Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning vs. Clinical Judgement for Extubation Failure: A Rigorous Re-evaluation

David René Rodríguez Lima
Hospital Universitario Mayor, Mederi
Cristhian Rubio Ramos
Hospital Universitario Mayor, Mederi
Esther María Campo-Bautista
La Cardio
Edith Elianna Rodríguez
edith.rodriguez@urosario.edu.co

Hospital Universitario Mayor, Mederi

Research Article

Keywords: Extubation failure, ABCDMed score, Machine learning models, Clinical judgment, Artificial intelligence

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11244575/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

In 2022 we derived and validated the ABCDMed score to predict extubation failure. Machine learning models (MLM) for the same outcome now report areas under the receiver operating characteristic curve (AUC) of up to 0.98. We asked whether flexible algorithms, given the same patients may improve on a model built from clinical judgement.

Methods.

A secondary analysis of a original prospective cohort of 800 consecutive adults extubated on a ICU in Bogotá, Colombia, at 2,600 m above sea level was performed. Extubation failure prediction was compared with two MLM, random forest, gradient boosting, and with nine further strategies, under 10 times repeated stratified 5-fold cross validation in which imputation, scaling, resampling, variable selection and hyperparameter search were performed inside training folds only, and every arm received the same candidate variables. Artificial intelligence was used in two further roles: to appraise 14 published machine learning studies of this outcome against the same reporting items, and to produce an adversarial peer review of this manuscript, every claim of which was verified by re-analysis.

Results.

Extubation failure occurred in 71 patients (8.9%). The published points score reproduced exactly, separating four risk strata with observed failure rates of 4.4%, 7.7%, 11.6% and 23.0%, but the score distributions of patients who failed and succeeded overlap almost completely. Given the same 12 candidate variables, random forest reached 0.630 and gradient boosting 0.627 against 0.655 for ordinary logistic regression, a spread of 0.028 in a design that cannot detect less than 0.068. Given all 42 variables, tuning, assembling and class rebalancing, no strategy exceeded the original score; SMOTE oversampling left discrimination unchanged while inflating predicted risk by 75% and driving net benefit below the treat-no strategy. Among the 14 published studies, 12 displayed an explainability figure, 7 reported calibration, 4 had any external validation, 3 reported net benefit, 3 shared code and none used repeated or nested cross validation.

Conclusions.

Artificial intelligence excels at extracting patterns from existing datasets, but it cannot generate information that is not there. In this cohort, neither advanced MLM nor extensive algorithmic tuning outperformed a parsimonious model based on clinical judgment.

Introduction

Endotracheal intubation and mechanical ventilation are the most frequently used organ support in critically ill patients, and its withdrawal is one of the few decisions in the intensive care unit (ICU) taken several times a day on essentially physiological grounds. Between 6% and 20% of patients judged ready for tube removal will require reintubation within 48 hours [1–3], and it is directly associated with a 50% increase in mortality, nosocomial pneumonia and a prolonged stay [4, 5]. Several scores have been developed to predict this extubation failure. The rapid shallow breathing index (RSBI) discriminates poorly on its own [6, 7], the cuff leak test addresses only upper airway obstruction, and cough strength and secretion burden are judged subjectively [8]. Also several ultrasound measurements such as the diaphragmatic excursion and thickening fraction has been elucidates as strategies for known the reediness from a patient to withdrawal of respiratory support [9]

In 2022 we derived and validated, in the cohort re-analyzed here, a five variable score for this purpose, the ABCDMed score, which reached an area under the receiver operating characteristic curve (AUC) of 0.687 and stratified patients into four risk groups [10]. Nevertheless, is not good enough. Recently, machine learning models (MLM) has been proposed as the missing ingredient. Published models for extubation or weaning outcome report AUCs of 0.79 to 0.98 [11–16], values that would transform bedside practice if they were reproducible. Those reports share features that deserve scrutiny: many validate on a single random partition, most compare an optimized MLM against an unoptimized or absent regression comparator, several include variables measured at or after the moment of extubation, and few report calibration or net benefit [17, 18]. Systematic reviews across clinical medicine have repeatedly found no performance benefit of MLM over logistic regression once those design differences are removed, and that most MLM prediction studies are at high risk of bias [19–22].

There is also an asymmetry that the usual head-to-head comparison hides. When a published clinical score is the comparator, its variables were chosen by investigators who had seen the data, while the MLM arm must discover its variables inside each resampling fold. The comparison then measures two things at once: the algorithm, and the provenance of the variable set. Since the contribution of clinical

judgement to a score is precisely the decision about what to measure and what to keep, that confounding is worth estimating rather than merely removing.

We propose a three-part assessment of this problem: We first present our own score and re-examine it, reporting the weaknesses that a rigorous re-analysis expose. We then ask whether two MLM that dominate this literature, random forest and gradient boosting, improve on it when both are held to the same protocol, and we extend the comparison to nine further strategies and to two further uses of artificial intelligence (AI): An assisted appraisal of the published literature, and an adversarial machine peer review of this manuscript. We end with the conclusion that the data force on us, which is not about algorithms at all. Reporting follows TRIPOD + AI [23] and the completed checklist is provided as Supplementary Table S5.

Methods

Study design, setting and participants

This is a secondary analysis of the prospective observational cohort used to derive and validate the ABCDMed score [10]. The study was conducted in the adult ICU of Hospital Universitario Mayor Méderi in Bogotá, Colombia, situated 2,600 m above sea level, with 71 adult ICU beds.

All consecutive patients older than 18 years who received invasive mechanical ventilation between June 2017 and April 2019 and were extubated on medical advice on a planned basis were included. Patients intubated outside the institution were excluded because the duration of ventilation could not be established reliably, as were unplanned and accidental extubations. The institutional weaning protocol required a checklist covering oxygenation, ventilatory mechanics, acid base status, hemodynamic stability and airway management, together with a spontaneous breathing trial of continuous positive airway pressure plus pressure support, with positive end expiratory pressure of 8 cmH2O or less and support of 8 cmH2O or less, sustained for at least 30 minutes.

The study was approved by the Institutional Human Research Ethics Committee of Universidad del Rosario (DVO005 929-CV904) and classified as minimum risk under Colombian legislation; general informed consent covered the use of medical record information. The present analysis used a de-identified extract and required no additional consent.

Outcome

The outcome was extubation failure, defined a priori as the need for reintubation within 48 hours of planned extubation. Reintubation is objective in its timing but not in its indication, and the clinicians who made that decision had seen the same variables the models use, including whether prophylactic noninvasive ventilation had been planned. This is unavoidable in any observational study of this outcome and is addressed in a sensitivity analysis.

Predictors

All predictors were recorded immediately before extubation and were therefore available at the moment of the decision. No variable measured at or after the time of failure was used at any stage.

Three predictor pools were specified in advance. The first is the five variable ABCDMed set: acid base status expressed as the exponential of arterial pH, the RSBI, the presence of effective cough, the predicted probability of death, and medical rather than surgical patient status. The second is the 12-candidate variable pool considered during the original derivation. The third is a 42 variable bedside pool comprising every pre-extubation variable in the database (Supplementary Table S1).

Two features of the predicted probability of death require disclosure because they bear on the interpretation of every model. It is derived from APACHE II in all patients except those undergoing cardiac surgery, in whom EuroSCORE II is used, so it is not a single instrument on a common scale and it partly encodes the admission category itself. And planned noninvasive ventilation after extubation, which appears in the larger pools, records the treating clinician's own risk judgement and is also the action a risk model would trigger; it was removed in a prespecified sensitivity analysis.

Missing data

Among variables used as candidate predictors, missingness exceeded 10% only for mixed venous oxygen saturation (34.8%), ideal body weight (29.8%), oxygenation index (16.8%) and cuff leak percentage (14.5%). The pools are therefore not equally exposed: 99.6% of patients

have complete values for the five ABCDMed variables, 83.3% for the 12 variable pool and 38.5% for the 42 variable pool. Median imputation was applied inside each training fold and applied unchanged to the held-out fold. Because differential exposure could favor the smaller pools, we repeated the comparison with iterative imputation and, for gradient boosting, the algorithm's native handling of missing values.

Prediction strategies compared

The two MLM placed in the foreground are random forest and gradient boosting, chosen because they are the methods most frequently reported in the published literature on this outcome and the ones a clinician is most likely to be asked to judge. Nine further strategies are reported alongside them: elastic net, a second gradient boosting implementation, a support vector machine, k nearest neighbours, a multilayer perceptron, a Super Learner stacked ensemble, gradient boosting preceded by SMOTE oversampling of the minority class [24, 25], and logistic regression using 12 and 42 variables.

Every arm was run under a matched protocol. Given the 12 variable pool and, separately, the 42 variable pool, each arm had to discover its own variables inside each training fold: purposeful selection emulating the strategy used to derive the published score (univariable screen at $p < 0.25$, backward elimination at $p > 0.05$, then refit), L1 selection followed by an unpenalized refit, penalization, or the implicit selection performed by a tree ensemble. The published five variable specification, whose variables are fixed, is reported alongside as the conventional comparator and labelled as such.

Three deliberately minimal benchmarks establish how much performance is available without a model: the RSBI alone, the predicted probability of death alone, and case-mix group alone. Hyperparameter grids are in Supplementary Table S2. The default logistic regression of the software used applies a ridge penalty; arms described as unpenalized were fitted with the penalty switched off.

Validation strategy

Every arm was evaluated by stratified 5 fold cross validation repeated 10 times, giving outer test folds containing a median of 14 events. Within each training fold, imputation, standardization, resampling, variable selection and hyperparameter selection were carried out in a nested inner 3 fold randomized search over 8 configurations, optimizing the logarithmic loss. Predictions for each held out fold were pooled, so each repeat produced 800 predictions from models that had never seen the patient being predicted.

Two summaries are reported and must not be confused. Variation across the 10 repeats measures only the instability introduced by partitioning. Uncertainty about performance is reported separately as a percentile bootstrap (2,000 resamples) of the repeat-averaged predictions; these intervals are roughly three times wider and are the ones that should be used to judge whether two strategies differ. Paired differences use the same bootstrap for point estimate and interval. The benchmark was repeated with an independent random seed.

Precision of the design

Because the paper makes a negative claim, we quantified what the design can detect. From the bootstrap standard error of the paired difference in AUC we computed the minimum difference detectable with 80% power at a two-sided alpha of 0.05, and the sample size needed to bound that difference within margins of 0.02, 0.03 and 0.05.

Performance measures and prespecified experiments

Discrimination was measured by the AUC and by the area under the precision recall curve, whose no-skill value is 0.089. Calibration was assessed by the calibration slope, the calibration intercept with the linear predictor as an offset, the expected to observed ratio, and the integrated calibration index from a locally weighted flexible calibration curve [26], reported across smoothing bandwidths. Overall accuracy used the Brier and scaled Brier scores. Clinical utility was assessed by decision curve analysis [27], reported as net benefit, as net benefit standardized by the maximum attainable value, and as paired bootstrap differences. Because no action has been validated against a model-defined threshold in this population, interpretation is restricted to thresholds of 5% to 15% and the implied harm ratio is stated explicitly.

Six prespecified experiments supported the critical reading: apparent performance with the same tuned configurations; 300 repetitions of a single 70/30 split; a learning curve over subsamples of 200 to 800 patients; stability of variable importance across 200 bootstrap refits; 50 repetitions of the whole cross validation with the outcome permuted; and a decomposition of discrimination by case mix, with stratum-specific AUCs and a within-stratum AUC weighted by discordant pairs [28].

