TYPE Mini Review
PUBLISHED 11 August 2022
DOI 10.3389/fmed.2022.961252

Check for updates

OPEN ACCESS

EDITED BY
Savino Spadaro,
University of Ferrara, Italy

REVIEWED BY
Warwick Wolf Butt,
Royal Children's Hospital, Australia
Robert Edward Kearney,
McGill University, Canada

*CORRESPONDENCE
Yutaka Igarashi
igarashiy@nms.ac.jp

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 04 June 2022
ACCEPTED 19 July 2022
PUBLISHED 11 August 2022

CITATION
Igarashi Y, Ogawa K, Nishimura K, Osawa S, Ohwada H and Yokobori S (2022) Machine learning for predicting successful extubation in patients receiving mechanical ventilation. *Front. Med.* 9:961252.
doi: 10.3389/fmed.2022.961252

Machine learning for predicting successful extubation in patients receiving mechanical ventilation

Yutaka Igarashi[1]*, Kei Ogawa[2], Kan Nishimura[2], Shuichiro Osawa[1], Hayato Ohwada[2] and Shoji Yokobori[1]

[1]Department of Emergency and Critical Care Medicine, Nippon Medical School, Tokyo, Japan,
[2]Department of Industrial Administration, Tokyo University of Science, Chiba, Japan

Ventilator liberation is one of the most critical decisions in the intensive care unit; however, prediction of extubation failure is difficult, and the proportion thereof remains high. Machine learning can potentially provide a breakthrough in the prediction of extubation success. A total of seven studies on the prediction of extubation success using machine learning have been published. These machine learning models were developed using data from electronic health records, 8–78 features, and algorithms such as artificial neural network, LightGBM, and XGBoost. Sensitivity ranged from 0.64 to 0.96, specificity ranged from 0.73 to 0.85, and area under the receiver operating characteristic curve ranged from 0.70 to 0.98. The features deemed most important included duration of mechanical ventilation, PaO_2, blood urea nitrogen, heart rate, and Glasgow Coma Scale score. Although the studies had limitations, prediction of extubation success by machine learning has the potential to be a powerful tool. Further studies are needed to assess whether machine learning prediction reduces the incidence of extubation failure or prolongs the duration of ventilator use, thereby increasing tracheostomy and ventilator-related complications and mortality.

KEYWORDS
extubation, intensive care unit, machine learning, mechanical ventilation, ventilator

Introduction

Invasive mechanical ventilation is an essential component of invasive care. The number of patients receiving ventilator care has been reported to be 270–314 per 100,000 population per year (1), and this number has increased dramatically due to coronavirus disease 2019 (COVID-19) pneumonia. Approximately half of the patients on ventilators died (2), and more than one million people have died of COVID-19 in the US (3).

Ventilator liberation comprises one of the most critical decisions in the intensive care unit (ICU). If ventilated patients are extubated too early, they may require reintubation, which prolongs hospital stay and increases mortality (4–6). On the other hand, longer ventilation increases complications such as ventilator-associated pneumonia and mortality; thus, it is critical to determine the optimal timing of extubation for patients receiving ventilation (7). Various predictors and protocols have been used to predict successful extubation (8), but single factors are not accurate (9, 10). Although the use of protocols has reduced the duration of ventilator management and ICU stay (11), the proportion of extubation failure remains ≥10% (12). There is no standardized protocol for extubation, and each facility has its own criteria (13).

In recent years, electronic health records have been collected in ICUs, and these data are publicly available. Simultaneously, the application of machine learning in the medical field has rapidly progressed, making it possible to make various predictions in real time regarding patients in ICUs (14). Machine learning can potentially be a breakthrough in the prediction of extubation success, for which many factors are involved in a complex manner. To date, several papers have been published on the prediction of extubation outcomes using machine learning. In this mini-review, the methods and results of previous studies are summarized, and current issues and future directions have been reviewed.

Protocols for scheduled extubation

Patients receiving ventilation are evaluated for extubation when the primary disease has improved, oxygenation and hemodynamics are stable, and the patient is breathing spontaneously and normally. A spontaneous breathing trial (SBT) is performed to assess whether the patient can tolerate minimal ventilatory support (15). SBT is performed with oxygenation <50%, low positive end-expiratory pressure (PEEP) (e.g., ≤5 cmH_2O), and low PS (e.g., ≤5 cmH_2O) for a certain period of time, and if there are no abnormalities in circulation or respiration, the patient is deemed to be ready for extubation. If the patient has upper airway problems (post upper airway surgery, positive cuff-leak test, history of difficult intubation), steroids may be administered before extubation (16). If the patient is at high risk for respiratory failure (chronic obstructive pulmonary disease, chronic respiratory failure, obesity, overhydration), prophylactic non-invasive positive pressure ventilation (non-invasive positive pressure ventilation and nasal high-flow) should be prepared after extubation.

