Diagnostic Accuracy Study

Medicine®

OPEN

Predicting ventilator-associated pneumonia with machine learning

Christine Giang, BA, Jacob Calvert, MSc, Keyvan Rahmani, PhD, Gina Barnes, MPH, Anna Siefkas, SM*, Abigail Green-Saxena, PhD, Jana Hoffman, PhD, Qingqing Mao, PhD, Ritankar Das, MSc

Abstract

Ventilator-associated pneumonia (VAP) is the most common and fatal nosocomial infection in intensive care units (ICUs). Existing methods for identifying VAP display low accuracy, and their use may delay antimicrobial therapy. VAP diagnostics derived from machine learning (ML) methods that utilize electronic health record (EHR) data have not yet been explored. The objective of this study is to compare the performance of a variety of ML models trained to predict whether VAP will be diagnosed during the patient stay.

A retrospective study examined data from 6126 adult ICU encounters lasting at least 48 hours following the initiation of mechanical ventilation. The gold standard was the presence of a diagnostic code for VAP. Five different ML models were trained to predict VAP 48 hours after initiation of mechanical ventilation. Model performance was evaluated with regard to the area under the receiver operating characteristic (AUROC) curve on a 20% hold-out test set. Feature importance was measured in terms of Shapley values.

The highest performing model achieved an AUROC value of 0.854. The most important features for the best-performing model were the length of time on mechanical ventilation, the presence of antibiotics, sputum test frequency, and the most recent Glasgow Coma Scale assessment.

Supervised ML using patient EHR data is promising for VAP diagnosis and warrants further validation. This tool has the potential to aid the timely diagnosis of VAP.

Abbreviations: ARDS = acute respiratory distress syndrome, AUROC = the area under the receiver operating characteristic curve, BUN = blood urea nitrogen, CAP = community acquired pneumonia, CDC = centers for disease control, CPIS = clinical pulmonary infection score, GCS = Glasgow Coma Scale, EHR = electronic health record, ICD = International Classification of Diseases, ICU = intensive care unit, MIMIC = Multiparameter Intelligent Monitoring in Intensive Care, ML = machine learning, MV = mechanical ventilation, PIRO = predisposition, insult, response, organ dysfunction, ROC = receiver operating characteristic, RR = respiratory rate, SHAP = Shapley additive explanation, SpO_2 = oxygen saturation, VAP = ventilator-associated pneumonia.

Keywords: logistic regression, machine learning, prediction, ventilator-associated pneumonia

1. Introduction

Pneumonia is one of the deadliest infections across the globe.[1] It can be acquired in the community or hospital settings, as well as through the use of invasive mechanical ventilation.[1,2] Ventilator-associated pneumonia (VAP) is notoriously difficult to diagnose due to the absence of a diagnostic gold standard, which can be attributed to the diversity of disease-causing pathogens, lung cultures that limit sampling to anatomical surfaces, clinical interpretation of pathogens extracted in lung samples, underlying health issues in individuals, and agreement of radiological scans with centers for disease control (CDC) diagnostic criteria.[1–5] VAP, which is defined as pneumonia that develops after more than 48 hours following the initiation of invasive mechanical ventilation,[4] is estimated to occur in 5% to 67% of intubated patients, with the highest rates among those who are hospitalized due to physical trauma.[6,7] Mortality estimates range from 24% to 76%.[8] VAP is also associated with an estimated $47,238 in additional healthcare costs per stay in the United States.[9]

Given the high mortality rate with VAP and the vulnerability of the patient population at risk for VAP, prevention is crucial for clinical management to reduce the prevalence of VAP.[6,7] The CDC's evidence-based recommendations to prevent VAP encompass in-hospital safety protocols that limit the introduction of microbes into the lungs of intubated patients, such as opting for non-invasive mechanical ventilation methods and avoiding intubation when feasible, keeping levels of subglottic secretions low by frequent secretion removal, and maintaining intubated patients at an elevated chest position.[10]

Diagnostic methods for VAP vary and rely on a combination of factors, including clinical assessment for infectious signs and

Editor: Kuang-Ming Liao.

All authors who have affiliations listed with Dascena (Houston, Texas, USA) are employees or contractors of Dascena.

The authors have no funding and conflicts of interest to disclose.

Supplemental Digital Content is available for this article.

The datasets generated during and/or analyzed during the present study are publicly available.

Dascena, Inc., Houston, TX, United States.

* Correspondence: Gina Barnes, 12333 Sowden Rd Ste B PMB 65148, Houston, Texas 77080-2059, United States (e-mail: gbarnes@dascena.com).

How to cite this article: Giang C, Calvert J, Rahmani K, Barnes G, Siefkas A, Green-Saxena A, Hoffman J, Mao Q, Das R. Predicting ventilator-associated pneumonia with machine learning. Medicine 2021;100:23(e26246).

Received: 10 December 2020 / Received in final form: 16 April 2021 / Accepted: 2 May 2021

http://dx.doi.org/10.1097/MD.0000000000026246

symptoms (eg, fever), chest radiology, lung biopsy, and/or quantitative microbiological testing of respiratory secretions.[8,11–13] However, diagnostic accuracy for VAP continues to be poor. A recent meta-analysis by Fernando et al found that eight common clinical criteria for VAP diagnosis lacked specificity, highlighting the need for diagnostic tools which can better inform the need for and timing of antibiotic treatment while respecting antibiotic stewardship.[14]

Despite the shortcomings of existing clinical indicators and scoring tools for VAP diagnosis, there is minimal research on the application of machine learning (ML) methods using electronic health record (EHR) data to diagnose VAP. Studies of ML applications for pneumonia have focused largely on mortality prediction among known pneumonia cases. ML has been proposed as a component of other diagnostic tools, such as to assist with chest radiography interpretation,[15] to predict pneumonia outcomes,[16] or work with electronic sensors to detect pneumonia in exhaled breath.[17] Yet ML has not been robustly investigated to develop stand-alone diagnostic tools.[18–22] The dearth of such ML research may reflect the very reason for its urgent need – the lack of a clear diagnostic gold standard. The absence of such a standard complicates the identification of pneumonia onset in retrospective EHR data, making it difficult to develop a supervised ML approach to pneumonia prediction. For VAP, the difficulty in identifying onset time in retrospective data is partially alleviated by the constraints on onset time relative to the start of mechanical ventilation, as per the definition of VAP. It is possible to determine when inpatients have reached the 48th hour from the initiation of mechanical ventilation, the minimum required time for pneumonia to be designated as "ventilator-acquired." Using this definition, it is possible to develop and assess the ability of ML technology to accurately diagnose VAP in retrospective intensive care unit (ICU) data.