Assisted appraisal of the published literature

We assembled the published studies developing a supervised machine learning model for extubation failure, reintubation after planned extubation, or weaning outcome in adults, from 2015 to 2026. The search strategy, screening decisions and studies identified but not appraisable are reported in Supplementary Table S6.

From each study we extracted sample size, events, candidate predictors, algorithms, the presence of a regression comparator, validation design, best reported AUC, and six elements that determine whether a reader can judge clinical usefulness: calibration, decision curve analysis or another utility measure, external or temporal validation, repeated or nested resampling, availability of code, and a properly specified regression comparator. We separately recorded the presence of an explain ability figure as a marker of presentation. The same sheet was applied to the present study. Extraction was assisted by a large language model, performed by a single appraiser from the accessible text, with fields that could not be located recorded as missing rather than inferred and an extraction confidence flag for each study. This is a focused appraisal, not a registered systematic review.

Adversarial artificial intelligence peer review

Artificial intelligence was used in a second role. We commissioned an adversarial peer review of this manuscript from a large language model (Claude, Anthropic), instructed to act as a hostile statistical reviewer for a high-impact critical care journal, given the full manuscript, the analysis code and the dataset, and permitted to execute code. Its output was treated as a reviewer report: every factual and numerical claim was independently re-derived by the authors before any change was made, and claims that did not reproduce were recorded as such. The review and our point-by-point responses are provided as Supplementary Material, and the analyses added in response are identified in the Results.

Sample size considerations

No new sample size was calculated, since the cohort was fixed. The sample size that would be required to develop each model reliably was computed with the criteria of Riley and colleagues [29], using the Cox and Snell R squared of the reference model observed here, an outcome proportion of 8.9%, an expected shrinkage of 10% and maximum optimism in the apparent R squared of 0.05. Because the R squared used is an apparent value, these requirements are, if anything, underestimates.

Software and reproducibility

Analyses were performed in Python 3.11 with scikit-learn 1.8, XGBoost 3.2, LightGBM 4.7, imbalanced-learn 0.14 and statsmodels 0.15. The seed was fixed at the level of the repeat, and the pipeline is reproducible from the deposited script and de-identified dataset.

Results

Cohort and outcome

Eight hundred patients were analyzed. Median age was 67 years IQR [58–75], 470 (58.8%) were men, 300 (37.5%) were classified as medical and 500 (62.5%) as surgical, of whom 304 had undergone cardiac surgery. Median duration of mechanical ventilation was 3 IQR [1 to 5]. Baseline characteristics are in Table 1 and in full in Supplementary Table S3.

Table 1
Baseline characteristics and pre-extubation physiology by extubation outcome (n = 800).

Variable	All patients (n = 800)	Extubation failure (n = 71)	Successful extubation (n = 729)	p value
Age, years	67.0 (57.8 to 75.0)	71.0 (58.5 to 77.5)	67.0 (57.0 to 75.0)	0.211
Male sex	470 (58.8%)	43 (60.6%)	427 (58.6%)	0.842
Medical patient	300 (37.5%)	41 (57.7%)	259 (35.5%)	< 0.001
Predicted probability of death, %	16.5 (5.9 to 32.2)	26.2 (16.5 to 42.4)	16.5 (5.1 to 29.7)	< 0.001
APACHE II score (n = 507)	17.0 (13.0 to 21.0)	18.0 (15.0 to 23.0)	17.0 (13.0 to 21.0)	0.082
EuroSCORE II (n = 293)	5.0 (3.0 to 7.0)	6.5 (5.2 to 8.5)	5.0 (3.0 to 7.0)	0.062
Duration of ventilation, days	3.0 (1.0 to 5.0)	4.0 (2.5 to 7.0)	2.0 (1.0 to 5.0)	< 0.001
Rapid shallow breathing index (n = 797)	36.0 (29.0 to 45.0)	40.0 (31.2 to 49.8)	36.0 (28.0 to 45.0)	0.034
Respiratory rate, breaths/min (n = 797)	18.0 (15.0 to 20.0)	18.0 (16.0 to 22.0)	18.0 (15.0 to 20.0)	0.036
Tidal volume, mL (n = 795)	475.0 (419.5 to 553.0)	457.5 (400.0 to 516.0)	479.0 (420.0 to 556.0)	0.046
PEEP, cmH2O	6.0 (6.0 to 8.0)	6.0 (6.0 to 8.0)	6.0 (6.0 to 8.0)	0.294
PaO2, mmHg	81.0 (70.0 to 94.0)	74.0 (65.0 to 87.0)	81.0 (71.0 to 94.0)	0.006
PaO2/FiO2 ratio	237.0 (203.0 to 282.0)	232.0 (199.0 to 261.0)	237.0 (205.0 to 282.0)	0.182
SaO2, %	96.0 (93.0 to 97.0)	94.0 (92.0 to 96.0)	96.0 (94.0 to 97.0)	0.011
Oxygenation index (n = 666)	4.0 (3.0 to 5.0)	4.2 (3.5 to 5.1)	4.0 (2.9 to 4.9)	0.158
Arterial pH	7.4 (7.3 to 7.5)	7.4 (7.4 to 7.5)	7.4 (7.3 to 7.4)	< 0.001
Bicarbonate, mmol/L	22.0 (19.0 to 26.0)	23.0 (20.0 to 27.0)	21.0 (19.0 to 26.0)	0.019
PaCO2, mmHg	36.0 (32.0 to 40.0)	35.0 (32.5 to 39.5)	36.0 (32.0 to 40.0)	0.719
Lactate, mmol/L	1.9 (1.4 to 2.4)	1.9 (1.5 to 2.4)	1.9 (1.4 to 2.4)	0.761
Mean arterial pressure, mmHg	90.0 (80.0 to 101.0)	92.0 (84.0 to 102.5)	89.0 (79.0 to 101.0)	0.088
Heart rate, beats/min	83.0 (75.0 to 96.0)	85.0 (77.0 to 96.0)	83.0 (74.0 to 96.0)	0.174
Vasopressor at extubation	243 (30.4%)	17 (23.9%)	226 (31.0%)	0.272
Inotrope at extubation	103 (12.9%)	8 (11.3%)	95 (13.0%)	0.812
Cuff leak, % (n = 684)	64.0 (50.0 to 76.0)	63.0 (50.5 to 78.0)	64.0 (50.0 to 76.0)	0.848
Abundant secretions	50 (6.2%)	8 (11.3%)	42 (5.8%)	0.116
Effective cough	788 (98.5%)	67 (94.4%)	721 (98.9%)	0.013
Delirium (CAM-ICU)	39 (4.9%)	5 (7.0%)	34 (4.7%)	0.549
Planned noninvasive ventilation after extubation	99 (12.4%)	18 (25.4%)	81 (11.1%)	0.001
ICU stay, days	6.0 (4.0 to 11.0)	16.0 (9.5 to 22.5)	6.0 (4.0 to 10.0)	< 0.001
Hospital stay, days	21.5 (14.0 to 32.0)	32.0 (20.0 to 49.0)	21.0 (14.0 to 31.0)	< 0.001
ICU mortality	90 (11.2%)	27 (38.0%)	63 (8.6%)	< 0.001
Mortality at 30 days	145 (18.1%)	35 (49.3%)	110 (15.1%)	< 0.001

Continuous variables are medians (interquartile range) and categorical variables are counts (%). p values are from the Mann Whitney U test or the chi squared test. APACHE II and EuroSCORE II are complementary by design and enter the models only through the derived probability of death. ICU stay, hospital stay and the two mortality rows are outcomes, not candidate predictors. PEEP, positive end expiratory pressure.

Extubation failure occurred in 71 patients (8.9%); 61 (7.6%) were reintubated within 24 hours and 10 (1.3%) between 24 and 48 hours. Failure was associated with a longer ICU stay (median 16 versus 6 days, $p < 0.001$), higher ICU mortality (38.0% versus 8.6%, $p < 0.001$),) and higher 30-day mortality (49.3% versus 15.1%, $p < 0.001$).

Part 1. The published score, presented and re-examined

The score as published

The ABCDMed score assigns points for five bedside variables: acid base status expressed through arterial pH, the RSBI, the presence of effective cough, the predicted probability of death, and medical patient status (Table 2) [10]. The underlying equation estimates the probability of extubation failure as 1 divided by [1 + exp(-(-6.6 + 0.0175 x RSBI + 0.0022 x exp(pH) − 1.2733 if cough is effective + 0.0110 x probability of death + 0.3873 if the patient is a medical patient))].

Table 2
The ABCDMed score as published, and its risk strata reproduced in the analysis dataset.

Component	Category	Points
A. Acid base status (arterial pH)	7.250 or less	0 *
	7.251 to 7.300	1.7
	7.301 to 7.350	3.5
	7.351 to 7.400	5.4
	7.401 to 7.450	7.4
	7.451 to 7.500	9.5
	7.501 to 7.550	11.7
	7.551 to 7.600	14.0
B. Rapid shallow breathing index	19 or less	0
	20 to 29	1.5
	30 to 39	3.0
	40 to 49	4.5
	50 to 59	6.25
	60 to 69	8.0
	70 to 79	9.5
	80 to 89	11.0
	90 to 99	12.75
	100 or more	14.25
C. Effective cough	Present	-11.5
D. Predicted probability of death, %	10 or less	0
	10.1 to 20	1.0
	20.1 to 30	2.25
	30.1 to 40	3.25
	40.1 to 50	4.5
	50.1 to 60	5.5
	60.1 to 70	6.5
	70.1 to 80	7.75
	over 80	8.75
Med. Medical patient	Yes	3.5
Risk group (total points)	**Patients in this study, n (%)**	**Observed failure, this study / as published**
Low, under 5% (0 or less)	342 (42.9%)	4.4% / 4.1%
Moderate, 5 to 10% (0.1 to 4.5)	196 (24.6%)	7.7% / 8.1%
High, 10 to 15% (4.6 to 9)	172 (21.6%)	11.6% / 11.5%
Very high, over 15% (over 9)	87 (10.9%)	23.0% / 22.7%

Points are summed across the five components; the total ranged from − 10.0 to 23.5 in this cohort. * pH of 7.250 or less scores zero, with the instruction not to continue weaning until the cause of the acidosis has been evaluated and corrected. The risk strata reproduce the published derivation closely: 342, 196, 172 and 87 patients here against 337, 198, 174 and 88 as published. The points score achieves an AUC of 0.677 and the underlying Eq. 0.685.

Applied to the analysis dataset, the published equation gave an AUC of 0.685 in the 797 patients with complete values, reproducing the originally reported 0.687, and the points score gave 0.677. The four published risk strata reproduced almost exactly (Fig. 1A, Table 2): observed failure rates of 4.4%, 7.7%, 11.6% and 23.0% against the published 4.1%, 8.1%, 11.5% and 22.7%, in groups of 342, 196, 172 and 87 patients against the published 337, 198, 174 and 88. Refitting the same specification gave an apparent AUC of 0.691 and coefficients that, after applying the shrinkage factor of 0.8478 used in the original derivation, agreed with the published values to within 6% (Supplementary Table S4).

