Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241289732
journals.sagepub.com/home/dhj
Sage

Multitask learning to predict successful weaning in critically ill ventilated patients: A retrospective analysis of the MIMIC-IV database

Ming-Yen Lin[1], **Hsin-You Chi**[1] **and Wen-Cheng Chao**[2,3,4,5]

Abstract

Objective: Weaning is an essential issue in critical care. This study explores the efficacy of multitask learning models in predicting successful weaning in critically ill ventilated patients using the Medical Information Mart for Intensive Care (MIMIC) IV database.

Methods: We employed a multitask learning framework with a shared bottom network to facilitate common knowledge extraction across all tasks. We used the Shapley additive explanations (SHAP) plot and partial dependence plot (PDP) for model explainability. Furthermore, we conducted an error analysis to assess the strength and limitation of the model. Area under receiver operating characteristic curve (AUROC), calibration plot and decision curve analysis were used to determine the performance of the model.

Results: A total of 7758 critically ill patients were included in the analyses, and 78.5% of them were successfully weaned. Multitask learning combined with spontaneous breath trial achieved a higher performance to predict successful weaning compared with multitask learning combined with shock and mortality (area under receiver operating characteristic curve, AUROC, 0.820 ± 0.002 vs 0.817 ± 0.001, $p < 0.001$). We assessed the performance of the model using calibration and decision curve analyses and further interpreted the model through SHAP and PDP plots. The error analysis identified a relatively high error rate among those with low disease severities, including low mean airway pressure and high enteral feeding.

Conclusion: We demonstrated that multitask machine learning increased predictive accuracy for successful weaning through combining tasks with a high inter-task relationship. The model explainability and error analysis should enhance trust in the model.

Keywords

Critical care, weaning, multitask learning, interpretability, error analysis

Submission date: 20 February 2024; Acceptance date: 17 September 2024

[1]Department of Information Engineering and Computer Science, Feng Chia University, Taichung
[2]Department of Critical Care Medicine, Taichung Veterans General Hospital, Taichung
[3]Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung
[4]Department of Automatic Control Engineering, Feng Chia University, Taichung
[5]Big Data Center, National Chung Hsing University, Taichung

Corresponding author:
Wen-Cheng Chao, Department of Critical Care Medicine, Taichung Veterans General Hospital, Taichung; Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung;
Department of Automatic Control Engineering, Feng Chia University, Taichung; Big Data Center, National Chung Hsing University, Taichung.
Email: cwc081@hotmail.com

Introduction

Mechanical ventilation is a fundamental organ support intervention in critically ill patients, and the number of patients requiring mechanical ventilation has increasingly risen in the past two decades, with approximately more than 40% of patients, who were admitted to intensive care units (ICUs), requiring mechanical ventilation.[1,2] The increased usage of mechanical ventilation underscores the need for advancements in patient management and outcome prediction strategies, particularly in the weaning from mechanical ventilation.[3–5] Given that delay in the discontinuation of mechanical ventilation increases the risk of ventilator-associated complications and hospitalisation costs, recent studies have explored various methodologies for predicting weaning mechanical ventilation, employing conventional techniques and advanced machine learning- or deep learning-based tools.[6–9] Our previous research has also contributed significantly to this field, utilising explainable machine learning algorithms for the prediction of weaning in those with prolonged mechanical ventilation and weaning attempts in critically ill ventilated patients.[10,11] The aforementioned extubation prediction systems are mainly used to aid comprehensive judgment when physicians are already evaluating the probability of weaning. The primary difference in our study is that we aim to use parameters available in daily ICU practice, such as ventilatory parameters, enteral feeding, and fluid status, and this approach allows us to provide a continuous prediction model.

Multitask learning (MTL) is emerging as a novel approach in artificial intelligence (AI), particularly for handling complex data, and is characterised by shared feature representation, improved generalisation, and resource efficiency.[12,13] MTL is becoming increasingly prominent in various medical fields due to its comprehensive approach to data analysis and interpretation, and accumulating evidence have shown the potential for the application of MTL in critical care medicine with multifaceted and interrelated data in critical care.[14,15] Notably, to discern the difference between clinically relevant outcomes and task relatedness is crucial for the application of MTL in critical care. Clinically relevant outcomes in critical care, such as respiratory failure, shock, or kidney failure, are highly correlated outcomes; however, their clinical relevance does not necessarily translate into task relatedness within the MTL framework. Task relatedness in MTL depends on the correlation among features used by distinct models rather than on the critically relevant outcomes.[16,17]

This study used the Medical Information Mart for Intensive Care (MIMIC) IV, a public critical care database with comprehensive ventilator-relevant data, in conjunction with MTL techniques to establish a prediction model for successful weaning in critically ill ventilated patients.[18] Our approach aims to establish not only a prediction model with high accuracy but also responsible AI practices, emphasising model interpretability and error analysis.

Methods

Database

MIMIC-IV, a publicly accessible electronic health record dataset, was established in collaboration with Beth Israel Deaconess Medical Center (BIDMC) and the Massachusetts Institute of Technology (MIT).[18] MIMIC-IV contains data from BIDMC on all patients who were admitted to either the emergency department or the ICU between 2008 and 2019. Data preprocessing steps include data cleaning, normalisation, and handling missing values. Relevant features for weaning prediction, such as demographic information, ventilatory parameters, vital signs, and laboratory test results, were extracted. All data were extracted using the Structured Query Language with PostgreSQL (version 13.3). This study was approved by the institutional review board of Taichung Veterans General Hospital (IRB number: CE23182A). The informed consent was waived because all data were anonymised.

Model establishment

The training-to-testing data proportion was set at 80/20, and the multilayer perceptron (MLP) was used to establish the model. Successful weaning was defined by weaning from mechanical ventilation for more than 48 h, whereas unsuccessful weaning was defined as any requirement to resume mechanical ventilation within 48 h after an initial attempt to wean the patient. Our aim was to predict weaning one day prior to successful weaning using data from two days prior to successful weaning. Therefore, both the feature window and the prediction window were set at 24 h (Supplemental Figure 1 provides details of the study's data time frame). The prediction model was established using an artificial neural network with a MLP architecture. An MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer consists of nodes that are fully connected to the nodes in the subsequent layer. The MLP processes input features, such as vital signs, through these interconnected layers, with the hidden layers performing transformations to learn complex patterns within the data. The final output layer generates the prediction of successful weaning. Our model is characterised by MTL, which concurrently learns patterns from multiple prediction tasks.

