COPD Exacerbation: Predictive Models

D. Rumschitzki [1]†, V.McGraw [2]‡, H. Walt [3]§ L. Jacobs [4]‖, H. Reed [5]¶, K. Huynh [6]††, I. Kemajou-Brown [7]‡‡, Z. Mohammadi [8]§§, S. Scruggs [9]‖‖,

[1] *City University of New York City*
[2] *Rochester Institute of Technology*
[3] *Mississippi State University*
[4] *University of Delaware*
[5] *University of Central Florida*
[6] *Virginia Commonwealth University*
[7] *Morgan State University*
[8] *University of Guelph*
[9] *Clemson University*

(*Communicated to* MIIR *on 6 March 2022*)

Study Group:

The 37th Annual Workshop on Mathematical Problems in Industry (MPI 2021) June 14–18, 2021 at The University of Vermont.

Communicated by: Taras I. Lakoba, University of Vermont

Industrial Partner: Vironix Health, Inc., `https://vironix.ai/`

Presenter: S. Swaminathan

Industrial Sector: Biomedical/Healthcare; Data Analysis

† `david@ccny.cuny.edu`
‡ `vm3258@rit.edu`
§ `hkw59@msstate.edu`
‖ `ljacobs@udel.edu`
¶ `hreed3@knights.ucf.edu`
†† `huynhk4@vcu.edu`
‡‡ `elisabeth.brown@morgan.edu`
§§ `zharam@uoguelph.ca`
‖‖ `srscrug@g.clemson.edu`

Tools: Multivariate models, neural networks, gradient boosting machine, and random forest model

Key Words: Prediction, COPD, generating realistic simulated data

MSC2020 Codes: 62-08, 62P10, 62R07, 65C10 (see MSC2020).

Summary

Chronic obstructive pulmonary disease (COPD) is a group of progressive lung diseases that cause airflow blockage and breathing related problem., It is the third leading cause of death globally. Most people with COPD are at least 40 years old and have at least some smoking history, although prolonged exposure to certain chemicals can also cause it. Air pollution, respiratory infections and other factors can cause critical acute conditions called COPD exacerbations.
In this MPI we attempted to use machine learning algorithms to study the correlations among different features and symptoms of COPD and their likelihood of presaging an acute exacerbation; this would alert the patient to take immediate action. Since real patient data - even anonymized - is not generally available, we used what little data we could find to generate fictitious patient data that we segregated into two groups, one to train and the other to test our correlation models. To do this, we relied upon a multivariate analysis based on correlations of particular symptoms or patient characteristics with known likelihoods of severe exacerbations assuming these symptoms and features were all independent of one another. We used and tested several types of correlations: two-hidden-level neural networks, logistic regression and gradient boosting machines and random forests. We found that our neural network model performed only marginally better than the other methods. Clearly an improvement in the availability of real patient data and an analysis that did not assume a priori that the symptoms and patient features were not correlated with each other would vastly improve the results.

Contents

1 Introduction

Chronic obstructive pulmonary disease (COPD) is a collection of diseases, mainly emphysema and chronic bronchitis, that cause often severe breathing difficulties. Symptoms include trouble breathing, excess wheezing, coughing and phlegm production and numerous others. In turn, patients have difficulties working, engaging in social activities, poor memory, depression, hospital visits, and more. Patients experiencing exacerbations experience severe symptoms and feel as if they cannot catch their breath. The focus of this paper lies in using a patient's acute symptoms to predict the likelihood of that patient experiencing a COPD flare-up or exacerbation that can quickly become life-threatening if not treated immediately, typically in a hospital emergency room. Unfortunately, different combinations of symptoms can indicate an exacerbation in different patients and similar symptoms in different patients can lead to different outcomes, thereby making predicting and preventing flare-ups a difficult task.

Our goal in this MPI study and report is to use data-driven probabilistic methods based on real patient symptom data to generate a large number of fictitious patient scenarios and outcomes in order to train and test several machine learning algorithms. The aim is to create an algorithm that a patient can use to input her/his instantaneous symptoms and to output the likelihood that those symptoms foreshadow an imminent exacerbation in this patient which, when high, indicates that the patient should seek immediate medical help.

The organization of this writeup is as follows: We begin with a literature review, first of the COPD conditions, including its pathology and current treatments, and then of the statistical and mathematical methods that we employ below. The subsequent sections describe in detail the methods and results for using the published likelihoods of each symptom correlating with an exacerbation to generate a set of fictitious patient profiles. These profiles include lists of patient features and symptoms and whether they will experience a severe or mild exacerbation signaled by that combination. We then describe the choice and implementation of correlation methods (e.g., neural networks) that we use to effectively link the combination of features and symptoms to predict with high accuracy those patients that will experience severe exacerbations. We then summarize our findings and suggest improvements for future work.

2 Literature Review

2.1 Pathology

COPD or chronic obstructive pulmonary disease is a collection of lung inflammatory diseases, most commonly chronic bronchitis and emphysema, that is the third leading cause of death in the world. In the US, there are 12-16 million COPD sufferers. A majority of COPD patients either have a history of smoking or have been exposed to noxious chemicals over long periods of time or to fumes from biofuels. Particulates and air pollution also contribute. COPD is associated with increased lung mucus, which causes cough, wheeze, labored breathing (dyspnoea), the destruction of vascular beds and lower oxygen levels in the lungs. Whereas patients can survive years with COPD at a clearly reduced

quality of life, an acute flare-up or exacerbation can quickly become critical and, when not rapidly treated, can cause the patient to expire [25].

In emphysema, the wall fibers of the alveolar sacks in which oxygen exchange occurs become damaged. On a cellular level, macrophages enter the alveoli and attract neutrophils, both of which secrete elastase that degrades sac elasticity. Drastically reduced alveolar elasticity reduces alveolar expansion needed to increase oxygen exchange during inhalation and reduces alveolar recoil needed for easy exhalations and CO_2 clearance. These changes result in labored breathing, mucus and cough. The heart must work harder, which can cause heart trouble and respiratory muscle fatigue, both of which make it more difficult to expel mucus. The resulting distal airway destruction can lead to permanent dilation and irreversible damage [7].

