Models to Predict Pulmonary Embolism for Patients with Asthma Exacerbation

David Hu

Department of Biostatistics, UNC Chapel Hill

Abstract

Pulmonary embolism (PE) is a medical condition characterized by arterial blockage in the lungs, primarily due to blood clots originating in the deep veins of the legs. PE remains challenging to diagnose due to non-specific nature of the symptoms, leading to potential underdiagnosis and adverse outcomes. This study aims to evaluate various modeling algorithms (logistic regression, random forest, support vector machines (SVM), and naive Bayes) for the predictions of PE in patients presenting with symptoms consistent with PE or asthma exacerbations using data from a retrospective study conducted at the University of Florida Health System. Feature importance analysis reveals that PE history, high BMI, and higher age are important predictors of PE. Despite limitations, these models are potential offer valuable insights into PE risk assessment, potentially guiding further diagnostic interventions.

Keywords: Logistic Regression, Naive Bayes, Pulmonary Embolism, Random Forest, Support Vector Machines

1 Introduction

Pulmonary Embolism (PE) is a serious medical condition when blood flow is blocked to an artery in the lungs [7]. The blockage is commonly caused by a blood clot that comes from the deep veins in the legs, called deep vein thrombosis (DVT), but the blockages in the blood vessels can be caused by other factors, like fat, tumors, or air bubbles [7]. Some common symptoms of PE include shortness of breath, chest pain, and fainting [7].

In the United States, an estimated 900,000 people each year are affected by DVT or PE and of those people, an estimated 60,000 to 100,000 people end up dying from DVT or PE [2]. 25% of those with PE will have sudden death as their first symptom and 33% of those who experience DVT will have long-term complications such as swelling and discoloration at the affected limb[2].

Since symptoms of PE are non-specific, it is difficult to diagnose PE when there are other medical conditions present. In one case, PE may be an underlying condition in patients with asthma exacerbation undergoing a computed tomographic pulmonary angiography (CTA)[1]. In another case, a retrospective study found that 30% of patients admitted to an emergency room for chest pain were ultimately diagnosed with PE [6].

Many factors have been listed as being associated with the risk of PE. Fracture and/or traumatic injuries in the lower limbs can lead to increased risk of PE because of immobilization and rest to recover from the injury, which means slow blood flow through the legs [9]. Personal history of venous thromboembolism (VTE), which includes DVT and PE, combined with suffering ankle and foot fractures can lead to occurrence of VTE [5]. Other factors include age, contraceptive use, and family history are risk factors for PE [12].

A multitude of diagnostic techniques are available for assessing the presence of pulmonary embolism (PE), including the D-Dimer test and computed tomographic pulmonary angiography (CTPA) [11]. In cases of PE, elevated levels of D-Dimer in plasma signify the activation of coagulation and fibrinolysis pathways [11]. However, the D-Dimer test exhibits low specificity, necessitating further diagnostic evaluation following a positive result [8].

CTPA is widely regarded as the gold standard for confirming PE; however, its utility may be limited in individuals with severe renal insufficiency, contrast media allergy, or during pregnancy. In such cases, alternative diagnostic modalities such as spiral computed tomography, high-probability ventilation–perfusion (V/Q) scintigraphy scans, or pulmonary angiography may be warranted [8].

Hence, there arises a crucial need to develop predictive models that can complement laboratory assays and imaging techniques to facilitate the accurate assessment of PE in clinical settings.

This paper will evaluate different modeling algorithms to predict pulmonary embolism in patients presenting with asthma exacerbation. The dataset comes from a retrospective study of adult patients presenting with asthma exacerbation who underwent CTA for suspected PE at the University of Florida (UF) Health System, Gainesville, Florida [1].

Previous studies using machine learning to predict pulmonary embolism have been conducted, but they have primarily been used to aid in analyzing chest imaging and/or include laboratory measures [10]. One such approach utilized and compared logistic regression, neural networks, and XGBoost [10], while the authors of the dataset used logistic regression and random forest [1]. This paper will use and compare the performance of logistic regression, random forest, support vector machines (SVM), and naive Bayes models in predicting PE while finding factors associated with PE.