The score therefore separates a group with a 4% risk from a group with a 23% risk, a fivefold gradient, using five things a clinician already has.

First weakness: the separation behind the risk gradient is small

The first weakness is the amount of separation behind the gradient. Under repeated cross validation the specification achieves an AUC of 0.663, and the distributions of the score in patients who failed and in patients who succeeded overlap almost completely (Fig. 1B): medians of 5.0 and 0.7 points on a scale running from − 10 to 23.5. In decision curve terms, at a 10% risk threshold the score delivers a net benefit of 0.018, which is 20% of the maximum attainable value of 0.089; expressed in patients, about 1.8 true positives per 100 extubations after subtracting the harm of the accompanying false positives. The Cox and Snell R squared is 0.041, so roughly 4% of the outcome variation is explained.

Second weakness: the five variables were chosen in this cohort

The second weakness is one we did not see when we published it. The five variables were selected using these same 800 patients, and that selection crosses every fold boundary in any subsequent validation. When the selection procedure itself is run inside the training folds, on the same 12 candidate variables, performance falls to 0.596 for an in-fold emulation of the original purposeful selection strategy and 0.651 for L1 selection (Table 3, Fig. 2). The advantage conferred by having chosen those five variables in this cohort is therefore between 0.012 and 0.067 of AUC. Part of that is genuine clinical judgement and part is in-cohort optimism, and internal data cannot separate them; only external validation can. The most relevant point for this paper is that the quantity is large, larger than any difference we will report between algorithms.

Table 3
Matched protocol: cross validated performance when every arm discovers its variables inside the training folds.

Strategy	Candidate pool	How variables were chosen	AUC (95% CI)	Calibration slope	Scaled Brier
Fixed ABCDMed variables (pre-selected in this cohort)	5 fixed	Chosen by clinicians, using this cohort	0.663 (0.591 to 0.732)	0.83	+ 0.036
Predicted probability of death alone	1 fixed	Single severity variable	0.656 (0.594 to 0.712)	0.93	+ 0.011
Logistic regression, all 12 (no selection)	12	None, all retained	0.655 (0.584 to 0.729)	0.70	+ 0.031
L1 selection then refit in-fold	12	L1 penalty, in-fold	0.651 (0.587 to 0.730)	0.74	+ 0.025
Purposeful selection in-fold	12	Purposeful selection, in-fold	0.596 (0.545 to 0.690)	0.64	+ 0.005
Elastic net	12	Penalisation, in-fold	0.644 (0.584 to 0.723)	0.97	+ 0.030
Random forest	12	Implicit, in-fold	0.630 (0.565 to 0.700)	1.08	+ 0.021
XGBoost	12	Implicit, in-fold	0.627 (0.569 to 0.694)	0.94	+ 0.017
Logistic regression, all 42, ridge at C = 1	42	None, all retained	0.640 (0.579 to 0.716)	0.46	-0.023
Logistic regression, all 42, truly unpenalised	42	None, all retained	0.629 (0.566 to 0.708)	0.36	-0.042
L1 selection then refit in-fold	42	L1 penalty, in-fold	0.636 (0.582 to 0.714)	0.54	-0.003
Purposeful selection in-fold	42	Purposeful selection, in-fold	0.576 (0.530 to 0.674)	0.50	-0.006
Neural network, corrected (no early stopping)	42	Implicit, in-fold	0.631 (0.586 to 0.716)	0.74	+ 0.016
XGBoost, prior-constrained search	42	Implicit, in-fold	0.617 (0.559 to 0.692)	1.00	+ 0.015
XGBoost, fixed regularised configuration (no search)	42	Implicit, in-fold	0.616 (0.557 to 0.682)	0.55	-0.002
Case-mix group alone (medical, surgical, cardiac surgery)	3 fixed	No physiology at all	0.610 (0.482 to 0.615)	0.85	+ 0.016

Every arm below the first two rows discovered its variables inside each training fold. Point estimates are means across 10 repeats of stratified 5 fold cross validation; intervals are percentile bootstraps of the repeat-averaged predictions. The spread across fitting strategies given the same 12 variable pool is 0.028 AUC; the spread attributable to how the variables were chosen is 0.067.

Third weakness: most of it is case mix, not weaning physiology

The third weakness is that most of what the score detects is case mix rather than weaning physiology. The cohort comprises three populations with very different event rates: medical patients (n = 300, 41 events, 13.7%), non-cardiac surgical patients (n = 196, 18 events, 9.2%) and patients after cardiac surgery (n = 304, 12 events, 3.9%). Within these strata the score discriminates at 0.622, 0.488 and 0.609 respectively, giving a within-stratum AUC weighted by discordant pairs of 0.595 against the pooled 0.666 (Fig. 3, Table 4). Two of the five variables encode that stratification directly, and the predicted probability of death is not one instrument but two. Consistent with this, the predicted probability of death alone achieves 0.656, so the other four variables add 0.007 of AUC, and case-mix group alone, three indicators containing no physiology whatever, achieves 0.610. A clinician at the bedside already knows which group the patient is in.

Table 4
Decomposition of discrimination by case mix, and single-variable benchmarks.

Population or benchmark	Patients	Events (%)	AUC of the five variable specification
Medical	300	41 (13.7%)	0.622
Surgical, non cardiac	196	18 (9.2%)	0.488
Cardiac surgery	304	12 (3.9%)	0.609
Within strata, weighted by discordant pairs	800	71 (8.9%)	0.595
Whole cohort, pooled	800	71 (8.9%)	0.666
Benchmark: predicted probability of death alone	800	71 (8.9%)	0.656
Benchmark: case-mix group alone	800	71 (8.9%)	0.610
Benchmark: RSBI alone	800	71 (8.9%)	0.566

The within-stratum value weights each stratum by its number of discordant pairs. A clinician at the bedside already knows the stratum, so the discrimination genuinely added by the model is closer to the within-stratum value than to the pooled value.

Fourth weakness: one failure in five had not happened yet

Among the 71 failures, the attributed causes were hypoxaemic respiratory failure in 25 (35.2%), cardiac arrest in 10 (14.1%), increased work of breathing in 9 (12.7%), inability to manage secretions in 8 (11.3%), altered consciousness in 8 (11.3%), stridor in 7 (9.9%) and return to the operating room in 4 (5.6%). Cardiac arrest and return to the operating room account for 14 failures, 19.7% of the total. These are new events, not errors in the assessment of readiness. A fifth of the outcome was determined after the prediction was made, and no model of pre-extubation physiology can reach it.

Part 2. Machine learning and artificial intelligence do not improve it

Two machine learning models under a matched protocol

Given the same 12 candidate variables and the same resampling discipline, the two MLM most used in this literature did not improve on regression. Random forest achieved a cross validated AUC of 0.630 and gradient boosting 0.627, against 0.655 for ordinary logistic regression using the whole pool, 0.651 for L1 selection with refitting and 0.644 for elastic net (Table 3, Fig. 2). The entire spread across five very different fitting strategies was 0.028 AUC. Neither MLM reached the fixed five variable specification, at 0.663, and neither reached the strongest regression strategy that had to select its own variables, at 0.655.

Given the full 42 variable pool, the ordering was the same and every value lower: ridge-penalized regression 0.640, L1 selection then refit 0.636, a neural network 0.631, unpenalized regression 0.629, and gradient boosting 0.614 to 0.617 depending on the search space. Adding 30 variables made every strategy worse, the signature of fitting noise at 1.7 events per candidate variable.

Every other strategy, given every advantage

Extending the comparison to every strategy we tried changes nothing (Table 5). Cross validated AUCs were 0.663 for the published specification, 0.636 for elastic net, 0.632 for a Super Learner stacked ensemble, 0.631 for a corrected neural network, 0.621 for random forest, 0.615 and 0.614 for the two gradient boosting implementations, 0.600 for a support vector machine, 0.599 for gradient boosting with SMOTE and 0.594 for k nearest neighbours. Using a single bootstrap estimand for both point estimate and interval, paired differences against the published specification ranged from − 0.006 to -0.060, and the interval excluded zero only for gradient boosting, k nearest neighbours and the RSBI alone (Table 6). The area under the precision recall curve told the same story against a no-skill value of 0.089: 0.185 for the published specification and 0.123 to 0.149 for the flexible methods.

Table 5
All strategies compared, including the two machine learning models given every advantage.

Strategy	AUC, mean of repeats (SD)	AUC of averaged predictions (95% CI)	AUPRC	Calibration slope	Calibration intercept	E:O	ICI	Scaled Brier	Standardised net benefit at 10%
Logistic regression, five expert variables	0.663 (0.010)	0.666 (0.591 to 0.732)	0.185	0.81	+ 0.00	1.00	0.011	+ 0.034	0.20
Logistic regression, 12 variables	0.655 (0.010)	0.660 (0.586 to 0.732)	0.183	0.68	+ 0.00	1.00	0.017	+ 0.027	0.18
Logistic regression, 42 variables, unpenalised	0.629 (0.021)	0.639 (0.567 to 0.707)	0.162	0.29	-0.07	1.05	0.052	-0.072	
Logistic regression, 42 variables, ridge	0.640 (0.023)	0.648 (0.579 to 0.714)	0.168	0.40	-0.06	1.05	0.045	-0.047	
Elastic net	0.636 (0.028)	0.656 (0.590 to 0.723)	0.145	0.74	-0.02	1.02	0.020	+ 0.002	0.15
Random forest	0.621 (0.017)	0.627 (0.562 to 0.696)	0.140	0.96	+ 0.01	0.99	0.012	+ 0.014	0.09
Gradient boosting (XGBoost)	0.614 (0.021)	0.618 (0.554 to 0.675)	0.124	0.89	+ 0.00	1.00	0.018	+ 0.004	0.11
Gradient boosting (LightGBM)	0.615 (0.013)	0.623 (0.561 to 0.687)	0.130	0.70	+ 0.12	0.92	0.022	-0.002	0.11
Support vector machine	0.600 (0.046)	0.645 (0.573 to 0.708)	0.128	0.73	-0.03	1.02	0.017	+ 0.006	0.11
k-nearest neighbours	0.594 (0.027)	0.606 (0.540 to 0.669)	0.128	0.73	+ 0.15	0.87	0.016	+ 0.006	0.07
Neural network (corrected)	0.631 (0.029)	0.650 (0.582 to 0.718)	0.149	0.61	-0.08	1.07	0.032	-0.018	
Super Learner	0.632 (0.027)	0.642 (0.576 to 0.707)	0.147	1.68	+ 0.00	1.00	0.024	+ 0.011	0.09
XGBoost with SMOTE	0.599 (0.030)	0.619 (0.553 to 0.684)	0.123	0.28	-0.72	1.75	0.086	-0.247	0.10
RSBI alone	0.566 (0.006)	0.568 (0.492 to 0.641)	0.138	0.80	+ 0.00	1.00	0.011	+ 0.007	0.09

The first numeric column is the mean of ten honest cross validations with its standard deviation across repeats; the second is the AUC of the repeat-averaged predictions with its bootstrap interval. They are different quantities and are reported separately so that no point estimate is paired with an interval belonging to something else. Calibration measures are means across repeats. E:O, expected to observed ratio; ICI, integrated calibration index; AUPRC, area under the precision recall curve (no-skill value 0.089). Standardised net benefit is net benefit at a 10% threshold divided by the maximum attainable value of 0.089.