In contrast to emphysema, in chronic bronchitis, the other major COPD affliction, the bronchioles, rather than the alveoli, are damaged. These airways become inflamed and produce excess mucus and loss of cilia function. Emphysema raises the amount of interleukin 8 and c-reactive proteins. The presence of mucus in the bronchioles induces the lungs to cough in an attempt to clear mucus. Mucus increases the viscosity of the fluid lining the brochioles. This increase and the loss of cilia function make mucus clearance far harder. The presence of mucus narrows or even block the air space in the bronchioles, which reduces both oxygen intake and carbon dioxide exhaled. Upper airway fibrosis and narrowing induces a feeling of chest tightness and causes wheezing. Lowered oxygen uptake can cause light-headedness, fatigue, blueness of the nails and lips, lower-body swelling and weight loss. It causes the heart to beat faster, which can lead to heart failure. Chronic bronchitis accelerates lung function decline and increases the risk of exacerbation. It correlates with a higher chance of death from respiratory effects as well as with higher mortality from all sources, likely due to the influence of chronic inflammation in the entire body. The risk of exacerbations is high [16].

As noted, the proximate cause of mortality for COPD patients is typically the occurrence of an acute exacerbation. Exposure to fungus or mold can raise antibody levels in the lungs, which can trigger an exacerbation. An exacerbation is typically a sudden increase in airway resistance due to severe outflow limitations or dynamic lung hyperinflation, which is an expiration flow limitation due to the failure to clear CO_2 completely from the lungs upon exhalation. At the end of exhalation, the pressure in the normal lungs is negative and small additional reductions in lung volume still significantly lower lung pressure. In contrast, under dynamic lung hyperinflation, lung pressure at exhalation remains positive. This means that full exhalation occurs at the flat portion of the volume vs pressure curve, where small volume changes do not reduce the pressure significantly. Residual positive pressure means that inhalation requires effort, which leads to rapid shallow breathing and a rapid heartbeat. Exacerbations further increase lung mucus and cough. Patients with worse base states or who have recently overcome an exacerbation are more susceptible to new exacerbations and tend to have worse outcomes; these features are therefore very good indicators or predictors of a patient's susceptibility to exacerbations. An exacerbation typically is accompanied by fever, a change in sputum, severe lip and nail blueness, a sharp rise in pulmonary artery pressure, i.e., pulmonary

hypertension, and even confusion. Exacerbations are often accompanied by a sharp rise in the COPD biomarker lactose dehydrogenase in the lungs [14].

Doctors typically strongly urge a COPD patient to quit smoking, which lessens lung injury and leads to better color and lower inflammation, but there is insufficient data to know if it reduces the risk of exacerbation. The goal of treatment is to control inflammation, lower mucus production and increase its transport by reinvigorating mucosiliary fibers to thereby reduce cough. Typical long-term treatments include the inhalation of brochodilators and expectorants, including isotonic saline, all of which increase mucus clearance and therefore reduce dyspnoea as well as reduce variability in smooth muscle cell tone. COPD is more complicated than asthma, since spirometry alone does not predict exacerbations in COPD. Successful treatment typically reduces the frequency of exacerbations. Should an excerbabation begin, one first treats it with inhaled steroids, which reduce inflammation and mucus and increases final exhalation volumes. Mechanical ventilation may be required if airway inflammation or lung resistance has spiked or if there is an obvious extreme expiration flow limitation. For acute treatment, beta androgenic receptor agonists and methylxanthines [28] quickly help with mucus clearance by increasing ciliary beating and mucus hydration. Anticholinurgics may help with mucus clearance, but since they may desiccate the lungs, they do not show a clear benefit. Since oxidative stress is a central pathogenesis of COPD, antioxidants that help reduce reactive oxygen species and lower mucus viscosity can reduce the frequency of exacerbations. The antibiotic erythromyocin reduces the frequency of exacerbations from emphysema, but not from chronic bronchitis.

It is worth briefly considering the likely mechanisms of bronchiole blockage. We model the brochioloe as an elastic tube lined with an annulus of a viscous fluid surrounding a core of moving air. For a laminar flow of a fluid filling a tube subject to no-slip boundary conditions on the tube wall, Poiseuille's law shows the fluid flow rate is proportional to the pressure drop times the radius to the fourth power [3]. This means that even a small decrease in radius of the core, the air flow region, due to a thickened very viscous mucus layer surrounding it lowers the air flow rate and/or raises the pressure drop needed for that flow drastically. This system has two potential major sources of instability that can lead to airway blockage:

(1) The Rayleigh Plateau surface tension instability can cause the liquid mucus to go from an annular lining to a lens shape that traverses the entire tube (bronchiole) cross section and blocks the flow of air; or
(2) The elasticity of the tube can cause the whole vessel to collapse or pucker, thereby blocking the airway;

or a combination of both of these mechanisms. Grotberg and coworkers (e.g., Halpern and Grotberg [6]) have investigated these mechanisms in great detail using real lung parameters. The critical amount of liquid for airway blockage (taking both effects into account) goes down as: (1) surface tension goes up; and (2) as wall elasticity goes up. So, emphysema, which stiffens alveoli and the nearby-terminal bronchi, may require

more mucus to cause an exacerbation than chronic bronchitis that affects only the larger bronchii.