The rest of the paper is as follows. First, an overview of the models followed by an overview of the evaluation metrics to evaluate them. Then, an overview of the data, summarizing both the demographic and clinical aspects. Next, present the results of each model and its metrics. Finally, conclude it with a discussion on the implications and generalizability of the results.

2 Methods and Subjects

This section describes the four models and the metrics that will be used to evaluate their performance.

2.1 Statistical Methods

2.1.1 Logistic Regression

Logistic regression can be used to model the probability of pulmonary embolism. The logistic function transforms a linear combination of predictor variables to a probability between 0 and 1. The logistic function is given below:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

In this equation, z represents a linear combination of input variables in the logistic regression model. The function $\sigma(z)$ maps the input z to the probability of the patient having PE.

2.1.2 Random Forest

Random Forest (RF) is a machine learning method that can be used for classifying PE. RF is built by an ensemble of decision trees by a method called bootstrapping aggregation (bagging) and randomly subsetting a set of predictors to build a tree. The final prediction is made by a majority vote by the ensemble of decision trees.

Bagging is a method where the training data is sampled with replacement to construct different training data sets. Each training set is fitted to a tree using a random subset of predictors, which then makes a prediction. These predictions are aggregated and for a classification problem, the majority class predicted by the trees is the predicted class. Using bagging reduces the variance between trees and randomly subsetting predictors to construct each tree decorrelates predictions of all the trees.

To achieve best performance while avoiding over-fitting, RF uses hyperparameters that can be tuned, such as the number of trees in the forest, the maximum depth of each tree, and the number of features considered at each split.

By using several decision trees to make its decision, RF is considered to have high predictive accuracy and robustness to noise and outliers. In addition, RF provides variable importance measures of each variable in context of all other variables.

2.1.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning algorithm that can be used for classification. To classify a patient with PE, SVM finds the optimal hyperplane that maximizes the margin, or the distance between the hyperplane and the nearest data points (support vectors) of each class. Support vectors are the data points that lie closest to the decision boundary (hyperplane). These points play a crucial role in defining the optimal hyperplane and determining the margin. By maximizing the margin, SVM achieves better generalization performance and is less sensitive to outliers.

To prevent overfitting, SVM incorporates a regularization parameter (C) that controls the trade-off between maximizing the margin and minimizing classification errors. A smaller value of C allows for a wider margin but may lead to misclassification of some data points, while a larger value of C reduces the margin in favor of minimizing classification errors.

While SVM is effective in datasets with a large amounts of variables, its decision boundary may not be easily interpretable, especially in higher-dimensional spaces or when using complex kernel functions. However, support vectors can help identify the most influential data points when making the decision boundary.

2.1.4 Naive Bayes

Naive Bayes (NB) is a probabilistic classification algorithm based on Bayes' theorem with an assumption of conditional independence among input variables given the outcome. It relies on Bayes' theorem, which describes the probability of a hypothesis given the evidence. NB calculates the posterior probability of each class given the input variables and selects the class with the highest posterior probability as the predicted class label. The posterior probability is computed using Bayes' theorem:

$$P(y|x_1, x_2, ..., x_n) \propto P(y) \times \prod_{i=1}^{n} P(x_i|y)$$

where $P(y|x_1, x_2, ..., x_n)$ is the posterior probability of class y given the input variables $x_1, x_2, ...x_n$. $P(y)$ is the prior probability of class y and $P(x_i|y)$ is is likelihood of variable x_i given y.

2.2 Evaluation Metrics

2.2.1 Sensitivity

Sensitivity estimates the conditional probability of correct positive predictions given the total number of actual positive instances. This metric is important in medical diagnoses because high sensitivity ensures that individuals with the disease are not missed, leading to early detection and treatment.

2.2.2 Specificity

Specificity estimates the conditional probability of correct negative predictions given the total number of actual negative instances. This metric is important in medical diagnoses because specificity ensures that healthy individuals are not wrongly diagnosed with a disease, reducing unnecessary treatments or interventions.

2.2.3 AUC-ROC

AUC-ROC evaluates the performance of a binary classification model. The Receiver Operating Characteristic Curve (ROC) is a set of data points on a plot where the x axis represents 1-specificity and y axis represents the sensitivity. Each of the points represents a distinct threshold value that produces different sensitivity and specificity values. The points form a curve, and the area under the curve of these points is called AUC (Area Under the Curve). The AUC is a number between 0 and 1 and is used to compare the performance of each binary classification model.