Table 6
Paired difference in the area under the curve against the published five variable specification.

Strategy	Difference in AUC	95% confidence interval	p value (bootstrap)
Logistic regression, 12 variables	-0.006	-0.035 to + 0.023	0.688
Logistic regression, 42 variables, unpenalised	-0.028	-0.094 to + 0.036	0.395
Logistic regression, 42 variables, ridge	-0.017	-0.076 to + 0.047	0.589
Elastic net	-0.009	-0.051 to + 0.033	0.660
Random forest	-0.038	-0.090 to + 0.009	0.131
Gradient boosting (XGBoost)	-0.047	-0.093 to -0.003	0.033
Gradient boosting (LightGBM)	-0.043	-0.097 to + 0.012	0.125
Support vector machine	-0.020	-0.070 to + 0.031	0.458
k-nearest neighbours	-0.060	-0.111 to -0.008	0.024
Neural network (corrected)	-0.016	-0.068 to + 0.037	0.553
Super Learner	-0.023	-0.075 to + 0.029	0.396
XGBoost with SMOTE	-0.047	-0.108 to + 0.016	0.132
RSBI alone	-0.098	-0.172 to -0.022	0.015

Point estimates and intervals come from the same 2,000 sample percentile bootstrap of the repeat-averaged predictions. Negative values favour the published five variable specification.

Calibration separated the strategies more sharply than discrimination (Fig. 4). The published specification had a calibration slope of 0.81, an intercept of 0.00, an expected to observed ratio of 1.00 and an integrated calibration index of 0.011, stable across smoothing bandwidths from 0.007 to 0.014. Random forest was well calibrated after tuning (slope 0.96), the Super Learner produced predictions that were too moderate rather than too extreme (slope 1.68), and the 42 variable regression was badly overfitted (slope 0.29 unpenalized, 0.40 with ridge).

The class rebalancing arm behaved worst, and it is the most widely practiced of these techniques. SMOTE oversampling left discrimination unchanged at 0.599 while inflating predicted risks by a factor of 1.75, with a calibration intercept of -0.72, a slope of 0.28 and an integrated calibration index of 0.086, eight times that of the published specification. Its scaled Brier score was − 0.247, meaning it was substantially less accurate overall than assigning every patient the average incidence of 8.9%, and on decision curve analysis it fell below the strategy of taking no action at all. An investigator reporting only the AUC would have seen nothing wrong.

Clinical utility

Decision curve analysis (Fig. 5) showed that a parsimonious regression gave the highest net benefit across the decision-relevant range. At a 10% risk threshold, which asserts that one unnecessary course of prophylactic noninvasive ventilation is worth one ninth of a missed reintubation, the published specification achieved 0.018, or 20% of attainable, against 0.011 for gradient boosting (11%) and 0.008 for random forest (9%). Paired bootstrap differences favored the published specification over random forest (0.016, 95% interval 0.005 to 0.026) and over gradient boosting (0.011, 0.001 to 0.022). Above a threshold of about 15% every strategy converged on zero and we do not interpret the remaining differences. Excluding planned noninvasive ventilation from every arm, because it records the clinician's own judgement and is also the action a model would trigger, changed nothing.

Two practices that would have reversed the conclusion

Two standard practices would have produced the opposite conclusion from the same data. Fitted and evaluated on the same 800 patients, a tuned random forest achieved an apparent AUC of 0.848 and a regularized gradient boosting model 0.859, against cross validated values of 0.621 and 0.616; the published specification lost only 0.028, from 0.691 to 0.663 (Fig. 6A). And across 300 random 70/30 partitions, with properly tuned models in every arm, the test AUC of the published specification had a standard deviation of 0.053 and ranged from 0.439 to 0.788, while the tuned random forest beat it in 15.7% of splits and gradient boosting in 12.7%, each by more than 0.05 of AUC in 2.3% (Fig. 6B). An investigator who split once could have published a machine learning victory from these data without any misconduct.

With randomly permuted outcomes the cross validated AUC was 0.497 (standard deviation 0.055) for the published specification and 0.499 (0.047) for gradient boosting, with maxima of 0.598 and 0.579 across 50 permutations (Supplementary Figure S3A), so values near 0.60 are obtainable from pure noise in a cohort of this size.

Neither the method of imputation nor the differential completeness of the pools accounted for any of this. Gradient boosting reached 0.636 with median imputation, 0.633 with iterative imputation and 0.638 using its native handling of missing values, and the published specification was unmoved (0.6632 and 0.6633). Variable importance, however, was unstable: across 200 bootstrap refits of the random forest the RSBI had a median importance rank of 2 with a 95% interval from 1 to 11, and only the RSBI appeared in the top five variables in more than 80% of resamples. Any explanation offered to clinicians would have changed had the cohort been resampled once.

What this design could have detected

The design cannot detect small differences, and this qualifies every negative statement above. The bootstrap standard error of the paired difference in AUC was 0.024 on average, so the minimum difference detectable with 80% power was 0.068 (range 0.061 to 0.073). The upper bound of the interval for the best machine learning strategy still admits an advantage of 0.032 AUC. Bounding the difference within 0.03 would require 2,001 patients (178 events) and within 0.02 would require 4,501 patients (400 events) (Table 7). Separately, the minimum sample size for stable development by the criteria of Riley and colleagues is 1,078 patients for a 5 variable model, 2,586 for a 12 variable model and 9,051 for the 42 variable pool; this cohort provides 74% of the first and 9% of the last.

Table 7
What this design can detect, and the sample sizes that would be required.

Quantity	Value
Standard error of the paired difference in AUC	0.024
Minimum difference detectable with 80% power	0.068 AUC
Largest machine learning advantage not excluded by the data	+ 0.032 AUC
Patients needed to bound the difference within 0.05	721 (64 events)
Patients needed to bound the difference within 0.03	2,001 (178 events)
Patients needed to bound the difference within 0.02	4,501 (400 events)
Riley minimum sample size, 5 candidate variables	1,078 (96 events)
Riley minimum sample size, 12 candidate variables	2,586 (230 events)
Riley minimum sample size, 42 candidate variables	9,051 (803 events)
Available	800 patients (71 events)

The minimum detectable difference is computed for 80% power at a two-sided alpha of 0.05 from the bootstrap standard error of the paired difference. Riley requirements use the apparent Cox and Snell R squared of the reference model (0.041) and are therefore, if anything, underestimates.

What the data do exclude decisively is the performance reported in the literature. The upper bootstrap limit for the best strategy in this cohort was 0.734, far below the 0.83 to 0.98 range of published machine learning models for this outcome.

Part 3. Artificial intelligence appraisal of the published literature, and an adversarial machine peer review of this manuscript

The published literature

Fourteen published machine learning studies of extubation failure, reintubation or weaning outcome were appraised with assistance from a large language model (Table 8, Fig. 7, Supplementary Table S6). Their median sample size was 818 patients (range 117 to 23,242) and their median best reported AUC was 0.874 (0.789 to 0.983). Where the number of candidate predictors could be established, the median number of events per candidate variable was 2.9; six of eight were below 10 and two were below 1.

Table 8
Appraisal of 14 published machine learning studies of extubation failure, reintubation or weaning outcome.

Study	Source	Patients	Events	Candidate variables	EPV	Best AUC	Validation	Calibration	Utility	Explainability figure	Code
Zhu 2026	Multicentre	1,371	204	47	4.3	0.972	Single split plus external	Unclear	Yes	Yes	No
Zhang 2026	Single centre	707	123	24	5.1	0.879	Single split	Yes	Yes	No	No
Fabregat 2021	Single centre	697	50			0.983	Single split	No	No	No	No
Jiang 2025	Registry	776	205			0.793	Cross validation	Yes	No	Yes	No
Wen 2025	Multicentre	861	107			0.969	Single split plus external	Yes	No	Yes	No
Lin 2023	Registry	1,789	1,024			0.830	Single split	Yes	No	Yes	Yes
Kim 2023	Registry	23,242	4,217	46	91.7	0.861	Cross validation	Yes	No	Yes	Yes
Park 2023	Single centre	138	35	28	1.2	0.912	Single split	No	No	Yes	No
Fenske 2025	Two centres	448		37		0.870	Temporal plus external	No	No	Yes	Yes
Chao 2021	Single centre	963	424	300	1.4	0.908	Single split	No	No	Yes	No
Gong 2022	Registry	8,599	2,609	89	29.3	0.829	Single split plus cross validation	Yes	Yes	Yes	No
Shen 2026	Registry plus local	5,895		48		0.859	Single split plus external	Qualitative only	No	Yes	No
Yang 2026	Single centre	413	45	59	0.8	0.789	Single split	Yes	No	Yes	No
Otaguro 2021	Single centre	117	13	57	0.2	0.950	Not reported	No	No	Yes	No
Present study	Single centre	800	71	42	1.7	0.663	Repeated nested cross validation	Yes	Yes	No	Yes

EPV, events per candidate variable, computed where both quantities were reported. Utility means decision curve analysis or another measure of clinical usefulness. Extraction was assisted by a large language model and performed by a single appraiser from the accessible text; blank cells are fields that could not be located rather than fields known to be absent. The complete extraction sheet, including extraction confidence for each study, is deposited.

Twelve of the 14 studies displayed an explainability figure, most commonly SHAP. Nine included a regression comparator, though none reported tuning it to the standard applied to the machine learning arm. Seven reported calibration in any form. Four had any external or temporal validation. Three reported decision curve analysis or another measure of clinical utility. Three made code available. None used repeated or nested cross validation, and six relied on a single random split as the only validation.

Reported discrimination fell as studies grew larger, the direction optimism predicts, though the association was not statistically significant across the 12 studies with both quantities available (Spearman rho − 0.39, p = 0.21). The four studies with fewer than 100 events reported a median AUC of 0.93; the three with 1,000 or more reported 0.83. The present study, with 71 events and repeated nested cross validation, sits at 0.663, below every published value (Fig. 7B). We cannot exclude that other cohorts are genuinely more predictable than ours; we note only that the studies least able to control optimism report the highest numbers.

Artificial intelligence as reviewer of our own work

In its third role, artificial intelligence reviewed this manuscript. The adversarial review produced ten major criticisms and eighteen minor ones. Nine of the ten major criticisms were confirmed when we re-derived them, and two changed our conclusions.

The first was the confounding of algorithm with variable provenance, which the matched protocol reported above exists to address, and which turned the paper's central claim from a statement about algorithms into a statement about variable selection. The second was that our own multilayer perceptron result, originally an AUC of 0.517, was a software artefact: the early stopping option of the software monitors classification accuracy, which at an event rate of 8.9% is maximized by predicting the majority class, so the fit was halted after a few dozen iterations. With early stopping disabled the same architecture reached 0.631. That figure had been supplying the most quotable number in our abstract. The review also correctly identified that three arms described as unpenalized were ridge-penalised by software default, that our point estimates and confidence intervals for paired differences referred to different stands, that our intervals measured partition variability rather than uncertainty, that we had described a calibration slope above 1 as overfitting when it indicates the opposite, that a claim to be the largest prospective series of planned extubations was unsupportable, that the area under the precision recall curve was specified and never reported, and that our sample size calculation used an apparent rather than an optimism-corrected R squared.