There are other variables that help to develop a precise application in predicting COPD. Clinicians have long been aware of the prevalence of neuropsychiatric conditions (cognitive disorders, impairment, depression and anxiety) in patients with COPD [20]. Antonelli-Incalzi et al. present the correlation between cognitive impairment and COPD [19]. Recent studies have investigated the link between chronic obstructive pulmonary disease and disability [9, 11, 17]. Nowadays tele-monitoring and mobile application-based tools are excellent nonpharmacologic strategies to diagnose early stages of COPD (and other chronic illnesses) and to improve home-based disease management [18, 15]. Swaminathan et.al have used machine learning techniques to triage patients with COPD [24]. They have thus far considered only the most relevant patient symptoms, vital signs and baseline characteristics in relation to COPD triage as features in their model. Our aim is to expand the symptoms and patient characteristics that enter into such an automatic diagnosis so as to make it far more accurate and predictive. In our analysis we have used their data-set as inputs, along with others that we have found or constructed as described below.

2.2 Staging COPD and predicting exacerbations: current status

According to the Global Initiative for Chronic Obstructive Lung Disease (GOLD), based on spirometry testing doctors classify patients into four grades: mild, moderate, severe and very severe. There is an existing Acute COPD Exacerbation Prediction Tool (ACCEPT) model that uses COPD trails data on patients with a history of exacerbations to provide a personalised risk profile that allows clinicians to tailor treatment regimens to the individual needs of the patient [1]. A patient's FEV1, the amount of air a person can force into her/his lungs in one second, is counted as one of the input variables in this method. Since it is hard to measure FEV1 accurately at home, we focus on other more easily at-home-collectible features from patients. Comparing data with the R (computer language) packages presented in this paper will serve as our validation tool to test our model.

2.3 Methods of analysis

2.3.1 *Individual and multivariate probabilities for realistic patient data generation*

We have implemented two methods for generating artificial patient data. We discuss the simple method in this section, and defer in depth discussion of how we implement the multivariate probabilities method to section 3.1. It is there that we compare the results of both methods and their performances in a neural network.

From the literature cited in Sec. 3, we were able to collect probability data that an individual with a severe or mild exacerbation was experiencing a particular symptom. The symptoms we chose to focus on were: wheezing, congestion, sore throat, headache, rhinorrhea, sputum, number of previous exacerbations, age, smoking status, and sex.

Based on these known probabilities, we wrote a python code that we now explain to randomly generate a number of fictitious patient profiles corresponding to individuals with severe and mild exacerbations. This simple method generates a table of binary (yes or no) symptoms from the data. Specifically, we use the reported probability that a patient who has a severe exacerbation has a particular symptom and assign that patient a '1' to indicate that the simulated patient has that symptom and a '0' if the patient does not have that symptom. The result is that the probability that a particular patient who experiences a severe exacerbation has a particular symptom is approximately equal to the percentage of the severe patients generated who have this symptom.

For non-binary symptoms such as the number of previous exacerbations a patient has suffered recently or her/his and age, we assign these values from a Gaussian distrubution whose means and standard deviations we have found in the literature cited in sec. 3. In practice, we use the random.normal function in Pythons' numpy library [5] to generate these data for each patient. Note that, in this simple methods, each variable correlates with the probability of an exacerbation independently of all other variables, i.e., without cross correlations. Figures 2-4 show the patient data breakdown for each symptom. The multivariate method outlined in sec. 3 allows for cross-correlations between symptoms.

2.3.2 *Correlation between symptoms and severity of excerbations*

To create simulated patients whose symptoms are correlated with each other, we use a branching algorithm based on correlations that we detect in the limited real patient data. We describe this in detail below. To assess inter-variable correlations between each feature and exacerabtions once we have created the simulated patients, we calculate the Pearson's correlation coefficient. This coefficient is a measure of linear correlation between two variables given by the ratio of the covariance between the two variables and the product of the two variables' standard deviations, that is commonly used with linear regression [2]. To do this we use the computer language R's "cor" function with default parameters [21].

2.3.3 *Methods of correlating symptoms with outcomes*

1. Neural networks [26]

A neural network is a correlation method that can be thought of as a network of "neurons" that are organised in layers, where each layer only communicates with the neighboring layers immediately above and below it. We arrange the network in a manner that the predictors (symptoms and characteristics) form the bottom layer, and the forecasts (output) form the top layer. Such models normally contain intermediate layers that contain "hidden neurons," and the choice of the number of such layers is a matter of experience. Each neuron-neuron connection has a strength parameter that is the scale of how that neuron's value influences the values of the strength parameters of the neurons in its neighboring layers to which it is connected. The training of a network amounts to a process of adjusting these strengths by comparison with the training data set. Basically, when a series of connections yields a result that agrees with the training