2. Objective

This exploratory analysis examined the suitability of ML methods for the prediction of VAP. To meet this goal, we developed and assessed a variety of ML approaches for the following two prediction tasks:

Intubation task. Among ICU encounters lasting at least 48 hours following the initiation of mechanical ventilation, predict whether or not the given encounter will be associated with a diagnosis of VAP at any later time during the patient's stay, with predictions generated at the 48th hour following intubation.
Admission task. Among all ICU encounters lasting at least k hours, classify whether or not the given encounter will be associated with a diagnosis of VAP at any time, with classifications made at the kth hour following admission. Note that, for this task, classifications could be made prior to the initiation of mechanical ventilation.

3. Materials and methods

3.1. Data processing

For all models, data were extracted from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III version 1.3 dataset collected from the ICU at Beth Israel Deaconess Medical Center in Boston, Massachusetts.[23] MIMIC-III contains EHR data (including lab results) and clinical notes on over 40,000 individual patient encounters. All MIMIC-III data were passively extracted from the patient EHR, were de-identified, and collected in compliance with the Health Insurance Portability and Accountability Act.

3.2. Intubation task

Data were included from encounters of patients aged 18 years or older, with a minimum of one

observation of each of the following vital signs and lab tests: diastolic blood pressure, creatinine, Glasgow Coma Scale (GCS), heart rate, oxygen saturation (SpO_2), platelet count, respiratory rate (RR), systolic blood pressure, temperature, hematocrit, and white blood cell count. Hematocrit has been shown to improve other pneumonia-related predictions.[24] Community-acquired pneumonia patients were identified by the presence of a pneumonia diagnosis at admission and were excluded. All encounters for this task were required to involve at least one period of invasive mechanical ventilation. As VAP is defined as pneumonia developing after 48 hours following intubation, encounters were required to last at least 48 hours after intubation. All mechanically ventilated patients in this dataset met this requirement. ML models were compared against the CURB-65,[25,26] VAP predisposition, insult, response, organ dysfunction (PIRO),[27] and clinical pulmonary infection score (CPIS) scoring systems.[28] To facilitate the comparison with CURB-65, we required encounters to include at least one measurement of blood urea nitrogen (BUN). These exclusion steps are summarized in Figure 1. For this task, the windows of data used to generate predictions were calculated backward from the 48th hour following the initiation of ventilation. That is, a 12-hour intubation task model used the 12 hours of data up to and including 48 hours after initiation of mechanical ventilation, or hours 37 to 48 after the initiation of mechanical ventilation. All windows for this intubation task included data from an identical number of patients.

3.3. Admission task

Identical exclusion criteria were applied for the admission task as were applied to the intubation task, with the exception of the initiation of mechanical ventilation requirement, which was not applied. For this task, the windows of data used to generate predictions were calculated forward from the time of ICU admission. For example, a 12-hour admission task model used the first 12 hours of data after a patient was admitted to the ICU, after which point a VAP risk prediction was generated. Patients were required to have a length of stay as long as or longer than the prediction window being examined; the number of patients included in the experiments therefore varied by prediction window (Fig. 1).

For both tasks, we extracted patient baseline and time-varying clinical measurements for each encounter. Baseline data included age, and a boolean value for the presence of any relevant comorbidities or symptoms at the time of admission (bacteremia, cirrhosis, congestive heart failure, fever, intracranial hemorrhage, renal failure, respiratory distress, respiratory failure, sepsis, subarachnoid hemorrhage, and shortness of breath). We additionally included an indicator for acute respiratory distress syndrome (ARDS), as pneumonia is associated with ARDS.[29] Time-varying clinical measurements included the required vital signs and laboratory tests, as well as urine output (evaluated as