Model interpretation

The black-box issue refers to the lack of transparency in how models make predictions. In other words, the prediction model might provide highly accurate predictions, but the processes of how the model predicts cannot be explained. To address concerns about the black-box issue in the model, we utilised the SHapley Additive exPlanations (SHAP) summary plot and partial dependence plot (PDP) for visualised interpretation of the model. We further used recursive feature elimination for a succinct model and used 18 features to establish the prediction model for successful weaning (Supplemental Figure 2 for the results of recursive feature elimination analysis). The SHAP summary plot provided a comprehensive representation of the direction and magnitude of correlations between features and weaning outcomes, while the PDP further illustrated the marginal effect of the feature on the weaning outcome.[11,19]

Multitask learning

MTL consists of two major modules, namely the bottom network and the tower network. The bottom network acquires common knowledge across all tasks.[12] The output of the bottom network serves as the input for the tower network. In our MTL framework, the shared bottom network was designed to prioritise the extraction of features that provide the greatest benefit to all tasks, adopting its focus based on the learning progress and the specific requirements of each task (Figure 1) (Supplemental methodology for details regarding the architecture of the multitask model).[13] This strategy not only improves the overall efficiency of the model but also enhances its adaptability, allowing it to effectively handle a wide range of tasks with varying degrees of complexity and interrelation.

Selection of tasks

The inter-task relationship is an essential issue in MTL.[20] However, clinically relevant outcomes in critical care might not reflect the task relatedness within the MTL framework. To address this issue, we tested the combination of successful weaning prediction with mechanical ventilation irrelevant tasks and mechanical ventilation relevant tasks. The first combination of tasks (MTL-1) consisted of two mechanical ventilator irrelevant tasks (i.e., shock and in-hospital mortality) and successful weaning. The other combination of tasks (MTL-2) was composed of three mechanical ventilation-related tasks: the initiation of spontaneous breathing trial (SBT), successful SBT, and successful weaning. In brief, SBT is one of the key components of protocol-based weaning, and the initiation of SBT was based on the met criteria, including low fraction of inspired oxygen (FiO_2) and positive end-expiratory pressure (PEEP) demand, no requirement of vasopressor, and adequate consciousness.[8,21] The failure of the SBT was defined by clinical intolerance, including tachypnea (respiratory rate >35 cycles/min), desaturation (oxygen saturation <90%), change of heart rate or blood pressure higher than 20%, cold sweating, and agitation of change of consciousness.[8,22] To assess the associations among tasks, we used the heat map to

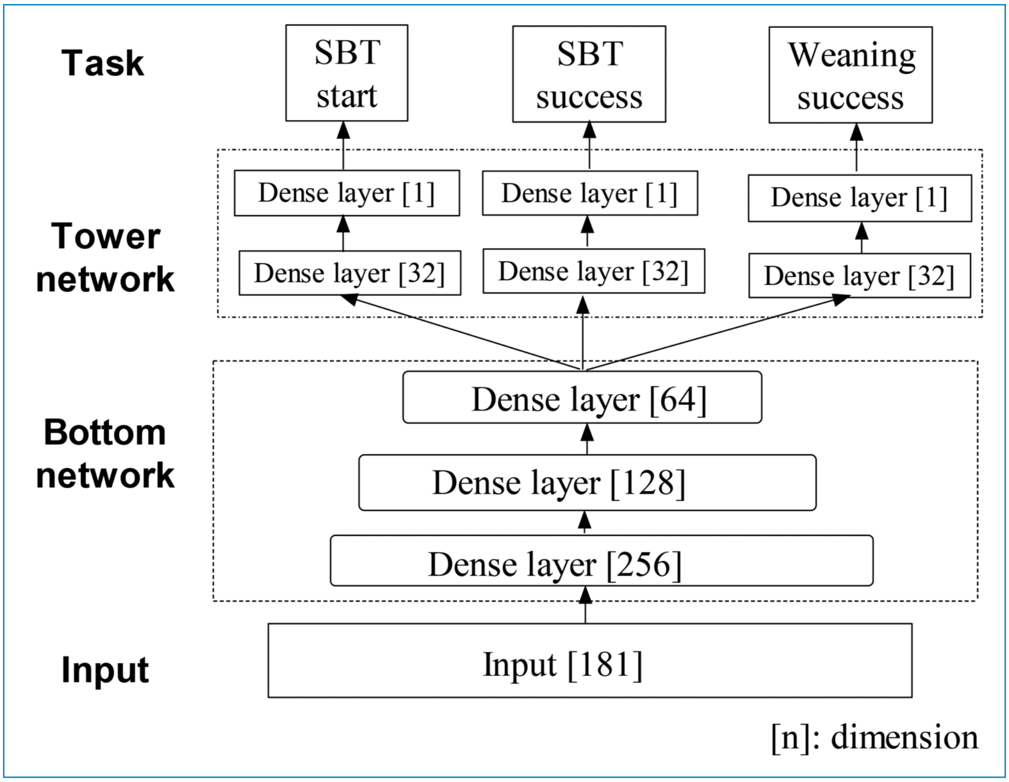

Figure 1. Architecture of the multitask model.

illustrate the association among features, with the high density of colour representing a high association determined by the ranks of features to predict distinct tasks.

Error analysis

The objective of error analysis is to identify the subgroup with high error rates and the factors contributing to these errors in our deep learning model. For this purpose, we use an easy-to-understand tree-based model, referred to as the surrogate model, which mimics the original black-box model, which was the teacher model. The surrogate model aims to replicate the output of the teacher model as closely as possible after proper training. Therefore, the explanations provided by the surrogate model can reveal the internal prediction mechanisms of the teacher model. In this study, we used the error analysis toolkit within the Responsible AI Widgets repository.

Statistical analysis

We presented the continuous data as means ± standard deviations, and categorical data were expressed as frequencies (percentages). The chi-square test and the Student's *t*-test were used to determine the difference between the two groups. The performance of the model was measured by the discrimination, accuracy across predictive probabilities and applicability of the models in the testing sets by using the area under receiver operating characteristic (AUROC) curve analysis, calibration plot and decision curve analysis, respectively.[23,24] We further used the DeLong test to determine the difference in performance among distinct models.[25] Python version 3.7.4 was applied in the present study.

Results

Patient characteristics

A total of 7758 independent critically ill patients requiring mechanical ventilation for more than three days during ICU admission were eligible for analyses (Figure 2). The median age of the enrolled subjects was 63.5 ± 16.7 years, and 55.7% of them were male (Table 1). We found that 78.5% (6091 of 7758) of them were successfully weaned from mechanical ventilation during ICU admission. Patients with and without successful weaning had similar distributions in ethnicity. Patients in the unsuccessful weaning group had a higher Acute Physiology and Chronic Health Evaluation (APACHE) III score (71.8 ± 25.9 vs 55.1 ± 23.1, $p < 0.01$). With regards to ventilator parameters, patients without successful weaning received a higher FiO2 (56.5 ± 20.6 vs 53.3 ± 17.0%, $p < 0.01$), PEEP (6.9 ± 3.2 vs 6.3 ± 2.5 cmH_2O, $p < 0.01$), peak airway pressure (22.0 ± 6.6 vs 20.6 ± 5.8 cmH_2O, $p < 0.01$), and mean airway pressure (10.3 ± 3.1 vs 9.6 ± 2.7 cmH_2O, $p < 0.01$) than those weaned from mechanical ventilation successfully.