Three criticisms we tested and could partly or wholly reject. Median imputation does not disadvantage the flexible arms. Enlarging the hyperparameter search does not materially change the result. And its specific claim that a fixed, regularized gradient boosting configuration reaches an AUC of 0.636, which would have halved the apparent deficit of the machine learning arm, did not reproduce in our hands: the same class of configuration reached 0.616.

That last observation is the general lesson. The structural criticisms were accurate and valuable, and several would have been difficult for a human reviewer to reach without re-running the analysis, yet roughly a third of the specific numerical claims did not reproduce when we recomputed them. An adversarial machine review is an efficient instrument for finding what is wrong with a prediction model paper, and an unreliable source of what the corrected numbers are. We would not have published this paper in its original form.

Discussion

Held to the same protocol and given the same variables, random forest and gradient boosting performed no better than ordinary logistic regression, and the spread across five fitting strategies, 0.028 of AUC, was less than half of what this design could have detected. Nine further strategies, including ensembling and class rebalancing, made no difference or made things worse. What did separate the strategies was not the algorithm but which variables entered the model, and the set chosen by clinicians outperformed every automated alternative applied to the same candidate pool. Meanwhile a single severity variable reproduced almost all of the published score's discrimination, and within case-mix strata that discrimination fell to 0.595. We conclude that all of these models, ours included, are detecting a modest amount of severity and case mix, and very little weaning physiology.

Why flexibility bought nothing

Four mechanisms explain why flexibility bought nothing. First, the signal is weak: the reference model explains about 4% of the outcome variation. When the total amount of predictive information is small, the difference between a linear and a flexible functional form is smaller still. Second, the events are few: 71 events against 42 candidate variables is 1.7 events per variable [30, 31], and a tree ensemble has more effective parameters than that. An apparent AUC of 0.85 against a cross validated 0.62 shows what the surplus capacity is spent on. Third, the relationships are close to linear and additive over the observed range. Penalized regression, gained nothing, and the tree ensembles found no interaction that generalized. And finally, a fifth of the failures had not happened yet. Cardiac arrest and return to the operating room are not failures of the readiness assessment; they are new events. No algorithm can predict what has not yet occurred and has no antecedent in the variables recorded.

Why published machine learning models report higher performance

Published MLM for extubation report AUCs of 0.79 to 0.98, we consider none of those models would help at an individual bedside because we proved no differences with the original score.

Sample size and data source may explain why the largest studies report the lowest values, a median of 0.83, and the smallest the highest, a median of 0.93. That is the signature of optimism, not of better models. Also predictor sets are mostly information by the bedside checklist [14–16], and in several reports the boundary between predictors and outcome is difficult to verify from the published text. Our own data show how easily this happens: one variable in our database, planned noninvasive ventilation, simultaneously encodes the clinician's risk judgement and the action a model would trigger.

Evaluation explains the rest. Six of 14 studies validated on a single random split and none used repeated or nested resampling. With these data and properly tuned models, a single partition would have favored machine learning in roughly one attempt in seven.

The pattern across the six reporting items is the most useful observation here for the reader of the next machine learning paper. The element most likely to be present is the one that makes a paper look sophisticated, the explainability figure, in 12 of 14. The elements that would let a clinician decide whether to use the model, calibration, net benefit, external validation and shared code, are present in 3 to 7 of 14. Explainability figures are of contested value for that purpose and our own stability analysis shows why: the rankings they display would have changed substantially had the cohort been resampled once and it has been described before [32, 33].

What artificial intelligence is very good at, and what it cannot do

Artificial intelligence was able to fit eleven model families under an identical resampling protocol, to appraise a literature against a structured instrument, and to review our own manuscript adversarial with access to the code, where it found an error that invalidated our first conclusion and a software artefact that had produced the worst result in our own paper. These methods were able to extract, compress and redistribute the information contained in a dataset. They do it faster and more thoroughly than we do, and they are not fooled by the habits that fool us. What they cannot do is add information that was never measured. An algorithm cannot know that the diaphragm of the patient in bed 12 is thin and slow, because nobody put a probe on it. That is not a limitation of the algorithm; it is a property of the dataset, and no amount of methodological sophistication converts one into the other.

What is missing from the data, and why it is not there yet

The variables in this cohort describe gas exchange, respiratory mechanics during a supported breathing trial, hemodynamics at rest and a global severity score. Set against the physiology of weaning failure, that list has conspicuous holes, and Table 9 sets out what is missing, what mechanism each measurement would capture, and why it is not yet routine.

Table 9

What is missing from the dataset: measurements that the physiology of weaning failure implies, and their status in routine practice.

Mechanism of weaning failure	Measurement that would capture it	Present in this dataset	Status in routine practice
Respiratory muscle weakness, low capacity	Diaphragmatic ultrasound (thickening fraction, excursion) [57 to 59]; maximal inspiratory pressure	No	Diaphragm dysfunction is common and associated with weaning failure [60,61], but a meta-analysis of 16 studies found sensitivity that was low and heterogeneous, and thresholds differ between studies [58]; the measurement is operator dependent [59]. Diaphragmatic function was named as a candidate in the original derivation [9]
High respiratory load, excessive effort	Oesophageal pressure and the pressure-time product [63]; airway occlusion pressure P0.1 [62]	No	Oesophageal manometry remains largely a research technique [63]; P0.1 is available on most modern ventilators as an estimate of drive and effort [62] and is rarely recorded systematically
Weaning-induced pulmonary oedema	Echocardiographic diastolic function and filling pressures during the breathing trial; extravascular lung water; natriuretic peptide trajectory [64,66]	No	The entity is well described and diagnosable during a spontaneous breathing trial [64], and natriuretic peptides identify cardiac-origin weaning failure [66] and can guide fluid management [67], but none of this is part of a routine weaning checklist
Loss of lung aeration	Lung ultrasound aeration score before and after the trial [65]	No	Aeration loss during a successful trial predicts post-extubation distress [65]; increasingly used, not routine before extubation, requires training
Fluid overload	Cumulative fluid balance [70]; natriuretic peptide-driven fluid management [67]	No, not collected in this cohort	Fluid balance is associated with weaning outcome [70] and is recorded in most units; it was simply not captured here, and was named as a candidate in the original derivation [9]
Impaired airway protection	Measured cough peak flow [68,69]; quantified secretion volume	Subjective judgement of cough only	Cough peak flow is an objective predictor of extubation outcome [68] and performs better than subjective assessment in meta-analysis [69]; a flowmeter measurement is not routine, and secretions are judged rather than measured
Dynamic deterioration during the trial	Ventilator pressure, flow and effort as a time series; asynchrony indices [71]	Single values at one moment	Asynchrony detected from continuously recorded waveforms is associated with mortality [71]; in our experience such data are generated by every ventilator and almost never stored
New events after extubation (cardiac arrest, return to the operating room)	None. These had not occurred when the prediction was made	Not applicable	Accounted for 19.7% of the failures in this cohort and are unreachable by any model of pre-extubation physiology

This table is the principal practical output of the study. None of the measurements listed is exotic, and several are performed daily somewhere in most large units; what none of them is, yet, is routine, standardised and recorded in a form a model can use.

Weaning failure of respiratory muscle origin is a mismatch between load and capacity, and we measured neither. Diaphragm dysfunction is common in ventilated patients and is associated with failure to liberate [34, 35]; diaphragmatic ultrasound with thickening fraction and excursion [36–38], maximal inspiratory pressure, airway occlusion pressure [39] and esophageal pressure with the pressure-time product [40] would address it directly, and diaphragmatic function was named as a candidate in our own original paper [10]. Weaning-induced pulmonary oedema is a cardiac event during the trial that can be identified while it happens [41], and we measured no diastolic function, no filling pressure, no lung aeration score [42] and no natriuretic peptide trajectory [43, 44]. Airway protection was represented by a subjective judgement of cough, where a measured cough peak flow exists and performs better [45, 46]. Fluid balance, which is associated with weaning outcome [70], was not recorded at all. And the ventilator produced a continuous stream of pressure, flow and effort data, from which asynchrony indices carry prognostic information [47], of which we retained a handful of numbers taken at one moment.

None of these is exotic and several are performed daily somewhere in most large units. What none of them is, yet, is routine, standardized and recorded in a form that a model can use: a meta-analysis of 11 diaphragm ultrasound studies found sensitivity for weaning outcome that was low and heterogeneous with thresholds that differ between studies [48], the measurements are operator dependent, and almost nothing about the spontaneous breathing trial is stored as a time series. Until that changes, every model of extubation failure, however it is fitted, will be built from the same impoverished description of the patient, and will land where ours did.

We would go further. The most valuable output of this analysis is not a model but a measurement agenda. The ceiling we ran into is not an argument for abandoning prediction; it is an argument for prospective, stratified, multicenter studies that measure the things the physiology says should matter, in enough patients to satisfy the sample size criteria, and only then ask which algorithm fits them best. On present

evidence the answer to that last question will again be that it hardly matters, and that is a comfortable position to be in, because it means the hard work is clinical rather than computational.

Why clinical knowledge remains decisive

That is why clinical knowledge and experience remain relevant, and our results let us say so quantitatively rather than sentimentally. Five variables chosen by clinicians outperformed every automated selection procedure applied to the same candidate pool, by more than any two algorithms differed from each other. The variables that survived, acid base status, the breathing pattern, cough, severity and the medical or surgical nature of the illness, are not a statistical accident; they are the shortlist a good intensivist would make. An algorithm searching the same 12 columns did not find them, and an algorithm searching 42 columns found something worse.

The clinician's role in a prediction model is not to be the baseline that the machine eventually beats. It is to decide what enters the dataset in the first place, which is the decision that determines the ceiling for everything that follows.

Limitations

This study has limitations. It is a single center cohort from a hospital at 2,600 m above sea level, where arterial oxygen tensions and the reference ranges of derived oxygenation indices differ from those at sea level; absolute performance should not be transported without recalibration. Altitude is unlikely to threaten the internal comparison between modelling strategies, but range restriction of the oxygenation variables on the steep part of the dissociation curve could in principle remove threshold effects that constitute a prior reason to expect tree ensembles to help.

The case mix, 38% of it after cardiac surgery with a 3.9% failure rate, gives an overall incidence of 8.9% at the low end of published estimates and limits the number of events; a unit with a different surgical fraction could reach a different answer. The comparison is underpowered for differences smaller than 0.068 of AUC, and we therefore report inconclusiveness rather than equivalence. Deep learning architectures beyond a multilayer perceptron were not tested, since they require sample sizes orders of magnitude larger. The literature appraisal was performed by a single assisted appraiser, is not a registered systematic review and does not include every study in the field; the extraction sheet is deposited so that it can be corrected and extended. Finally, we performed internal validation only, and none of the models studied here, including our own score, is ready for autonomous use.

Conclusion

Artificial intelligence excels at extracting patterns from existing datasets, but it cannot generate information that is not there. In this cohort, neither advanced machine learning nor extensive algorithmic tuning outperformed a parsimonious model based on clinical judgment. In real-world critical care, models are fundamentally limited because crucial diagnostic data is often unmeasured or absent. Clinical judgment determines which variables matter and when additional context is required; while AI can optimize predictions within historical data, it cannot replace the clinical insight needed to identify and integrate missing clinical signals. True progress in predicting extubation failure will not arise from applying complex algorithms to identical feature sets, but from clinical research that evaluates novel physiological predictors such as diaphragmatic and lung ultrasound, inspiratory effort, and cardiac function at the bedside.