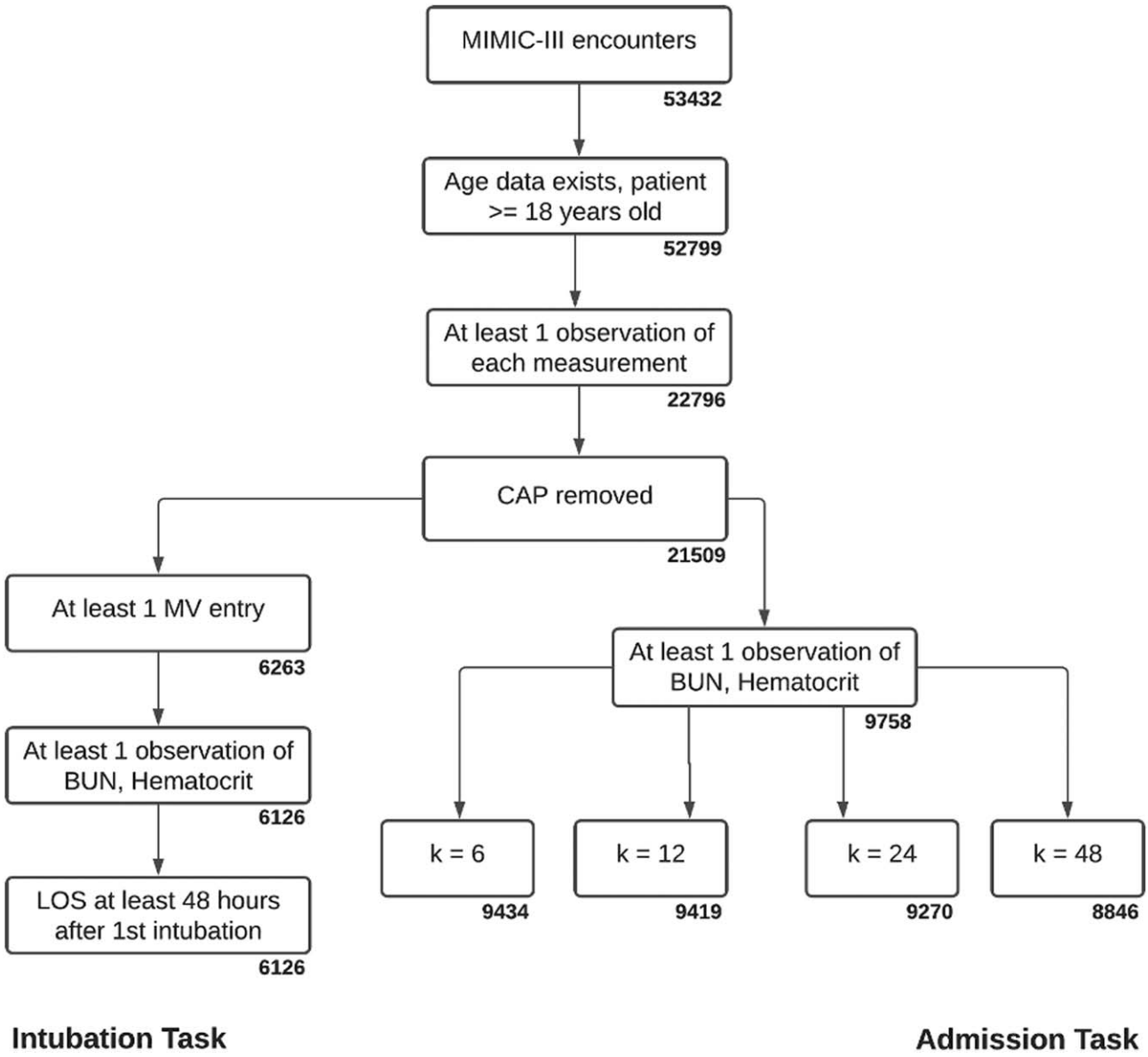

Figure 1. Patient exclusion numbers for the intubation task. BUN = blood urea nitrogen, CAP = community acquired pneumonia, MV = mechanical ventilation.

the number of urine measurements over the duration of stay) and blood culture information (evaluated as the order of any tests during the relevant window, and as the total test count during the window). We further included the hour of the initiation of mechanical ventilation and the number of accumulated mechanical ventilation hours at the time of prediction.

Raw measurements were binned into 1-hour intervals and averaged within bins to produce a single, representative value for each hour. Missing values were imputed based on median values, which were determined using only data in a training set. This process did not allow information from the hold-out test set to influence the imputation. We calculated six summary statistics (minimum, maximum, median, first, last, and average) of each vital sign and laboratory test over a variable-length window (Table 1). Specifically, for each window length $k \in \{6,12,24,48\}$, we calculated the statistics for the k hours preceding and including the 48th hour after the initiation of mechanical ventilation (in the case of the intubation task) or the 48th hour after admission. Age, the number of total urine output events, and the number of blood culture tests were kept in their raw form. Boolean indicators were added for the presence of antibiotics, sputum labs, blood culture labs, comorbidities and symptoms listed above, and ARDS (Table 1). All variables were then concatenated into one vector for each encounter.

3.4. Gold standard

The International Classification of Diseases (ICD) Revision 9 code 997.31 was the gold standard definition of VAP. Literature assessing the accuracy of ICD codes for VAP identification remains limited.[30] However, studies have suggested that, while the sensitivity of administrative coding may be only moderate for

Table 1

Data included as input to the algorithm.

Required vitals and labs	Boolean indicators	Optional measures
– Systolic BP	– Antibiotics	– Age
– Diastolic BP	– Sputum labs	– Total urine output events
– HR	– Blood culture labs	– Number of blood culture tests
– Respiratory rate	– Any of cirrhosis, congestive heart failure, fever, bacteremia, intracranial hemorrhage, renal failure, respiratory distress, respiratory failure, sepsis, subarachnoid hemorrhage, shortness of breath	– Number of sputum tests
– Temperature		– Number of MV hours
– Hematocrit		
– SpO_2		
– BUN		
– GCS		
– Platelet count		
– WBC		
– Creatinine	– ARDS	

ARDS = acute respiratory distress syndrome, BP = blood pressure, BUN = blood urea nitrogen, GCS = Glasgow Coma Scale, HR = heart rate, MV = mechanical ventilation, SpO_2 = oxygen saturation, WBC = white blood cell.

VAP identification, specificity, and negative predictive value are quite high.[31,32]

3.5. ML methods and comparators

Due to the absence of VAP prediction literature, a variety of ML methods were chosen to identify the most appropriate ML methods to address these novel tasks. The ML methods evaluated included both simple linear models and complex, non-linear models, to evaluate the potential effectiveness of different ML models for VAP prediction.

For each prediction task and each window length, we evaluated 5 ML methods: logistic regression, multilayer perceptron, random forest, support vector machines, and gradient boosted trees. The logistic regression and support vector machines methods were chosen as representative linear models and the random forest and gradient boosted trees methods were chosen as representative ensemble learning and tree-based methods. The multilayer perceptron model was included in lieu of neural network methods with more layers, as there were too few training examples to effectively train such models. Except for the gradient boosted trees model, which was created using the XGBoost Python package, the models were implemented using the scikit-learn Python package.

For both the admission and intubation prediction tasks, each of the above ML methods was used to train prediction models for each predefined window length (6, 12, 24, and 48 hours). The use of five methods at each of four prediction windows resulted in a total of 20 models being evaluated for each prediction task.

We compared the performance of the ML algorithm to the CURB-65, VAP PIRO, and CPIS scores for evaluating pneumonia severity. CPIS performance was estimated as described in the literature[14] as it could not be calculated in our dataset. We implemented CURB-65 and VAP PIRO in our dataset.[25,33] CURB-65 values were calculated for each hour according to the number of the following which was true: BUN > 19 mL/dL, RR ≥ 30, systolic BP < 90 mmHg/diastolic BP ≤ 60 mmHg, and age ≥65. We tried several variations of assigning a CURB-65 score to a temporal window, including using its maximum, average, and last values over the window. As the results were similar in each case, we reported its average over the window. PIRO is a four-variable score based on predisposition, insult, response, and organ dysfunction. The score is measured by assigning 1 point in 4 areas: detection of a comorbidity (chronic obstructive pulmonary disease, immunocompromised, heart failure, cirrhosis, or chronic renal failure), bacteremia, a systolic BP < 90 mmHg, and ARDS.

