Journal of Microbiology, Immunology and Infection 58 (2025) 94–102

Contents lists available at ScienceDirect

Journal of Microbiology, Immunology and Infection

journal homepage: www.e-jmii.com

Prediction model, risk factor score and ventilator-associated pneumonia: A two-stage case-control study

Hua Meng [a,1], Yuxin Shi [a,1], Kaming Xue [b,1], Di Liu [c,1], Xiongjing Cao [a], Yanyan Wu [a], Yunzhou Fan [a], Fang Gao [a], Ming Zhu [a], Lijuan Xiong [a,*]

[a] Department of Nosocomial Infection Management, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[b] Department of Integrated Traditional Chinese and Western Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[c] Interventional Diagnostic and Therapeutic Center, Zhongnan Hospital of Wuhan University, Wuhan, China

ARTICLE INFO

Keywords:
Ventilator-associated pneumonia
Prediction model
Risk factor scoring system
Machine-learning
Case-control study

ABSTRACT

Background: Ventilator-associated pneumonia (VAP) is one of the most important hospital acquired infections in patients requiring mechanical ventilation (MV) in the intensive care unit, but the effective and robust predictable tools for VAP prevention were relatively lacked.
Methods: This study aimed to establish a weighted risk scoring system to examine VAP risk among a two-stage VAP case-control study, and to evaluate the diagnostic performance of risk factor score (RFS) for VAP. We constructed a prediction model by least absolute shrinkage and selection operator (LASSO), random forest (RF), and extreme gradient boosting (XGBoost) models in 363 patients and 363 controls, and weighted RFS was calculated based on significant predictors. Finally, the diagnostic performance of the RFS was testified and further validated in another 177 pairs of VAP case-control study.
Results: LASSO, RF and XGBoost consistently revealed significant associations of length of stay before MV, MV time, surgery, tracheotomy, multiple drug resistant organism infection, C-reactive protein, PaO_2, and APACHE II score with VAP. RFS was significantly linearly associated with VAP risk [odds ratio and 95 % confidence interval = 2.699 (2.347, 3.135)], and showed good discriminations for VAP both in discovery stage [area under the curve (AUC) = 0.857] and validation stage (AUC = 0.879).
Conclusions: Results of this study revealed co-occurrence of multiple predictors for VAP risk. The risk factor scoring system proposed is a potentially useful predictive tool for clinical targets for VAP prevention.

1. Introduction

Ventilator-associated pneumonia (VAP) is one of the major causes of hospital acquired infections developing 48 h or more after receiving mechanical ventilation (MV), with high morbidity and mortality.[1] Previous clinical and histologic studies have shown about 5 %–40 % of patients put on MV developing VAP, and the mortality rate of VAP has been estimated at 10 %.[2,3] Compared to patients with similar diseases without VAP, patients with VAP face a longer duration of MV, thereby increasing hospital course and healthcare costs.[4,5] The incidence of VAP has become an important indicator of the quality of clinical diagnosis and treatment, infection prevention and control in intensive care units

Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; AUC, area under the curve; BMI, body mass index; CI, confidence interval; COPD, chronic obstructive pulmonary disease; CRAB, carbapenem-resistant Acinetobacter baumannii; CRE, carbapenem-resistant Enterobacteriaceae; CRP, C-reactive protein; CRPA, carbapenem-resistant *Pseudomonas aeruginosa*; DCA, decision curve analysis; ICU, intensive care unit; IQR, interquartile range; LASSO, least absolute shrinkage and selection operator; LOS, length of stay; LOS before MV, length of stay before mechanical ventilation; MDRO, multiple drug resistant organism; MRSA, methicillin-resistant *Staphylococcus aureus*; MSE, mean squared error; MV, mechanical ventilation; OR, odds ratio; $PaCO_2$, arterial partial pressure of carbon dioxide; PaO_2, arterial partial pressure of oxygen; RF, random forest; RFS, risk factor score; ROC, receiver operator characteristic curve; SaO_2, arterial oxygen saturation; SD, standard deviation; VAP, ventilator-associated pneumonia; VIF, variance inflation factor; XGBoost, extreme gradient boosting.

* Corresponding author. Department of Nosocomial Infection Management, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, 1277 JieFang Avenue, Wuhan, 430022, China.
E-mail address: lijuanxiong2016@126.com (L. Xiong).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jmii.2024.11.005
Received 4 June 2024; Received in revised form 21 September 2024; Accepted 12 November 2024
Available online 13 November 2024

(ICU).[6]

Although some VAP-related risk factors such as MV time and tracheotomy have been widely studied,[7,8] there is no established risk scoring system that integrates these factors to guide for VAP prevention. In addition, there is no VAP screening tool to facilitate patient risk stratification during the MV period. Identification of risk factors and establishment of prediction models associated with VAP could permit clinical staff to provide close infection surveillance and timely care for VAP.

Current studies on exploring risk factors of VAP mainly use traditional logistic regression, which neglects the mutual effect of multiple stressors.[9,10] With the development of the statistical methodology for multi-exposure models, some novel statistical methods have been proposed in epidemiologic studies for estimating the healthy effects, such as least absolute shrinkage and selection operator (LASSO) penalized regression,[11] random forest (RF), and extreme gradient boosting (XGBoost).[12] As the risk factors of VAP are not single, there is a need for identifying the independent predictors of VAP from a more comprehensive view of exposure-effect association, to capture and characterize the complexity of multiple risk factors.

In the present study, we first established a VAP prediction model by using three machine learning methods (LASSO, RF and XGBoost) in 363 pairs of VAP case-control study, and was presented as a nomogram. We also examined the association of a weighted risk factor score (RFS) based on significant predictors selected by the three models with VAP risk. Finally, the diagnostic performance of the weighted RFS for VAP was testified and further validated in another 177 pairs of VAP case-control study.

2. Materials and methods

2.1. Study population

We performed a two-stage VAP case-control study in a total of 1080 Chinese participants. The discovery stage comprised 363 VAP patients and 363 age- (±3 years-old) and gender-matched controls. The validation stage consisted of 177 VAP cases and 177 age- (±3 years-old) and gender-matched normal controls. All subjects were collected during January 2019 to December 2019 and January 2022 to December 2022, respectively, in general ICUs of two tertiary hospitals in Wuhan, China. Patients and controls who were at least 18 years-old, had been admitted to hospital and received MV for at least 48 h were included in the study. However, patients who developed VAP infection 48 h within intubation and those who with VAP at admission were excluded from the study. VAP was suspected in a patient that developed a new or persistent infiltrate on chest radiography, and with one of the following criteria: (1) purulent tracheal secretions, (2) fever, (3) increased or decreased white blood cell count. The diagnosis of VAP was confirmed by positive quantitative culture of a respiratory sample, and was confirmed both by hospital infection management physicians and clinicians.[13,14] For each case, we matched one control that underwent MV at the same ICUs with cases during the study period and without postoperative VAP, thus ensuring that controls were drawn from the same source population as cases. Matching of controls was performed addressing each patient only once, if more than one control was available per case, we randomly selected one out of the matched eligible population.