The performance of MTL using related and unrelated tasks

We then compared the performance of the single-task learning with that of the two MTLs (Table 2). We found that the MTL-1, consisting of shock, mortality and successful weaning, outperformed the single task learning of successful weaning (0.814 ± 0.002 vs 0.817 ± 0.002, $p < 0.01$ determined by the DeLong test). Notably, the MTL-2, weaning-related tasks, slightly outperformed the performance of MTL-1 (0.820 ± 0.002 vs 0.817 ± 0.001, $p < 0.001$). Supplemental Table 1 shows the performance of single-task and MTL to predict distinct outcomes, including shock, mortality, initiation of SBT, and successful SBT. The AUROC analysis was used to assess the discriminative ability of the MTL-2 to predict successful weaning (Figure 3(A)). We also illustrated the calibration plot to address the reliability of the model by comparing predicted probabilities with actual outcomes across different risk thresholds (Figure 3(B)). We also plotted the decision curve analysis that assessed the trade-offs between the benefits of true positive results and the harms of false positive results (Figure 3(C)).

Interpretability of the model

We then employed the SHAP plot to illustrate how these key features affect the probability of successful weaning (Figure 4). Using the SHAP summary plot, not only the strength but also the direction of each feature were clearly illustrated. For example, increased enteral feeding, a low APACHE III score, low peak airway pressure, improved consciousness status determined by Richmond Agitation Sedation Scale (RASS) score, low FiO_2, and a negative fluid balance were positively associated with a higher probability of successful weaning one day later. To further illustrate how each feature affects the probability of successful weaning, we used a PDP plot of the six key features (Figure 5). Taking peak airway pressure as an example, we found that the probability of successful weaning decreased gradually among patients whose peak airway pressure was higher than 16–20 cmH_2O. Taken together, these visualised interpretations should allow clinicians for a straightforward understanding of the model. To further elaborate on the related and unrelated tasks, we used the heat map to illustrate the association of ranks of features among distinct tasks (Figure 6). We found that the ranks of features appear to be consistent in related tasks (Figure 6(A)), whereas some features had discordant ranks among the unrelated tasks (Figure 6(B)). For examples, peak airway pressure was an essential feature in predicting successful weaning, but not for the prediction of shock and mortality.

Error analysis

We then conducted the error analysis and found that the error rate tended to be relatively high in patients with low mean airway pressure and high enteral feeding (35.1%, 220 of 626)

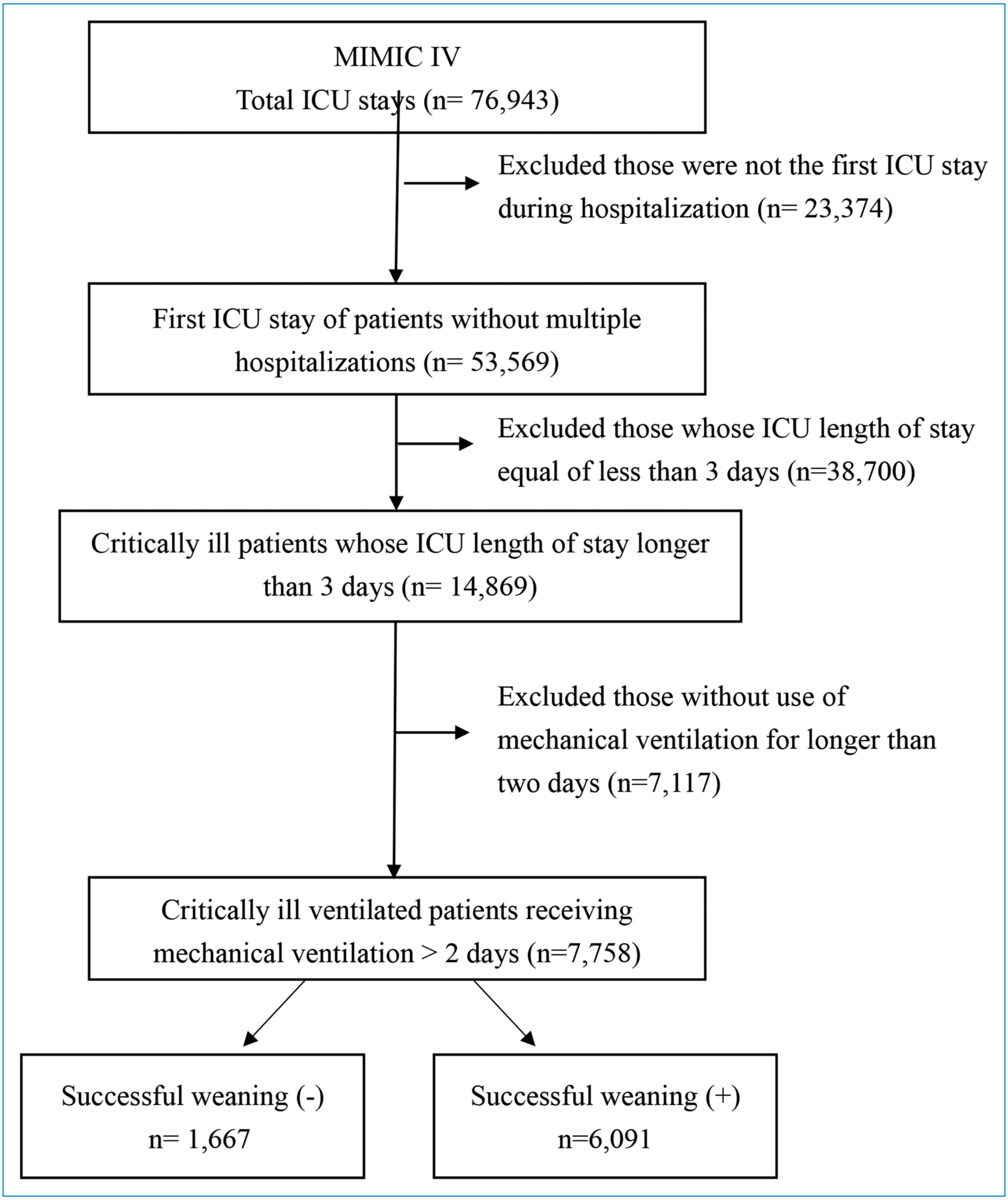
MIMIC IV
Total ICU stays (n= 76,943)
Excluded those were not the first ICU stay during hospitalization (n= 23,374)
First ICU stay of patients without multiple hospitalizations (n= 53,569)
Excluded those whose ICU length of stay equal of less than 3 days (n=38,700)
Critically ill patients whose ICU length of stay longer than 3 days (n= 14,869)
Excluded those without use of mechanical ventilation for longer than two days (n=7,117)
Critically ill ventilated patients receiving mechanical ventilation > 2 days (n=7,758)
Successful weaning (-) n= 1,667
Successful weaning (+) n=6,091

Figure 2. Flowchart of subject enrollment.