2.3 Choosing a threshold value

Each of the four methods outputs an estimated probability of PE. A threshold value is used to convert the estimated probability of PE into a binary yes/no prediction. Since high sensitivity is essential in the context of predicting PE, the threshold value was chosen to generate a sensitivity of 0.80 in the training set.

2.4 Subjects

The data are derived from a retrospective study conducted by Alzgoul et al., which included 763 patients treated for asthma exacerbation and subjected to CTA between June 2011 and October 2018 [1]. A total 22 demographic and clinical variables were recorded for each patient.

Sixty-three out of 763 patients (8%) records had one or more missing values. The assumption is that these incomplete cases are attributed to missing completely at random and will not be used in the analysis. All analyses were done using R version 4.3.2.

3 Results

3.1 Clinical and Demographic Characteristics

After dropping the incomplete observations, there were 700 records, with 564 diagnosed with acute PE. Of those with PE, they had a median age of 55 and BMI of 31. Of those without PE, they had a median age of 54 and BMI of 29. In both groups, 70% of the patients were women. The table below summarizes the demographic and clinical variables for the PE and non PE group.

Table 1: Patient Demographic Characteristics

	PE (n=136)	**No PE (n=564)**
Age in Years (Median)	55	54
Sex, n (%)		
Male	40 (29%)	172 (31%)
Female	96 (71%)	392 (69%)
Race, n (%)		
Black	79 (42%)	214 (38%)
Other	57 (58%)	350 (62%)

Table 2: Patient Clinical Characteristics

n,(%) unless specified	**PE (n=136)**	**No PE (n=564)**
Atrial Fibrillation	28 (21%)	69 (12%)
Body Mass Index kg/m^2 (Median)	31.2	29.3
Cancer History	58 (43%)	185 (33%)
Cerebrovascular Disease	11 (8%)	22 (4%)
Congestive Heart Failure	34 (25%)	97 (17%)
Contraceptive Use	2 (1%)	11 (2%)
Coronary Artery/Peripheral Vascular disease	46 (34%)	132 (23%)
Diabetes	62 (46%)	173 (31%)
DVT History	16 (12%)	22 (4%)
Fractures or General Anesthesia in prior month	2 (1%)	2 (0.4%)
Heart Rate bpm (Median)	101	101
Hemoptysis	1 (0.8%)	6 (1%)
Hyperlipidemia	67 (49%)	122 (21.6%)
Hypertension	106 (78%)	355 (63%)
Inhaled Corticosteroid	75 (55%)	281 (50%)
Number of ER Visits Prior Year (Median)	0	0
PE History	60 (44%)	15 (3%)
Prednisone Use Greater than 30 Days	49 (36%)	119 (21%)
Smoking History	85 (63%)	347 (62%)

3.2 Prediction Modeling

A total of 760 patients were employed to assess and validate the performance of the four models. Before commencing model training, the dataset underwent a random split to 80% for training and 20% for testing. Since about all the values for the contraceptive use, fractures or general anesthesia in prior month, and hemoptysis were all the same, they were dropped before training the model. The models were trained using 10-fold cross validation.

When fitting a backward step-wise feature selection for logistic regression fitting on the training data, the selected features were previous history of PE, cerebrovascular disease, and hypertension. All of these variables had only two levels, which means the model can only estimate 8 probabilities. To make the logistic model more useful, the variables selected using RF were added to the logistic model because it had continuous variables. All the other models were fitted with default tuning parameters. The threshold values were determined by iterating through a range of threshold values and selected the value that achieved .8 sensitivity of the training data.

Among all models employed, logistic regression had the highest sensitivity at 0.81. Naive Bayes had the highest specificity at .6. While the models had high sensitivity, the low specificity means the performance is poor and would not be ideal to be used by itself. A summary of each model's performance and AUC-ROC figure is shown in Table 3 and Figure 1.