The data were partitioned uniformly at-random into a set for training and hyperparameter tuning (80%) and a 20% hold-out test set, against which all trained models were evaluated for final performance metrics in the last step. For each task and for each window length k, each model was trained using four-fold grid search cross-validation on the 80% training set. After searching the space of model hyperparameter values, the hyperparameters that produced the best cross-validation performance in terms of area under the receiver operating characteristic (AUROC) curve were chosen. Class weights were used as a hyperparameter in all ML algorithms to improve the performance on the imbalanced dataset. Final hyperparameter ranges used for all models are presented in Supplementary Table 1, http://links.lww.com/MD/G175. Each model was then tested on the 20% hold-out test set. Feature importance was measured through Shapley additive explanation (SHAP) values to assess similarities or differences in the features used to generate predictions across model types.

3.6. Minimal input models

In both admission and intubation tasks, we conducted feature selection on the overall best performing model (ie, combination of ML method and time window) for each task using the full set of features. As the best performing model was XGBoost in both cases, we selected the most important features from the SHAP plots to eliminate a large subset of features. In the first step, the top 10 most important features as measured by SHAP values were kept and the performance of the models were re-evaluated on these features. We observed that the performance of the models remained relatively unchanged using only the top 10 features in each case. A further reduction of the number of features was also done by assessing the correlations among the remaining 10 features and removing the features that are highly correlated. This resulted in a five feature model for the intubation task and a nine feature model for the admission task.

For each minimal input model, we assessed model AUROC. We additionally assessed model specificity when sensitivity was fixed at 0.80.

4. Results

In total, 6126 patients were included in the intubation task experiments. Of those, 524 received a diagnosis of VAP during their stay, resulting in a VAP prevalence in the intubation task patient population of 8.55% and of 4.97% in the admission task patient population. Those who were diagnosed with VAP had a higher prevalence of ARDS, a greater number of sputum labs performed and were on average older when compared to those without VAP (Table 2). Additional comorbidity information for the intubation task population is presented in Supplementary Table 2, http://links.lww.com/MD/G175.

4.1. Prediction of VAP 48 hours after intubation

For this task, the gradient boosted trees models demonstrated better performance than other model types, except when using summary statistics from a 24-hour window (Table 3), in which case a random forest model demonstrated the highest perfor-

Table 2

Demographic and comorbidity information for the experimental population.

Characteristic	VAP positive n = 524	VAP negative n = 5602
Age		
<30	27 (5.15%)	215 (3.84%)
30–49	81 (15.46%)	735 (13.12%)
50–59	105 (20.04%)	972 (17.35%)
60–69	120 (22.90%)	1383 (24.69%)
70–79	107 (20.42%)	1185 (21.15%)
80+	71 (13.55%)	984 (17.57%)
ARDS		
Yes	35 (6.68%)	284 (5.07%)
No	489 (93.32%)	5318 (94.93%)
Sputum test performed		
Yes	497 (94.85%)	2644 (47.20%)
No	27 (5.15%)	2958 (52.80%)
Gender		
Male	313 (59.73%)	3172 (56.62%)
Female	211 (40.27%)	2430 (43.38%)
Ethnicity		
White	354 (5.15%)	4072 (72.69%)
Black/African-American	44 (8.40%)	484 (8.64%)
Asian	16 (3.05%)	134 (2.39%)
Hispanic/Latino	11 (2.10%)	209 (3.73%)
Unknown/other	99 (18.89%)	703 (12.55%)

ARDS = acute respiratory distress syndrome, VAP = ventilator-associated pneumonia.

mance. The best AUROC was recorded by XGBoost using summary statistics from a 6-hour window. Multilayer perceptron models demonstrated lower performance than all other models, particularly when using 48 hours of data. All models outperformed the CURB-65 and PIRO scores at all prediction times.

Receiver operating characteristic (ROC) curves for all intubation task models at the time windows $k = 6$ and $k = 48$ are presented in Figure 2; these models are annotated with the performance of several common clinical criteria in the diagnosis of VAP, as determined by a meta-analysis.[14] All models meet or exceed the reported performance of all of the clinical criteria. ROC curves for the remaining time points are presented in Supplementary Figure 1, http://links.lww.com/MD/G175.

An assessment of SHAP plots for the intubation task models showed a high degree of overlap in the features identified as most important for generating predictions (Supplemental Figures 2–6, http://links.lww.com/MD/G175), with key recurrent features including length of time for mechanical ventilation, sputum culture measures, and clinical measures related to SpO_2 and GCS.

Table 3

AUROC results on the hold-out test set of models trained to predict VAP 48 hours after intubation, using summary statistics from the previous *k* hours of patient data.

k (hours)	6	12	24	48
Logistic regression	0.744	0.751	0.739	0.776
Multilayer perceptron	0.731	0.740	0.722	0.741
Random forest	0.771	0.767	0.780	0.777
Support vector machines	0.765	0.769	0.764	0.775
XGBoost	0.799	0.794	0.775	0.791
CURB-65	0.503	0.498	0.498	0.498
PIRO	0.565	0.555	0.566	0.557

AUROC = the area under the receiver operating characteristic curve, ICU = intensive care unit, PIRO = predisposition, insult, response, organ dysfunction, VAP = ventilator-associated pneumonia.

4.2. Prediction of VAP after admission

Of the models trained for the admission task, each performed best when using summary statistics from 48 hours between admission and the time of prediction (Table 4). Overall best performance was obtained by the XGBoost model using 48 hours of data. Figure 3 provides comparisons of ROC curves for these windows, with annotated comparisons to common clinical criteria. Most models meet or exceed the reported performance of existing clinical criteria at 6 hours, and all did at 48 hours. The $k = 12$ and $k = 24$ cases are shown in Supplementary Figure 7, http://links.lww.com/MD/G175.

As with the intubation task, SHAP plots demonstrated that features most important for generating VAP predictions were similar across model types and prediction windows (Supplementary Figures 8–12, http://links.lww.com/MD/G175), with mechanical ventilation hours, GCS, and sputum again being important features across many models. This finding further supports that a wide range of models may be suitable for this prediction task.