Furthermore, VAP infections were defined as early if they occurred within 7 days after hospitalization, and as late if they occurred after 7 days.[13] The Ethics Committee of Union hospital, Tongji Medical College, Huazhong University of Science and Technology approved the study (Identifier: 2023 No. 0792), and all participants provided informed consent.

2.2. Data collection and definition

Data were retrospectively collected from the hospital's complete electronic medical record and the surgical paper chart during the general ICU stay. Variables assessed included basic demographics, lifestyle factors, medical history, surgery, medication and laboratory measurements. Baseline clinical and biochemical indices were recorded as the first examination after admission. Surgery, tracheostomy, microbiological and antibiotic treatment information were collected before the diagnosis of VAP for case groups. APACHE II score was measured by the clinician within 24 h of patient's admission to ICU, and it is also before the diagnosis of VAP. To assess length of stay (LOS) and MV time as risk factors, we compared the interval from admission to MV (i.e., LOS before MV) for all subjects, and the duration from the start of MV to VAP diagnosis among cases to the overall duration of ventilation among controls, respectively. Multiple drug resistant organism (MDRO) was defined as non-susceptibility to three or more antimicrobial categories,[15] and included methicillin-resistant *Staphylococcus aureus* (MRSA), carbapenem-resistant Enterobacteriaceae (CRE), Vancomycin-resistant Enterococcus faecium, carbapenem-resistant *Pseudomonas aeruginosa* (CRPA), and carbapenem-resistant Acinetobacter baumannii (CRAB). More detailed definition or basis of variables were described in Table S1.

2.3. VAP prediction model development

We first conducted a systematic literature review of previous published VAP prediction models and evidence for their application in VAP diagnosis, aiming to identify candidate predictor variables. Then, the available clinical and laboratory data of all subjects were collected. Variables with missing value $>$ 30 % were excluded in the database. Finally, we identified 30 possible variables to be included in the modeling process.

In the discovery stage, variance inflation factor (VIF) analysis was first used to analyze the collinearity of thirty variables, and the most colinear factor was deleted until no collinearity existed. As Zhang et al. suggested,[16] factors with VIF $>$5 were excluded. Second, three different models (LASSO regression, RF, and XGBoost) were used for selection of significant variables.

2.3.1. LASSO regression

LASSO penalized regression analysis with a logit link and the Gaussian family was used to select independent predictors highly associated with risk of VAP. The optimal value of λ was chosen by using 10-fold cross-validation.[17] LASSO was used to improve the accuracy of the logistic model and avoid over-fitting by penalizing coefficients with large values. The logistic model derived from LASSO was reduced most of the coefficients to zero, and the features with non-zero coefficients were essential for predicting the target variables or patient labels.[18]

2.3.2. RF regression

RF uses bagging with random feature selection, and the randomness in the structure is useful for decreasing the variance of the model and the prediction is made by averaging the forecasts of the trees employed.[19] The training procedure was employed as follows: (1) from the training dataset (70 % of all subjects in the discovery stage), a bootstrap sample was drawn as a randomized subset; (2) each individual tree was grown using the randomized subset of predictor variables; (3) repeat the step (2) until the number of trees was grown. Then the predicted results were aggregated by averaging them. The top 10 features were identified as significant predictors. Discriminations of the models both for training set and testing set were assessed using the area under the curve (AUC).

2.3.3. XGBoost

XGBoost is an extension of the gradient boosting algorithm that uses a different objective function and regularization techniques to improve generalization performance and prevent overfitting.[20] As the tree structure, the final prediction was calculated by summing up the scores across all leaves. All subjects were randomly split into a training set (70

%) and an internal validation set (30 %) to execute XGBoost, which was the same as RF regression. Meanwhile, the top 10 features were recognized as significant predictors. Discriminations of the models both for training set and testing set were also assessed using the AUC.

Finally, the intersection predictors of the above three models were used to establish prediction model of VAP. We also used conditional logistic regression model to assess the variable effects of selected predictors for the 1:1 paired case-control design.[21] Then, VAP prediction nomogram was used to visually assess the probability of independent risk factors for predicting the occurrence of VAP. The calibration curve and the Harrell C-index were used to evaluate the calibration of VAP prediction model and the discrimination performance, respectively. Furthermore, decision curve analysis (DCA) was used to estimate the clinical usefulness of the VAP prediction model by quantifying the net benefits at different threshold probabilities.[22]

Table 1
Characteristics of the subjects included in the discovery and validation stage.