(Supplemental Figure 3). We hence excluded the 626 patients with low mean airway pressure as well as high enteral feeding and ran the same model in the remaining 1291 patients in the test dataset (Table 3). Notably, after excluding the 626 patients, we observed an increase in the AUROC and accuracy but a slight decrease in precision and an obvious decrease in recall. This finding reflects that the model in the remaining subgroup tended to be conservative, with high discrimination between positive and negative cases despite a slight decrease in true positive and a relative increase in false negative cases (Table 3).

Table 1. Characteristics of the 7758 critically ill patients receiving mechanical ventilator categorised by weaning outcome.

	All	Not successfully weaned	Successfully weaned	*p*-value
	N = 7758	N = 1667	N = 6091	
Demographic data				
Age (years)	63.5 ± 16.7	64.6 ± 16.7	63.2 ± 16.6	<0.01
Male	4466 (55.7%)	928 (55.7%)	3538 (58.1%)	<0.01
Female	3292 (42.3%)	739 (44.3%)	2553 (41.9%)	<0.01
Body weight (kg)	85.9 ± 24.5	85.2 ± 26.3	86.0 ± 24.0	0.22
Height (cm)	169.4 ± 9.3	169.1 ± 9.6	169.5 ± 9.3	0.13
APACHE III	58.7 ± 24.7	71.8 ± 25.9	55.1 ± 23.1	<0.01
Ethnicity				0.52
Caucasian	4784 (53.8%)	897 (53.8%)	3887 (63.8%)	
African American	626 (8.4%)	140 (8.4%)	486 (8.0%)	
Hispanic	263 (3.4%)	56 (3.4%)	207 (3.4%)	
Asian	199 (2.6%)	44 (2.6%)	155 (2.5%)	
Others/unknown	1886 (31.8%)	530 (31.8%)	1356 (22.3%)	
Ventilatory parameters				
FiO_2 (%)	54.0 ± 17.7	56.5 ± 20.0	53.3 ± 17.0	<0.01
PEEP (cmH_2O)	6.4 ± 2.7	6.9 ± 3.2	6.3 ± 2.5	<0.01
V_T (mL)	465.3 ± 78.6	454.2 ± 82.3	468.3 ± 77.3	<0.01
P_{peak}, (cmH_2O)	20.9 ± 6.0	22.0 ± 6.6	20.6 ± 5.8	<0.01
Pmean, (cmH_2O)	9.7 ± 2.8	10.3 ± 3.1	9.6 ± 2.7	<0.01
Laboratory data				
White blood cell count (count/μl)	12.23 ± 5.08	12.35 ± 5.25	12.2 ± 5.03	0.28
Hemoglobin (g/dL)	10.77 ± 2.02	10.82 ± 2.13	10.75 ± 2.0	0.26
Platelet (10^3/μL)	200.73 ± 92.05	203.21 ± 100.36	200.05 ± 89.64	0.21
Total bilirubin (mg/dL)	1.62 ± 0.66	1.61 ± 0.63	1.62 ± 0.67	0.48
HCO_3 (mmol/L)	22.74 ± 4.04	22.28 ± 4.52	22.87 ± 3.89	<0.01
$PaCO_2$ (cmH_2O)	41.51 ± 7.21	41.7 ± 8.09	41.46 ± 6.95	0.24
Outcome				

(continued)

Table 1. Continued.

	All N = 7758	Not successfully weaned N = 1667	Successfully weaned N = 6091	*p*-value
ICU length of stay (day)	7.66 ± 4.27	8.43 ± 4.83	7.44 ± 4.08	<0.01
Ventilator-day (day)	5.57 ± 5.54	8.51 ± 7.58	4.77 ± 4.51	<0.01
Hospital-stay (day)	14.31 ± 12.79	10.65 ± 10.16	15.31 ± 13.25	<0.01

Data are presented as mean ± standard deviation and number (percentage). Data were analysed by the Student's *t* test for continuous variables and chi-square test for categorical variables.
APACHE IV: acute physiology and chronic health evaluation IV; FiO2: the fraction of inspired oxygen; PEEP: positive end-expiratory pressure; V_T: tidal volume; P_{peak}: peak airway pressure; Pmean: mean airway pressure; HCO3: bicarbonate; PaCO2: partial pressure of carbon dioxide; ICU: intensive care unit.

Table 2. Performance of single-task and multitask learning to predict successful weaning.

	AUROC	Accuracy	Precision	F1-score	Recall
Single task learning[a]	0.814 ± 0.002	0.752 ± 0.005	0.673 ± 0.024	0.558 ± 0.037	0.481 ± 0.065
Multi-task learning-1[b]	0.817 ± 0.001*	0.754 ± 0.001	0.670 ± 0.042	0.576 ± 0.041	0.513 ± 0.079
Multitask learning-2[c]	0.820 ± 0.002#	0.755 ± 0.004	0.661 ± 0.010	0.582 ± 0.025	0.522 ± 0.043

AUROC: area under the receiver operating characteristics.
[a] Successful weaning.
[b] Tasks included shock, mortality, and successful weaning.
[c] Tasks included the start of spontaneous breathing trial, successful spontaneous breathing trial., and successful weaning.
*<0.001 between single task learning and multi-task learning-1; #<0.001 between multi-task learning-1 and multi-task learning-2, determined by Delung test.

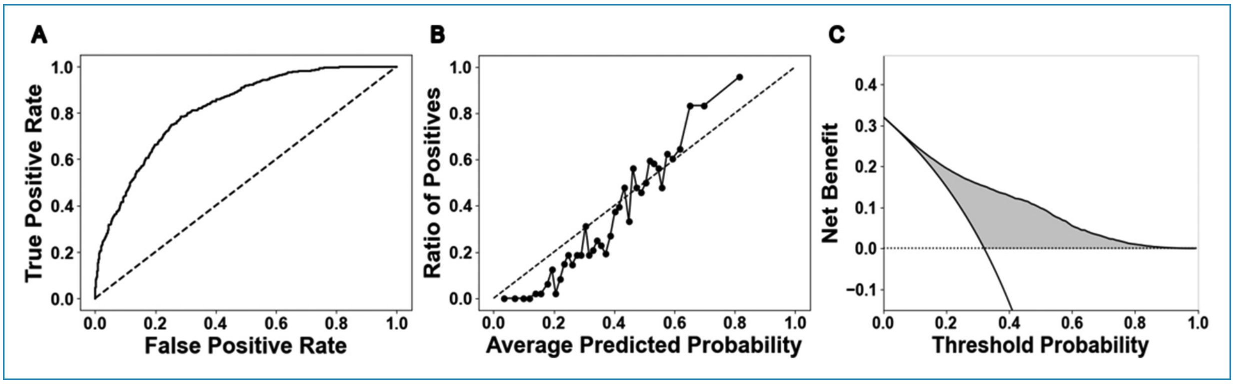

Figure 3. Performance evaluation of the proposed model for predicting successful weaning. (A) Receiver operating characteristic plot, (B) calibration curve, (C) decision curve analysis.