Thus, during SBT and weaning, vital signs related to respiration and circulation (blood pressure, pulse rate, respiratory rate, SpO_2, and boosting agent levels), arterial blood gas tests (PaO_2, pH), and ventilator measurements (tidal volume, minute volume, and rapid shallow breathing index) have been used as criteria for extubation. Since all machine learning studies are backward-looking studies, the data are, in principle, obtained by performing SBT. Therefore, the process of performing SBT and that of determining extubation from the collected data remains the same when predicting extubation by machine learning.

Aim of machine learning prediction

The most important goal of machine learning is to improve the prediction of the success or failure of extubation. In particular, it is expected that machine learning models can predict extubation failure and thereby reduce its incidence. Conversely, if a machine learning model is too conservative, in order to reduce the number of extubation failures, the period of ventilator management may be prolonged and tracheotomies may increase. Therefore, it is important to maintain a balance between these two factors.

Further, in some machine learning models, feature importance may represent objective indicators to explain the model. It is possible to compare the important features with the relevant guidelines to ensure that there is validity in the model's explanation. It is also possible to identify unknown factors that have not been used to determine extubation but are of high importance.

However, machine learning models cannot predict the causes of extubation failure, although reasons for extubation failure, such as laryngeal edema, are important. The purpose of machine learning models is not to predict the causes of extubation failure but to support decision-making.

Methods of previous studies

A total of seven studies on the prediction of extubation using machine learning were published between 2015 and 2021 (Table 1). We compared the differences between the methods of each study; studies on ventilator weaning were excluded (24).

Dataset

Some studies used single-center ICU databases with hundreds to thousands of cases, while others used nationwide databases for COVID-19 or large free-access databases, with thousands to tens of thousands of cases. For example, the Medical Information Mart for Intensive Care (MIMIC) is a large, single-center, free-access database comprising information relating to patients admitted to the ICU of a large tertiary care hospital with a variety of diseases, such as myocardial infarction, postoperative complications, medical complications, and trauma (25). Machine learning using such a database may be adaptable to patients with a wide variety of diseases.

The dataset is divided into training data for model development and test data for evaluation. If multiple data from one patient are used to increase the sample size, data from the same patient may be included in both training and test data. If a test is performed on the data of the patient for whom the model was developed, it will be a so-called cheat, resulting in a higher

TABLE 1 Summary of the seven studies.

Author, year	Algorithm with the highest accuracy	Dataset (Training and test) (N, patients)	Dataset (External validation) (N, patients)	Number of features	Definition and rate of extubation failure	n-fold cross-validation	Accuracy	Sensitivity	Precision	Specificity	F1 score	AUROC	Strengths	Weakness
Kuo (17)	ANN	Two hospitals ($N =$ 121)	No	8	<48 h, 26%	5	0.80	0.82	N/A	0.73	N/A	N/A	To determine the optimal number of hidden-layer perceptron, it was set from 10 to 39. Balanced data because there was failed extubation in 26% of patients.	Small number of cases. Fewest predictors: vital signs and laboratory results were not used. Undersampling and external validation were not performed.
Hsieh (18)	ANN	Single hospital ($N =$ 3,602)	No	37	<72 h, 5%	10	N/A	0.822	0.939	N/A	0.867	0.85	Better prediction compared with other weaning parameters.	Undersampling was not performed, although there was failed extubation in only 5% of the patients. External validation was not performed.
Chen et al. (19)	LightGBM	MIMIC-III ($N =$ 3,636)	No	68	<48 h, 17%	5	0.8020	0.8394	N/A	0.7477	N/A	0.8198	After developing a model with all features, the model was created again with only the important features. The synthetic minority oversampling technique (SMOTE) was used, but results with or without SMOTE were not obvious. Feature importance and SHAP value were obtained.	External validation was not performed.