Variables	Discovery stage			Validation stage		
	Cases (n = 363)	Controls (n = 363)	P^a	Cases (n = 177)	Controls (n = 177)	P^a
Age, years-old	58.5 ± 13.9	58.5 ± 13.8	0.928	50.3 ± 15.9	50.4 ± 15.7	0.946
Gender			1.000			1.000
Males	239 (65.8)	239 (65.8)		133 (75.1)	133 (75.1)	
Females	124 (34.2)	124 (34.2)		44 (24.9)	44 (24.9)	
Smoking status			0.637			0.512
Ever smokers	125 (34.4)	118 (32.5)		64 (36.2)	71 (40.1)	
Never smokers	238 (65.6)	245 (67.5)		113 (63.8)	106 (59.9)	
Alcohol drinking status			1.000			0.109
Ever drinkers	110 (30.3)	109 (30.0)		63 (35.6)	48 (27.1)	
Never drinkers	253 (69.7)	254 (70.0)		114 (64.4)	129 (72.9)	
BMI, kg/m^2	23.6 ± 4.3	24.6 ± 9.9	0.076	24.1 ± 4.0	23.8 ± 4.4	0.614
Chronic disease						
COPD	8 (2.2)	12 (3.3)	0.496	6 (3.4)	4 (2.3)	0.748
Emphysema	6 (1.7)	10 (2.8)	0.448	4 (2.3)	1 (0.6)	0.368
CHD	63 (17.4)	63 (17.4)	1.000	22 (12.4)	22 (12.4)	1.000
Hypertension	187 (51.5)	167 (46.0)	0.158	81 (45.8)	81 (45.8)	1.000
Diabetes mellitus	60 (16.5)	52 (14.3)	0.472	29 (16.5)	41 (23.2)	0.149
Cancer	29 (8.0)	36 (9.9)	0.435	15 (8.5)	11 (6.2)	0.541
LOS before MV, days	3.6 (0.7, 10.9)	6.3 (0.9, 13.1)	**0.009**	1.3 (0.2, 9.3)	4.3 (0.7, 14.2)	**0.001**
MV time, days	7.5 (5.8, 11.1)	5.0 (3.0, 7.0)	**<0.001**	5.1 (3.4, 7.8)	4.0 (3.0, 7.0)	**0.011**
Surgery			**0.047**			0.573
Yes	260 (71.6)	234 (64.5)		115 (65.0)	121 (68.4)	
No	103 (28.4)	129 (35.5)		62 (35.0)	56 (31.6)	
Tracheotomy			**<0.001**			**0.002**
Yes	110 (30.3)	49 (13.5)		65 (36.7)	37 (20.9)	
No	253 (69.7)	314 (86.5)		112 (63.3)	140 (79.1)	
Combination of antibiotics			**<0.001**			0.572
Yes	349 (96.1)	316 (87.1)		172 (97.2)	169 (95.5)	
No	14 (3.9)	47 (12.9)		5 (2.8)	8 (4.5)	
MDRO infection			**<0.001**			**<0.001**
Yes	276 (76.0)	102 (28.1)		132 (74.6)	69 (39.0)	
No	87 (24.0)	261 (71.9)		45 (25.4)	108 (61.0)	
Clinical biochemical index						
Hemoglobin, g/L	118.00 (98.00, 134.00)	116.00 (95.00, 132.00)	0.188	112.00 (89.00, 130.00)	114.00 (93.00, 127.00)	0.961
Platelet, g/L	166.00 (117.50, 215.00)	175.00 (128.00, 231.00)	0.213	166.00 (110.00, 227.00)	169.00 (118.00, 222.00)	0.717
CRP, mg/L	53.60 (8.78, 143.00)	49.85 (3.91, 128.00)	0.101	69.70 (15.20, 155.00)	51.40 (11.10, 127.00)	0.073
Total protein, g/L	59.30 (53.70, 64.35)	59.60 (54.40, 65.25)	0.385	58.60 (50.10, 64.90)	58.40 (53.30, 63.40)	0.627
Albumin, g/L	34.70 (29.60, 38.35)	35.10 (29.15, 38.70)	0.676	32.80 (27.80, 38.00)	33.70 (29.30, 37.40)	0.441
Creatinine, μmol/L	77.65 (61.33, 120.62)	74.00 (61.10, 103.00)	0.244	78.90 (64.70, 121.00)	78.10 (60.70, 120.80)	0.705
Fasting blood-glucose, mmol/L	6.15 (4.90, 8.20)	6.16 (4.97, 7.54)	0.847	6.65 (5.10, 8.28)	6.10 (4.90, 8.00)	0.096
Na, mmol/L	140.80 (138.40, 142.95)	140.80 (138.10, 143.00)	0.902	140.10 (137.70, 143.20)	140.90 (138.00, 143.30)	0.385
K, mmol/L	3.99 (3.64, 5.25)	4.01 (3.71, 4.36)	0.265	4.04 (3.69, 4.36)	3.94 (3.66, 4.38)	0.443
Arterial blood gas						
pH	7.40 (7.35, 7.45)	7.41 (7.34, 7.45)	0.595	7.38 (7.34, 7.45)	7.40 (7.34, 7.45)	0.423
$PaCO_2$, mmHg	36.60 (32.00, 42.40)	35.70 (31.05, 41.70)	0.264	38.10 (32.83, 43.93)	36.20 (32.00, 42.75)	0.158
PaO_2, mmHg	97.15 (74.95, 141.75)	111.00 (83.22, 155.00)	**0.001**	100.00 (82.75, 135.50)	99.55 (78.28, 142.00)	0.732
SaO_2, %	98.40 (95.40, 100.00)	99.00 (96.10, 100.00)	**0.032**	99.00 (97.00, 100.00)	98.60 (96.00, 100.00)	0.614
Serum bicarbonate, mmol/L	22.35 (18.90, 26.00)	21.90 (18.35, 25.70)	0.974	22.95 (19.27, 24.95)	21.00 (16.80, 25.10)	0.133
APACHE II score	24.00 (20.00, 29.00)	22.00 (15.25, 26.00)	**0.006**	23.00 (20.00, 27.00)	18.00 (13.75, 23.00)	**0.002**
Following outcome						
LOS, days	30.0 (19.8, 41.6)	19.9 (11.3, 31.8)	**<0.001**	25.8 (16.8, 42.9)	21.0 (15.2, 35.0)	**0.027**
Type of VAP						
Early VAP	120 (33.1)			69 (39.0)		
Late VAP	243 (66.9)			108 (61.0)		

Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; BMI, body mass index; CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; CRP, C-reactive protein; LOS, length of stay; LOS before MV, length of stay before mechanical ventilation; MDRO, multiple drug resistant organism; MV, mechanical ventilation; $PaCO_2$, arterial partial pressure of carbon dioxide; PaO_2, arterial partial pressure of oxygen; SaO_2, arterial oxygen saturation; VAP, ventilator-associated pneumonia.

Note.

[a] *P* values were calculated by using Student's t-test or nonparametric test for continuous variables and Chi-square test for categorical variable.

2.4. Diagnostic performance of RFS for VAP risk

We estimate the joint effect of the above selected VAP risk factors based on weighted RFS. For all predictors with significant association with risk of VAP, the weighted RFS was calculated as the sum of their coefficients with VAP risk multiplied by values. The calculation formula was listed as follows:

$$\text{weighted RFS} = \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \ldots + \beta_nX_n$$

where β represented the estimated regression coefficient of the predictors derived from the conditional logistic regression analysis, and X represented the value of each predictor. The weighted RFS were $\log_2$-transformed to approximate normal distribution. The association between RFS and risk of VAP was assessed by conditional logistic regression model and restricted cubic spline analysis, with adjustment for age, gender. Furthermore, to test the robustness of the association, we conducted stratification analyses according to age (≤60 years-old, >60 years-old), gender, and BMI (≤25 kg/m^2, >25 kg/m^2), and interaction effects were carried out by including a multiplicative interaction term (RFS × stratification variable) into the above model.

Finally, the diagnostic performance of the weighted RFS for all, early, and late VAP was further assessed based on the AUC in receiver operator characteristic (ROC) curve in the discovery stage, and further validated in the validation stage.

2.5. Statistical analysis

Data were presented as mean ± standard deviation (SD) or median (interquartile range, IQR) for continuous variables and number (percent) for categorical variables. Differences between VAP cases and controls were compared using T test or Mann-Whitney *U* test for continuous variables, and Fisher's exact test or χ^2 test for categorical variables, as appropriate. The features were considered as odds ratio (OR) having 95 % confidence interval (CI) and a *P* value. Differences with 2-sided *P* values < 0.05 were considered as statistically significant.

In this study, the statistical analyses were performed with R software (version 3.6.0, The R Foundation for Statistical Computing), and SAS program (version 9.4, SAS Institute, Carry, NC).

3. Results

3.1. Characteristics of study populations

Among this study, 726 and 354 subjects were included in the discovery and validation stage, respectively. The general characteristics of all participants were presented in Table 1. In the discovery stage, the mean age for the VAP cases and controls were both 58.5 years-old, the demographic and baseline characteristics, chronic comorbidities, and clinical biochemical index were comparable between VAP cases and controls. The exceptions were a significantly shorter LOS before MV (3.6 *vs.* 6.3 days, $P = 0.009$), but longer MV time (7.5 *vs.* 5.0 days, $P < 0.001$), and higher APACHE II score (24.00 *vs.* 22.00, $P = 0.006$). The case group had higher proportion of surgery, tracheotomy, MDRO infection, combination of antibiotics, and lower baseline $PaO_2$and SaO_2 (all $P < 0.05$). As a result, the LOS was significantly longer in the case group than in the control group (30.0 *vs.*19.9 days, $P < 0.001$). Among 363 VAP cases, the number of early VAP and late VAP were 120 (33.1 %) and 243 (66.9 %), respectively.