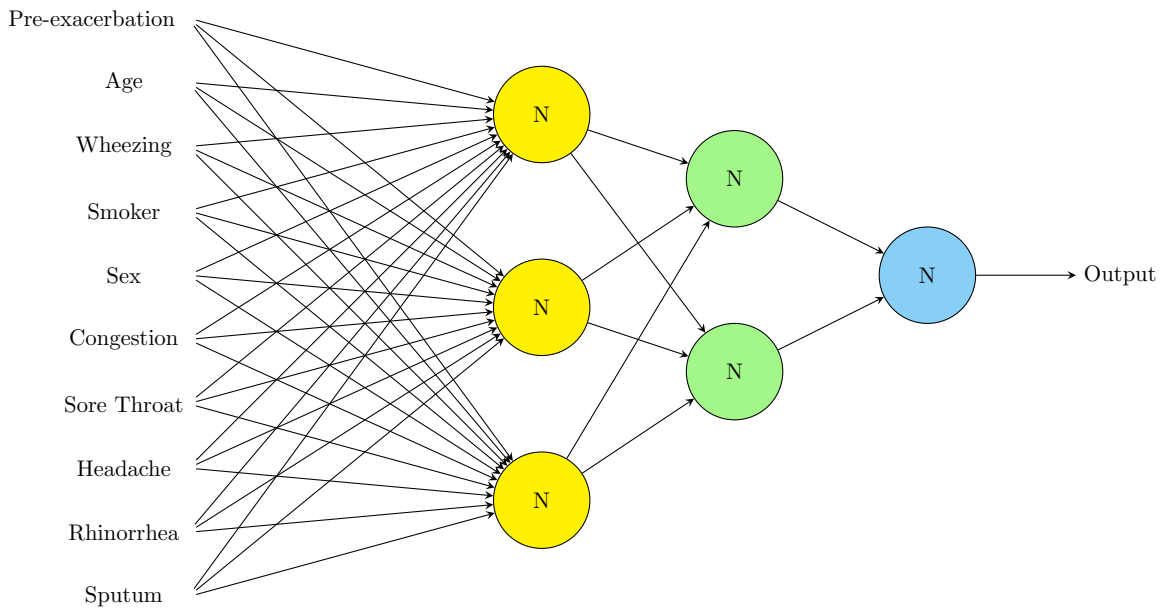


Figure 1. Neural network with 2 hidden layer

data, those strengths are increased. When a set of connections between neurons yields an incorrect results, the strengths of those connections are reduced. This yields a network with a set of strengths based on the network architecture chosen and the training set used. In our case, as noted, the predictors were pre-exacerbation, age, wheezing, smoker, sex, congestion, sore throat, headache, rhinorrhea and sputum. The output from our neural network was either a severe or mild exacerbation. Choosing parameters for deep neural networks is in no way rigorous or unique and the results are very dependent on the chosen network architecture. Many industrial applications of neural networks posit a network that has two intermediate layers. In practice, this is often enough for binary classification problems [8], a choice that we have adopted, as figure 1 illustrates. We split our generated data into training and testing sets with 90% in the former and 10% in the latter since a larger training set should increase the robustness of the resulting network parameters; we recognize that, since both the training and the test data sets were generated by the same algorithm, there is a bias towards the method working better on these data than on real patient data. This said, the next section reports which predictors correlate best with outcomes, which show little or no correlation and, finally, the method's resulting accuracy on the test data set.

2. Logistics regression [23]

Logistic regression is a technique for assessing the association of categorical and/or continuous variables with a variable that can have two discrete values, i.e., it is a classification algorithm that predicts a binary outcome based on a series of independent variable inputs. As with the neural network approach above, we chose the predictors as pre-exacerbation, age, wheezing, smoker, sex, congestion, sore throat, headache, rhinor-

rhea and sputum, and the output from our logistics regression was a severe (1) or mild (0) exacerbation. We again split the data into a training data and a test data set. The method first correlates the training set of data to a linear function of several variables (using linear regression to the form of a constant plus a sum of constants times the values of each of the input variables) with output data of 0 and 1 corresponding to mild and sever exacerbations. One then sets the probablity of each outcome to the inverse of 1+ the exponential of this determined linear function to produce a sigmoid function with range between 0 and 1. If the predicted probability of the resulting sigmoid function is greater (less) than 0.5, we classify the patient as a severe (mild) case. After this regression on our training data we observe how the model performs on the test data set.

3. Gradient boosting machines and random forests [12, 4]

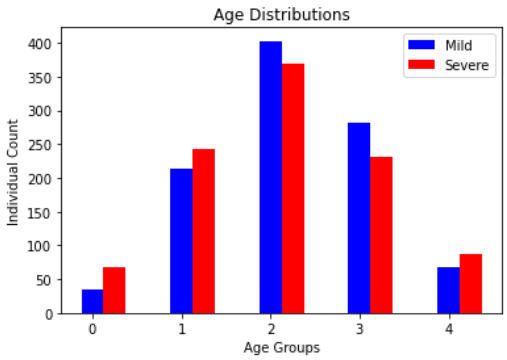
Gradient boosting machines and random forests or random decision forests constitute an ensemble of a machine learning methods for classification, regression and other tasks that operate by constructing a multitude of decision trees during model training. A random decision forest simply returns the average of the outputs of all of the decision trees as its predictions., Gradient boosting is a method that generally improves on the results of random forests by working in an iterative fashion, i.e., when a decision tree yields an imperfect results, one constructs a new decision tree to correlate the residuals created by the imperfections of the initial tree. After several iterations, one usually obtains a correlation with a better predictive ability than a simple random forest. In this study we first tested several random tree structures with varying the numbers of trees, where each tree is assigned, i.e., correlated to a random subset of the patient data. Random forests use fully grown decision trees (that is, decision trees with low bias and high variance). It uses the training data to reduce the variance and the error. Since our data has several binary variable inputs we consider both random forest models and gradient boosting machines in order to specify which decision would more reliably predict severe cases. [Unfortunately, neither the details of the tree structures used for the random forest method or the gradient boosting methods, nor which residuals were corrected, nor how many iterations were used, was reported in this report.] Again, our predictors were previous exacerbation number, age, wheezing, smoker, sex, congestion, sore throat, headache, rhinorrhea and sputum and outputs are either severe or mild exacerbation.

3 Generation of Realistic Simulated Patient Data

Our starting approach for generating realistic patient information was simplistic. We first read through several studies and extracted relevant data, i.e., probabilities. From Ref. [10], we obtained the probabilities of a severe/non-severe exacerbation given a certain feature or symptom. The features used in this study include sex, smoker status, and the presence of wheezing, congestion, sore throat, headache, rhinorrhea, and sputum. This study contained no usable data on age. We therefore took age and number of previous exacerbation data from Ref. [13]. We then wrote a Python script, which is found posted alongside this report, using all our acquired data to generate patient profiles. This first approach is simple; it assumes that all variables are independent of each other. For instance, we assume that the probability that a patient has rhinorrhea is not contingent

on if the patient also has congestion. For each patient, we select non-binary numerical features (age and previous exacerbation) for which we have limited real data from a normal distribution. The "patient" is then put into a bin based on their age. We assigned the patient's other features according to the corresponding probabilities extracted from the studies mentioned above. The output from our algorithm were patients with features and their corresponding classification of severe or non-severe exacerbation.

