# Navigating Lung Surgery Complications: Data-Driven Exploration of Demographic, Environmental and Intraoperative Factors

CS4991 Capstone Report, 2023

Keivon Chamanara
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
kkc4me@virginia.edu

## ABSTRACT

An anesthesiologist at the University of Virginia's hospital wanted to determine if certain factors affected lung surgery patients' chances of post-operative complications in order to better evaluate future patients' pre-surgery risk levels. To discover these potential relationships, I cleaned and explored a dataset containing information on hundreds of lung surgery patients and employed several machine learning classifiers to form informative predictive models. I used Python, along with several libraries made for data manipulation and machine learning, such as Pandas, NumPy, and Scikit-learn, for my initial analyses. With these libraries, I also utilized different algorithms, like random forest, logistic regression, and support-vector machines, to create the models. Preliminary findings indicated that the amount of time a patient is under anesthesia or where they were discharged to after their surgery may impact the chance of post-operative complications. Further research and analysis are required to make more sense of the data and draw meaningful conclusions.

## 1. INTRODUCTION

Can previous health complications or living next to a park or grocery store affect one's chance of experiencing post-surgical complications? Previous research on this topic exists, but the leaders of this project wanted to expand their understanding and learn how factors inside and outside the operating room affected the likelihood of lung surgery patients developing complications. For example, they aimed to see how, or if, lifestyle choices and geographic variables played a role in the recovery of patients after their surgeries. Prior prediction models primarily focused on demographic and intraoperative factors. Considering this, a major goal of the project was to extend the models beyond these factors and to include geospatial components to determine if model prediction would improve.

To this end, I engaged in a joint effort between doctors at the University's hospital and researchers at the Biocomplexity Institute, a research center that seeks to answer questions at the intersection of biology and a range of diverse disciplines. Our primary objective beyond our research was to provide the doctors with meaningful results and real-life applications that could improve their current and future patient outcomes.

## 2. RELATED WORKS

Research on the link between certain demographic and intraoperative factors already exists. The prior research piqued our interest and provided a baseline for our further exploration into this topic.

Shiono, et. al. (2013) studied the risk factors regarding post-operative complications for elderly lung surgery patients. They focused on the intraoperative circumstances of these patients, such as surgical time, blood loss, and date of surgery. Their project offers valuable insight into surgical factors that play a significant role in the development of post-operative complications in elderly patients. My project is similar but looks at all lung surgery patients instead of solely elderly ones to see if age is also a factor in complications after lung surgery. A reason for this change is that age is widely assumed to affect the likelihood of general surgical complications, so my team wanted to consider this as an independent variable.

Additionally, Bhiken Naik, the anesthesiologist working with my team, published multiple papers along with several other researchers analyzing the effect of a lower tidal volume regimen during lung surgery on post-operative pulmonary complications. A very common approach to facilitate surgical exposure during lung surgeries is one-lung ventilation: the mechanical separation of the two lungs to allow ventilation of only one of the lungs, while the other is compressed by the surgeon or allowed to passively deflate. Colquhoun, et. al. (2021) discuss the impact of using a one-lung ventilation regimen with low tidal volume ventilation, which means decreasing the amount of air that moves in or out of the lungs with each respiratory cycle, on post-pulmonary surgery complications. Their analysis of a vital intraoperative factor provided a valuable framework for my project and gave me a good sense of the potential outcomes of interest, including some independent variables that should be taken into consideration.

## 3. PROJECT DESIGN

To analyze the provided data and draw meaningful conclusions from it, I followed most of the generic machine learning process: preparing the data, performing exploratory data analyses, choosing the models, training the models, and evaluating their performance.

### 3.1 Dataset

The doctors provided our team with a dataset including a host of sources providing information on lung surgery patients throughout the hospital's history, such as their addresses, race, discharge location, and severity of pre-surgical health conditions. The "Primary Outcome" variable is a binary variable that serves as the focal point of the dataset: 0 means that the patient did not experience post-surgical complications, while 1 means that they did.