In the validation stage, the mean age for the VAP cases and controls were 50.3 and 50.4 years-old, respectively. Compared to the control group, VAP patients also had significantly shorter LOS before MV (1.3 *vs.* 4.3 days, $P = 0.001$), but longer MV time (5.1 *vs.* 4.0 days, $P = 0.011$) and longer LOS (25.8 *vs.*21.0 days, $P = 0.027$), higher proportions of tracheotomy, MDRO infection, and higher APACHE II score (all $P < 0.05$, Table 1), and marginally higher baseline CRP ($P = 0.073$). Among 177 VAP cases, the number of early VAP and late VAP were 69 (39.0 %) and 108 (61.0 %), respectively. The flowchart of the study was shown in Fig. S1.

3.2. VAP prediction model development

Considering the multicollinearity of certain variables (Fig. S2), four factors (albumin, fasting blood-glucose, $PaCO_2$, and serum bicarbonate) were excluded owing to collinearity existed (VIF >5). 26 variables were initially included to perform further analysis, and their VIFs were shown at Table S2. As shown in Fig. 1A and B, LASSO regression revealed that 0.016 was the optimal value of λ, at which MDRO infection, surgery, tracheotomy, combination of antibiotics, APACHE II score, MV time, hemoglobin, creatinine, CRP, PaO_2, and LOS before MV were identified as candidate predictors. As shown in Fig. 1C, the top 10 essential features in RF regression were APACHE II score, MV time, MDRO infection, PaO_2, surgery, tracheotomy, CRP, LOS before MV, SaO_2, and Na; the RF model reached an AUC of 0.981 in training set and 0.975 in the testing set (Fig. S3A). Fig. 1D revealed that the 10 most essential features in XGBoost were APACHE II score, MV time, MDRO infection, tracheotomy, CRP, pH, LOS before MV, surgery, platelet, and PaO_2; the XGBoost reached an AUC of 1.000 in training set and 0.983 in the testing set (Fig. S3B). Eight variables (LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO_2, and APACHE II score) in the intersection set were finally selected as significant predictors for VAP risk (Fig. 1E). The model that incorporated the above independent predictors was developed and was presented as a risk nomogram (Fig. 2).

Furthermore, the calibration curve of the nomogram for predicting the risk of VAP in ICU patients demonstrated good agreement in this trial (Fig. S4A). The C-index for the prediction nomogram was 0.857 (95 % CI = 0.830, 0.884) for this clinical dataset, suggesting proper discrimination by the model. A DCA analysis for the VAP prediction nomogram was presented in Fig. S4B, which showed that if the threshold probability of a patient is set between 10 % and 93 %, then the use of this VAP prediction nomogram is more beneficial to patients compared with the extreme situation of diagnosing VAP in all patients or none. These findings indicated that the present model provided a higher net benefit across a reasonably wide range of threshold probabilities for predicting VAP development, and thus had good clinical utility.

Then, we fitted conditional logistic regression model to estimate the variable effects between the above significant predictors and VAP risk (Table S3). In the conditional logistic analysis, the positive associations of MV time, surgery, tracheotomy, MDRO infection, APACHE II score [OR (95 % CI) = 2.134 (1.410, 3.242), 2.568 (1.666, 4.001), 1.690 (1.049, 2.738), 6.642 (4.538, 9.827), and 1.399 (1.280, 1.539), respectively], and the negative associations of LOS before MV, PaO_2 [OR (95 % CI) = 0.636 (0.420, 0.960), 0.993 (0.989, 0.997), respectively] remained significant with VAP. However, the association of CRP with VAP risk was marginally significant in the conditional logistic analysis [(OR (95 % CI) = 1.002 (0.999, 1.003), $P = 0.081$].

3.3. Association of RFS with VAP risk

To estimate the joint effects of LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO_2, and APACHE II score on VAP, we calculated weighted RFS for the eight predictors. Each one-fold increase in RFS was significantly associated with 1.699-fold increased risk of VAP [(OR (95 % CI) = 2.699 (2.347, 3.135), $P < 0.001$, Table 2]. BMI marginally modified this association ($P_{interaction} = 0.082$). Restricted cubic spline regression analysis indicated a significant linear relationship between RFS and VAP risk ($P < 0.001$, Fig. 3), with a *p*-value of nonlinearity of 0.891.

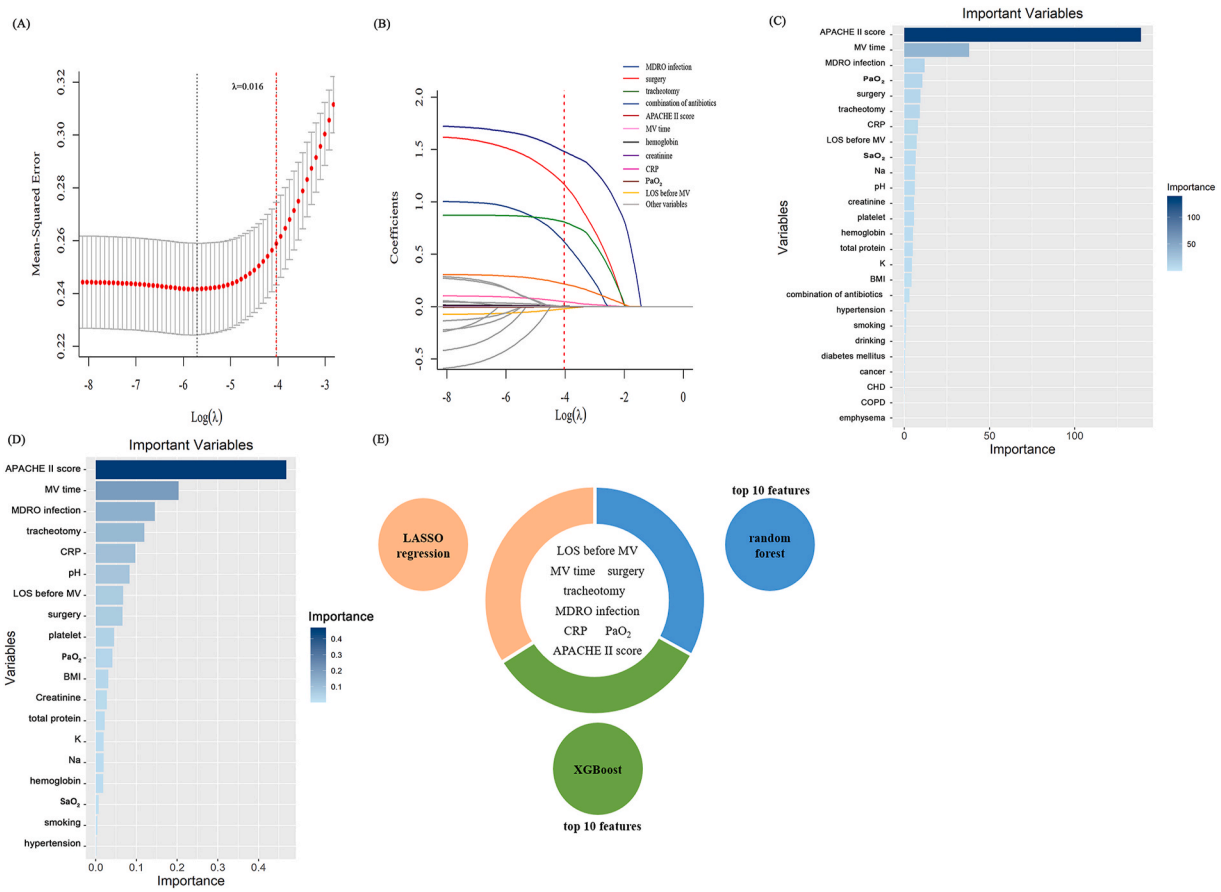

Fig. 1. The significant variables chosen by three models (A–D) and the intersection set of these variables (E). (A–B) The LASSO regression analysis for screening the risk factors of VAP. Note: The black dotted and its error bars represent the cross-validation curve at different values of λ. The red solid line represented the optimal value of λ. Adjustment factors included age, and gender. (C) The importance ranking of variables identified by RF analysis. (D) The importance ranking of variables identified by XGBoost analysis. (E) The intersection set of these models revealed eight variables (LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO_2, APACHE II score) were associated with VAP risk.

Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; CRP, C-reactive protein; LASSO, least absolute shrinkage and selection operator; LOS before MV, length of stay before mechanical ventilation; MDRO, multiple drug resistant organism; MSE, mean squared error; MV, mechanical ventilation; $PaCO_2$, arterial partial pressure of carbon dioxide; RF, random forest; VAP, ventilator-associated pneumonia; XGBoost, extreme gradient boosting.

3.4. Diagnostic performance of RFS for VAP risk

In the discovery stage, the RFS consistently demonstrated good predictive performance for all, early and late VAP patients [AUC (95 % CI) = 0.857 (0.830, 0.884), 0.877 (0.832, 0.922), 0.847 (0.813, 0.882), Fig. 4A], and there was no difference of discrimination for early and late VAP ($P = 0.299$). Furthermore, we validated the diagnostic value of RFS among another independent VAP case-control study, and the model had similar classification performance for all, early, and late VAP [AUC (95 % CI) = 0.879 (0.843, 0.914), 0.879 (0.822, 0.937), 0.876 (0.830, 0.922), Fig. 4B], and there was no difference of discrimination for early and late VAP ($P = 0.930$).

4. Discussion

To our knowledge, this is the first study to develop a weighted risk scoring system to examine VAP by LASSO regression, RF regression, and XGBoost model among a two-stage VAP case-control design. The RFS comprised eight reproducible and part of modifiable predictors (identified by the above three models together) for VAP risk, which included LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO_2, and APACHE II score, and was significantly linearly related to VAP risk. More importantly, the prediction model of RFS showed good discriminative power for all, early, and late VAP, as indicated by the AUC of 85.7 %, 87.7 %, and 84.7 % in the discovery stage and replicated successfully with AUC of 87.9 %, 87.9 %, and 87.6 % among the validation stage.

VAP is a clinically important, potentially preventable complication of MV, with various related risk factors.[13,23] Each factor is not only an independent risk factor of VAP, but also has an influence on each other. Therefore, it is necessary to identify risk factors of VAP by using the models applicable with multiple exposures, which can account for potential confounding effects of multiple variables co-exposure. In the current study, LASSO, RF and XGBoost consistently indicated LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO_2, and APACHE II score as independent predictors for risk of VAP. Compared with the traditional logistic regression method that neglects the mixed and complex relationships between the variables, the novel LASSO regression is an advanced and robust variable selection algorithm, since it adds a penalization on the effect sizes in the objective function to perform the variable selection by shrinking irrelevant variables to zero;[18] RF is an integrated learning algorithm belonging to the tree model, which makes predictions and assesses feature importance by constructing multiple decision trees; while XGBoost has received widespread attention in clinical prediction model research due to its fast computational speed, strong generalization ability, and high prediction performance. Our study highlighted the importance of applying different and novel statistical methods for hazard assessment of complex

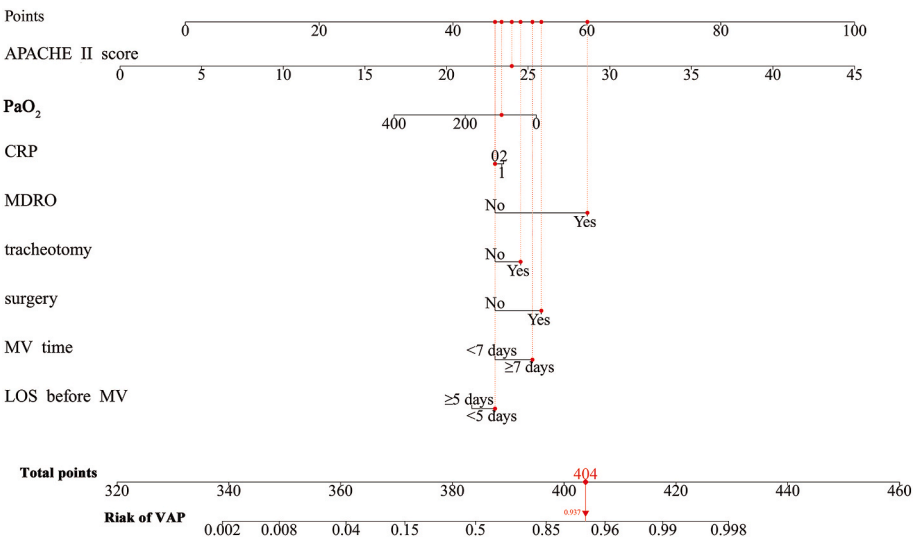

Fig. 2. A nomogram for predicting the risk of VAP.
Note: LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO2, and APACHE II score are eight variable axes, the sum of each variable axis is the total points, which correspond to the risk of VAP. The categories of CRP in "0", "1", "2" represented "<10 mg/L", "10–100 mg/L", and "≥100 mg/L", respectively. For example, a mechanically ventilated subject with LOS before MV < 5 days, MV time ≥7 days, MDRO infection, CRP <10 mg/L, APACHE II score of 24, PaO_2 of 97.4 mmHg, receiving tracheotomy and surgery, then the total score of VAP was 404, corresponding to the probability of VAP is 93.7 %.

Table 2
Associations of RFS with risk of VAP in the discovery stage.

Variables	n (cases/controls)	OR (95 % CI)[a]	*P*
All subjects	363/363	2.699 (2.347, 3.135)	<0.001
Age, years-old			
≤60	194/191	2.770 (2.291, 3.414)	<0.001
>60	169/172	2.619 (2.145, 3.269)	<0.001
$P_{interaction}$[b]			0.371
Gender			
Males	239/239	2.692 (2.273, 3.237)	<0.001
Females	124/124	2.727 (2.147, 3.568)	<0.001
$P_{interaction}$[b]			0.964
BMI, kg/m^2			
≤25	244/243	2.524 (2.144, 3.016)	<0.001
>25	119/120	3.132 (2.421, 4.199)	<0.001
$P_{interaction}$[b]			0.082

Abbreviations: CI, confidence interval; OR, odds ratio; RFS, risk factor score.
[a] Conditional logistic regression model with the log_2-transformed RFS as an independent variable and adjusted for age, gender (variables that were stratified were not adjusted).
[b] Interaction effects were carried out by including a multiplicative interaction term into the models for the stratification analysis.