(Continued)

TABLE 1 Continued

Author, year	Algorithm with the highest accuracy	Dataset (Training and test) (N, patients)	Dataset (External validation) (N, patients)	Number of features	Definition and rate of extubation failure	n-fold cross-validation	Accuracy	Sensitivity	Precision	Specificity	F1 score	AUROC	Strengths	Weakness
Fabregat (20)	SVM	Single hospital (*N* = 1,108)	No	19	<48 h, 9%	7	0.946	N/A	N/A	N/A	N/A	0.983	Highest accuracy and AUROC. Undersampling was performed.	Although having the highest accuracy and AUROC, data from the same patient was included in both training and test data sets, making cheating possible. External validation is required. Laboratory results were not used for prediction. Predictive performance was not obtained without accuracy and AUROC.
Otaguro (21)	LightGBM	Single hospital (*N* = 117)	No	58	<72 h, 11%	5	0.9265	0.9602	0.9146	N/A	0.9369	0.9502	Undersampling was performed. Feature importance was obtained.	Small number of cases. Lowest precision, but more than 0.90. Although having the highest sensitivity, data from the same patient were included in both training and test data sets, making cheating possible.

(Continued)

TABLE 1 Continued

Author, year	Algorithm with the highest accuracy	Dataset (Training and test) (N, patients)	Dataset (External validation) (N, patients)	Number of features	Definition and rate of extubation failure	n-fold cross-validation	Accuracy	Sensitivity	Precision	Specificity	F1 score	AUROC	Strengths	Weakness
Zhao et al. (22)	CatBoost	MIMIC-IV (*N* = 16,189)	Single hospital (*N* = 502)	78	<48 h, 17[a] and 11%[b]	N/A	N/A	0.64	0.97	0.85	0.77	0.80	Highest specificity and precision. Largest number of cases and features. Clinical scores were not used because they make the models inconvenient in clinical settings. Eleven models were developed and compared with other predictive factors commonly used in the ICU. External validation was performed. Feature importance and SHAP value were obtained.	Lowest sensitivity and F1 score. Clinical scores were commonly used for developing models, but this study did not use them.
Fleuren (23)	XGBoost	COVID-19 database (*N* = 883)	No	20	<48 h, 13%; <7 days 19%	5	N/A	N/A	N/A	N/A	N/A	0.70	The most detailed information on sedative and analgesic dosages. SHAP value was obtained.	Lowest AUROC. Predictive performance was not obtained without AUROC. Not generalizable because it only involved patients with COVID-19. External validation was not performed.

ANN, artificial neural network; AUROC, area under the receiver operating characteristic curve; COVID, coronavirus disease; MIMIC, Medical Information Mart for Intensive Care; N/A, not available; SHAP, shapley additive explanations; SVM, support vector machine. Because Chen and Zhao's study was about extubation failure prediction, the sensitivity and specificity of the original data were replaced. [a]Training and test dataset, [b]External validation dataset.

than actual accuracy. Therefore, one patient's data must be used in either the training or test data.

Features

The features used for prediction can be categorized as follows: demographic information, vital signs, laboratory results, ventilator information, clinical intervention, and clinical scores (Table 2). Features that are clinically important and are frequently used in clinical practice should be included in the protocols.

The largest number of features that have been used for the prediction of extubation success thus far is 78. Using a larger number of features may result in more accurate predictions and may help identify unknown features that affect predictions. On the other hand, as the number of features increases, the proportion of missing values also increases because they may not have been measured; if this proportion exceeds a certain percentage, the missing values are excluded from the list of features. Missing values can be compensated for to a certain extent, but an increase in the number of missing values affects the development and prediction of the machine learning model. Some studies have developed models using all possible features, as well as new models using only features of high importance. For example, in the study by Chen et al. (19) features with missing data of ≥40% were not used, and missing values were compensated for by using the average values of the other patients. The model was initially developed with 68 features, and finally, a prediction model was developed with 36 of the most important features. A comparison of the accuracy based on the number of features using LightGBM, which had the highest accuracy, revealed that the accuracy ranged from 0.8023 to 0.8020, sensitivity from 0.7485 to 0.7477, specificity from 0.8327 to 0.8394, and area under the curve from 0.8130 to 0.8198; all these values were almost equivalent.

Outcomes (labeling)

Supervised learning involves labeling the correct answer (success or failure), called annotation. Failed extubation was defined as reintubation within 48 or 72 h of extubation in previous studies. The proportion of extubation failure ranged from 5 to 26%.

In general, most studies on ventilator liberation define extubation failure as reintubation within 24–72 h (26, 27), or up to 7 days (28, 29). It is difficult to define patients who require non-invasive positive pressure ventilation (NPPV) or nasal high-flow as successful extubations. However, studies have shown that post-extubation NPPV or nasal high-flow reduces reintubation compared to standard oxygen therapy, and clinical practice guidelines also recommend nasal high-flow for high-risk patients (30–34). Therefore, the use of NPPV or nasal high flow cannot be defined as extubation failure as they are sometimes used routinely.