Declarations

Ethics approval and consent to participate

Ethics approval and consent to participate: the parent cohort was approved by the Institutional Human Research Ethics Committee of Universidad del Rosario (DVO005 929-CV904) and classified as minimum risk; general informed consent covered the use of medical record information. This secondary analysis used a de-identified extract. The study was conducted in accordance with the Declaration of Helsinki

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests. The authors derived and published the ABCDMed score that is re-examined here

Use of artificial intelligence

Statistical analysis scripts and a first draft of the manuscript were prepared with the assistance of a large language model (Claude, Anthropic). The same system was used for the assisted appraisal of the published literature and, separately, to produce the adversarial peer review reported in the Results. Every criticism it made was verified by re-analysis before being acted upon, and claims that did not reproduce are reported as such. All analyses were specified, inspected and verified by the authors, who take full responsibility for the content.

Funding

None

Author Contribution

DRRL and CRR collected the original data and contributed to the analysis. EER and EMCP wrote the first draft of the original manuscript and contributed to the analysis and design. DRRL and EER conceived the study, directed the design and analysis, and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgement

Not applicable

Data Availability

Statistical analysis scripts and a first draft of the manuscript were prepared with the assistance of a large language model (Claude, Anthropic). The same system was used for the assisted appraisal of the published literature and, separately, to produce the adversarial peer review reported in the Results. Every criticism it made was verified by re-analysis before being acted upon, and claims that did not reproduce are reported as such. All analyses were specified, inspected and verified by the authors, who take full responsibility for the content.

References

1. Esteban A. Characteristics and Outcomes in Adult Patients Receiving Mechanical VentilationA 28-Day International Study. JAMA. 2002;287(3):345. 10.1001/jama.287.3.345.
2. Béduneau G, Pham T, Schortgen F, Piquilloud L, Zogheib E, Jonas M, et al. Epidemiology of Weaning Outcome according to a New Definition. The WIND Study. Am J Respir Crit Care Med. 2017;195(6):772–83. 10.1164/rccm.201602-0320OC.
3. Thille AW, Boissier F, Ben Ghezala H, Razazi K, Mekontso-Dessap A, Brun-Buisson C. Risk factors for and prediction by caregivers of extubation failure in ICU patients: a prospective study. Crit Care Med. 2015;43(3):613–20. 10. 1097/CCM.0000000000000748 PubMed PMID: 25479115.
4. Epstein SK, Ciubotaru RL, Wong JB. Effect of failed extubation on the outcome of mechanical ventilation. Chest. 1997;112(1):186–92. 10.1378/chest.112. .1.186 PubMed PMID: 9228375.
5. Frutos-Vivar F, Ferguson ND, Esteban A, Epstein SK, Arabi Y, Apezteguía C, et al. Risk Factors for Extubation Failure in Patients Following a Successful Spontaneous Breathing Trial. Chest. 2006;130(6):1664–71. 10.1378/chest.130.6.1664.
6. Yang KL, Tobin MJ. A Prospective Study of Indexes Predicting the Outcome of Trials of Weaning from Mechanical Ventilation. N Engl J Med. 1991;324(21):1445–50. 10.1056/NEJM199105233242101.
7. Torrini F, Gendreau S, Morel J, Carteaux G, Thille AW, Antonelli M, et al. Prediction of extubation outcome in critically ill patients: a systematic review and meta-analysis. Crit Care. 2021;25(1):391. 10.1186/s13054-021-03802-3.
8. Khamiees M, Raju P, DeGirolamo A, Amoateng-Adjepong Y, Manthous CA. Predictors of Extubation Outcome in Patients Who Have Successfully Completed a Spontaneous Breathing Trial. Chest. 2001;120(4):1262–70. 10.1378/chest.120.4.1262.
9. Paiva JA, Pereira JM. Measuring intensive care performance to attain collaborative quality improvement: Work in progress but much to be done. J Crit Care. 2025;90:155209. 10.1016/j.jcrc.2025.155209 . PubMed PMID: 40737761.
10. Rodriguez Lima DR, Navarrete Aldana N, Roncallo Valencia VE, Rubio Ramos C, Ibáñez Pinilla M, Torres Suarez Y et al. Derivation and validation of the ABCDMed clinical score to estimate the probability of extubation failure in intensive care units. 30. 2022;30(2):58–65.

11. Hsieh MH, Hsieh MJ, Cheng AC, Chen CM, Hsieh CC, Chao CM, et al. Predicting weaning difficulty for planned extubation patients with an artificial neural network. Medicine. 2019;98(40):e17392. 10.1097/MD.0000000000017392.
12. 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. 10.3389/fmed.2021.676343.
13. 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(5):408–17. 10.1272/jnms.JNMS.2021_88-508.
14. Huang KY, Hsu YL, Chen HC, Horng MH, Chung CL, Lin CH, et al. Developing a machine-learning model for real-time prediction of successful extubation in mechanically ventilated patients using time-series ventilator-derived parameters. Front Med. 2023;10:1167445. 10.3389/fmed.2023.1167445.
15. 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(11):1560–9. 10.4187/respcare.03648.
16. 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(1):448. 10.1186/s13054-021-03864-3.
17. 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, On behalf of Topic Group, Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. 10.1186/s12916-019-1466-7
18. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. 2022;22(1):101. 10.1186/s12874-022-01577-x.
19. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
20. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, Van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol. 2020;122:95–107. 10.1016/j.jclinepi.2020.03.005.
21. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021;375:n2281. 10.1136/bmj.n2281 . PubMed PMID: 34670780; PubMed Central PMCID: PMC8527348.
22. Andaur Navarro CL, Damen JAA, Van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2023;154:8–22. 10.1016/j.jclinepi.2022.11.015.
23. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
24. Van Den Goorbergh R, Van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022;29(9):1525–34. 10.1093/jamia/ocac093.
25. Carriero A, Luijken K, De Hond A, Moons KGM, Van Calster B, Van Smeden M. The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study. Stat Med. 2025;44(3–4):e10320. 10.1002/sim.10320.
26. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019;38(21):4051–65. 10.1002/sim.8281.
27. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Mak. 2006;26(6):565–74. 10.1177/0272989X06295361.
28. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biometrical J. 2023;65(8):2200302. 10.1002/bimj.202200302.
29. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38(7):1276–96. 10.1002/sim.7992.
30. Van Smeden M, De Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16(1):163. 10.1186/s12874-016-0267-3.
31. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455–74. 10.1177/0962280218784726 . PubMed PMID: 29966490; PubMed Central PMCID: PMC6710621.
32. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. 10.1038/s42256-019-0048-x.

33. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50. 10.1016/S2589-7500(21)00208-9.
34. Demoule A, Jung B, Prodanovic H, Molinari N, Chanques G, Coirault C, et al. Diaphragm Dysfunction on Admission to the Intensive Care Unit. Prevalence, Risk Factors, and Prognostic Impact—A Prospective Study. Am J Respir Crit Care Med. 2013;188(2):213–9. 10.1164/rccm.201209-1668OC.
35. Dres M, Dubé BP, Mayaux J, Delemazure J, Reuter D, Brochard L, et al. Coexistence and Impact of Limb Muscle and Diaphragm Weakness at Time of Liberation from Mechanical Ventilation in Medical Intensive Care Unit Patients. Am J Respir Crit Care Med. 2017;195(1):57–66. 10.1164/rccm.201602-0367OC.
36. DiNino E, Gartman EJ, Sethi JM, McCool FD. Diaphragm ultrasound as a predictor of successful extubation from mechanical ventilation. Thorax. 2014;69(5):431–5. 10.1136/thoraxjnl-2013-204111.
37. Le Neindre A, Philippart F, Luperto M, Wormser J, Morel-Sapene J, Aho SL, et al. Diagnostic accuracy of diaphragm ultrasound to predict weaning outcome: A systematic review and meta-analysis. Int J Nurs Stud. 2021;117:103890. 10.1016/j.ijnurstu.2021.103890.
38. Goligher EC, Fan E, Herridge MS, Murray A, Vorona S, Brace D, et al. Evolution of Diaphragm Thickness during Mechanical Ventilation. Impact of Inspiratory Effort. Am J Respir Crit Care Med. 2015;192(9):1080–8. 10.1164/rccm.201503-0620OC.
39. Telias I, Junhasavasdikul D, Rittayamai N, Piquilloud L, Chen L, Ferguson ND, et al. Airway Occlusion Pressure As an Estimate of Respiratory Drive and Inspiratory Effort during Assisted Ventilation. Am J Respir Crit Care Med. 2020;201(9):1086–98. 10.1164/rccm.201907-1425OC.
40. Mauri T, Yoshida T, Bellani G, Goligher EC, Carteaux G, Rittayamai N, et al. Esophageal and transpulmonary pressure in the clinical setting: meaning, usefulness and perspectives. Intensive Care Med. 2016;42(9):1360–73. 10.1007/s00134-016-4400-x.
41. Dres M, Teboul JL, Anguel N, Guerin L, Richard C, Monnet X. Extravascular lung water, B-type natriuretic peptide, and blood volume contraction enable diagnosis of weaning-induced pulmonary edema. Crit Care Med. 2014;42(8):1882–9. doi:10.1097/CCM.0000000000000295 PubMed PMID: 24717458.
42. Soummer A, Perbet S, Brisson H, Arbelot C, Constantin JM, Lu Q, et al. Ultrasound assessment of lung aeration loss during a successful weaning trial predicts postextubation distress*. Crit Care Med. 2012;40(7):2064–72. 10.1097/CCM.0b013e31824e68. ae PubMed PMID: 22584759.
43. Zapata L, Vera P, Roglan A, Gich I, Ordonez-Llanos J, Betbesé AJ. B-type natriuretic peptides for prediction and diagnosis of weaning failure from cardiac origin. Intensive Care Med. 2011;37(3):477–85. 10.1007/s00134-010-2101-4.
44. Dessap AM, Roche-Campo F, Kouatchet A, Tomicic V, Beduneau G, Sonneville R, et al. Natriuretic Peptide–driven Fluid Management during Ventilator Weaning: A Randomized Controlled Trial. Am J Respir Crit Care Med. 2012;186(12):1256–63. 10.1164/rccm.201205-0939OC.
45. Smina M, Salam A, Khamiees M, Gada P, Amoateng-Adjepong Y, Manthous CA. Cough Peak Flows and Extubation Outcomes. Chest. 2003;124(1):262–8. 10.1378/chest.124.1.262.
46. Duan J, Zhang X, Song J. Predictive power of extubation failure diagnosed by cough strength: a systematic review and meta-analysis. Crit Care. 2021;25(1):357. 10.1186/s13054-021-03781-5.
47. Blanch L, Villagra A, Sales B, Montanya J, Lucangelo U, Luján M, et al. Asynchronies during mechanical ventilation are associated with mortality. Intensive Care Med. 2015;41(4):633–41. 10.1007/s00134-015-3692-6.
48. Qian Z, Yang M, Li L, Chen Y. Ultrasound assessment of diaphragmatic dysfunction as a predictor of weaning outcome from mechanical ventilation: a systematic review and meta-analysis. BMJ Open. 2018;8(9):e021189. 10.1136/bmjopen-2017-021189.