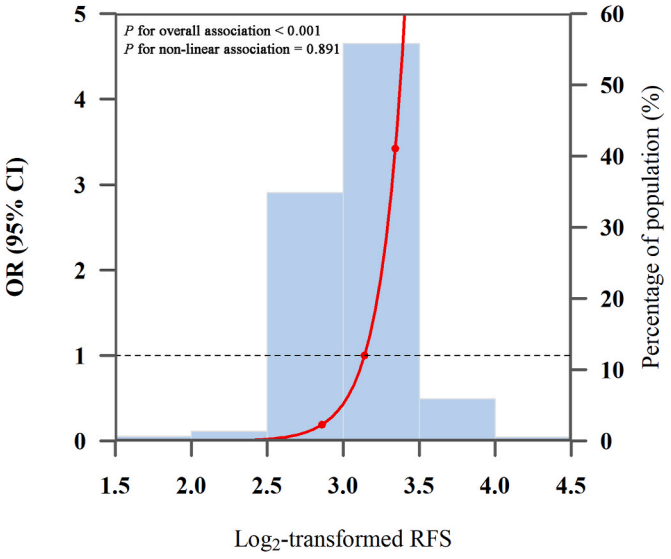

Fig. 3. The restricted cubic spline for the association between RFS and VAP risk.
Abbreviations: CI, confidence interval; OR, odds ratio; RFS, risk factor score.
Note: The red line represents adjusted OR, based on restricted cubic splines for the log2-transformed RFS in the logistic regression model. Knots were placed at the 25th, 50th, and 75th percentiles of the RFS distribution, and the reference value was set at the 50th percentile. Adjustment factors included age, gender. The bars represent histograms of RFS among the discovery stage.

exposures in epidemiological researches.

In this study, we calculated weighted RFS of the above nine significant predictors, and then evaluated its diagnostic performance for VAP risk. Of which, increased MV time and tracheotomy, as the internationally recognized factors,[24,25] are also identified to be associated with increased risk of VAP in our and other previous studies. However, Nseir et al. revealed tracheotomy as an independent protective factor for VAP, they found that tracheotomy was associated with 82 % reduction in risk of VAP [OR (95 % CI) = 0.18 (0.10, 0.30)],[26] which was contrary to our result, the reason may be that different timing of tracheotomy leads to difference in risk of VAP. Previous studies usually revealed positive relationship between total LOS with VAP risk, however in this study, we first found significant effect of LOS before MV on reducing risk of VAP, which means that it is very important to define ventilation duration when considering it as a predictor for VAP development. Few studies have explored the association of LOS before MV with VAP risk. Tobin et al. thought that selection of timing of intubation needed caution and it was not always better to intubate earlier.[27] Additionally, earlier need for intubation was more likely precipitated by severer underlying illness. In deciding when to intubate, clinicians must balance the risk of premature exposure to ventilation against the potential harms of unassisted breathing, including disease progression and worsening of multiorgan failure.[28]

MDRO, such as MRSA, CRE, CRAB and CRPA, have been categorized as urgent threats for antibiotic therapy by The World Health Organization.[29,30] MDRO infection may result in difficulty for VAP treatment and increased mortality.[2,31] Few previous studies have found that patients who have undergone surgery usually need long-time MV, and they

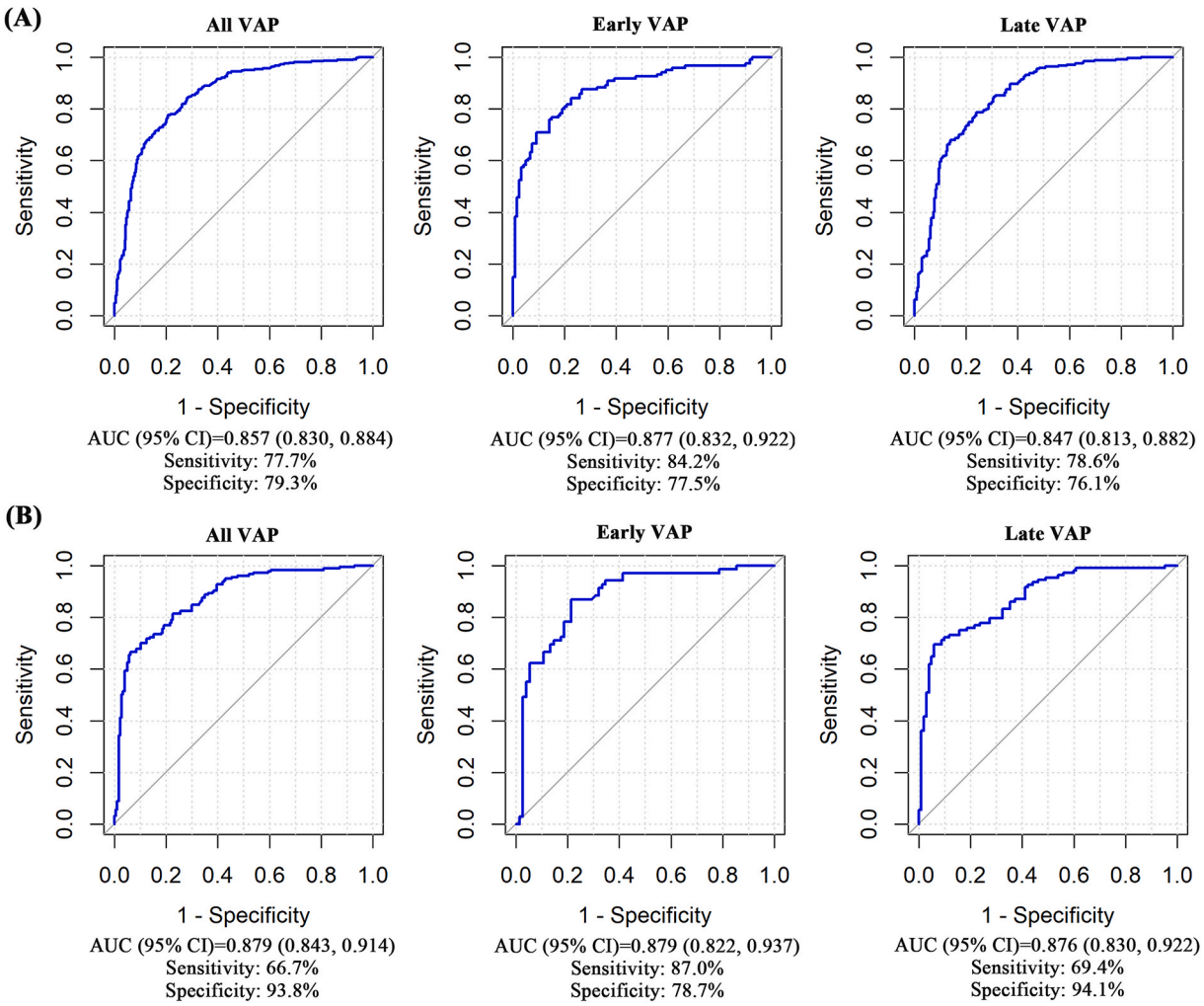

Fig. 4. Diagnostic performance of RFS for all, early, and late VAP in the discovery stage (A) and validation stage (B).
Note: AUC, area under the curve.