Machine learning algorithms

In general, artificial neural networks are models that mimic parts of the neural circuits of the brain, which are highly accurate but cannot be explained, like a black box. On the other hand, since medicine emphasizes causal relationships and accountability to patients, models that can be explained by the importance of features, Shapley additive explanations (SHAP) values, and decision trees are sometimes preferred. This method is often used to improve performance by creating multiple weak models (weak learners), called boosting, where the previous learner is repeatedly modified by the next learner.

Predictive models were built using multiple algorithms and compared for accuracy. No model was consistently accurate, and results varied across the studies. It is difficult to determine which algorithm is most appropriate by comparing the values of accuracy across studies because some studies do not list all scores.

Validation

When machine learning models are developed at a single institution or with a small number of cases, there is a possibility that biased models will be created. Therefore, it is necessary to perform external validation to evaluate the accuracy using data that were not used for model construction. By evaluating the accuracy using external data, it is possible to know if the machine learning model is generalizable. However, only one study conducted external validation (22).

Performance evaluation

The common performance metrics used to evaluate the predictive performance of the model are accuracy, sensitivity (recall), precision (positive predictive value), specificity, and F1 score. In the equations given below, successful extubation is described as positive and failed extubation as negative. Note that in the opposite case, sensitivity and specificity are opposite (Table 1).

Accuracy is the ratio of correct predictions to the total number of predictions.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

TP, true positive; TN, true negative; FP, false positive; FN, false negative.

TABLE 2 Classification and listing of features.

Features	Kuo	Hsieh	Chen	Fabregat	Otaguro	Zhao	Fleuren
Demographic							
Age	X	X	X	X	X	X	X
Gender		X	(X)	X	X	X	X
Ethnicity						X	
Weight			X		X		
Height			X		X		
Body mass index		X		X	X	X	X
Weight loss			(X)				
Past medical history		X	(X)			X	X
Charlson index						X	
Reasons for respiratory failure	X						
Vital signs							
Heart rate		X	X	X	X	X	X
Respiratory rate		X	X	X	X	X	X
Body temperature			X		X	X	
Systolic blood pressure			X		X		
Diastolic blood pressure			X		X		
Mean arterial pressure		X	X		X	X	
Glasgow coma scale			X	X	X	X	X
Richmond agitation-sedation scale			X	X			X
SpO_2				X		X	
O_2 saturation to inspired fraction ratio				X		X	
SpFiO2/RR				X			
End-tidal carbon dioxide					X		
Number of premature ventricular contraction					X		
Laboratory results							
White blood cell			X		X	X	X
Red blood cell					X	X	
Hemoglobin		X	X		X	X	
Hematocrit		X			X	X	X
Platelet			X		X	X	X
Mean corpuscular volume						X	
Mean corpuscular hemoglobin						X	
Mmean corpuscular hemoglobin concentration						X	
Red cell distribution width						X	
Arterial pH		X	X		X	X	
$PaCO_2$		X	X		X	X	X
PaO_2		X	X		X	X	
P/F ratio		X			X	X	X
SaO2			X		X	X	
Base excess					X	X	
Na^+		X	X		X	X	
K^+		X	X		X	X	
Ca^+		X	X		X	X	

(Continued)

TABLE 2 Continued

Features	Kuo	Hsieh	Chen	Fabregat	Otaguro	Zhao	Fleuren
P^{+}		X					
Cl^{-}		X	X		X	X	
$HCO3^{-}$					X	X	
Anion gap					X	X	
Lactate			X		X	X	
Carboxyhemoglobin					X		
Methemoglobin					X	X	
Alveolar-arterial oxygen gradient			(X)				
Central venous oxygen saturation			X			X	X
Glucose		X	X			X	X
Creatinine		X	X		X	X	
Blood urea nitrogen (BUN)			(X)				
Troponin			(X)		X		
Total protein			(X)				
B-type natriuretic peptide			(X)		X		X
C-reactive protein			(X)		X		
Aspartate aminotransferase			(X)		X	X	
Alanine transaminase					X	X	
Lactate dehydrogenase (LDH)					X	X	
Alkaline phosphatase					X		
Creatine phosphokinase			(X)		X	X	
Total bilirubin		X	X		X	X	
Albumin					X		
Amylase			(X)			X	
Prothrombin time			X		X	X	
Activated partial thromboplastin time			X		X	X	
PT/INR					X	X	
Fibrinogen							
Ventilator information							
Number of previous mechanical ventilation events				X			
Time under mechanical ventilation (T_{MV})	X	X	X	X	X	X	X
Hours since last controlled mode							X
Ventilation mode				X			
Fraction of inspired oxygen		X			X	X	X
Tidal volume	X		X	X	X	X	
Tidal volume per kg ideal body weight							X
Minute volume		X	X		X		
Mean airway pressure			X		X	X	
Peak inspiration pressure			X	X	X		
Plateau pressure				X		X	
PEEP		X				X	X