Shown in Figures 2, 3, and 4 are graphs of the distributions of features we considered that the simple method above generated.



Group	0	1	2	3	4
Age	<50	50-60	60-70	70-80	>80

Figure 2. Distributions of demographic features for simple patient generation model

3.1 Multivariate Probability Analysis

We found in the literature that when running experiments, symptoms were assumed independent of each other. While this assumption makes data analysis easier, it ignores the correlation between some of the symptoms. Also, under this assumption we get results that are unrealistic. For example, with assumed independence there could be patients who are five years of age that show signs of smoking for 30 years. Thus, our goal is to remove the assumption of feature independence to generate a more realistic patient. To do this, we create a branching algorithm where, given a list of features, we can generate a realistic patient. The branching algorithm indicates the influence from the previous features in our branch. For this project, we assume a multivariate normal distribution for the correlated features. We also used the assumptions and distributions from the Simple Method's statistics.

We begin the process by dividing age groups $40-90$ into five age groups of $40-49$, $50-59$, $60-69$, $70-79$, $80-89$. We then divide the next branches of the process based on whether or not the patient is smoking. Then given whether or not the patient is smoking, we form the next branch based on the patient's previous exacerbation with COPD. We

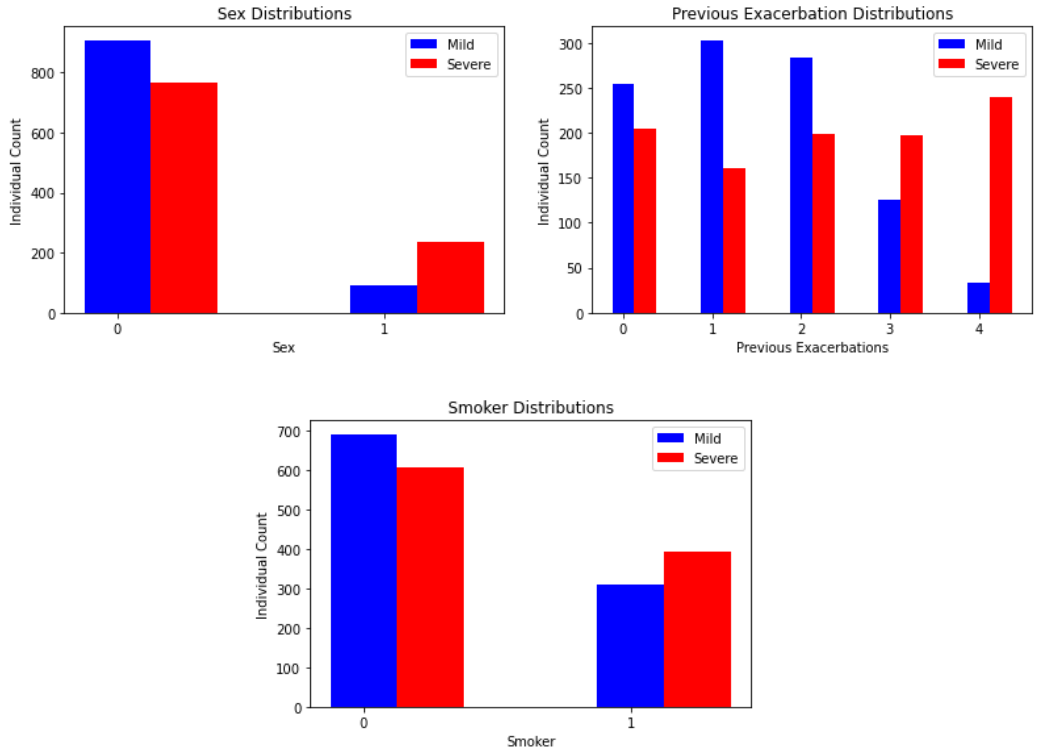


Figure 3. In top-left panel, '0' represents male and '1' represents female. In lower panel, '0' represents non-smokers and '1' represents smokers

divided the branches in this section of the process according to the number of previous exacerbations, ranging from 0 to 4. Next, we form the next group with branches if the patient experienced wheezing, congestion, sore throat, headache, runny nose, or sputum. Since these features were the most correlated subgroup of the features considered, we decided to group them in the multivariate normal distribution because they are generally dependent on each other. The final step in the branching process was to determine the potential severity of an exacerbation. Based on all the previously generated features, we calculated an impact factor that weights the relative probabilities of a severe or mild exacerbatione. A branch of the tree is shown in Figure 5.

We chose our primary algorithm to take the form of a branching process because we noticed in our initial study of the data that certain features of COPD patients are more influenced by other features in determining if the patient will have a more severe case of COPD. For example, the age of the admitted COPD patient influences the chance that they had a previous exacerbation. With this observation, we decided that a branching process will better model the influence of features with each other. The data used for this report suggest which features influence which other features; this dictated the ordering of the different subgroups in our model. The arrows shown in Figure 5 illustrate the impact