represent a special subpopulation at high risk for infection.[32] However, limited information exists regarding the clinical characteristics of VAP in this setting. APACHE II is a severity-of-disease classification system, and it was designed to measure the severity of disease for adult patients admitted to the ICU.[33,34] Zhou et al. revealed that APACHE II scores were significantly higher in VAP patients compared with non-VAP subjects (23.1 ± 4.8 *vs.* 16.7 ± 4.6; $P < 0.001$).[35] CRP belongs to a family of proteins named pentraxins, suggesting a central role in immunological response. CRP was a helpful biomarker for early identification of infection of VAP.[36,37] A multicenter retrospective study revealed increased CRP as a potential risk factor of pneumonia.[38] PaO_2 in arterial blood gas analysis represents the degree of hypoxia directly, and hypoxia state usually means more susceptible to lung infections by weakening immunity. Previous studies revealed that hypoxia was associated with VAP, and was one of the most frequently used clinical symptom and diagnostic criteria of VAP.[37,39]

Current practice in VAP prediction and prevention is reactive. We advocate for a more proactive approach that includes risk-based screening and early prevention to mitigate the development of VAP. For those patients receiving MV, the VAP risk scoring system serves as a functional predictive tool that provides guidance for early intervention to reduce VAP risk,[40] by methods such as trying to delay using intubation, shortening MV time, avoiding unnecessary tracheotomy, timely treatment of inflammation or hypoxia, and so on. Early preventive and anticipatory guidance may be the key to addressing the VAP epidemic and decreasing VAP-related mortality. An excellent prediction model would have beneficial consequence for clinical decision making by assisting in early identifying the risk of VAP.[41] In our current study, the prediction model established by RFS showed uniformly good discriminative power for both early and late VAP, with AUC of 87.7 % and 84.7 % in the discovery study and of 87.9 % and 87.6 % among the replicated study. Our findings provide convincing evidence that the prediction model improved prediction of VAP risk.

The current study has several strengths. First, this study was the first to establish a novel and modifiable risk factor scoring system by three models (LASSO, RF and XGBoost) in a two-stage VAP case-control study. Since the use of the weighted risk score method revealed results that considers the weights of different variables, improvements for the larger weights will yield better preventive effect. The association between a higher RFS and a higher VAP risk suggests the relevance and potential use of this risk scoring assessment tool in those receiving MV. On the other hand, the applications of novel machine learning models could overcome the statistical limitations of the traditional logistic regression approach. More importantly, the diagnostic performance of the prediction model was further validated successfully in another independent VAP case-control study. However, several limitations should also be acknowledged. First, the cross-sectional design of the study could not establish the causal inferences between risk factors and VAP risk. Future prospective studies are warranted to disclosure the biological causality. Second, previous studies have also predicted the risk of VAP, and some

of the predictors identified in this study have previously been associated with VAP.[32,42] However, our design based on two-stage case-control study and usage of three machine learning models are novel and robust. Third, VAP is known to lead to severe lung injury, other related factors that could reflect lung injury, such as driving pressure and mechanical power,[43] were not analyzed in the current research but need to be accessed in the further studies. Finally, our study only collected data for the duration of inpatient stay, which ignored relevant data related to patient outcomes after discharge, such as functional outcomes, 30-day morbidity, and mortality or related re-admissions.

5. Conclusion

In this two-stage VAP case-control study, we proposed a novel risk scoring tool for VAP that was based on eight risk factors, including LOS before MV, MV time, surgery, tracheotomy, MDRO infection, CRP, PaO_2, and APACHE II score, which was selected by LASSO, RF, and XGBoost models. The RFS was linearly associated with increased risk of VAP and could predict VAP risk, as indicated by the AUC of 85.7 % in the discovery stage and replicated successfully with AUC of 87.9 % among the validation stage. Our results may gain insights into the occurrence of VAP and reveal potential clinical targets for VAP prevention.

CRediT authorship contribution statement

Hua Meng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yuxin Shi:** Writing – original draft, Methodology, Formal analysis, Data curation. **Kaming Xue:** Writing – original draft, Conceptualization. **Di Liu:** Formal analysis, Data curation. **Xiongjing Cao:** Methodology, Formal analysis. **Yanyan Wu:** Investigation, Conceptualization. **Yunzhou Fan:** Investigation, Conceptualization. **Fang Gao:** Investigation, Conceptualization. **Ming Zhu:** Investigation, Conceptualization. **Lijuan Xiong:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Ethical approval statement

This study was approved by the Ethics Committee of Union hospital, Tongji Medical College, Huazhong University of Science and Technology (Identifier: 2023 No. 0792). All participants provided informed consent.

Funding source

This study was supported by the funds from the Hubei Provincial Public Health Leading Talent Construction Project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to appreciate all participants in this study as well as all volunteers for the effort they made in contributing to the current study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmii.2024.11.005.