Figures

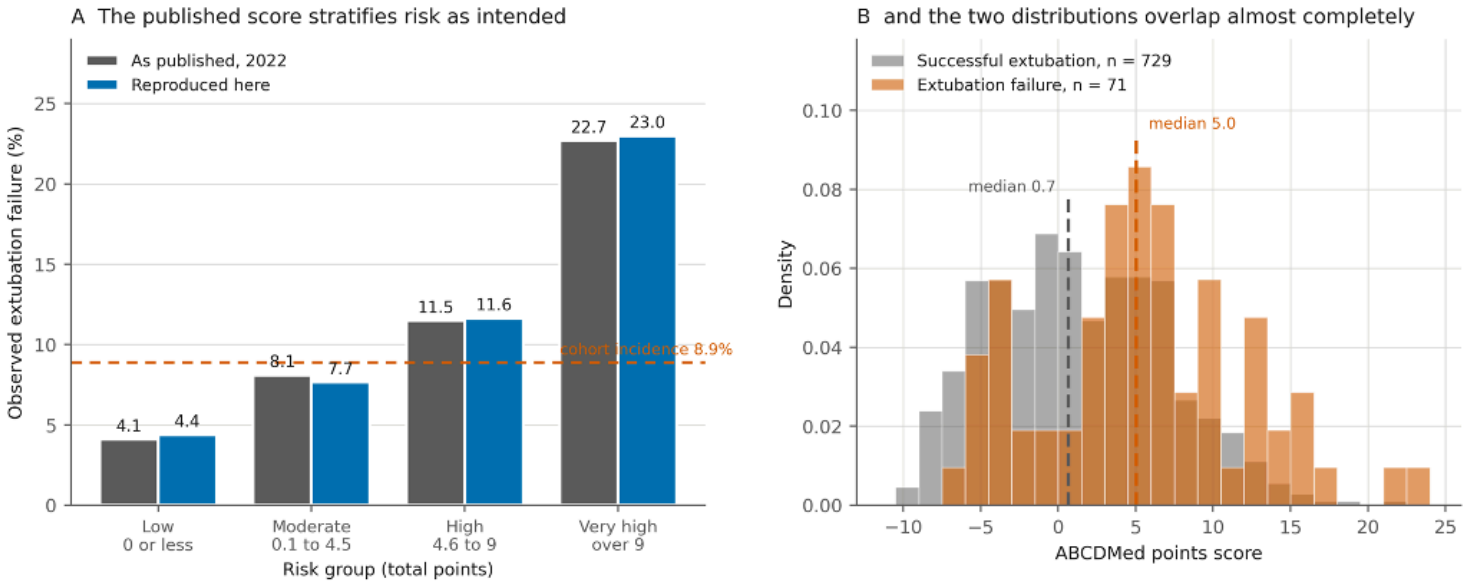


Figure 1

The published ABCDMed score. (A) Observed extubation failure in the four published risk strata, as reported in the original derivation and as reproduced in the analysis dataset. (B) Distribution of the points score in patients whose extubation failed and in those whose extubation succeeded, with medians marked. The score stratifies risk as intended, and the distributions overlap almost completely.

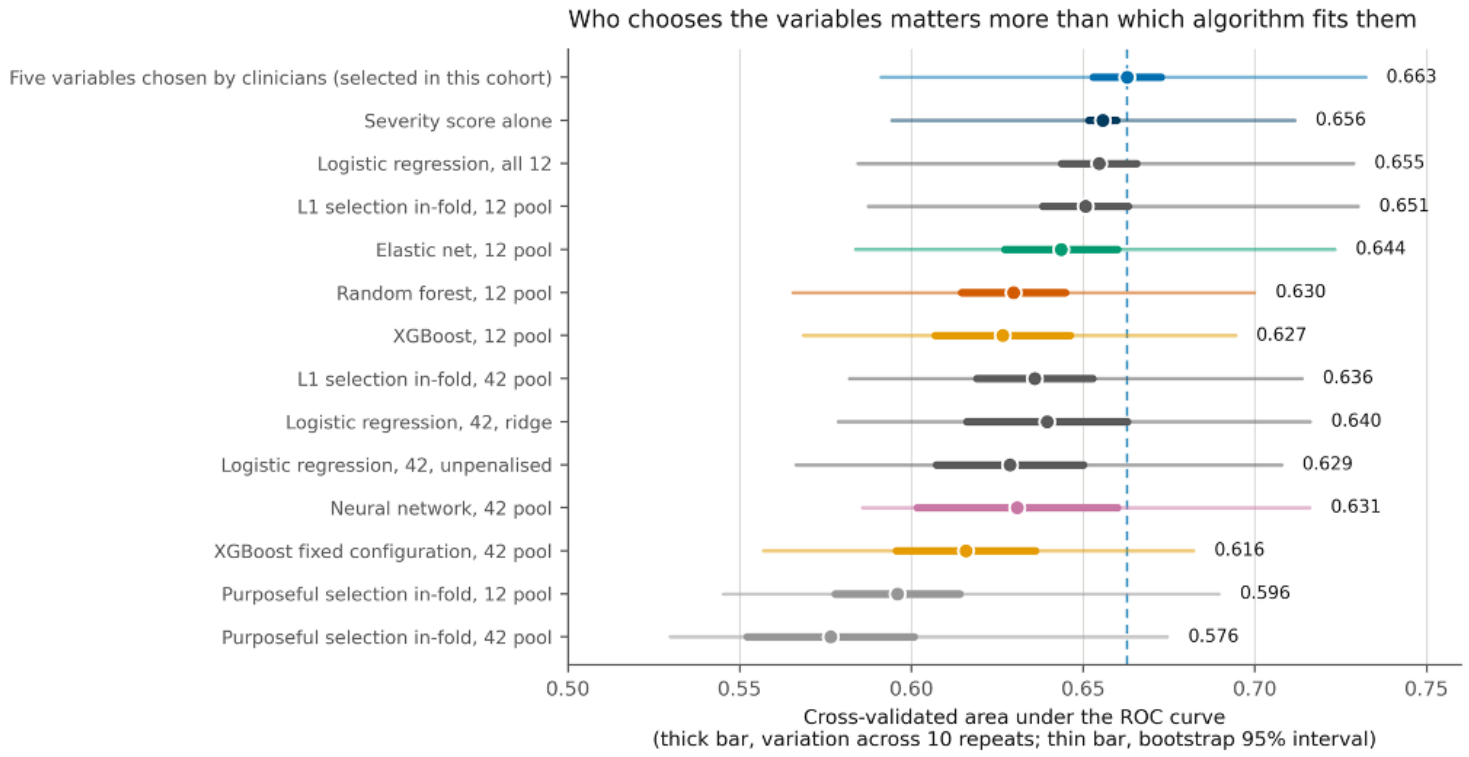


Figure 2

Cross-validated discrimination under the matched protocol, in which every arm discovers its variables inside the training folds from the same candidate pool. Points are the mean AUC across 10 repeats of stratified 5 fold cross validation; the thick bar is one standard deviation across repeats and the thin bar is the bootstrap 95% interval. Blue denotes the clinician-specified set and the single severity variable, grey the regression strategies with in-fold selection, and colour the flexible methods.

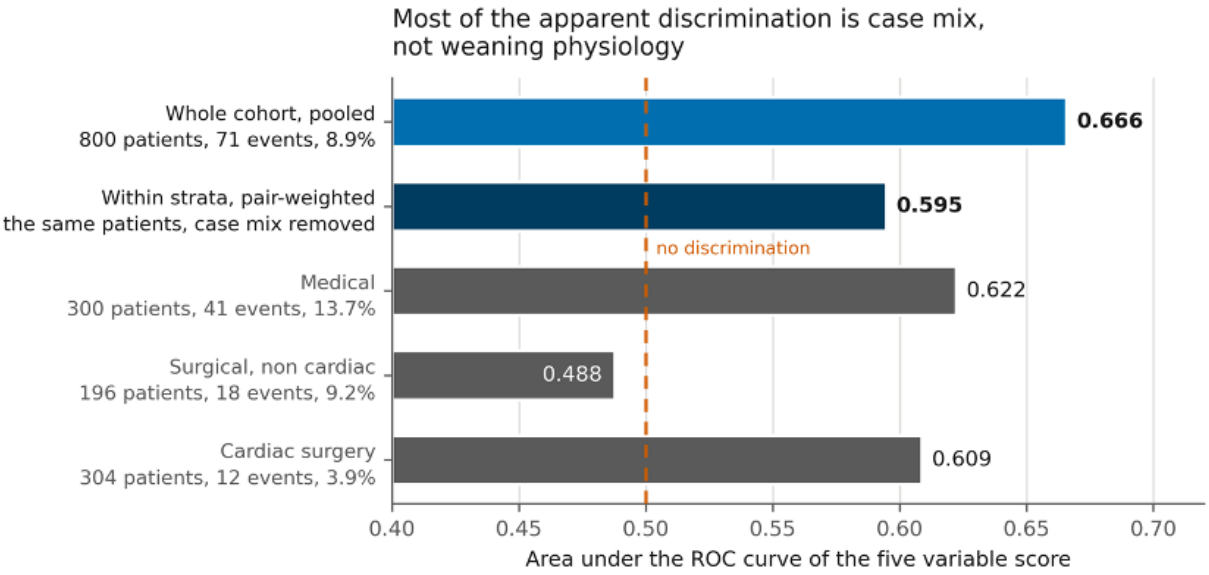


Figure 3

Discrimination of the published specification within case-mix strata and in the pooled cohort. The within-stratum value is weighted by the number of discordant pairs in each stratum. A clinician at the bedside already knows the stratum.

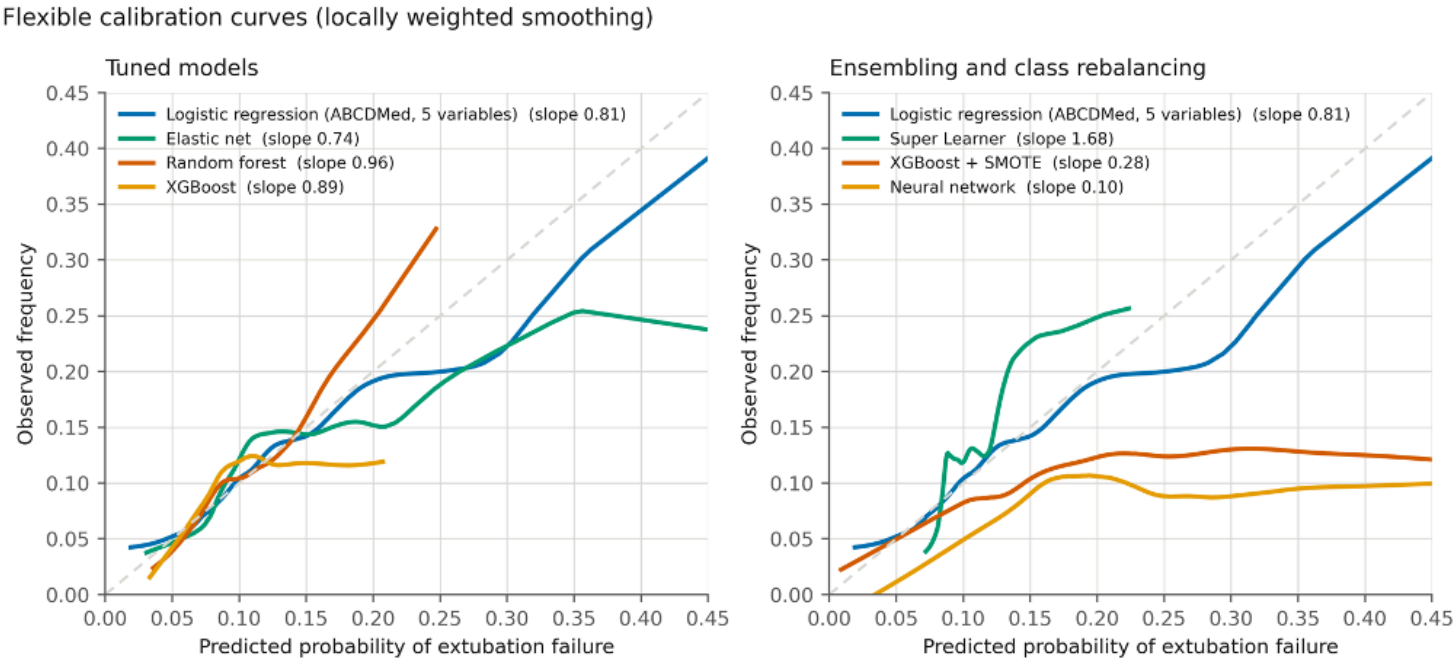


Figure 4

Flexible calibration curves obtained by locally weighted smoothing of out of fold predictions. Left: the published specification and three tuned models. Right: ensembling and class rebalancing. The diagonal is perfect calibration; calibration slopes are given in the legend. Curves below the diagonal indicate predicted risk that is too high.