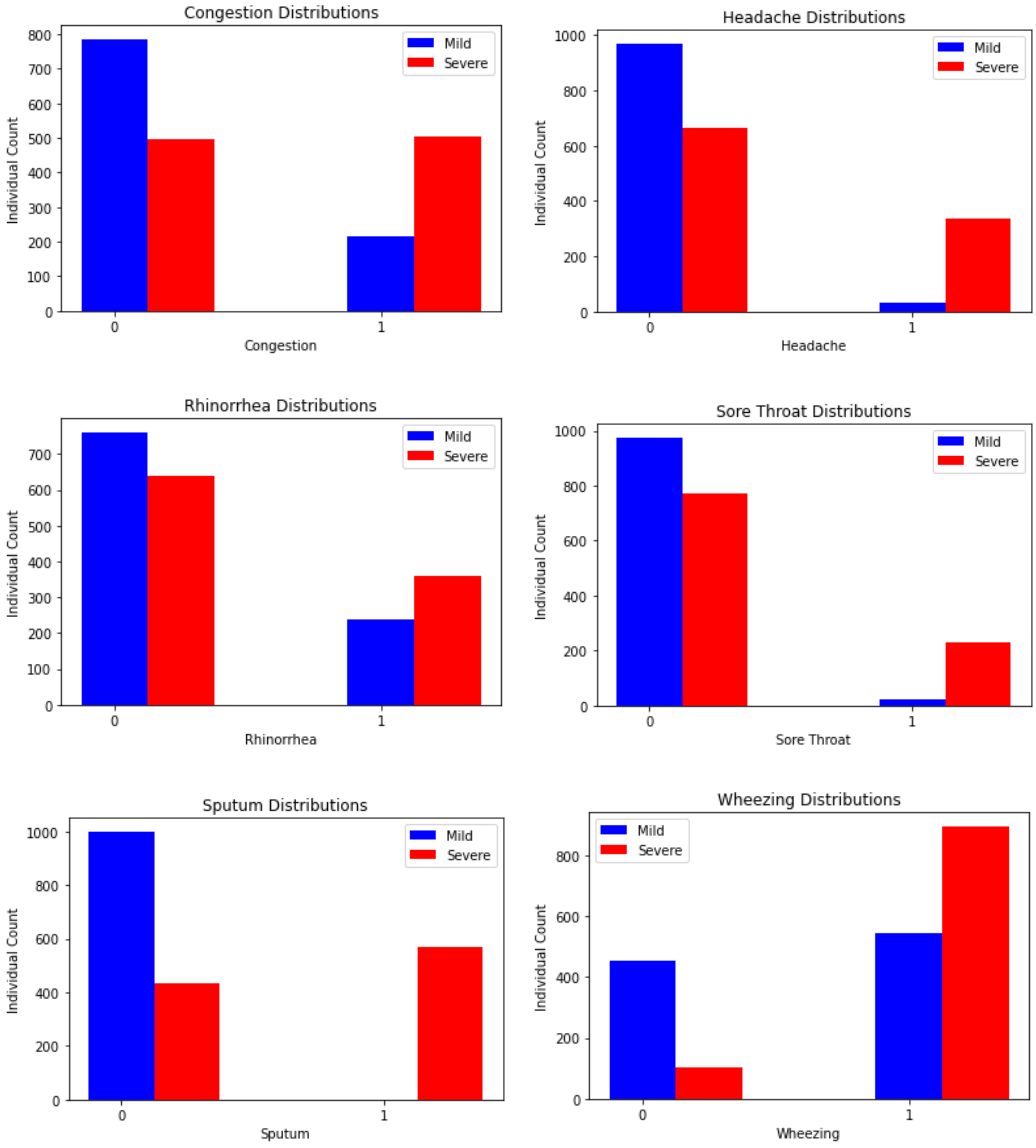


Figure 4. In these figures, '0' represents the absence of the symptom and '1' represents presence of the symptom in question

of each feature on the next feature. The initial parameters and distributions used were the results of the Simple Method for generating fictitious patient data.

We wrote a Python code (posted alongside this paper) to generate a more realistic patient data set given this list of symptoms. Figure 6 illustrates the example of the symptom congestion within our current model.

For each symptom our program generated a similar graph. These results contrast

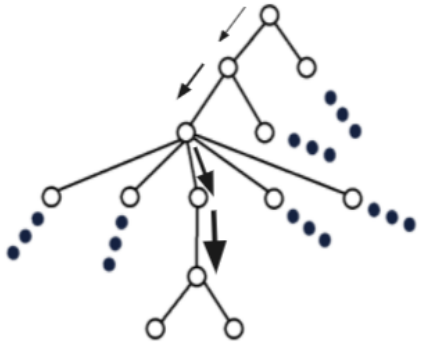


Figure 5. Illustration of branching diagram

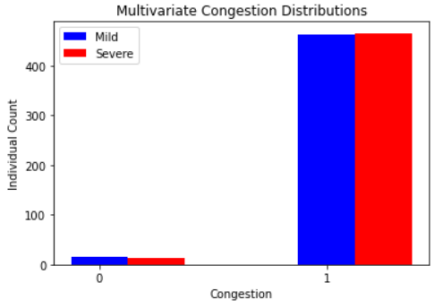


Figure 6. Nasal Congestion with the Branching Process

with the distribution from the Simple Method and show a major limitation in this method as it currently stands. The limited real patient data provided have a limitation with respect to our model: the ER only reported the chief symptom and rarely any other overlapping symptom. As a result our model produced results such as Figure 6. The covariance matrix utilized to define the relationship between biologically linked features is thus not likely representative of the likely true correlation. Since the multi-variate distribution comprises over half the features considered for a particular patient, these incomplete data per patient result in generated patients with odd arrays of symptoms. That is, a small portion of their features are realistic, but their overall set of symptoms contain either wheezing, congestion, sore throat, headache, runny nose, and sputum features or none of them. As a result these data lack any subtly of subsets of these correlated features.

For the purposes of this study and given how little real data we have to work with, the multi-variate approach to patient construction is too detailed for what is needed. As a result we just use the Simple Method's patient generation in the balance of this report. However, the multivariate process is able to generate more realistic patients based on the given data and can potentially replace the Simple Method for constructing more realistic patients if a more accurate correlation relationship of these features can be found in the literature or through new studies. Another advantage of our algorithm is that it can

include more patient-specific data such as allergies, medications, etc. A possible extension of this portion of the project is to verify the algorithm using cleaner data and to train the model to better generate more realistic patients.

4 Results

4.1 Generated fictitious patient data

[At this point, it would be appropriate to present a table of generated patient data for both the simple and the multivariate generation methods. Unfortunately, these data were not submitted for this presentation.] We now analyze the fictitous patient data that we generated.