(Continued)

TABLE 2 Continued

Features	Kuo	Hsieh	Chen	Fabregat	Otaguro	Zhao	Fleuren
Positive end-expiratory pressure		X			X	X	X
Maximum inspiratory pressure		X					
Maximum expiratory pressure		X					
Airway occlusion pressure							X
Ventilatory ratio							X
Inspiratory time	X						
Expiratory time	X						
Spontaneous breathing trial success times						X	
Clinical intervention							
Hospital stay	X						
ICU stay	X						
Sedation day			X				
Sedatives and analgesics dose				X			X
Total cumulative dose (sedatives and analgesics)							X
Vasopressor			(X)			X	
Antibiotic type (ABX)						X	
Fluid balance							X
Urine output						X	
Continuous renal replacement therapy						X	
Crystalloid and colloid amount						X	
Transufusion (RBC, FFP, PLT)						X	
Hours since last proning session						X	
Central venous pressure						X	
Rapid shallow breathing index		X	(X)	X		X	
Clinical scores							
Sequential organ failure assessment			X			X	
Simplified acute physiology score ==						X	
Acute physiology and chronic health evaluation-II	X			X			X
SEMICYUC code				X			
ROX index						X	
Top 5 important features							
1	N/A	N/A	T_{MV}	N/A	T_{MV}	Strokes	N/A
2			PaO2		Age	RR	
3			PaCO2		PEEP	ABX	
4			pH		LDH	T_{MV}	
5			BUN		APTT	SpO2	

FiO2, fraction of inspired oxygen; PT/INR, Prothrombin time and international normalized ratio; N/A, not applicable; P/F, arterial oxygen partial pressure to fractional inspired oxygen; ROX, respiratory rate-oxygenation; SEMICYUC, Sociedad Española de Medicina Intensiva, Crítica y Unidades Coronarias; SpFiO2/RR, respiratory rate-oxygen index. Sedatives and analgesics include benzodiazepine, clonidine, dexmedetomidine, fentanyl, haloperidol, midazolam, propofol, and quetiapine. (X) was included in the first 68 features but was not in the top 36, so it was not used to develop the second model.

Sensitivity (recall) is the ratio of true positives to total actual positives. It represents the proportion of patients who underwent successful extubation and in whom extubation was predicted to be successful.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

Precision (positive prediction value) is the ratio of true positives to total predicted positives.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

Specificity is the ratio of true negatives to total negatives. It indicates the proportion of patients who underwent unsuccessful extubation and in whom extubation was predicted to be unsuccessful.

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN})$$

$\text{F1} - \text{score}$ is the harmonic mean of precision and recall.

$$\text{F1 score} = 2\times(\text{recall}\times\text{precision})/(\text{recall} + \text{precision}).$$

Among the seven studies, sensitivity ranged from 0.64 to 0.96, precision ranged from 0.91 to 0.97, specificity ranged from 0.73 to 0.85, and the area under the receiver operating characteristic curve ranged from 0.70 to 0.98 (Table 1).

Feature importance

Three studies showed the importance of features and two studies showed SHAP values. In the study by Chen et al., the top 10 most important features were duration of mechanical ventilation, PaO_2, $PaCO_2$, arterial pH, blood urea nitrogen (BUN), mean heart rate, weight, age, creatine, and Glasgow Coma Scale (GCS) score (19). In the study by Otaguro et al. (21) the top 10 most important features were duration of mechanical ventilation, age, PEEP, lactate dehydrogenase, activated partial thromboplastin time, alveolar arterial oxygen tension difference gradient, BUN, GCS score, C-reactive protein (CRP), and albumin. In the study by Zhao et al. (22) the top 10 most important features were stroke, respiratory rate, antibiotic type, pressure support ventilation level, central venous pressure, mechanical ventilation duration, SpO_2, PEEP, heart rate, and number of successful SBTs.

