicians prior to extubation. The significant number of missing data points and outliers reduce this number to 2392 of which 192 correspond to failed events and 2200 to successful extubations.

Ultimately, the predictor set used to train the three classifiers contains the following data:

- Binned averaged values of Type I and II variables in the time span $t^k_{0,i} - \Delta L \leq t^k \leq t^k_{0,i}$ for each extubation $i = \{1 \dots M^k\}$ for patient k. The time span ΔL has been set to 2 hours to ensure that any

SBT procedure impact on the 'patient state' previous to extubation noticeable on the monitor signals (Type I and II variables) is accounted.

- Number of previous events for each extubation $i = \{1 \dots M^k\}$.
- The value of each other Type III variable that is closest to each corresponding extubation time stamp.
- Patient k demographic data (Type IV variables).

The response dataset contains the outcome of the corresponding extubation, either *succesful* or *failed*.

For each class, Tab. 2 shows the median and inter quartile range (IQR, in parenthesis) for each continuous predictor and Wilcoxson's rank-sum test p-values for both groups.

Predictor	Failed $N = 192$	Successful $N = 2200$	Wilcoxson's rank-sum test
HR	88.15 (27.77)	85.27 (23.60)	0.20
RR	20.20 (6.14)	19.00 (5.90)	< 0.001
V_T	0.54 (0.11)	0.54 (0.15)	0.70
P_{IP}	19.65 (8.75)	16.00 (6.92)	< 0.001
P_{PLAT}	18.40 (6.53)	15.00 (5.10)	< 0.001
SpFiO2	317.64 (57.79)	325.92 (53.42)	< 0.001
Δt	204.17 (203.9)	131.50 (159.6)	< 0.001
TCD	0.00 (0.00)	0.00 (42.22)	0.44
TGD	0.00 (0.00)	0.00 (0.00)	0.81
NPE	1.00 (1.00)	0.00 (0.00)	< 0.001
GCS	11.00 (1.00)	12.00 (4.00)	< 0.001
RASS	0.00 (1.00)	0.00 (1.00)	< 0.001
AGE	61.00 (19.00)	60.00 (21.00)	0.98
APACHEII	25.00 (10.00)	25.00 (12.00)	0.005
BMI	26.12 (3.81)	26.12 (5.51)	0.16
ROX	15.09 (3.71)	16.80 (6.64)	< 0.001
RSBI	35.37 (15.86)	35.45 (16.64)	0.015

Table 2: Successful and failed extubation groups p-values for the continuous predictors.

Similarly, Tab. 3 shows the count and percentage (in parenthesis) for each class for the categorical predictors and the Fisher exact test p-value for both groups.

Predictor	Failed	Succesful	Fisher exact test
GENDER (Male)	138 (72%)	1408 (64%)	0.03
ICUAR (1)	24 (12%)	467 (21%)	0.003
V-Mode (CV)	80 (42%)	612 (28%)	< 0.001

Table 3: Successful and failed extubation groups p-values for the categorical predictors.

3.1. Training of ML classifiers

Three different ML classifiers have been compared in this work: SVM with radial basis, GBM with Bernoulli loss and Linear Discriminant Analysis (LDA). Collected performance metrics include the Mean Accuracy ('acc') defined as the fraction of correctly predicted classifications over the total number of data points in the testing set and the Area under the Receiver Operating Characteristics Curve (AUROC).

4. Results

The entire pre-processing and data analysis presented in this study were performed using R version 3.6.1 [20] using the *mlr* package version 2.15.0 for pre-processing, resampling, hyperparameter tuning and learner benchmarking.

As shown in Tab. 4, SVM exhibits excellent predictive capabilities in terms of mean accuracy and AUROC for the test dataset with values of 94.6% and 98.3% respectively. Scores for GBM are 87% and 96%. With the lowest performance capabilities, LDA scored 72% and 79%. These results suggest that f can not be satisfactorily represented by additive contributions of the predictors and instead exhibits a strong non-linear behaviour.

Classifier	% Accuracy	% AUROC
SVM	94.6	98.3
GBM	89.6	96.1
LDA	72.4	79.4

Table 4: Mean accuracy and AUROC for each classifier using a classification threshold of 0.5 and undersampling for class imbalance.