Table 3: Methods' Performance on Test Set

	Sensitivity	**Specificity**	**AUC-ROC**
Logistic Regression	.81	.56	.73
Random Forest	.78	.44	.71
SVM	.79	.56	.75
Naive Bayes	.7	.6	.76

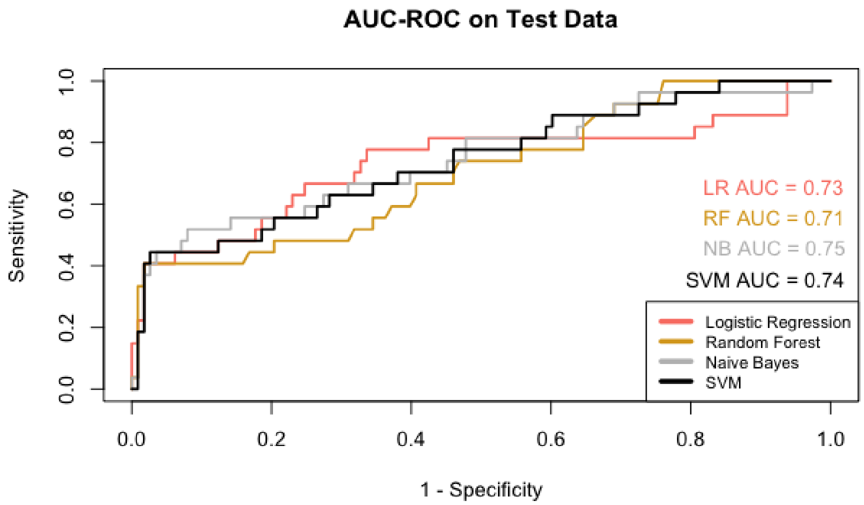

Figure 1: AUC-ROC on Test Data

Furthermore, the feature importance plot for RF was generated via the VarImp function within the caret package. Based on the importance values, the top four features were PE history, BMI, heart rate, and age.

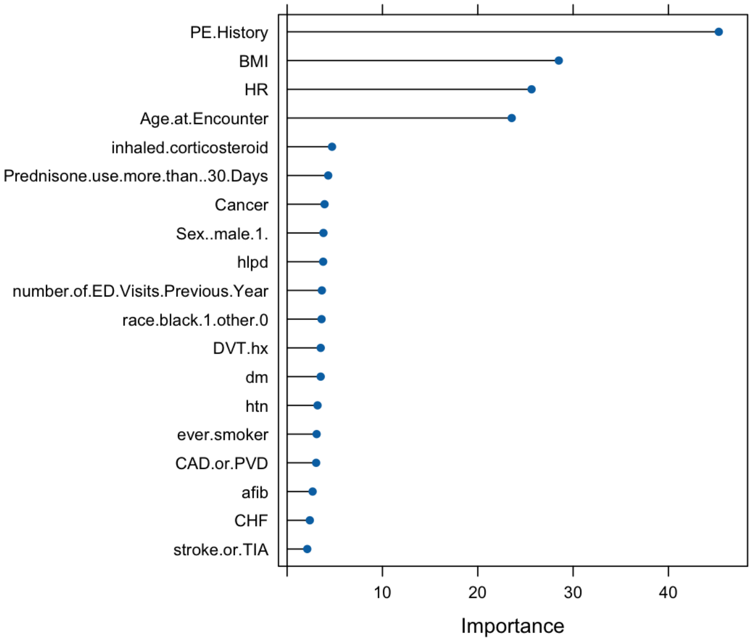

Figure 2: Feature Importance Random Forest computed on training set

4 Discussion

This project demonstrated the predictive capabilities of these models in identifying PE among patients experiencing asthma exacerbation. The timely identification of patients potentially harboring PE is crucial due to the potentially life-threatening nature of this condition. Thus, early detection is paramount for optimizing patient care and mitigating the adverse effects of PE. In such scenarios, employing predictive methods can facilitate the timely recommendation of a CTA scan or other diagnostic tests for confirming PE.

The strengths of these models lie in their ability to identify PE-associated features with a robust sample size. Both logistic regression and Random Forest models highlighted PE history as being associated with PE, alongside influential variables such as BMI, age, heart rate, cerebrovascular disease, hypertension, and fractures or general anesthesia in the prior month. Notably, these variables are readily obtainable and eliminate the need for laboratory tests, which consume valuable time and resources. Given the comparable performance of all models, the use of complex models lacking interpretability is deemed unnecessary; hence, logistic regression suffices.