We used the generated simulated patient data to calculate the Pearson correlation coefficients, which gives the linear correlation between one feature and every other feature, with '1' being perfectly correlated, '0' being uncorrelated and '-1' being inversely correlated. We then produced and a heat map of the resulting matrices using R's corrplot package [27]. The resulting heat maps are shown in **Figures 7 and 8**. These figures show that the features sputum and headache are strongly positively correlated, meaning that they commonly co-occur, and, similarly, wheezing and congestion are also strongly positively correlated with each other. The two matrices that resulted from the simple model and the multivariate model were very similar, indicating that there was not much improvement using the multivariate model.

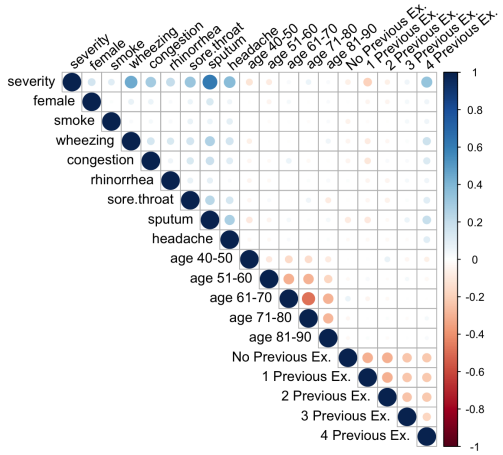


Figure 7. Heat map of Pearson's correlation coefficient matrix after simulating 1000 patients using the simple model.

4.2 Comparison of Data Generation Methods

As explained in the methods section, we test three types of machine learning correlation methods to connect the fictitious patient data - symptoms and outcome. These methods are a two-hidden level neural network, logistic regression, random forest

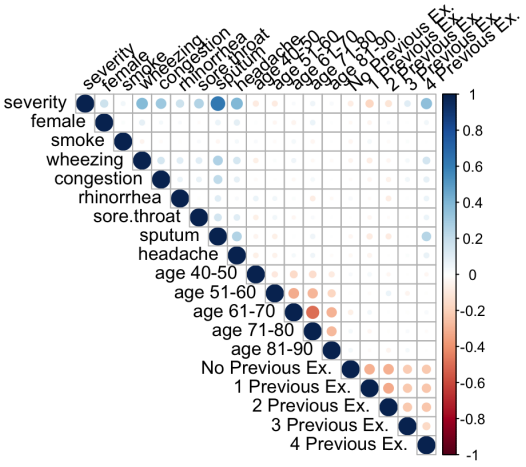


Figure 8. Heat map of Pearson's correlation coefficient matrix after simulating 1000 patients using the multivariate model.

and gradient boosting machines. These parameters in the correlation methods are fit using a subset of the fictitious patient data called the training set and the results are tested by comparing their predictions with the corresponding outcomes from the subset of the fictitious data called the test set. We realize that since our generated patient data, which is distinct from actual patient data, has only limited reliability, the predictions of these methods is likely limited by the level to which the data are realistic. Nevertheless, our team has used each model to predict which patients were about to undergo severe or mild exacerbations.

In our results, the gradient-booster had the highest accuracy in the test data (88%) compared to random forest (85.5%) and neural-network (86%) for correctly classifying severe 90% and mild 82%[1] exacerbations based on our fictitious patient data.[2] As it turned out, the top 5 features that correlate best with correctly predicting patient outcomes in our model were sputum, headache, previous exacerbations, wheezing, and sore throat; the other initial predictors showed no significant correlation.

5 Limitations of the current work and suggested future Work

The most severe limiting factor in models that are directed for clinical purposes is

[1] It is unclear what is meant by these latter two percentages since they do not add up to 100%

[2] It would have been appropriate for the group to redo these trials with several partitions of the generated fictitious data into training and test sets. This would have allowed an assessment of the standard deviation of the resulting percentage accuracy of each correlation method and thus would clarify if the differences presented are significant or not. Unfortunately the group did not do this. It would also have been interesting to compare the predictive correlations with the available real patient data that is available but the limitations of those data listed above made such a comparison less attractive.

finding high quality public clinical data. A large clinical data set was found [22], with data from over 500,000 patients with 972 columns of patient histories. Unfortunately, the only features that this data set reports are the patients' single chief complaint; so no useful information about co-occurring features could be gleaned. This data set listed all the patients' preexisting conditions, which is useful for understanding if any of these preexisting conditions could play a role in acute COPD exacerbation. According to the 2021 GOLD report, some of the most frequent chronic medical conditions that co-occur with COPD are asthma, heart failure and chronic kidney disease. Of the 560,486 patients in the dataset, 44,343 had only COPD, 16,347 had COPD and asthma, 9,425 had COPD and heart failure, 5,885 had COPD and chronic kidney disease, and 773 had all the previously listed preexisting conditions. Of these patients, 54.6%, 48.4%, 68.5%, 68.0%, and 68.7% , respectively, were admitted to the emergency room, which is a likely indicator of an acute exacerbation. Although these data are lacking in feature information, which was the main focus of this model, they could provide useful insight to what pre-existing conditions play a role in acute COPD exacerbation.

Clearly the addition of personalized factors such as allergies, prescriptions, exacerbation history, and genetic factors to each patient's data set, would improve their usefulness in these models since these factors may be important predictors of acute exacerbations. To find how correlated these features are with an acute exacerbation, one would certainly begin with a far more expansive literature search that either finds a far larger trove of real patient data or presents real correlation and cross-correlation results based on real patient data..

Along with a lack of data, much of the GOLD Standard symptoms that are indicative of severe exacerbations are hard to measure at home (spirometry, O_2 saturation, etc.) and when done so yield unreliable results; thus these measurements are, ultimately not useful for this model's goal of only using features that can be accurately and easily measured. Moreover, even the categorization of a patient's episode as severe (or acute) or mild is subjectivie. More precise uniformity in these definitions would certainly help hone in on what the most important features are for predicting the onset of a truly severe case.