Mechanical ventilation duration was included in three studies; PaO_2 (or SpO_2), BUN, heart rate, GCS, age, and PEEP in two studies; BUN, creatinine, CRP, and albumin were not included in the protocols, but may have influenced the success or failure of extubation because these features reflect the patient's general condition (BUN and creatinine reflect renal function, CRP reflects inflammatory status, and albumin reflects nutritional status).

Discussion

We reviewed studies that used machine learning to predict the success or failure of extubation. Protocols determine extubation based on whether a patient meets certain defined conditions, but machine learning makes predictions based on certain features. Currently, it remains unclear whether protocols or machine learning contributes to a higher extubation success rate.

Performance evaluation

A model with high specificity is preferable for predicting extubation failure. On the other hand, there exists a trade-off relationship: the higher the specificity, the lower the sensitivity. Therefore, the number of patients in which extubation is actually possible but expected to fail increases. This increases the risk of prolonging the duration of ventilator management and increasing tracheostomies.

For example, Zhao's study has the highest specificity of 0.85 but the lowest sensitivity of 0.64. Therefore, this model was able to predict the highest number of patients who failed extubation, so if this model predicts successful extubation, the probability of successful extubation is quite high, with a precision of 0.97. This model could reduce the rate of extubation failure from its current value of 11% to a theoretical 3%. In contrast, a model with high sensitivity can predict more patients who will succeed extubation. However, this model cannot be used for monitoring for possible early extubation. These models were trained on patients who underwent SBT and were deemed extubatable, and did not include patients whose final decision was not to be extubated. As a result, these models cannot be automatically used for all patients in the ICU to alert the timing of extubation, and must be considered as a support tool aimed to validate the physician's decision upon extubation. As the sensitivity increases, the specificity decreases, which leads to an aggressive model that predicts success in patients who cannot be extubated. A mode with high sensitivity and low specigicity is not suitable for this purpose. Therefore, a model with high specificity that can predict as many extubation failures as possible is preferred. If a conservative model for extubation is used in clinical practice, it is important to use the neative predictive valie as an indicator to avoid prolonged ventilatory management.

Problems with machine learning methods

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) serves as a guideline for the development, validation, and update of clinical prediction models (35). The TRIPOD statement aims to increase the transparency of reporting on prediction models, regardless of the research methodology. We believe that all studies on prediction models using machine learning should be reported according to the TRIPOD checklist.

Several of the reviewed studies included issues related to imbalanced data and validation. In many databases, the failure rate of extubation was 10–20%. The number of successful and failed extubations was 9:1 to 8:2 in imbalanced data, and even if there was a model that predicted all successful extubations, it would have an accuracy of 80–90%, which could lead to a model that excessively predicts successful extubations. In this case, balancing the number of samples by under or oversampling can prevent the model from being overly predictive of successful extubation. Under or oversampling should be done in models that predict successful or failed extubation because the dataset is likely to be unbalanced. Some of the studies used data from multiple time points prior to extubation to increase the sample size for extubation failure, whereas some used patient data that were used as training data during cross validation as test data, which may have resulted in a higher than actual accuracy. External validation should be performed on databases that were not used as training data. In a single-center dataset, the unique extubation process of the institution may have influenced the development of the machine learning model. ICU datasets are freely available [Amsterdam UMCdb (36), eICU-CRD (37), HiRID (38), and MIMIC-IV (25)]; it is possible to perform external validation using these datasets. Conversely, it is also possible to build a machine learning model using these datasets and perform external validation using the dataset of a single institution.

Future prospects

All the studies reviewed here were retrospective, and no prospective studies have been conducted thus far. It is necessary to evaluate the accuracy of machine learning via prospective studies and to compare it with the physicians' decisions. However, since it is ethically questionable to make an extubation decision based solely on the results of machine learning, it is necessary to evaluate the predictions and outcomes of machine learning when the physician makes the decision to extubate, assuming that the results of machine learning are blinded to the physician. Furthermore, it is necessary to consider how synergy between using the protocol and the predictions of machine learning can be obtained.

Finally, to verify whether machine learning predictions are useful in clinical practice, a study is required in which a group of patients are randomly divided into two groups: those who underwent extubation using machine learning predictions and those who underwent extubation without using machine learning predictions. Even if machine learning reduces the incidence of extubation failure, outcomes need to be evaluated to assess whether it prolongs the duration of ventilator use and increases performance of tracheostomies, ventilator-related complications, and mortality.