4.3. Minimal input models

The best performing model for the intubation task was XGBoost using six hours of data and for the admission task was XGBoost using 48 hours of data. After all feature selection based on feature importance (Supplementary Figures 6 and 12, http://links.lww.com/MD/G175) and correlation, the minimal input intubation task model included five features (last GCS, last temperature, mechanical ventilation duration, antibiotics indication, and count of sputum cultures). The minimal input admission task model included nine features (last measure of GCS, last systolic blood pressure, last RR, last white blood cell count, mechanical ventilation duration, count of sputum cultures, count of blood cultures, count of urine cultures, and age).

The AUROC of both of these minimal input models were very close to the initial all feature models, with AUROC of 0.80 and 0.83 for the intubation and admission tasks, respectively. With sensitivity fixed at 0.80, the minimal input intubation and admission task models achieved specificity values of 0.69 and 0.73, respectively.

5. Discussion

Our retrospective results demonstrate the success of ML models trained to predict VAP in use cases (Tables 3 and 4). Due to the novelty of the task, a variety of models and prediction windows were explored, all of which demonstrated strong predictive performance. In both use cases, the test set performance of ML models significantly exceeds the reported performance of classic clinical indicators of VAP[14] and does so with the potential of advance warning (Figures 1 and 3, Supplementary Figures 1 and 4, http://links.lww.com/MD/G175). Additionally, simple, interpretable models such as logistic regression demonstrate strong performance for both tasks. Given the morbidity and mortality associated with VAP[34] and with VAP treatment delays,[35] these models may have important implications for improving patient care and outcomes, subject to future external and prospective validation.

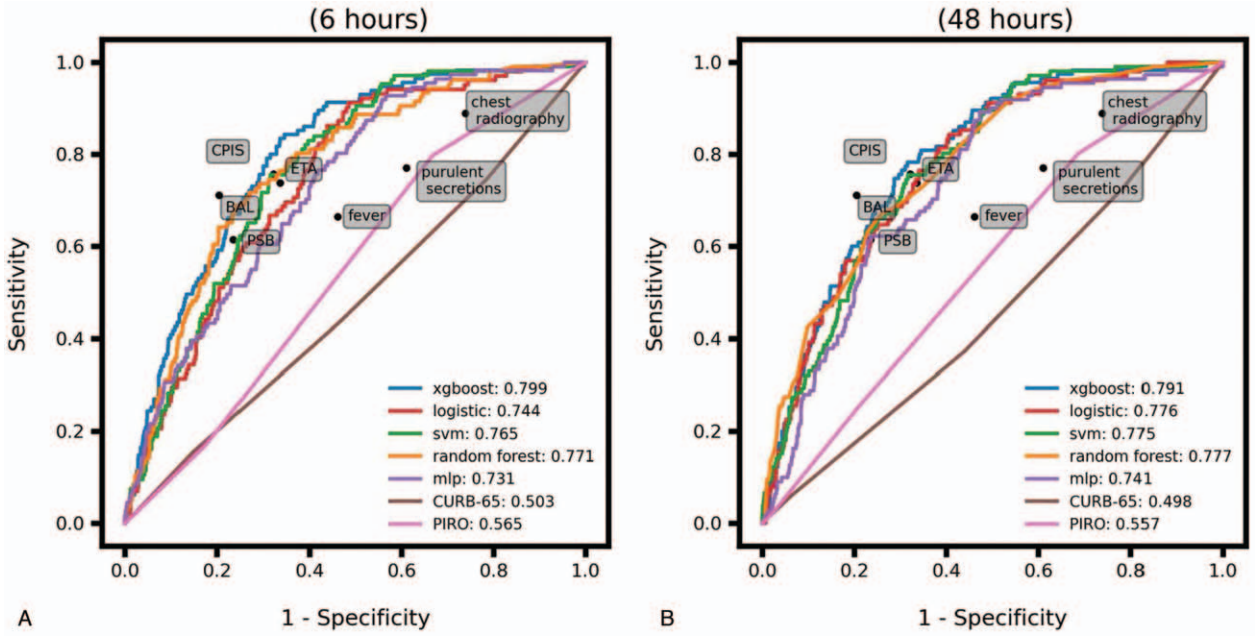

Figure 2. ROC curve comparison for intubation task models, with summary statistics calculated from the (A) 6 hours and (B) 48 hours of data preceding the time of prediction. K = number of hours used to make prediction, MIMC = multiparameter intelligent monitoring in intensive care, ROC = receiver operating characteristic.

ML methods were explored due to the need for methods that improve upon the discriminatory ability of existing VAP prediction and detection methods. Existing methods include the PIRO score, which was originally developed for sepsis and has since been utilized for VAP risk stratification in hospitalized patients.[27,36] The CPIS tool is also used as a VAP diagnostic tool for hospitalized patients.[28] However, prior research has shown these tools exhibit poor diagnostic performance. In a recent study, PIRO achieved an AUROC of 0.605.[33] A recent meta-analysis found that CPIS demonstrated a sensitivity and specificity of 73.8% and 66.4%, respectively, for VAP detection.[14] The CURB-65 score, utilized for risk stratification and to predict 30-day mortality, has better accuracy than PIRO and CPIS with a c-statistic of 0.761.[25,26] However, the sensitivity of CURB-65 varies with the severity of pneumonia. A 2016 study showed that 36% of patients who were classified as low risk based on their CURB-65 score were ultimately hospitalized as a result of pneumonia.[26]

Despite the urgent need for better VAP diagnostics and the popularity of ML applications in healthcare, relatively little effort has been devoted to the application of ML to EHR data for the purpose of predicting VAP. Several methods have been developed for identifying community-acquired pneumonia using neural networks and genetic algorithms[37–39] and one study predicted hospital-acquired pneumonia in patients with schizophrenia.[40] Concerning VAP, studies have examined the accuracy of electronic nose (e-nose) sniffers for screening potential VAP cases. These devices use ML methods to analyze exhaled breath for metabolites that may be suggestive of VAP, and some have demonstrated strong discrimination for identifying the presence of VAP.[41,42] However, prospective validation of an e-nose tool found that sensitivity and specificity were insufficient for general clinical use.[43] In this context, our study provides a valuable characterization of a variety of ML methods applied to two VAP prediction use cases.