Due to the nature of this data, I was required to obtain a Health Science Research certification from the Collaborative Institutional Training Initiative Program (CITI) to ensure its proper handling and use. Acquiring the certification was necessary to prove I am capable of responsibly dealing with the sensitive medical data.

### 3.2 Data Preparation

Before the data was ready to be used for training, it had to be cleaned and formatted properly.

The first step was to check for any duplicate rows to remove. I used the Pandas library in Python for this task, as it is well suited for data manipulation and analysis. After removing duplicate rows, I imputed, or replaced, any missing data with alternate values. Some columns contained null values that had to be given actual numerical ones, so I replaced these with their respective column's mean values. Mean imputation can provide a good estimate of the missing values for most datasets. However, I could only use this technique on numerical variables, so the categorical variables were imputed later during the feature engineering step explained in section 3.4. Also, I removed the columns with over 95% missing data, which is an arbitrary cutoff, but these were highly unlikely to provide meaningful insight.

### 3.3 Data Exploration

After cleaning the data, I performed exploratory data analysis. I created visualizations to glance at possible relationships between variables and better understand the underlying structure of the data.

Figure 1 is one such graph, showing two side-by-side boxplots of the distribution of patients' ages for each "Primary Outcome" result. Overall, the ages of patients with outcome 1 are much more condensed than the ages of patients with outcome 0, although outcome 1 has considerably more outliers. Also, the median age of patients with outcome 1 is slightly higher than the median age of patients with outcome 0.

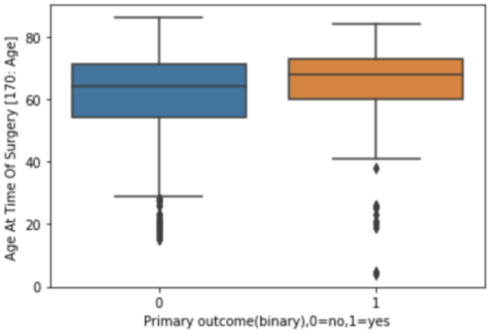

Figure 1: Age Distribution by Primary Outcome

Figure 2 displays the percentage breakdown of "Primary Outcome" by race. For most of the categories, there is no significant distinction between the percentages of each racial group. However, around 40% of Asian-Americans have an outcome of 1, which is much greater than the 10-20% range for other races. Also, American Indians and Alaska Natives have no instances of outcome 1, which can most likely be explained by the relatively small number of American Indians and Alaska Natives in the dataset compared to other races.

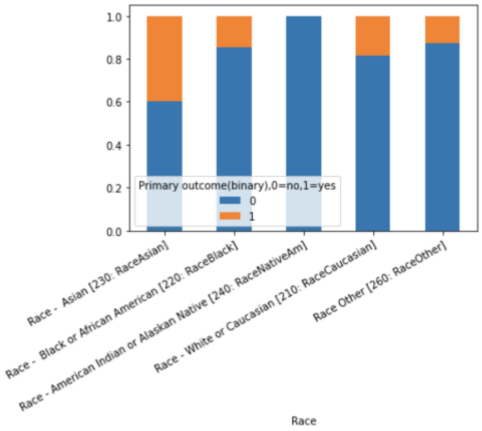

Figure 2: Percentage Breakdown of Primary Outcome by Race

### 3.4 Feature Engineering

Next, I performed feature engineering, which turns variables into useful features that can be inputted into the created models to make the results more understandable.

Regarding the categorical variables, I first imputed the missing values with the mode of their respective columns as opposed to their mean, since it is not possible to find the mean value of a categorical column. Then I one-hot encoded the categorical variables. The purpose of one-hot encoding is to represent the categorical variables in a numerical form that can be digested by the models, since they technically require numeric input. I normalized all of the numerical variables to represent values ranging from zero to one. I did this so that all the features are on a similar scale, rather than all of them having widely different ranges.