Author contributions

Conception and design: YI. Acquisition of data: YI, KO, KN, and SO. Analysis and interpretation of data and drafting or revising the article: YI, KO, KN, SO, HO, and SY. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by KAKENHI 20K17876, Grant-in-Aid for Young Scientists.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Carson SS, Cox CE, Holmes GM, Howard A, Carey TS. The changing epidemiology of mechanical ventilation: a population-based study. *J Intensive Care Med.* (2006) 21:173–82. doi: 10.1177/0885066605282784

2. Lim ZJ, Subramaniam A, Ponnapa Reddy M, Blecher G, Kadam U, Afroz A, et al. Case fatality rates for patients with COVID-19 requiring invasive mechanical ventilation. a meta-analysis. *Am J Respir Crit Care Med.* (2021) 203:54–66. doi: 10.1164/rccm.202006-2405OC

3. WHO Coronavirus (COVID-19) Dashboard (2022). Available online at: https://covid19.who.int/ (accessed June 1, 2022).

4. Melsen WG, Rovers MM, Groenwold RH, Bergmans DC, Camus C, Bauer TT, et al. Attributable mortality of ventilator-associated pneumonia: a meta-analysis of individual patient data from randomised prevention studies. *Lancet Infect Dis.* (2013) 13:665–71. doi: 10.1016/S1473-3099(13)70081-1

5. Rengel KF, Hayhurst CJ, Pandharipande PP, Hughes CG. Long-term cognitive and functional impairments after critical illness. *Anesth Analg.* (2019) 128:772–80. doi: 10.1213/ANE.0000000000004066

6. Frutos-Vivar F, Esteban A, Apezteguia C, González M, Arabi Y, Restrepo MI, et al. Outcome of reintubated patients after scheduled extubation. *J Crit Care.* (2011) 26:502–9. doi: 10.1016/j.jcrc.2010.12.015

7. Esteban A, Anzueto A, Frutos F, Alía I, Brochard L, Stewart TE, et al. Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study. *JAMA.* (2002) 287:345–55. doi: 10.1001/jama.287.3.345

8. Baptistella AR, Sarmento FJ, da Silva KR, Baptistella SF, Taglietti M, Zuquello RÁ, et al. Predictive factors of weaning from mechanical ventilation and extubation outcome: a systematic review. *J Crit Care.* (2018) 48:56–62. doi: 10.1016/j.jcrc.2018.08.023

9. Heunks LM, van der Hoeven JG. Clinical review: the ABC of weaning failure–a structured approach. *Crit Care.* (2010) 14:245. doi: 10.1186/cc9296

10. 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) 161:97–111. doi: 10.1016/j.chest.2021.06.030

11. Blackwood B, Alderdice F, Burns K, Cardwell C, Lavery G, O'Halloran P. Use of weaning protocols for reducing duration of mechanical ventilation in critically ill adult patients: cochrane systematic review and meta-analysis. *BMJ.* (2011) 342:c7237. doi: 10.1136/bmj.c7237

12. Thille AW, Richard JC, Brochard L. The decision to extubate in the intensive care unit. *Am J Respir Crit Care Med.* (2013) 187:1294–302. doi: 10.1164/rccm.201208-1523CI

13. 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) 325:1173–84. doi: 10.1001/jama.2021.2384

14. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med.* (2018) 6:905–14. doi: 10.1016/S2213-2600(18)30300-X

15. 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 Fail Collab Group N Engl J Med.* (1995) 332:345–50. doi: 10.1056/NEJM199502093320601

16. Fan T, Wang G, Mao B, Xiong Z, Zhang Y, Liu X, et al. Prophylactic administration of parenteral steroids for preventing airway complications after extubation in adults: meta-analysis of randomised placebo controlled trials. *BMJ.* (2008) 337:a1841. doi: 10.1136/bmj.a1841

17. 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) 60:1560–9.

18. 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) 7.

19. Chen T, Xu J, Ying H, Chen X, Feng R, Fang X, et al. Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access.* (2019) 7:150960–8. doi: 10.1109/ACCESS.2019.2946980

20. Fabregat A, Magret M, Ferre JA, Vernet A, Guasch N, Rodriguez A, et al. A machine learning decision-making tool for extubation in intensive care unit patients. *Comput Methods Programs Biomed.* (2021) 200:105869.