The first use case corresponds to the intubation task, which predicted VAP at the first time VAP can be diagnosed, the 48th hour following the initiation of mechanical ventilation. While the highest overall AUROC was demonstrated by XGBoost using data from the 6-hour window leading up to and including the prediction time, all models met or exceeded the performance of existing VAP identification methods. An alert at or before VAP onset by any of these methods is likely to improve identification of VAP, potentially overcoming limitations in diagnostic criteria that may lead to both under- and over-treatment with antibiotics.[14,44,45] It is worth noting that XGBoost demonstrated decreasing performance over longer data collection windows for the intubation task. This may be due to the fact that, with increasing time from the point at which predictions are made, many of the model inputs lose clinical relevance to the current patient state, decreasing the overall relevance of the inputs. However, other methods, such as random forest, demonstrated increased performance with longer data windows. The relatively low performance of the multilayer perceptron model may be because a single layer perceptron can only classify linearly separable sets of vectors. Since the data used here have at least 82 dimensions, it is likely that the classes are not linearly separable.

Table 4

AUROC results on the hold-out test set of models trained to predict VAP *k* hours after ICU admission.

k (hours)	6	12	24	48
Logistic regression	0.772	0.788	0.812	0.822
Multilayer perceptron	0.587	0.712	0.753	0.820
Random forest	0.706	0.762	0.822	0.838
Support vector machines	0.766	0.799	0.812	0.829
XGBoost	0.733	0.796	0.820	0.854
CURB-65	0.481	0.496	0.506	0.517
PIRO	0.584	0.595	0.599	0.622

AUROC = the area under the receiver operating characteristic curve, PIRO = predisposition, insult, response, organ dysfunction, VAP = ventilator-associated pneumonia.

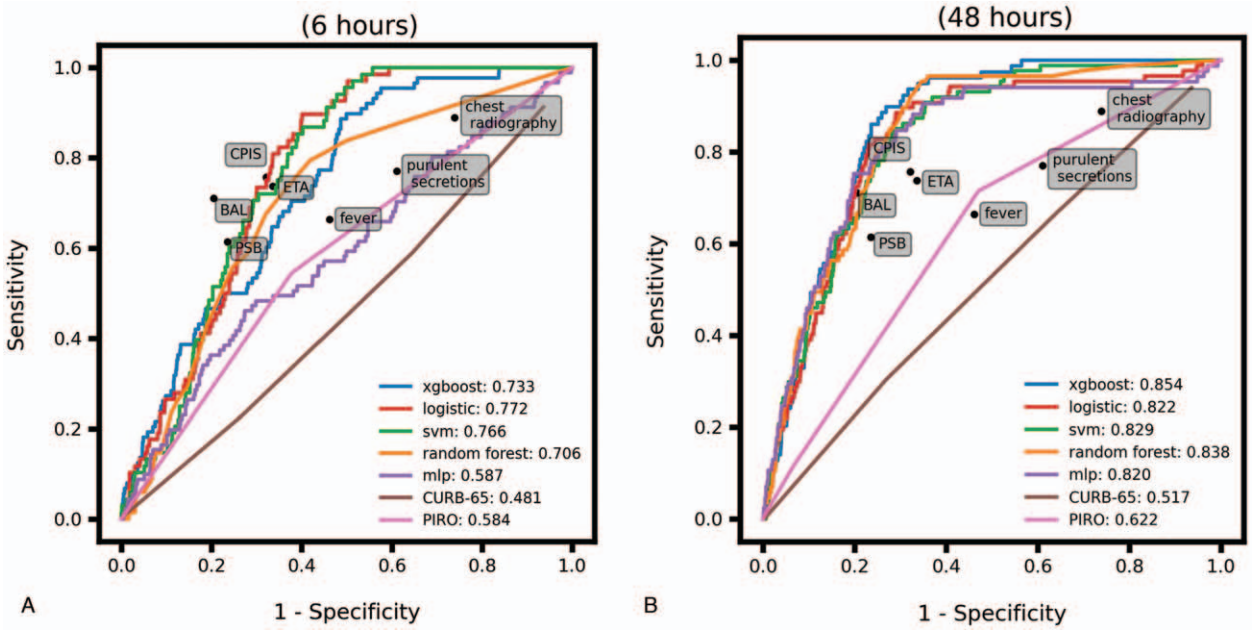

Figure 3. ROC curve comparison for admission task models with summary statistics calculated from the (A) 6 hours and (B) 48 hours of data preceding the time of prediction.

Although we tried a multilayered neural network, we had insufficient training data.

The second use case corresponds to the admission task, which aims to predict VAP a fixed number of hours (6, 12, 24, or 48) following admission to the ICU. In the first three of these cases, alerts generated by the models give advance warning of VAP, as patients cannot have been ventilated for the 48 hours required for a VAP diagnosis. Further, these models have no requirement that patients be ventilated for algorithm alerts to be generated, although they incorporate information about current ventilation status. Therefore, these models may be able to identify patients at high risk of developing VAP should they become ventilated before ventilation occurs. For this task, the best-performing model was XGBoost using summary statistics generated from the entire 48-hour window after admission (Table 4, Figure 3). All models demonstrated performance meeting or exceeding methods current methods for VAP identification. For this task, models generally demonstrated increasing performance when using larger windows of data and when generating predictions further into the patient stay.

Models addressing these two use cases exhibit different strengths, with potential implications for future clinical use. The intubation task is, by definition, applied to a high-risk population of patients. In many clinical settings, the positive predictive value of an intubation task alert is therefore likely to be high, and an intubation-based ML system likely to identify the vast majority of VAP cases. Such a system could therefore meaningfully improve the timeliness of antimicrobial administration, improving patient outcomes. In contrast, the admission task is designed for application to all ICU patients, regardless of ventilation status. Therefore, in addition to providing early identification of patients for whom additional monitoring is warranted, the admission task alert may provide clinicians with the opportunity to consider non-invasive methods of ventilation for high-risk patients,[46,47] preventing VAP entirely. The success of both approaches in this retrospective study supports the potential for ML methods to meet a wide range of clinical needs relating to VAP treatment, identification, and prevention.