Discussion

Weaning from mechanical ventilation is a substantial issue in patients requiring mechanical ventilation. In this study, we used the MIMIC IV database and employed MTL with weaning-relevant tasks to establish an explainable model with high accuracy, with an AUC of 0.820. Moreover, the error analyses demonstrate that the model had higher accuracy after excluding those with low mean airway pressure and good enteral nutrition, indicating that the model might be too conservative among these patient groups and has high practical value in patients with high disease severities. We believe the established model, by using big data and

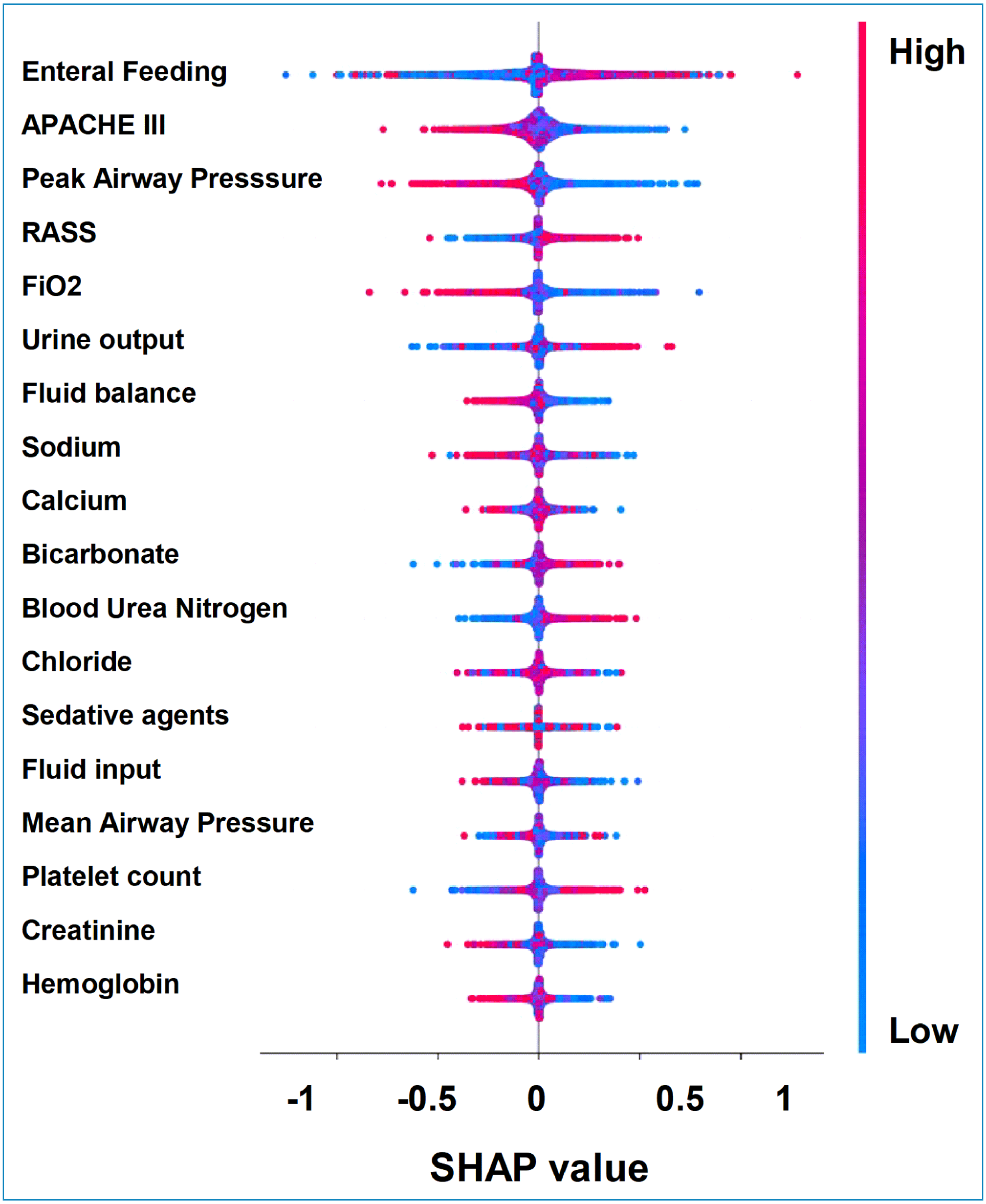

Figure 4. SHapley Additive exPlanation (SHAP) to illustrate the key features to predict successful weaning. Each point on the plot represented a Shapley value for one feature and subject.

employing MTL, explainable AI, and error analysis, should be a practical and responsible AI model to predict successful weaning in critically ill ventilated patients.

Notably, our proposed weaning prediction model can provide daily predictions. A number of studies have established weaning prediction systems, including ML models, based on the Rapid Shallow Breathing Index (RSBI).[6] However, RSBI is not a respiratory parameter measured daily, and RSBI is mainly measured when physicians attempt to assess the possibility of extubation or after a successful SBT.[7] Therefore, the aforementioned extubation prediction systems are mainly used to aid comprehensive judgment when physicians are already evaluating the probability of weaning. The primary difference in our study is that we used daily