With a 94.6% mean accuracy, the learner based on SVM demonstrates to have a notable potential for reducing the failed extubation rate in critical patients. While a 'successful' prediction would reinforce the decision to extubate, a 'failed' prediction would indicate the need to reassess the clinical decision and reconsider the suitability of the extubation. Moreover, the classifier can be tuned to minimize the Type I error associated to the False Positive Rate. While wrongly predicting a 'failed' extubation would result in an unnecessary prolongation of the time under IMV, a wrongly predicted 'successful' extubation may potentially lead to a 'failed' extubation. Therefore, it may be beneficial to adopt a more conservative protocol by increasing the threshold value above 0.5. The dependence of the performance on the classification threshold is shown in Fig. 2 that presents the mean ROC curves for each classifier.

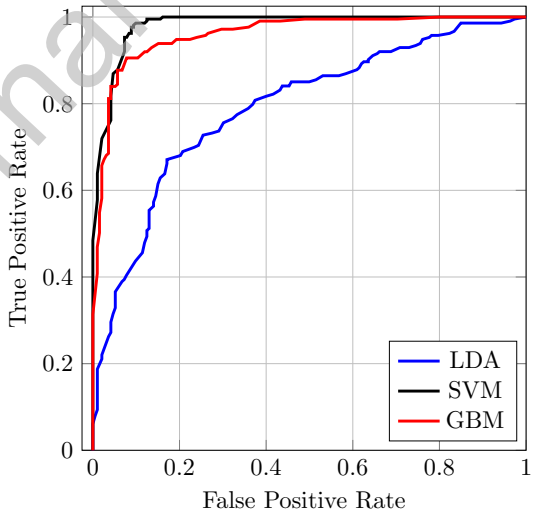


Figure 2: Mean ROC curve for SVM, GBM and LDA classifiers.

GBM provides information about the importance of each predictor in correctly classifying each extubation into each outcome class. These results suggest that the top five predictors in descending order of importance

are Δt, GCS, BMI, ROX and P_{PLAT}. On the other hand, the least relevant predictors in descending order of importance are ICUAR, GENDER, TCD, TGD and V-Mode.

5. Discussion

Weaning is a complex clinical procedure that begins by analyzing a series of selected clinical variables that allow to determine the adequacy for withdrawal of IMV. If the patient is found to meet the criteria, intensivists proceed to perform the SBT using a set of up to 60 different predictors. Only if the SBT is successful, the IMV support is removed. In the event of a patient needing IMV within the 48 hours following an extubation, weaning is considered a failure. Given the clinical risks associated with reintubation, a failed extubation necessarily implies an increased risk of complications for the patient that may have a significant impact on his/her outcome from the ICU. Despite the protocols for deciding on weaning mentioned above, correctly predicting the outcome of an extubation is still a daily challenge for thousands of intensivists. Current evidence shows that between 3-23% of patients who pass a SBT require reintubation [4, 21, 22]. In the present study, the UCI reintubation rate is 9%, a value in accordance with the quality standards defined by the Spanish Society of Critical Intensive Care Medicine and Coronary Units [23].

In this work, we used ML tools to build an accurate model capable of predicting the outcome of an extubation in ICU patients under IMV in order to reduce the current rate of failed extubations.

Until recently many predictive models have been based on linear regressions obtained by fitting modestly large datasets often gathered manually [24, 25, 26]. Current ML efforts in medical sciences largely benefit from a combination of CIS that automatically collect, process and store large amounts of data [27] and advanced Statistical Learning techniques that allow for complex predictor interactions including Multiple non-Linear Regression, Tree-based models and SVM to mention a few [28, 18].