For both use cases, the strong performance of simple linear models (eg, logistic regression) has important implications. Logistic regression models are readily interpretable, with the relative importance of each input feature measurable as the relative magnitude of its coefficient. While many ML methods are viewed as "black boxes," linear models are in contrast far more transparent, with clear similarities to tabular scores already commonly used in clinical practice. This feature of simple models may increase trust in the model, increasing its utility.[48] Additionally, the importance of commonly collected clinical features, simple indicator variables, and readily available information such as length of ventilation support the practicality of a clinical application of these models. ML models may be challenging to implement in clinical settings when specialized testing is required. However, the strong performance found in this study supports further work to develop models that use accessible measurements. Such work may help enable the practical implementation of ML tools in a clinical setting.

There are several limitations to this study. Because the exact onset time of VAP could not be determined retrospectively from this dataset, it was not possible to determine the degree of advance warning provided by the models. For the best-performing models, the cumulative duration of mechanical ventilation, the presence of antibiotics administration, and the ordering of a sputum test were the most important features, along with statistics of GCS and creatinine (Supplementary Figures 2 and 4, http://links.lww.com/MD/G175). It may be that, while by definition VAP may not be diagnosed until 48 hours after the initiation of mechanical ventilation, clinicians may suspect VAP is developing and order sputum tests or administer antibiotics within this time window. In these cases, while the classifications made by the models may be technically considered predictions, the alerts they would provide would not necessarily lead to a significant change in care. Given this, it may be valuable to

conduct future research on prediction tasks using data obtained immediately after mechanical ventilation, so physicians have a greater lead time to intervene prior to the 48th hour. These results may also be limited by the use of ICD codes for the VAP gold standard, which may fail to accurately capture all patients who experienced VAP during their hospitalization. However, the generally high specificity of codes for hospital-acquired infections supports that our positive class consisted of true positive VAP cases. Another limitation is that these models were trained and tested on data from a single institution, which may limit generalizability. Model performance on novel patient populations or specific subpopulations cannot be inferred. Finally, due to the retrospective nature of this study, the impact that these algorithms may have on patient care and outcomes in a live clinical setting cannot be determined. These limitations underscore the need for future additional and prospective validation. In addition to prospective validation of model performance, additional work is needed to determine the ideal data collection and prediction windows. Incorporation of clinician feedback will be essential to ensure that a prospectively implemented model appropriately balances the need for early prediction with the collection of sufficient patient data.

6. Conclusion

The development of accurate and timely diagnostic tools for ventilator-associated pneumonia has been limited, despite the prevalence, mortality, and costliness associated with VAP. ML may be a key contributor for future management of VAP risk associated with mechanical intubation, with a variety of ML methods demonstrating suitability for this prediction task. Future work is necessary for further validation of ML algorithms for VAP prediction.

Author contributions

Conceptualization: Jacob Calvert.
Data curation: Christine Giang.
Formal analysis: Christine Giang, Keyvan Rahmani.
Methodology: Jacob Calvert.
Project administration: Ritankar Das.
Software: Christine Giang.
Supervision: Jacob Calvert, Jana Hoffman, Qingqing Mao, Ritankar Das.
Validation: Jacob Calvert, Jana Hoffman, Qingqing Mao.
Visualization: Christine Giang.
Writing – original draft: Gina Barnes, Anna Siefkas.
Writing – review & editing: Christine Giang, Jacob Calvert, Keyvan Rahmani, Gina Barnes, Anna Siefkas, Abigail Green-Saxena, Jana Hoffman.

References

[1] Wagh H, Acharya D. Ventilator associated pneumonia – an overview. Br J Med Pract 2009;2:16–9.
[2] Klompas M. Interobserver variability in ventilator-associated pneumonia surveillance. Am J Infect Control 2010;38:237–9.
[3] Depuydt P, De Bus L. Controversies in ventilator-associated pneumonia diagnosis. ICU Manag Pract [Internet] 2016;16(3.): Available from: https://healthmanagement.org/c/icu/issuearticle/controversies-in-ventilator-associated-pneumonia-diagnosis.
[4] Bergin SP, Coles A, Calvert SB, et al. PROPHETIC: prospective identification of pneumonia in hospitalized patients in the ICU. Chest 2020;158:2370–80.
[5] Wunderink RG. Mortality and the diagnosis of ventilator-associated pneumonia: a new direction. Am J Respir Crit Care Med 1998;157:349–50.
[6] Torres A, Niederman MS, Chastre J. 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). Eur Soc Intensive Care Med ESICM Eur Soc Clin Microbiol Infect Dis ESCMID Asoc Latinoam Tórax ALAT 2017;50(3.):
[7] Timsit J-F, Esaied W, Neuville M, Bouadma L, Mourvillier B. Update on ventilator-associated pneumonia. F1000Research 2017;6.
[8] Miller F. Ventilator-Associated Pneumonia [Internet]. WFSA Resource Library: Anesthesia Tutorial of the Week. [cited 2021 Apr 7]. Available from: https://resources.wfsahq.org/atotw/ventilator-associated-pneumonia/.
[9] Estimating the Additional Hospital Inpatient Cost and Mortality Associated With Selected Hospital-Acquired Conditions|Agency for Health Research and Quality. Accessed Oct [Internet]. 2020; 6. Available from: https://www.ahrq.gov/hai/pfp/haccost2017-results.html.
[10] American Thoracic Society; Infectious Diseases Society of America; American Thoracic Society; Infectious Diseases Society of AmericaGuidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am J Respir Crit Care Med 2005;171:388–416.
[11] Wang G, Ji X, Xu Y, Xiang X. Lung ultrasound: a promising tool to monitor ventilator-associated pneumonia in critically ill patients. Crit Care 2016.
[12] Mayhall CG. Ventilator-associated pneumonia or not? Contemp Diagn 2001;7:
[13] Kalanuria AA, Zai W, Mirski M. Ventilator-associated pneumonia in the ICU. Crit Care 2014;18(2.):
[14] Fernando SM, Tran A, Cheng W. Diagnosis of ventilator-associated pneumonia in critically ill adult patients—a systematic review and meta-analysis. Intensive Care Med 2020;Publ Online. 2020 Apr 18;.
[15] Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJPC. Identifying pneumonia in chest X-rays: a deep learning approach. Measurement 2019.
[16] Kang SY, Cha WC, Yoo J. Predicting 30-day mortality of patients with pneumonia in an emergency department setting using machine-learning models. Clin Exp Emerg Med 2020;7:197–205.
[17] Wilson AD. Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath. Metabolites 2015;5:140–63.
[18] Cooper GF, Aliferis CF, Ambrosino R. An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 1997;9:107–38.
[19] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery, Data Mining, KDD, '15, ACM. Press, 2015.
[20] Rajpurkar P, Irvin J, Zhu K, et al. In: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning Computer Vision and Pattern Recognition 2017 [Internet]. Available from: https://arxiv.org/abs/1711.05225.
[21] Toğaçar M, Ergen B, Cömert Z. A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models. IRBM Publ Online Novemb. 2019; 1.
[22] Chandra TB, Verma K. Pneumonia Detection on Chest X-Ray Using Machine Learning Paradigm. In: Chaudhuri BB, Nakagawa M, Khanna P, Kumar S, editors. Proceedings of 3rd International Conference on Computer Vision and Image Processing Advances in Intelligent Systems and Computing. Springer; 2020.
[23] Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
[24] Ahn JH, Choi EY. Expanded A-DROP score: a new scoring system for the prediction of mortality in hospitalized patients with community-acquired pneumonia. Sci Rep 2018;8:1–9.
[25] Lim WS, Eerden MM, Laing R. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 2003;58:377–82.
[26] Sharp AL, Jones JP, Wu I. CURB-65 performance among admitted and discharged emergency department patients with community-acquired pneumonia. Acad Emerg Med 2016;23:400–5.