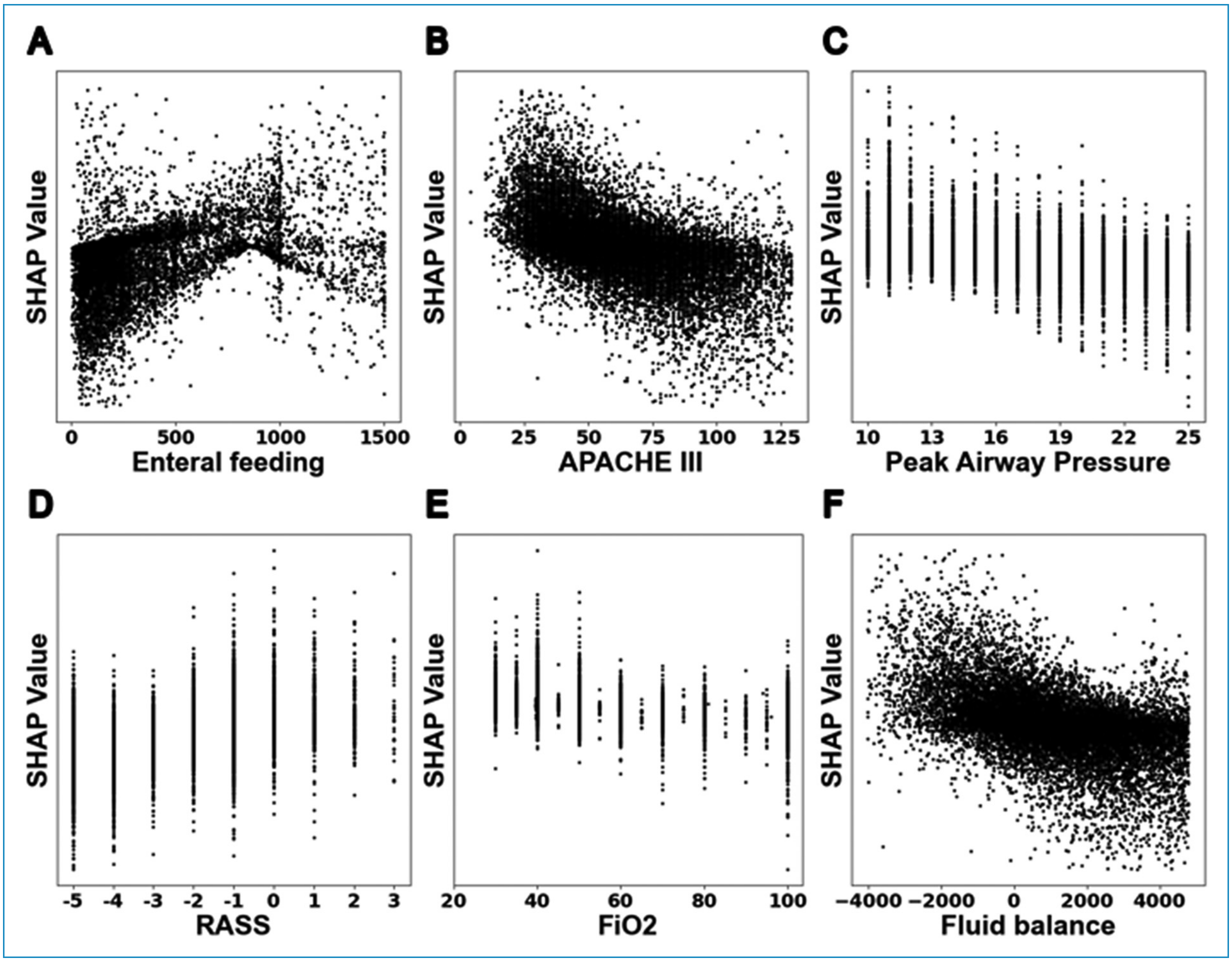

Figure 5. Partial dependence plots illustrating the effects of key features on the prediction model. (A) Enteral feeding, (B) APACHE III, (C) peak airway pressure, (D) RASS, (E) FiO_2, (F) fluid balance.
APACHE III: acute physiology and chronic health evaluation III; RASS: Richmond Agitation Sedation Scale; FiO_2: fraction of inspired oxygen.

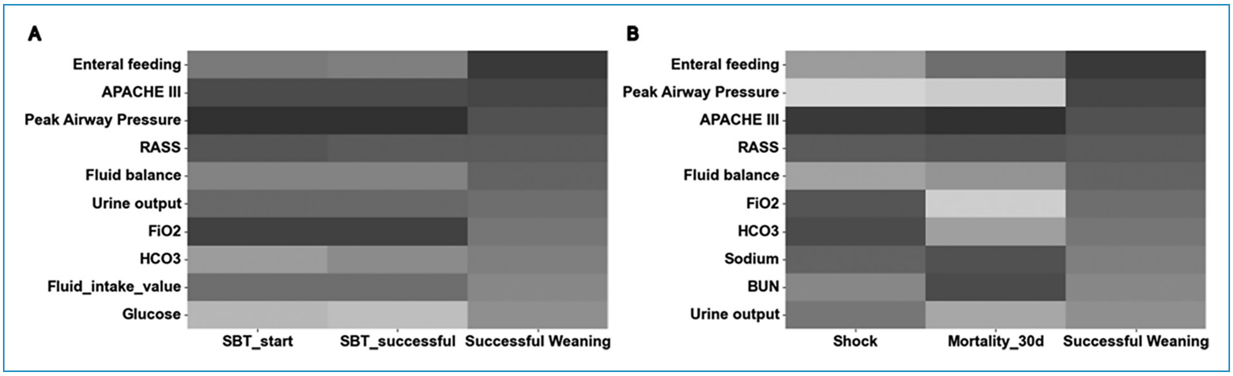

Figure 6. Heat map illustrating the relationships among features in related MTL (A) and non-related MTL (B). Darker regions on the greyscale indicate stronger associations, as determined by feature ranks in predicting distinct tasks.

parameters, such as peak airway pressure, FiO2, mean airway pressure, enteral feeding and fluid status, making the proposed model a daily prediction model. Additionally, we used a feature window of one day; therefore, the model can start to provide the real-time probability of weaning on the second day of ventilator use.

Table 3. Performance of original test population and sub-group analysis of multitask learning to predict successful weaning.

	AUROC	Accuracy	Precision	F1-score	Recall
Original population (n = 1917)	0.821	0.758	0.660	0.598	0.546
Sub-group (n = 1291)[a]	0.827	0.812	0.657	0.487	0.387

[a]Exclusion of 626 patients whose Pmean was lower than 9.5 cmH_2O and enteral feeding was higher than 589 mL/day.
AUROC: area under the receiver operating characteristics.

MTL, a subfield of AI, can exploit commonalities and differences across tasks, and the approach can lead to improved learning efficiency and prediction accuracy compared to training separate models for each task.[12] In critical care medicine, MTL is particularly relevant due to the complex, multifaceted nature of critical care and the need for a holistic patient assessment.[14,15] Critical care patients often present with multiple interrelated health issues, making it challenging to diagnose and treat them based on a single parameter or prediction model.[14,15] Harutyunyan et al., using the MIMIC-III database, benchmarked MTL against single-task learning across various clinical prediction tasks.[14] Harutyunyan et al. reported that MTL models outperformed single-task models in predicting in-hospital mortality (AUROC, 0.861 vs 0.855), decompensation (AUROC, 0.904 vs 0.892), and length of stay (Kappa, 0.450 vs 0.438).[14] Chi CY et al., using a nationwide database with 168,693 patients who had experienced in-hospital cardiac arrest (IHCA) in Taiwan, found that MTL outperformed single-task learning in predicting 30-day mortality (AUROC, 0.752 vs 0.658) and 30-day readmission (AUROC, 0.889 vs 0.872) among IHCA survivors.[15] Taken together, MTL allows for the integration of various data types and sources, such as demographic data, complex hemodynamic data, laboratory results, and ventilator parameters, to make more accurate and comprehensive predictions of patient outcomes in critical care.