In training the classification models considered in this work, a combination of different types of predictors have been used including monitor signal data streams up to 2 hours before an extubation, discrete event information collected by intensivists and medical staff during the patient ICU stay and patient demographics at ICU admission. The target events are programmed extubations that are decided according to the weaning protocol. Importantly, identifying the programmed extubation events has been a major challenge given the significant lack of agreement between the time stamps in IMV events in the medical records and interruptions/resumptions in the monitor signal associated to extubations/reintubations. As a result, a significant fraction of legit programmed extubations has been discarded to ensure the quality of the working dataset. This rises concerns about the need for better data curation and quality control protocols. In this context of rapid penetration of ML techniques for clinical and medical applications, efforts aimed at obtaining high quality and reliable data sets are essential to ensure the proper development and implementation of better decision-making tools. In this context, any initiative directed to involve medical, auxiliary and administrative staff of ICU in the improvement of data collection/curation practices can significantly influence our ability to obtain better ML products. A set of basic strategies aimed to maximize the quality of ICU data collection may include (i) promote a culture of quality among clinical professionals by implementing concise data gathering protocols, (ii) ensure CIS integration with relevant clinical units (e.g., labs, patient clinical history) and (iii) test communications between bedside monitor equipment (ventilators, dialysis machines...) and centralized information systems.

Among the three different tested classifiers, Support Vector Machines (SVM) and Gradient Boosting Machine (GBM) have demonstrated to have superior prediction capabilities in comparison to Linear Discriminant Analysis (LDA).

The effectiveness of our best ML model, based on SVM, outperformed the previously achieved by [14] (Acc: 94.6% vs 86.0%), which was based on backpropagation neural network (BPN). However, in that study, their BPN model outperformed their SVM model (Acc: 86.0% vs 75.5%), which suggests that BPN could be an interesting approach to improve our decision support tool in further studies. The models derived here have been trained with predictors that essentially capture the state of the patient over the 2 hour previous to an extubation decided upon by the medical ICU staff according to a IMV weaning protocol that included SBT tests. Therefore, the training stage does not include data from prospective weaning events

that were finally identified as not feasible by the medical staff. As a result, the models can not be thought as general-purpose predictor of success for programmed extubations or as a monitoring alarm system that identifies favourable conditions for extubation during the patient ICU stay.

Instead, the models must be considered as a support tool aimed to validate the medical staff decision upon extubation. With 94.6% predictive accuracy, the SVM classifier could significantly reduce the rate of failed extubation from its current value of 9% to a theoretical 1%, thus minimizing the incidence of clinical complications associated with reintubation procedures.

In practice, after deciding to remove the invasive respiratory ventilation for a critically ill patient, the intensivists can use the model to predict the extubation outcome and assess their decision. The model computational requirements and predictor dataset size are both very modest and any regular computer available in the ICU should be enough to perform the prediction. Since all software used to develop the model is open source, no additional costs are expected. To facilitate the process, on-going efforts are directed to include the current model predictive capabilities into the existing user-friendly interface [29] allowing on-the-fly automatic predictor data import and predicted response display. Once deployed, this clinical practice decision-making tool will be tested to show its performance in reducing reintubation rate and its impact on the patient clinical safety.

6. Conclusions

Data from Intensive Care Unit monitor systems have been combined with patient demographics, medical records and respiratory event logs to assemble a working dataset used to train several Machine Learning learners to predict the outcome of an extubation. To ensure that only data corresponding to proper programmed extubations were included in the training dataset, data pre-processing methodology compared interruptions in monitor signals and record entries in respiratory event logs to discard non relevant respiratory events.

Three different type of classifiers have been used to predict the outcome of extubations: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and Gradient Boosting Machine (GBM). Standard R packages have been used to perform hyper-parameter tuning, re-sampling and performance estimations.

The results suggest that Machine Learning tools are especially well suited to evaluate the decision making protocol based on Spontaneous Breathing Trials used by intensivists to decide whether to withdraw mechanical ventilation. Specifically, our classifier based on SVM have been found to correctly predict the outcome of a Invasive Mechanical Ventilation weaning with a 94.6% accuracy. This unprecedented predictive capabilities should lead to a reduction in failed extubation rates and incidence of complications associated to reintubations.

7. Declarations of interest

None.

8. Acknowledgements

The authors would like to acknowledge financial support from the Spanish Ministry of Economy and Competitiveness under the project RTI2018-100907-A-I00 (MCIU/AEI/FEDER, UE), the FIS grant PI PI16/00491 (Carlos III Institute of Health, FEDER, Spain) and the Catalan Government for the quality accreditation of the research groups 2017-SGR-1409, 2017-SGR-1234 and 2017 SGR 127.