21. Otaguro T, Tanaka H, Igarashi Y, Tagami T, Masuno T, Yokobori S, et al. Machine learning for prediction of successful extubation of mechanical ventilated patients in an intensive care unit: a retrospective observational study. *J Nippon Med Sch.* (2021) 88:408–17. doi: 10.1272/jnms.JNMS.2021_88-508

22. Zhao QY, Wang H, Luo JC, Luo MH, Liu LP Yu SJ, et al. Development and validation of a machine-learning model for prediction of extubation failure in intensive care units. *Front Med.* (2021) 8:676343. doi: 10.3389/fmed.2021.676343

23. Fleuren LM, Dam TA, Tonutti M, de Bruin DP, Lalisang RCA, Gommers D, et al. Predictors for extubation failure in COVID-19 patients using a machine learning approach. *Crit Care.* (2021) 25:448.

24. Kwong MT, Colopy GW, Weber AM, Ercole A, Bergmann JH. The efficacy and effectiveness of machine learning for weaning in mechanically ventilated patients at the intensive care unit: a systematic review. *Bio-Des Manufact.* (2019) 2:31–40. doi: 10.1007/s42242-018-0030-1

25. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 1.0). In: *PhysioNet.* (2021). doi: 10.13026/7vcr-e114

26. Smina M, Salam A, Khamiees M, Gada P, Amoateng-Adjepong Y, Manthous CA. Cough peak flows and extubation outcomes. *Chest.* (2003) 124:262–8. doi: 10.1378/chest.124.1.262

27. Martinez A, Seymour C, Nam M. Minute ventilation recovery time: a predictor of extubation outcome. *Chest.* (2003) 123:1214–21. doi: 10.1378/chest.123.4.1214

28. Epstein SK, Ciubotaru RL, Wong JB. Effect of failed extubation on the outcome of mechanical ventilation. *Chest.* (1997) 112:186–92. doi: 10.1378/chest.112.1.186

29. Esteban A, Alía I, Gordo F, Fernández R, Solsona JF, Vallverdú I, et al. Extubation outcome after spontaneous breathing trials with T-tube or pressure support ventilation. The Spanish lung failure collaborative group. *Am J Respir Crit Care Med.* (1997) 156:459–65. doi: 10.1164/ajrccm.156.2.9610109

30. Maggiore SM, Idone FA, Vaschetto R, Festa R, Cataldo A, Antonicelli F, et al. Nasal high-flow versus Venturi mask oxygen therapy after extubation. Effects on oxygenation, comfort, and clinical outcome. *Am J Respir Crit Care Med.* (2014) 190:282–8. doi: 10.1164/rccm.201402-0364OC

31. Hernández G, Vaquero C, Colinas L, Cuena R, González P, Canabal A, et al. Effect of post-extubation high-flow nasal cannula vs noninvasive ventilation on reintubation and post-extubation respiratory failure in high-risk patients: a randomized clinical trial. *JAMA.* (2016) 316:1565–74. doi: 10.1001/jama.2016.14194

32. Hernández G, Vaquero C, González P, Subira C, Frutos-Vivar F, Rialp G, et al. Effect of post-extubation high-flow nasal cannula vs conventional oxygen therapy on reintubation in low-risk patients: a randomized clinical trial. *JAMA.* (2016) 315:1354–61. doi: 10.1001/jama.2016.2711

33. Thille AW, Muller G, Gacouin A, Coudroy R, Decavèle M, Sonneville R, et al. Effect of post-extubation high-flow nasal oxygen with noninvasive ventilation vs. high-flow nasal oxygen alone on reintubation among patients at high risk of extubation failure: a randomized clinical trial. *JAMA.* (2019) 322:1465–75. doi: 10.1001/jama.2019.14901

34. Rochwerg B, Brochard L, Elliott MW, Hess D, Hill NS, Nava S, et al. Official ERS/ATS clinical practice guidelines: non-invasive ventilation for acute respiratory failure. *Eur Respir J.* (2017) 50:1602426. doi: 10.1183/13993003.02426-2016

35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* (2015) 350:g7594. doi: 10.1136/bmj.g7594

36. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU patient data responsibly under the society of critical care medicine/European society of intensive care medicine joint data science collaboration: the Amsterdam university medical centers database (AmsterdamUMCdb) example. *Crit Care Med.* (2021) 49:e563–77. doi: 10.1097/CCM.0000000000004916

37. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data.* (2018) 5:180178. doi: 10.1038/sdata.2018.178

38. Faltys M, Zimmermann M, Lyu X, Hüser M, Hyland S, Rätsch G, et al. HiRID, a high time-resolution ICU dataset (version 1.1.1). In: *PhysioNet.* (2021).