The selection of tasks with high inter-task relationships is an essential issue in MTL.[20] We found that incorporating weaning-related tasks, rather than shock- or mortality-related tasks, improved the accuracy of the MTL model for predicting successful weaning. Additionally, MTL enhanced predictive performance in accordance with the clinical workflow sequence of SBT initiation, successful SBT, and successful weaning. We found that the tasks that precede and are weaning-related may improve the accuracy of subsequent task prediction with successful weaning, whereas MTL cannot improve the accuracy of the early tasks, such as the initiation of SBT. If the auxiliary tasks are not sufficiently related to the primary task, they may not contribute useful information and could even introduce noise into the model (Supplemental Table 1).[17,26] A number of studies have shown that the inclusion of too many loosely related or unrelated tasks in MTL can be detrimental to the performance of the primary task.[16,17] Similar to our findings on the application of MTL among unrelated tasks in critically ill ventilated patients, Roy et al. used MIMIC III to explore the use of shared-bottom MTL in predicting unrelated critical care outcomes, including acute kidney injury, continuous renal replacement therapy, vasoactive medication, use of mechanical ventilation, mortality, and remaining length of stay.[27] Similar to our finding, Roy *et al.* demonstrated that shared-bottom MTL tends to show a performance drop relative to single-task learning.[27] In contrast, related tasks can lead to a richer representation of the data, as they often contribute complementary information.[26,28] The combinations of correlated tasks may result in more robust and accurate predictive models. Moreover, learning related tasks together may contribute to the generalisation of the model through reducing the risk of overfitting by leveraging shared information among distinct tasks.[29] Collectively, the use of MTL in critical care offers the potential for personalised medicine by integrating data from various sources in related tasks in patients admitted to the ICU.

Notably, the key issue of the AI model in medical fields, particularly critical care medicine, relies on the potential to put the AI model into the clinical workflow, the so-called human-in-the-loop design.[30,31] Our approach in this study of linking the prediction of successful weaning with SBT aligns with the clinical workflow of weaning in critical care. The proposed model can provide the probability of successful weaning daily, and the prediction window of one day aligns with clinical workflow in the ICUs and allows physicians to consider whether to extubate on the next day and to arrange needed managements, such as measurement of weaning index and arrangement of family meeting, today. Collectively, the real-world application of the proposed AI model is to spontaneously integrate the data and present the probability of successful weaning to support the physicians in making decisions of weaning.

Responsible AI, especially in healthcare, prioritises transparency, fairness, and accountability in the design and application, ensuring the decision support system to be trustworthy and to enhance human judgment in clinical settings.[32] In the context of our study, transparency was addressed through explainable AI methods.[33] SHAP values, for instance, offer an understanding of feature contributions to model output, thus at least partly demystifying the black box. This is crucial in a clinical setting, where understanding the rationale behind AI predictions has an essential impact on the trust of the user in the AI systems.[34] Furthermore, we performed the error analyses and found that the model has higher accuracy after excluding those with low Pmean and good enteral nutrition. We explored the data between the 626 and 1291 patients and found that the 626 patients with high error rates were those with low severities of critical illness, with a low APACHE III, low ventilatory demand, including peak airway pressure, FiO2 and mean airway pressure, alert consciousness status, as well as high enteral feeding (Supplemental Table 2). Hence, the model tends to be relatively conservative and underestimates the probability of successful weaning in patients with low disease severities, and the patient can still be extubated successfully by the professional judgment of the intensivist. The error analysis points out the limitations and strengths of the model, such as the high performance among critically ill ventilated patients with high disease severities. Instead, the error analysis also provides the direction of improving the model in patient subgroups, such as those with low disease severities. Further advanced approaches, such as synthetic data generation, enhanced feature engineering and dynamic re-weighting method, and fairness-achieving algorithm among distinct populations, might be considered for individualised prediction models in this subgroup population.[35,36]

Indeed, error analysis is particularly vital in healthcare AI due to the potential for harm.[37] Incorrect predictions can lead to inappropriate clinical decisions, such as prematurely weaning a patient from ventilation, which can have severe or even fatal consequences, and rigorously analysing errors, as we have shown in this study, can identify and address the sources of the error.[38] Through the error analysis, the user, intensivist or respiratory therapist can know the potential limitation of the model in the practical application of the model. Furthermore, the parameter might be adjusted based on the continuous error analyses after the practical application. In summary, we employed SHAP and PDP to interpret the model and conducted error analysis to identify its limitations, ensuring the development of transparent decision support systems in critical care.[39]

In brief, we developed a practical and easy-for-practical application multitask ML model to predict successful weaning with at least five advantages, including using only 18 common features, which are available in critically ill ventilated patients, a moving feature window to provide continuous prediction daily, a prediction window of one day, which allows physicians to consider extubation or not based on both the clinical information as well as the suggestion of a model, the visualised interpretation of the model to gain the trust of the user, and error analysis to demonstrate the strength and limitation in sub-group patients.

There are limitations that merit discussion. First, the single hospital retrospective nature of this study and further studies are warranted for validation. Second, we used a moving feature window with a daily data approach, and the predictive performance might be compromised. Although using cumulative data from ICU admission may improve the predictive performance due to comprehensive data, it is difficult to land in a real-world setting.[40] In contrast, the moving feature window with the daily data is practically to be landed to provide the daily prediction. Third, our model interpretation relied on post-hoc analyses and cannot fully resolve the black-box nature inherent in the model.[41]

Conclusion

Weaning is an essential issue in critically ill ventilated patients. The present study demonstrates that MTL can effectively enhance predictive accuracy in critical care settings with complex, multifaceted and interrelated data. We established an MTL model with high accuracy to predict successful weaning through integrating weaning-relevant tasks. Moreover, we conducted the visualised the interpretability of the model and error analysis, and these approaches enable the user to realise the model, to know limitations, and to trust the model. More prospective studies are warranted to validate our findings and to identify issues on the practical application of the weaning prediction model.

Availability of data and material: All of the data and materials are provided in the manuscript and the supplemental data. The code has been put in public Github, and is available via https://github.com/GitEricLin/MTL_Weaning.

Contributorship: Study concept and design: MYL and WCC. Acquisition of data: MYL, HYC, and WCC. Analysis and interpretation of data: MYL, HYC, and WCC. Drafting the manuscript: MYL and WCC. All authors agreed and are responsible for the content of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This study was approved by the institutional review board of Taichung Veterans General Hospital (IRB number: CE23182A). The informed consent was waived because all of the data were di-identified data. All the methods were performed in accordance with the relevant guidelines and regulations.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the

Veterans General Hospitals and University System of the Taiwan Joint Research Program (VGHUST113-G2-1-2) and the National Science and Technology Council Taiwan (NSTC 112-2314-B-075A-001-MY2).

Guarantor: WCC.