Appendix A. Extubation event detection

The strategy to identify extubations (and reintubations) for each patient follows these steps:

1. Find the time stamps of IMV data stream interruptions.
2. Compare the IMV data stream interruption time stamps with those in the entries for programmed extubations.
3. Generate the list of extubation times by selecting only the coincident time stamps up to a threshold.
4. Repeat steps 1 to 3 for IMV data resumptions (associated to reintubations).
5. If any, determine the length of the 'gap' size formed by an extubation/reintubation pair of events.
6. Classify the extubations as *failed* if the time stamp 'gap' is smaller or equal to 48 hour or as *successful* otherwise.

An extubation, regarding of its nature (programmed or accidental), leaves a footprint in the form of an interruption in the IMV monitor data stream. Similarly, a resumption in this IMV data stream is associated to a reintubation. Therefore, for each patient k and IMV variable i, we collected every data stream interruption time stamp, $t^k_{0,i}$ and every data stream resumption, $t^k_{1,i}$.

To be considered a potential extubation (or reintubation) event, the interruption (and resumption) time stamps must be simultaneous across the M variables up to a prescribed threshold α. In order to discard spurious results, the gap in the IMV data stream, linked to an extubation followed by a reintubation, is required also to be larger than a prescribed threshold β.

Formally, for each patient k, we define the averaged extubation time stamp $\langle t^k_0 \rangle$ and the difference between the latest and the earliest extubation time stamps Δt^k_0 across all four IMV variables as

$$\langle t^k_0 \rangle = \sum_{i=1}^{4} t^k_{0,i}/4, \tag{A.1}$$

$$\Delta t^k_0 = \max\{t^k_{0,i}\} - \min\{t^k_{0,i}\}. \tag{A.2}$$

Similarly for reintubations,

$$\langle t^k_1 \rangle = \sum_{i=1}^{4} t^k_{1,i}/4, \tag{A.3}$$

$$\Delta t^k_1 = \max(t^k_{1,i}) - \min(t^k_{1,i}). \tag{A.4}$$

The set of legit extubation and reintubations time stamps for each patient k, t^k_0 and t^k_1 respectively, can be defined as:

$$\left\{\langle t^k_0 \rangle \in \tilde{t}^k_0 : \Delta t^k_0 \leq \alpha\right\} \tag{A.5}$$

$$\left\{\langle t^k_1 \rangle \in \tilde{t}^k_1 : \Delta t^k_1 \leq \alpha\right\} \tag{A.6}$$

$$\left\{\tilde{t}^k_0 \in \hat{t}^k_0 : \tilde{t}^k_1 - \tilde{t}^k_0 \geq \beta\right\} \tag{A.7}$$

$$\left\{\tilde{t}^k_1 \in \hat{t}^k_1 : \tilde{t}^k_1 - \tilde{t}^k_0 \geq \beta\right\} \tag{A.8}$$

Although extubations (and reintubation) events generate a footprint in the form of interruptions (and resumptions) in the IMV data stream, other non-relevant events may also generate IMV data gaps (medical procedures that require extubation, IMV monitor malfunction...). The distribution of the length of IMV data stream gaps that may potentially be associated to extubation/reintubation is shown in Fig. A.3. In order to discriminate legit extubations and reintubations, $\hat{t}^k_0$ and $\hat{t}^k_1$ were compared to the extubation and reintubation time stamps entries manually recorded by the medical staff (T^k_0 and T^k_1 respectively). Since these records may also contain errors due to errors in data entry, the final set of extubations and reintubations events is restricted to those for which there is a coincidence between these IMV data stream time stamps and the entries in the manual records up to another prescribed tolerance γ.