Unfortunately, it was not possible to convert the location-based variables, such as address, to their coordinates, so I had to exclude these from the models.

### 3.5 Model Selection

After the data was properly processed and ready to be used by the models, I chose several classification models for this task, including logistic regression, gradient boosting, random forest, support vector machine, and XGBoost, and later selected the most suitable model for the given problem.

### 3.5 Model Training

I split the data into training and testing sets. With the Scikit-learn library from Python, I created the five classifiers to be used for training. Using the training data, I first fit the models and then utilized these trained models to make predictions on the testing set.

### 3.6 Performance Evaluation

With the predicted results, I evaluated the performance of the models using accuracy scores and ROC-AUC curves. Also, using the model with the optimal performance, I carried out exhaustive feature selection on the dataset by evaluating all feature combinations. This process helped to determine the group of predictor variables so the resulting model would be expected to have the highest performance metrics, such as accuracy, recall, and precision.

## 4. RESULTS

Of the five tested classification models, logistic regression performed the best.

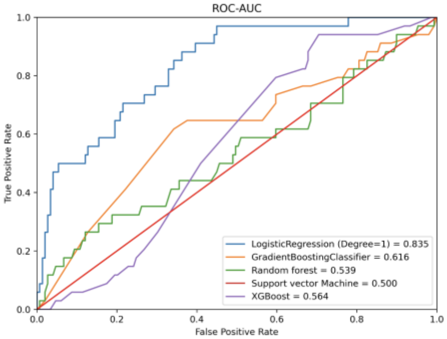

Figure 3 : ROC-AUC Curves of Models

Figure 3 shows the ROC-AUC curves for each of these models. The ROC-AUC curve visually illustrates a binary classification model's performance by plotting the trade-off between true positive rate (sensitivity) and false positive rate. A higher AUC value indicates superior prediction capabilities. The logistic regression model seems to have the largest AUC, making it optimal in this case.

After selecting logistic regression and performing exhaustive feature selection to see which features produce the most accurate model, I concluded that the following features do the best job at predicting whether a patient will have post-surgical complications: “Discharged with chest tube”; “Discharged to healthcare facility (HCF) or not”; and “Anesthesia total time.” Interestingly, these features are all either intraoperative or health-related, meaning that the demographic features, such as race and sex, are not as impactful in predicting post-operative complications.

## 5. CONCLUSION

These findings offer valuable insights for anesthesiologists at the University of Virginia hospital and can benefit other medical professionals and researchers seeking a deeper understanding of the factors contributing to complications in patients undergoing lung surgery. The machine learning models effectively identified key factors contributing to complications by leveraging the provided data. As medical knowledge advances, the application of these insights may contribute to improved patient outcomes and more effective surgical interventions in the future.

## 6. FUTURE WORK

For the future, further analyses on the data can be carried out to find even more relationships between the variables and the possibility of post-surgical complications.

Additional datasets with other variables can be used to extend the scope of the project. Also, more work can be done to properly convert the geospatial variables to useable data to input them into the models.

## REFERENCES

Shiono, S., Abiko, M., & Sato, T. (2013). Postoperative complications in elderly patients after lung cancer surgery. *Interactive cardiovascular and thoracic surgery*, *16*(6), 819–823. https://doi.org/10.1093/icvts/ivt034

Colquhoun, D. A., Leis, A. M., Shanks, A. M., Mathis, M. R., Naik, B. I., Durieux, M. E., Kheterpal, S., Pace, N. L., Popescu, W. M., Schonberger, R. B., Kozower, B. D., Walters, D. M., Blasberg, J. D., Chang, A. C., Aziz, M. F., Harukuni, I., Tieu, B. H., & Blank, R. S. (2021). A Lower Tidal Volume Regimen during One-lung Ventilation for Lung Resection Surgery Is Not Associated with Reduced Postoperative Pulmonary Complications. *Anesthesiology*, *134*(4), 562–576. https://doi.org/10.1097/ALN.0000000000003729