ORCID iDs: Ming-Yen Lin https://orcid.org/0000-0003-3180-3132
Wen-Cheng Chao https://orcid.org/0000-0001-9631-8934

Supplemental material: Supplemental material for this article is available online.

References

1. Walter K. Mechanical ventilation. *JAMA* 2021; 326: 1452–1452.
2. Wunsch H, Wagner J, Herlim M, et al. ICU occupancy and mechanical ventilator use in the United States. *Crit Care Med* 2013; 41: 2712–2719. 2013/08/22.
3. Boles JM, Bion J, Connors A, et al. Weaning from mechanical ventilation. *Eur Respir J* 2007; 29: 1033–1056. 2007/05/02.
4. Beduneau G, Pham T, Schortgen F, et al. Epidemiology of weaning outcome according to a new definition. The WIND study. *Am J Respir Crit Care Med* 2017; 195: 772–783. 2016/09/15.
5. Akella P, Voigt LP and Chawla S. To wean or not to wean: a practical patient focused guide to ventilator weaning. *J Intensive Care Med* 2022; 37: 1417–1425. 2022/07/12.
6. Baptistella AR, Sarmento FJ, da Silva KR, et al. Predictive factors of weaning from mechanical ventilation and extubation outcome: a systematic review. *J Crit Care* 2018; 48: 56–62. 2018/09/02.
7. Baptistella AR, Mantelli LM, Matte L, et al. Prediction of extubation outcome in mechanically ventilated patients: development and validation of the Extubation Predictive Score (ExPreS). *PLoS One* 2021; 16: e0248868. 2021/03/19.
8. Menguy J, De Longeaux K, Bodenes L, et al. Defining predictors for successful mechanical ventilation weaning, using a data-mining process and artificial intelligence. *Sci Rep* 2023; 13: 20483. 2023/11/23.
9. Igarashi Y, Ogawa K, Nishimura K, et al. Machine learning for predicting successful extubation in patients receiving mechanical ventilation. *Front Med (Lausanne)* 2022; 9: 961252. 2022/08/30.
10. Pai K-C, Su S-A, Chan M-C, et al. Explainable machine learning approach to predict extubation in critically ill ventilated patients: a retrospective study in central Taiwan. *BMC Anesthesiol* 2022; 22: 351.
11. Lin MY, Li CC, Lin PH, et al. Explainable machine learning to predict successful weaning among patients requiring prolonged mechanical ventilation: a retrospective cohort study in central Taiwan. *Front Med (Lausanne)* 2021; 8: 663739. 2021/05/11.
12. Crawshaw M. Multi-task learning with deep neural networks: a survey. arXiv:2009.09796. DOI: 10.48550/arXiv.2009.09796.
13. Wallingford M, Li H, Achille A, et al. Task adaptive parameter sharing for multi-task learning. In: *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 18-24 June 2022, 2022, pp.7551–7560.
14. Harutyunyan H, Khachatrian H, Kale DC, et al. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019; 6: 96. 2019/06/19.
15. Chi CY, Ao S, Winkler A, et al. Predicting the mortality and readmission of in-hospital cardiac arrest patients with electronic health records: a machine learning approach. *J Med Internet Res* 2021; 23: e27798. 2021/09/14.
16. Zhang Y and Yang Q. A survey on multi-task learning. arXiv:1707.08114. DOI: 10.48550/arXiv.1707.08114.
17. Mahony N O', Campbell S, Krpalkova L, et al. Regressing relative fine-grained change for sub-groups in unreliable heterogeneous data through deep multi-task metric learning. arXiv:2208.05800. DOI: 10.48550/arXiv.2208.05800.
18. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023; 10: 1.
19. Chan MC, Pai KC, Su SA, et al. Explainable machine learning to predict long-term mortality in critically ill ventilated patients: a retrospective study in central Taiwan. *BMC Med Inform Decis Mak* 2022; 22: 75. 2022/03/27.
20. Guo M, Haque A, Huang D-A, et al. Dynamic task prioritisation for multitask learning. In: *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp.270–287.
21. Perren A and Brochard L. The importance of timing for the spontaneous breathing trial. *Ann Transl Med* 2019; 7: S210. 2019/10/28.
22. Tonnelier JM, Prat G, Le Gal G, et al. Impact of a nurses' protocol-directed weaning procedure on outcomes in patients undergoing mechanical ventilation for longer than 48 hours: a prospective cohort study with a matched historical control group. *Crit Care* 2005; 9: R83–R89. 2005/03/19.
23. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017; 318: 1377–1384. 2017/10/20.
24. Vickers AJ and Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26: 565–574. 2006/11/14.
25. DeLong ER, DeLong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44: 837–845. 1988/09/01.
26. Hur K, Oh J, Kim J, et al. GenHPF: general healthcare predictive framework for multi-task multi-source learning. *IEEE J Biomed Health Inform* 2023: 1–12. DOI: 10.1109/JBHI.2023.3327951.
27. Roy S, Mincu D, Loreaux E, et al. Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. *J Am Med Inform Assoc* 2021; 28: 1936–1946. 2021/06/22.
28. Feng R, Cao Y, Liu X, et al. Chronet: a multi-task learning based approach for prediction of multiple chronic diseases. *Multimed Tools Appl* 2022; 81: 41511–41525.
29. Yang C, Westover MB and Sun J. ManyDG: Many-domain generalization for healthcare applications. arXiv:2301.08834. DOI: 10.48550/arXiv.2301.08834.

30. *Good machine learning practice for medical device development: guiding principles*. The U.S. Food and Drug Administration, 2021.
31. Loh HW, Ooi CP, Seoni S, et al. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). *Comput Methods Programs Biomed* 2022; 226: 107161. OI:
32. Kundu S. AI In medicine must be explainable. *Nat Med* 2021; 27: 1328. 2021/07/31.
33. Hatherley J, Sparrow R and Howard M. The virtues of interpretable medical AI. *Camb Q Healthc Ethics* 2023: 1–10. 2023/01/11. DOI: 10.1017/S0963180122000664.
34. Dragoni M, Donadello I and Eccher C. Explainable AI meets persuasiveness: translating reasoning results into behavioral change advice. *Artif Intell Med* 2020; 105: 101840. 2020/06/09.
35. Li C, Ding S, Zou N, et al. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *J Biomed Inform* 2023; 143: 104399.
36. Li C, Lai D, Jiang X, et al. FERI: a multitask-based fairness achieving algorithm with applications to fair organ transplantation. *AMIA Jt Summits Transl Sci Proc* 2024: 593.
37. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digit Med* 2022; 5: 2.
38. Cho MK. Rising to the challenge of bias in health care AI. *Nat Med* 2021; 27: 2079–2081.
39. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI Tool for quantitative evaluation of AI studies for clinical decision support. *JAMA Netw Open* 2023; 6: e2335377. 2023/09/25.
40. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020; 46: 383–400. 2020/01/23.
41. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; 1: 206–215. 2019/05/01.