Formally,

$$\left\{\hat{t}_0^k \in t_0^k : \left|\hat{t}_0^k - T_0^k\right| \leq \gamma\right\} \tag{A.9}$$

$$\left\{\hat{t}_1^k \in t_1^k : \left|\hat{t}_1^k - T_1^k\right| \leq \gamma\right\} \tag{A.10}$$

$$\tag{A.11}$$

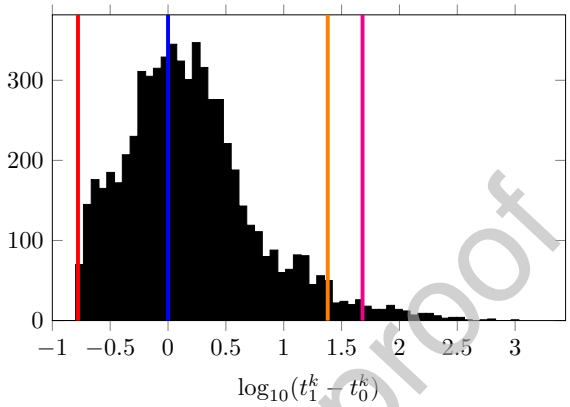


Figure A.3: Distribution of data stream gaps $t_1^k - t_0^k$ associated to potential extubation/reintubation event pairs (logarithmic scale). Red, blue, orange, magenta vertical lines indicate 0.17, 1, 24, 48 hours respectively. Note that the magenta line, located at $t_1^k - t_0^k = 48$ hours, respresents the limit above which a data stream gap cannot be associated to an extubation/reintubation pair. The red line indicates the location of the threshold (0.17 hours) below which an extubation/reintubation pair is discarded.

Appendix B. Pre-processing details

Appendix B.1. Resampling

Following the A-test [19] criteria, the structural risk of a classification method $\Gamma_{\zeta,z}$ over a $z-$fold range of $z = 2 \ldots 15$ is shown in Fig. B.4. Results suggests that a 7-fold cross-validation resampling strategy guarantees the stability of the classification model with new test data.

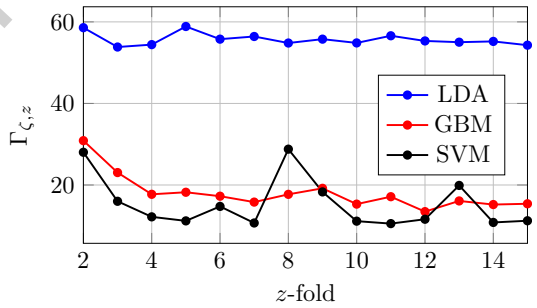


Figure B.4: A-test results for all three learners over a $z = 2 \ldots 15$-fold Cross-Validation resampling.

Appendix B.2. Hyperparameter tuning

The LDA learner has no hyperparameters suitable for optimization. Two hyperparameters were optimized for the SVM learner, namely, the cost C associated to the regularization term in the Lagrange

C	γ	Mmce	Acc	AUROC	iteration	CPU Time(s)
47.35	9.79e+02	0.467	0.533	0.569	1	0.945
1552.42	2.46e+04	0.474	0.526	0.489	2	0.953
292.54	4.50e+00	0.173	0.827	0.983	3	0.959
17.48	1.08e-01	0.074	0.926	0.975	4	0.967
1739.38	8.60e+03	0.474	0.526	0.484	5	0.905
10985.82	3.16e-01	0.049	0.946	0.983	6	0.864

Table B.5: SVM hyperparameter tuning results. *Mmce* stands for test mean classification error, *Acc* stands for test mean accuracy and *AUROC* stands for test mean area under the curve

formulation and the γ value of the radial kernel used in this work. The optimization strategy has been implemented using the native *makeParamSet* function and a search space of the type *makeNumericParam* with a 10-th power transformation taking values between -2 and 5 combined with a random control approach with 100 maximum iterations:

```
ps_svm = makeParamSet(
     makeNumericParam("cost", lower = -2, upper = 5, trafo = function(x) 10^x),
     makeNumericParam("gamma",lower = -2, upper = 5, trafo = function(x) 10^x)
     )
ctrl_svm = makeTuneControlRandom(maxit = 100L)
```

The optimal values for SVM were found to be $C^* = 10985$ and $\gamma^* = 0.316$.

To illustrate the search history for optimal values and the hyperparameters tuning effects on both the accuracy and AUROC, Tab. B.5 shows the *mlr* package version 2.15.0 results including the CPU time (in seconds) for each iteration.