References

1. Pozuelo-Carrascosa DP, Torres-Costoso A, Alvarez-Bueno C, Cavero-Redondo I, López Muñoz P, Martínez-Vizcaíno V. Multimodality respiratory physiotherapy reduces mortality but may not prevent ventilator-associated pneumonia or reduce length of stay in the intensive care unit: a systematic review. *J Physiother*. 2018;64(4):222–228.
2. Papazian L, Klompas M, Luyt CE. Ventilator-associated pneumonia in adults: a narrative review. *Intensive Care Med*. 2020;46(5):888–906.
3. Vacheron CH, Lepape A, Savey A, et al. Attributable mortality of ventilator-associated pneumonia among patients with COVID-19. *Am J Respir Crit Care Med*. 2022;206(2):161–169.
4. Thatrimontrichai A, Rujeerapaiboon N, Janjindamai W, et al. Outcomes and risk factors of ventilator-associated pneumonia in neonates. *World J Pediatr : WJP*. 2017;13(4):328–334.
5. Wang G, Ji X, Xu Y, Xiang X. Lung ultrasound: a promising tool to monitor ventilator-associated pneumonia in critically ill patients. *Critical care (London, England)*. 2016;20(1):320.
6. Nair GB, Niederman MS. Ventilator-associated pneumonia: present understanding and ongoing debates. *Intensive Care Med*. 2015;41(1):34–48.
7. Frost P, Wise MP. Tracheotomy and ventilator-associated pneumonia: the importance of oral care. *Eur Respir J*. 2008;31(1):221–222, 221; author reply.
8. Pozuelo-Carrascosa DP, Herráiz-Adillo Á, Alvarez-Bueno C, Añón JM, Martínez-Vizcaíno V, Cavero-Redondo I. Subglottic secretion drainage for preventing ventilator-associated pneumonia: an overview of systematic reviews and an updated meta-analysis. *Eur Respir Rev : An Off J European Respir Soc*. 2020;29(155).
9. Bayman EO, Dexter F. Multicollinearity in logistic regression models. *Anesth Analg*. 2021;133(2):362–365.
10. Yu L, Liu W, Wang X, et al. A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. *Environ pollut (Barking, Essex : 1987)*. 2022;306, 119356.
11. Dai L, Koutrakis P, Coull BA, Sparrow D, Vokonas PS, Schwartz JD. Use of the adaptive LASSO method to identify PM2.5 components associated with blood pressure in elderly men: the veterans affairs normative aging study. *Environ Health Perspect*. 2016;124(1):120–125.
12. Li S, Li M, Wu J, et al. Developing and validating a clinlabomics-based machine-learning model for early detection of retinal detachment in patients with high myopia. *J Transl Med*. 2024;22(1):405.
13. Kalil AC, Metersky ML, Klompas M, et al. Management of adults with hospital-acquired and ventilator-associated pneumonia: 2016 clinical practice guidelines by the infectious diseases society of America and the American thoracic society. *Clin Infect Dis : An Off Publ Infect Dis Soc America*. 2016;63(5):e61–e111.
14. Ibn Saied W, Merceron S, Schwebel C, et al. Ventilator-associated pneumonia due to Stenotrophomonas maltophilia: risk factors and outcome. *J Infect*. 2020;80(3):279–285.
15. Fernández J, Bert F, Nicolas-Chanoine MH. The challenges of multi-drug-resistance in hepatology. *J Hepatol*. 2016;65(5):1043–1054.
16. Zhang Y, Xiang T, Wang Y, et al. Explainable machine learning for predicting 30-day readmission in acute heart failure patients. *iScience*. 2024;27(7), 110281.
17. Le KT, Chaux C, Richard FJP, Guedj E. An adapted linear discriminant analysis with variable selection for the classification in high-dimension, and an application to medical data. *Comput Stat Data Anal*. 2020;152, 107031.
18. Tang G, Qi L, Sun Z, et al. Evaluation and analysis of incidence and risk factors of lower extremity venous thrombosis after urologic surgeries: a prospective two-center cohort study using LASSO-logistic regression. *Int J Surg*. 2021;89, 105948.
19. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794.
21. Cheng S, Lyu J, Shi X, et al. Rare variant association tests for ancestry-matched case-control data based on conditional logistic regression. *Briefings Bioinf*. 2022;23(2).
22. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj*. 2016;352, i6.
23. Fan Y, Gao F, Wu Y, Zhang J, Zhu M, Xiong L. Does ventilator-associated event surveillance detect ventilator-associated pneumonia in intensive care units? A systematic review and meta-analysis. *Critical care (London, England)*. 2016;20(1):338.
24. Hassoun-Kheir N, Hussein K, Abboud Z, et al. Risk factors for ventilator-associated pneumonia following cardiac surgery: case-control study. *J Hosp Infect*. 2020.
25. Chorath K, Hoang A, Rajasekaran K, Moreira A. Association of early vs late tracheostomy placement with pneumonia and ventilator days in critically ill patients: a meta-analysis. *JAMA Otolaryngol Head Neck Surg*. 2021;147(5):450–459.
26. Nseir S, Di Pompeo C, Jozefowicz E, et al. Relationship between tracheotomy and ventilator-associated pneumonia: a case control study. *Eur Respir J*. 2007;30(2):314–320.
27. Tobin MJ, Laghi F, Jubran A. Caution about early intubation and mechanical ventilation in COVID-19. *Ann Intensive Care*. 2020;10(1):78.
28. Lee KG, Roca O, Casey JD, et al. When to intubate in acute hypoxaemic respiratory failure? Options and opportunities for evidence-informed decision making in the intensive care unit. *Lancet Respir Med*. 2024;12(8):642–654.
29. Kanj SS, Bassetti M, Kiratisin P, et al. Clinical data from studies involving novel antibiotics to treat multidrug-resistant Gram-negative bacterial infections. *Int J Antimicrob Agents*. 2022;60(3), 106633.
30. Tacconelli E, Carrara E, Savoldi A, et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis*. 2018;18(3):318–327.

31. Kao HH, Peng CK, Sheu CC, et al. Mortality and ventilator dependence in critically ill patients with ventilator-associated pneumonia caused by carbapenem-resistant Acinetobacter baumannii. *J Microbiol Immunol Infect = Wei mian yu gan ran za zhi*. 2023;56(4):822–832.
32. He S, Chen B, Li W, et al. Ventilator-associated pneumonia after cardiac surgery: a meta-analysis and systematic review. *J Thorac Cardiovasc Surg*. 2014;148(6): 3148–3155. e3141-3145.
33. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Apache II: a severity of disease classification system. *Crit Care Med*. 1985;13(10):818–829.
34. Ishikawa D, Takehara Y, Takata A, Takamura K, Sato H. Combination of dirty mass volume and Apache II score predicts mortality in patients with colorectal perforation. *World J Emerg Surg*. 2021;16(1):17.
35. Zhou XY, Ben SQ, Chen HL, Ni SS. A comparison of Apache II and CPIS scores for the prediction of 30-day mortality in patients with ventilator-associated pneumonia. *Int J Infect Dis*. 2015;30:144–147.
36. Vigushin DM, Pepys MB, Hawkins PN. Metabolic and scintigraphic studies of radioiodinated human C-reactive protein in health and disease. *J Clin Invest*. 1993;91 (4):1351–1357.
37. Howroyd F, Chacko C, MacDuff A, et al. Ventilator-associated pneumonia: pathobiological heterogeneity and diagnostic challenges. *Nat Commun*. 2024;15(1): 6447.
38. Shi T, Yu Y, Shen Y, et al. Serum C-reactive protein to albumin ratio as a potential risk indicator of pneumonia caused by Chlamydia psittaci: a multicenter retrospective study. *Front Cell Infect Microbiol*. 2024;14, 1371625.
39. Fally M, Haseeb F, Kouta A, et al. Unravelling the complexity of ventilator-associated pneumonia: a systematic methodological literature review of diagnostic criteria and definitions used in clinical research. *Critical care (London, England)*. 2024;28(1):214.
40. Torres A, Niederman MS, Chastre J, et al. International ERS/ESICM/ESCMID/ALAT guidelines for the management of hospital-acquired pneumonia and ventilator-associated pneumonia: Guidelines for the management of hospital-acquired pneumonia (HAP)/ventilator-associated pneumonia (VAP) of the European Respiratory Society (ERS), European Society of Intensive Care Medicine (ESICM), European Society of Clinical Microbiology and Infectious Diseases (ESCMID) and Asociación Latinoamericana del Tórax (ALAT). *Eur Respir J*. 2017;50(3).
41. Ranzani OT, Motos A, Chiurazzi C, et al. Diagnostic accuracy of Gram staining when predicting staphylococcal hospital-acquired pneumonia and ventilator-associated pneumonia: a systematic review and meta-analysis. *Clin Microbiol Infect*. 2020;26 (11):1456–1463.
42. Wu Z, Liu Y, Xu J, et al. A ventilator-associated pneumonia prediction model in patients with acute respiratory distress syndrome. *Clin Infect Dis*. 2020;71(Suppl 4): S400–s408.
43. Cressoni M, Gotti M, Chiurazzi C, et al. Mechanical power and development of ventilator-induced lung injury. *Anesthesiology*. 2016;124(5):1100–1108.