[27] Marshall JC. The PIRO (predisposition, insult, response, organ dysfunction) model. Virulence 2014;5:27–35.
[28] Zilberberg MD, Shorr AF. Ventilator-associated pneumonia: the clinical pulmonary infection score as a surrogate for diagnostics and outcome. Clin Infect Dis 2010;51(Supplement_1):
[29] ARDS – Symptoms and causes [Internet]. Mayo Clinic. [cited 2020 Oct 26]. Available from: https://www.mayoclinic.org/diseases-conditions/ards/symptoms-causes/syc-20355576.
[30] Goto M, Ohl ME, Schweizer ML, Perencevich EN. Accuracy of administrative code data for the surveillance of healthcare-associated infections: a systematic review and meta-analysis. Clin Infect Dis 2014;58:688–96.
[31] Stevenson KB, Khan Y, Dickman J. Administrative coding data, compared with CDC/NHSN criteria, are poor indicators of health care-associated infections. Am J Infect Control 2008;36:155–64.
[32] Verelst S, Jacques J, Van Den Heede K, et al. Validation of hospital administrative dataset for adverse event screening. Qual Saf Heal Care 2010;19(5.):
[33] Furtado GH, Wiskirchen DE, Kuti JL, Nicolau DP. Performance of the PIRO score for predicting mortality in patients with ventilator-associated pneumonia. Anaesth Intensive Care 2012;40:285–91.
[34] Monegro AF, Muppidi V, Regunath H. Hospital Acquired Infections [Internet]. In: StatPearls. StatPearls Publishing; 2020. Available from: http://www.ncbi.nlm.nih.gov/books/NBK441857/.
[35] Grief SN, Loza JK. Guidelines for the evaluation and treatment of pneumonia. Prim Care 2018;45:485–503.
[36] Lisboa T, Diaz E, Sa-Borges M, et al. The ventilator-associated pneumonia PIRO score: a tool for predicting ICU mortality and health-care resources use in ventilator-associated pneumonia. Chest 2008;134:1208–16.
[37] Heckerling PS, Gerber BS, Tape TG, Wigton RS. Selection of predictor variables for pneumonia using neural networks and genetic algorithms. Methods Inf Med 2005;44:89–97.
[38] Heckerling PS, Gerber BS, Tape TG, Wigton RS. Use of genetic algorithms for neural networks to predict community-acquired pneumonia. Artif Intell Med 2004;30:71–84.
[39] Er O, Sertkaya C, Temurtas F, Tanrikulu AC. A comparative study on chronic obstructive pulmonary and pneumonia diseases diagnosis using neural networks and artificial immune system. J Med Syst 2009;33:485–92.
[40] Kuo KM, Talley PC, Huang CH, Cheng LC. Predicting hospital-acquired pneumonia among schizophrenic patients: a machine learning approach. BMC Med Inf Decis Mak 2019;19(1.):
[41] Liao Y-H, Shih C-H, Abbod MF, Sheih JS, Hsiao YJ. Development of an E-nose system using machine learning methods to predict ventilator-associated pneumonia. Microsyst Technol 2020;16Publ Online March.
[42] Liao YH, Wang ZC, Zhang FG, Abbod MF, Shih CH, Shieh JS. Machine learning methods applied to predict ventilator-associated pneumonia with *Pseudomonas aeruginosa* infection via sensor array of electronic nose in intensive care unit. Sensors 2019;19(8.):
[43] Schnabel RM, Boumans MLL, Smolinska A. Electronic nose analysis of exhaled breath to diagnose ventilator-associated pneumonia. Respir Med 2015;109:1454–9.
[44] Camargo LFA, De Marco FV, Barbas CSV. Ventilator associated pneumonia: comparison between quantitative and qualitative cultures of tracheal aspirates. Crit Care 2004;8(6.):
[45] Nussenblatt V, Avdic E, Berenholtz S. Ventilator-associated pneumonia: overdiagnosis and treatment are common in medical and surgical intensive care units. Infect Control Hosp Epidemiol 2014;35:278–84.
[46] Brochard L. Mechanical ventilation: invasive versus noninvasive. Eur Respir J 2003;22(47 suppl.):
[47] Makhabah DN, Martino F, Ambrosino N. Noninvasive mechanical ventilation in patients with high-risk infections in intermediate respiratory care units and on the pneumology ward. Noninvasive Vent High-Risk Infect Mass Casualty Events 2013;329–32.
[48] Shortliffe EH, Sepulveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018;320:2199–200. PMID:30398550.