Similarly, the search and control strategies for the GBM hyperparamteres, namely, the Number of Trees N_t and the Interaction Depth i_d, using the *mlr* package version 2.15.0, have been implemented as:

```
ps_gbm = makeParamSet(makeDiscreteParam("n.trees",           values = seq(1000,10000,1000)),
                      makeDiscreteParam("interaction.depth", values = seq(5,45,5))
                      )
ctrl_gbm = makeTuneControlGrid()
```

Optimal values were found to be $N_t^* = 6000$, $i_d^* = 20$ for a Shrinkage value of $S^* = 0.001$. Figure B.5 shows the iterative evolution of the GBM learner mean accuracy.

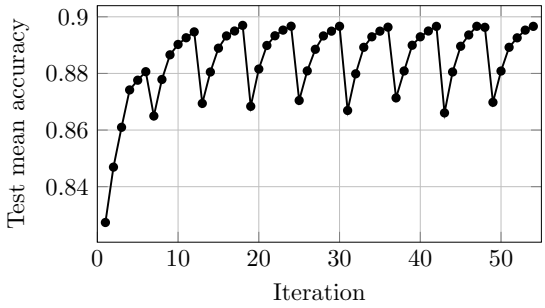


Figure B.5: Test mean accuracy iterative search history for the GBM learner.

References

[1] A. S. Slutsky, V. M. Ranieri, Ventilator-induced lung injury, N. Engl. J. Med. 369(22) (2013) 2126–2136.

[2] S. Desai, T. Law, N. Dale, Long-term complications of critical care, Critical Care Medicine 39 (2) (2011) 371–379.

[3] J.-M. Boles, J. Bion, A. Connors, M. Herridge, B. Marsh, C. Melot, R. Pearl, H. Silverman, M. Stanchina, A. Vieillard-Baron, T. Welte, Weaning from mechanical ventilation, European Respiratory Journal 29 (5) (2007) 1033–1056.

[4] S. K. Epstein, R. L. Ciubotaru, J. B. Wong, Effect of failed extubation on the outcome of mechanical ventilation, Chest 112 (1) (1997) 186 – 192.

[5] F. Frutos-Vivar, A. Esteban, C. Apezteguia, M. González, Y. Arabi, M. Restrepo, F. Vidal, C. Santos, J. Alhashemi, F. Pérez, O. Peñuelas, A. Anzueto, Outcome of reintubated patients after scheduled extubation, Journal of critical care 26 (2011) 502–9.

[6] O. Peñuelas, F. Frutos-Vivar, C. Fernández, A. Anzueto, S. Epstein, C. Apezteguia, M. González, N. Nin, K. Raymondos, V. Tomicic, P. Desmery, Y. Arabi, P. Pelosi, M. Kuiper, M. Jibaja-Vega, D. Matamis, N. Ferguson, A. Esteban, Characteristics and outcomes of ventilated patients according to time to liberation from mechanical ventilation, American journal of respiratory and critical care medicine 184 (2011) 430–7.

[7] F. Frutos-Vivar, N. Ferguson, A. Esteban, S. Epstein, Y. Arabi, C. Apezteguia, M. González, N. Hill, S. Nava, G. D'Empaire, A. Anzueto, Risk factors for extubation failure in patients following a successful spontaneous breathing trial, Chest 130 (2007) 1664–71.

[8] J. Gowardman, D. Huntington, J. Whiting, The effect of extubation failure on outcome in a multidisciplinary australian intensive care unit, Critical care and resuscitation : journal of the Australasian Academy of Critical Care Medicine 8 (2007) 328–33.

[9] M. Kuhn, K. Johnson, Applied predictive modeling, Springer, 2013.

[10] J. Kim, H. Chang, D. Kim, D.-H. Jang, I. Park, K. Kim, Machine learning for prediction of septic shock at initial triage in emergency department, Journal of Critical Care 55 (2020) 163 – 170.

[11] M. Komorowski, L. Celi, O. Badawi, A. Gordon, A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature Medicine 24 (2018) 1716–1720.

[12] M. Mueller, J. Almeida, R. Stanislaus, C. Wagner, Can machine learning methods predict extubation outcome in premature infants as well as clinicians?, Journal of Neonatal Biology 2. doi:10.4172/2167-0897.1000118.