However, this study is not without limitations. First, these results are only generalizable to patients with known history of asthma. Also, the retrospective nature of the study, conducted using data from a single hospital, limits its generalizability to diverse patient populations or demographic regions. Second, all models exhibited low specificity, with the high being .6, which can lead to over diagnoses and unnecessary diagnostic scans.

Future endeavors could explore other resampling methods to address the inherent imbalance in the dataset, where non-PE to PE patient ratio was at 4 to 1 to improve performance. Downsampling and upsampling techniques could be evaluated to gauge their efficacy in improving model performance. A cost function could be implemented with or without resampling methods to guide the model to focus on making better predictions for someone having PE.

Other models like weighted random forest has been shown to improve performance through cost sensitive learning and sampling technique for imbalanced classification data [3]. XGBoost has been shown to predict PE well albeit using laboratory tests results in its model, which are not present in the dataset [10].

In summary, this study endeavors to develop and assess models for predicting PE among patients experiencing asthma exacerbation. The models provide adequate sensitivity estimates for medical diagnosis. However, the low specificity means the models would be over-diagnosing those without PE as having PE. Logistic regression emerges as a viable option due to its feature interpretability. Despite their limitations, these models serve as valuable tools in aiding clinical decision-making regarding the necessity for further PE testing.

References

[1] Bilal N Alzghoul, Rajat Reddy, Mercy Chizinga, et al. "Pulmonary Embolism in Acute Asthma Exacerbation: Clinical Characteristics, Prediction Model and Hospital Outcomes". In: *Lung* 198 (2020), pp. 661–669. DOI: 10.1007/s00408-020-00363-0.

[2] Centers for Disease Control and Prevention. *Data and statistics on venous thromboembolism.* Published April 25, 2022. Accessed March 24, 2024. Centers for Disease Control and Prevention. 2022. URL: https:https://www.cdc.gov/ncbddd/dvt/data.html.

[3] Chao Chen and Leo Breiman. "Using Random Forest to Learn Imbalanced Data". In: *University of California, Berkeley* (Jan. 2004).

[4] Yonathan Freund, Fanny Cohen-Aubart, and Benjamin Bloom. "Acute Pulmonary Embolism: A Review". In: *JAMA* 328.13 (2022), pp. 1336–1345. DOI: 10.1001/jama.2022.16815.

[5] Michael J Gouzoulis and et al. "Risk factors for venous thromboembolism following fractures isolated to the foot and ankle fracture". In: *PloS One* 17.10 (Oct. 2022), e0276548. DOI: 10.1371/journal.pone.0276548.

[6] Antoine Lefevre-Scelles et al. "Investigation of Pulmonary Embolism in Patients with Chest Pain in the Emergency Department: A Retrospective Multicenter Study". In: *European Journal of Emergency Medicine* 27.5 (Oct. 2020), pp. 357–361. DOI: 10.1097/MEJ.0000000000000680.

[7] Mayo Clinic. *Pulmonary Embolism: Symptoms & Causes.* Published December 1, 2022. Accessed on March 24, 2024. Mayo Clinic. 2022. URL: https://www.mayoclinic.org/diseases-conditions/pulmonary-embolism/symptoms-causes/syc-20354647.

[8] F Piovella and DI Iosub. "Acute pulmonary embolism". In: *The Clinical Respiratory Journal* 10 (2016), pp. 545–554. DOI: 10.1111/crj.12264.

[9] Vikas Rajpurohit and et al. "Metatarsal Fracture Leading to Massive Pulmonary Embolism". In: *Indian Journal of Critical Care Medicine* 21.6 (2017), pp. 401–403. DOI: 10.4103/ijccm.IJCCM_125_17.

[10] Lucas Ryan et al. "Predicting Pulmonary Embolism Among Hospitalized Patients With Machine Learning Algorithms". In: *Pulm Circ* 12.1 (2022), e12013. DOI: 10.1002/pul2.12013.

[11] V Vyas and A Goyal. "Acute Pulmonary Embolism". In: *StatPearls [Internet]* (2024). [Updated 2022 Aug 8]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560551/.

[12] K Walter. "What Is Pulmonary Embolism?" In: *JAMA* 329.1 (2023), p. 104. DOI: 10.1001/jama.2022.17782.