Clearly the main limitation of the proposed model is the lack of available complete real patient data. Since robust clinical patient data - even anonymized - are not generally available to the public, making accurate predictions and determining the significance of co-occurring predictors is limited. The data upon which one must then draw conclusions is limited to that found by extensive literature searches that pull data from multiple studies of varying consistency and biases. Simply put, a correlation cannot be better than the data upon which it is based. Clearly, having a large dataset of real COPD patient data would make our model more uniform and likely more realistic and reliable. It would allow one to achieve a statistical understanding of which features that can be measured at home are most important in an acute COPD exacerbation. If such data were available, one would use them to directly correlate of features with outcomes and co-occurring features with each other. Even if such data sets were not large enough in and of themselves to accomplish this, one could use them to generate a far better and

more realistic set of fictitious patient data. All of these improvements would likely greatly improve model efficacy.

References

[1] Adibi A, Sin D D, Safari A, Johnson K M, Aaron S D, FitzGerald J M, and Sadatsafavi M. The acute copd exacerbation prediction tool (accept): development and external validation study of a personalised prediction model. *bioRxiv*, 2019.

[2] Profillidis V A and Botzoris G N. Chapter 5 - statistical methods for transport demand modeling. In Profillidis V A and Botzoris G N, editors, *Modeling of Transport Demand*, pages 163–224. Elsevier, 2019.

[3] Bird R B, Stewart W E, and Lightfoot E N. *Transport Phenomena*. Wiley, 2002.

[4] Sheppard C. *Tree-based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting*. Createspace Independent Publishing Platform, 2017.

[5] Beazley D. *Python Distilled*. Addison-Wesley Professional, 2021.

[6] Halpern D and Grotberg J B. Fluid-elastic instabilities of liquid-lined flexible tubes. *Journal of Fluid Mechanics*, 244(1):615–632, 1992.

[7] Shapiro S D. The macrophage in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 160(5 Pt 2):S29–S32, 1999.

[8] Stathakis D. How many hidden layers and nodes? *International Journal of Remote Sensing*, 30(8):2133–2147, 2009.

[9] Locke E, Thielke S, Diehr P, Wilsdon A G, Barr R G, Hansel N, Kapur V K, Krishnan J, Enright P, Heckbert S R, and et al. Effects of respiratory and non-respiratory factors on disability among older adults with airway obstruction: the cardiovascular health study. *COPD*, 10(5):588–596, 2013.

[10] Rhode G, Wiethege A, Borg I, Kauth M, Bauer T T, Gillissen A, Bufe A, and Schultze-Werninghaus G. Respiratory viruses in exacerbations of chronic obstructive pulmonary disease requiring hospitalisation: a case-control study. *British Medical Journal*, 2002.

[11] Martinez C H, Richardson C R, Han M K, and Cigolle C T. Chronic obstructive pulmonary disease, cognitive impairment, and development of disability: the health and retirement study. *Annals of the American Thoracic Society*, 11(9):1362–1370, 2014.

[12] Brownlee J. A gentle introduction to the gradient boosting algorithm for machine learning. 2016.

[13] Leidy N K, Wilcox T K, Jones P W, Powers J H, and Sethi S. Standardizing measurement of chronic obstructive pulmonary disease exacerbations reliability and validity of a patient-reported diary. *American Journal of Respiratory and Critical Care Medicine*, 2010.

[14] Mise K, Ivancevic Z, Gudelj I, Kotarac S, and Svalina-Grmusa J. Lung and serum biomarkers of tissue lesions due to acute exacerbation of copd. *European Respiratory Journal*, 40(56), 2012.

[15] Sohrabi K, Mursina L, Seifert O, Scholtes M, Hoehle L, Hildebrandt O, and et al. Telemonitoring and medical care supporting of patients with chronic respiratory diseases. *Stud Health Technol Inform*, 212:141–145, 2015.

[16] Campbell M and Sapra A. Physiology, airflow resistance. *StatPearls [Internet]*, 2021.

[17] Saglam M, Vardar-Yagli N, Savci S, Inal-Ince D, Calik-Kutukcu E, Arikan H, and Coplu L. Functional capacity, physical activity, and quality of life in hypoxemic

patients with chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis*, 10(4):423–428, 2015.

[18] Ambrosino N, Vagheggini G, Mazzoleni S, and Vitacca M. Telemedicine in chronic obstructive pulmonary disease. *Breathe*, 12(4):351–356, 2016.

[19] Antonelli-Incalzi R, Corsonello A, Trojano L, Acanfora D, Spada A, Izzo O, and Rengo F. Correlation between cognitive impairment and dependence in hypoxemic copd. *J Clin Exp Neuropsychol*, 30(2):141–150, 2008.

[20] Ouellette D R and Lavoie K L. Recognition, diagnosis, and treatment of cognitive and psychiatric disorders in patients with copd. *Int J Chron Obstruct Pulmon Dis*, 12:639–650, 2017.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[22] Hong W S, Haimovich A D, and Taylor R A. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7):e0201016, 2018.

[23] Menard S. *Applied Logistic Regression Analysis*. SAGE Publications, Thousand Oaks, CA, 2002.

[24] Swaminathan S, Qirko K, Smith T, Corcoran E, Wysham N G, Bazaz G, Kappel G, and Gerber A N. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PloS one*, 12(11):e0188532, 2017.

[25] Mayo Clinic Staff. Chronic obstructive pulmonary disease (copd). *Mayo Clinic*, 2020.

[26] Hagan M T, Demuth H B, and Beale M. *Neural Network Design*. PWS Publishing Co., USA, 1997.

[27] Wei T and Simko V. *R package "corrplot": Visualization of a Correlation Matrix*, 2021. (Version 0.89).

[28] Healthwise Content Development Team. Healthwise helps you make better health decisions. *Kaiser Permanente*, 2018.