[13] W. Shalish, L. Kanbar, S. Rao, C. Robles-Rubio, L. Kovacs, S. Chawla, M. Keszler, D. Precup, K. Brown, R. Kearney, G. Sant'Anna, Prediction of extubation readiness in extremely preterm infants by the automated analysis of cardiorespiratory behavior: Study protocol, BMC Pediatrics 17. doi:10.1186/s12887-017-0911-z.

[14] T.-L. Tsai, M.-H. Huang, C.-Y. Lee, W.-W. Lai, Data science for extubation prediction and value of information in surgical intensive care unit, Journal of Clinical Medicine 8 (2019) 1709. doi:10.3390/jcm8101709.

[15] A. Esteban, F. Frutos, M. Tobin, I. Alía, J. Solsona, I. Valverdú, R. Fernandez, M. A. de la Cal, S. Benito, R. Tomás, A comparison of four methods of weaning patients from mechanical ventilation, The New England journal of medicine 332 (1995) 345–50.

[16] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, Springer Publishing Company, Incorporated, 2014.

[17] F. Lessa, C. Paes, R. Tonella, S. Araújo, Comparison of the rapid shallow breathing index (rsbi) calculated under direct and indirect form on the postoperative period of cardiac surgery, Revista brasileira de fisioterapia (São Carlos (São Paulo, Brazil)) 14 (2010) 503–9.

[18] A. Mikhno, C. Ennett, Prediction of extubation failure for neonates with respiratory distress syndrome using the mimic-ii clinical database, Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 2012 (2012) 5094–7.

[19] A. Gharehbaghi, M. Lindén, A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network, IEEE Transactions on Neural Networks and Learning Systems 29 (9) (2018) 4102–4115.

[20] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013).
URL http://www.R-project.org/

[21] A. Esteban, I. Alía, M. Tobin, A. Cano, F. Vidal, I. Vallverdú, L. Blanch, A. Bonet, A. Vázquez, R. Pablo, A. Torres, M. A. de la Cal, S. Macias, G. Hernandez Poblete, Effect of spontaneous breathing trial duration on outcome of attempts to discontinue mechanical ventilation, American journal of respiratory and critical care medicine 159 (1999) 512–8. doi:10.1164/ajrccm.159.2.9803106.

[22] S. K. Epstein, Decision to extubate, Intensive Care Med. 28 (5) (2002) 535–546.

[23] S. española de medicina intensiva crítica y unidades coronarias (Semicyuc), Indicadores de calidad del enfermo crítico actualización, iSBN: 978-84-941142-4-3 Depósito Legal: M-21564-2017 (2017).

[24] E. Fujii, K. Fujino, S. Tanaka-Mizuno, Y. Eguchi, Variation of risk factors for cause-specific reintubation: A preliminary study, Canadian Respiratory Journal 2018 (2018) 1–6.

[25] A. Thille, F. Boissier, H. Ghezala, K. Razazi, A. Mekontso-Dessap, C. Brun-Buisson, Risk factors for and prediction by caregivers of extubation failure in icu patients: A prospective study, Critical care medicine 43.

[26] K. Asehnoune, P. Seguin, S. Lasocki, A. Roquilly, A. Delater, A. Gros, F. Denou, P.-J. Mahé, N. Nesseler, D. Demeure-dit Latte, Y. Launey, K. Lakhal, B. Rozec, Y. Mallédant, V. Sébille, S. Jaber, A. Le Thuaut, F. Feuillet, R. Cinotti, A. group, Extubation success prediction in a multicentric cohort of patients with severe brain injury, Anesthesiology: The Journal of the American Society of Anesthesiologists 127 (2) (2017) 338–346.

[27] A. Johnson, T. Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 160035.
[28] A. Beam, I. Kohane, Big data and machine learning in health care, JAMA 319.
[29] L. Claverías, J. Gómez, A. Rodríguez, J. Albiol, F. Esteban, M. Bodí, Soporte a la organización de las unidades de cuidados intensivos durante la pandemia, a través de mapas creados a partir de los sistemas de información clínica, Medicina Intensivadoi:https://doi.org/10.1016/j.medin.2020.08.006.
URL http://www.sciencedirect.com/science/article/pii/S0210569120302709