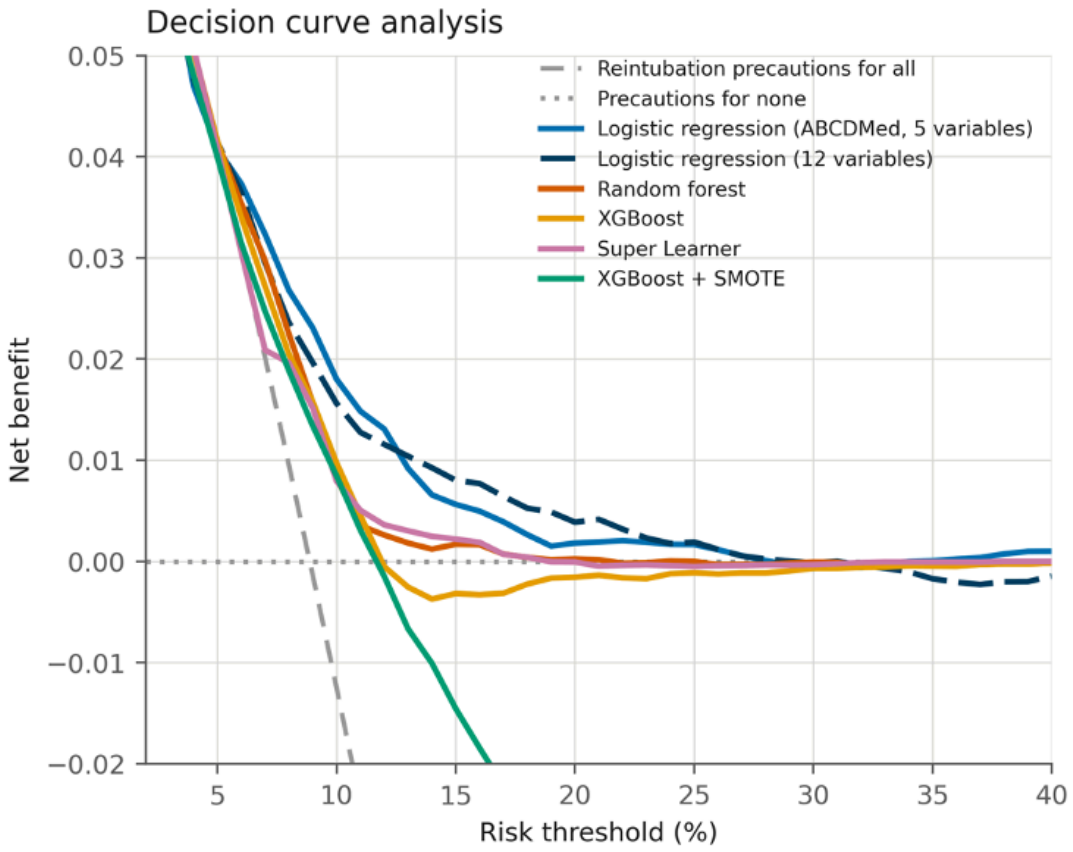


Figure 5

Decision curve analysis. Net benefit of each strategy relative to applying post-extubation precautions to all patients (dashed grey) or to none (dotted grey). Interpretation is restricted to thresholds of 5% to 15%, the decision-relevant range at an incidence of 8.9%.

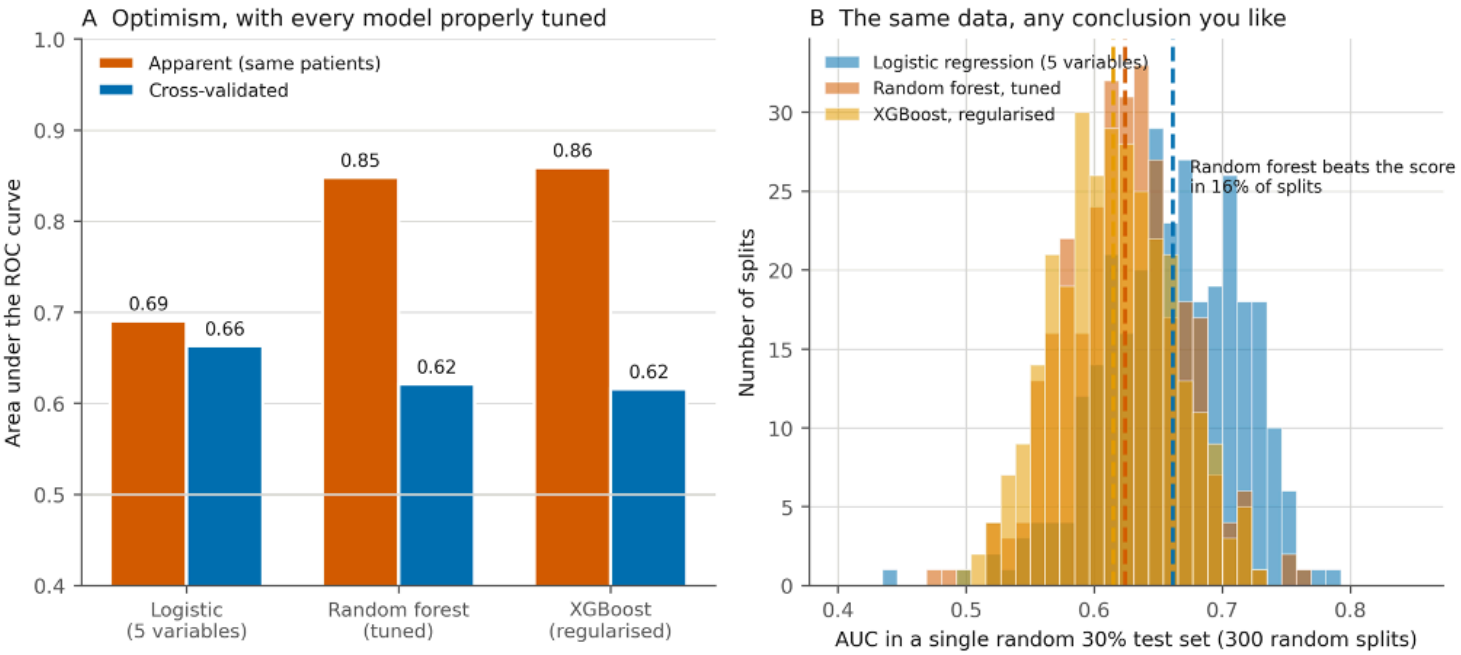


Figure 6

Two standard practices that would have reversed the conclusion, with every model properly tuned. (A) Apparent performance, obtained by evaluating each model on the same patients used to fit it, against cross validated performance. (B) Distribution of the test set AUC across 300

random 70/30 stratified splits of the same cohort. Dashed lines mark the means.

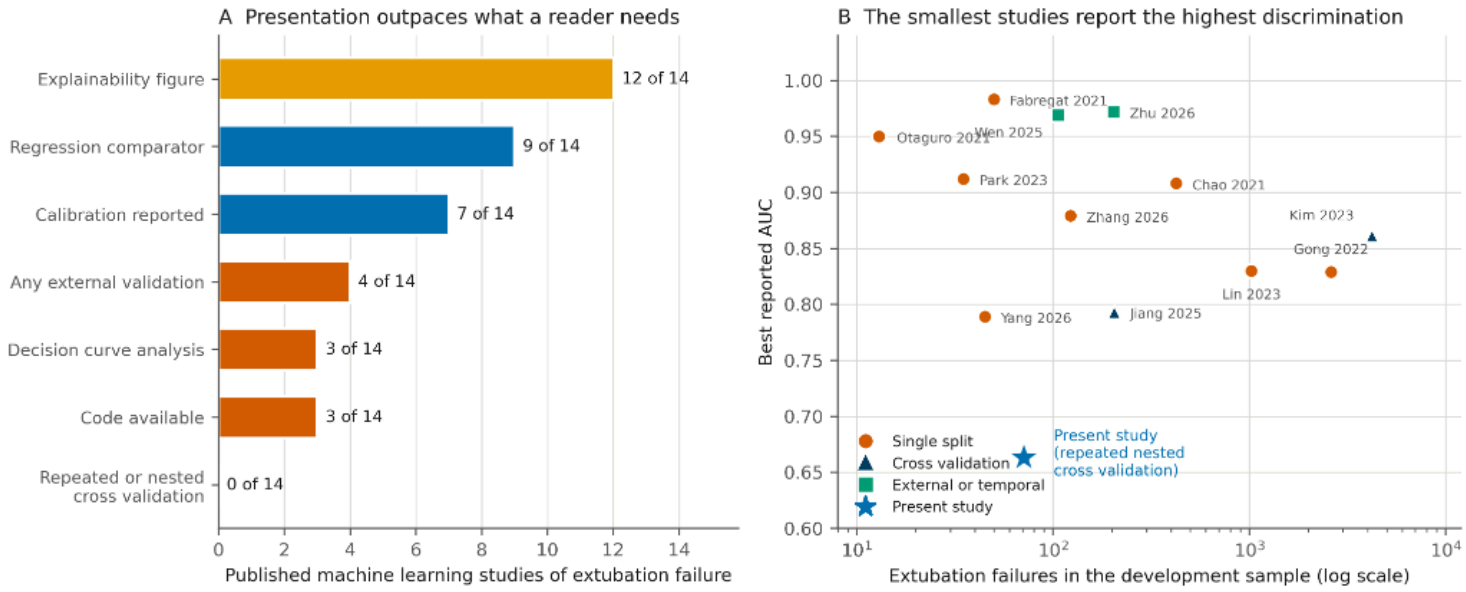


Figure 7

The published machine learning literature on extubation failure (14 studies). (A) Presence of an explainability figure against the six elements that allow a reader to judge clinical usefulness. (B) Best reported AUC against the number of extubation failures in the development sample, by validation design, with the present study marked.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DatabaseAnonymized.xlsx
- SA.docx