An Effective Random Generalised Linear Model to Predict COPD

Linah Saraireh
CDT, School of Architecture
Computing and Engineering, UEL,
University way, London, UK.
u1638295@uel.ac.uk

Mhd Saeed Sharif
CDT, School of Architecture
Computing and Engineering, UEL,
University way, London, UK.
s.sharif@uel.ac.uk

Muna Alsallal
Electronics and Communication Dept,
College of Engineering, Al-Muthanna
University, Education Zone, AL-
Muthanna
aa7095@mu.edu.iq

***Abstract*—Chronic obstructive pulmonary disease (COPD) is a type of chronic lung illness that worsens with time and leads to a restriction in the outflow of air from the lungs. According to the World Health Organisation, The World Health Organization ranks COPD as the third leading cause of death. Clinically, the diagnosis of this disease is relatively difficult; therefore, early identification of individuals at risk of developing COPD is vital for implementing preventative strategies. This research work has developed a generalised linear model (GLM) to predict the COPD status of the patients. A dataset of 1262 patients (688 COPD cases and 574 controls) was used. Exploratory data analysis (EDA) was utilised to observe how potential covariates were related to the response variable (COPD status). By employing rigorous model selection techniques (forward selection and backwards elimination) according to (AIC) which stand from Akaike information criterion and (BIC) which stand from Bayesian information criterion (BIC), a consensus was reached that the most suitable model is a binomial logistic regression model which includes the smoking history, gender, and age. The model was validated using an independent test set with an accuracy of 73%. Such a model, once fully validated, has the ability for predicting the risk of developing COPD in patients with existing lung conditions, including but not limited to, asthma.**

Keywords—Chronic obstructive pulmonary disease (COPD), Generalized linear model (GLM), Akaike information criterion (AIC), Bayesian information criterion (BIC).

I. INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a type of chronic lung illness that worsens with time and leads to reduced lung capacity and restricted airflow. It is recognized as the third largest cause of death worldwide [1], affecting more than 250 million people across the world [2].

The main pathological features of COPD are obstructive bronchiolitis and emphysema. It was also found that there is evidence of mucus hypersecretion in many cases. CT imaging showed that a slight airway loss is an early feature of COPD that might progress to severe airway obstruction [3]. COPD is also linked to chronic inflammation that impacts peripheral airways and lung parenchyma. [3]. Early diagnosis and treatment reduce symptoms and flare-ups [1]. Yet, diagnosing this disease clinically is relatively difficult; therefore, early identification of individuals at risk of developing COPD is vital for implementing preventative strategies.

Sample classification, and more specifically the prediction of disease state, is an important field of research for COPD investigations. There have only been a limited number of approaches, that have been developed to deal with this issue, such as machine learning based approaches [25]. In this research work, we focus on one sub-challenge which is the COPD and its prediction method. The researcher outlined critical pre-processing steps to make training and test data comparable. A generalized linear model (GLM) was developed to predict the COPD status of patients using several predictors including age, gender, smoking status, and ethnicity.

This research study aims to develop a model that predicts the probability of patients developing COPD, considering disease status as a predicted outcome and age, gender, smoking status, and ethnicity as predictors.

II. LITERATURE REVIEW

COPD is a chronic lung condition that worsens with time and is defined by ongoing respiratory symptoms and a diminishment in lung function. It has a transformative and disruptive effect on life, and it is associated with a low life quality and a high mortality rate [2,9]. In addition, COPD has been linked to a number of different disorders, including lung cancer [4]. A primary concern of COPD is that exacerbations become more frequent as COPD progresses [5], usually triggered by rhinovirus infection and increased growth of colonizing bacteria [6].

Increased control over COPD symptoms can be achieved when patients quit smoking, exercise regularly, and get pneumonia, influenza, and coronavirus vaccines [1]. Moreover, long-term adherence to pharmacotherapies is associated with less need for hospitalization and less mortality rate [6-8].

The COVID-19 pandemic has transformed COPD management. It induced changes in the provision of COPD healthcare services. Also, it reduced the specialist access and shifted the service and treatment cost onto community care resources [7].

Numerous studies were conducted to assess the correlation between smoking and COPD risks [7]. The correlation between smoking and the risk of COPD has mostly been investigated through studies that have focused on summary measures, such as smoking status [7]. Other studies investigated different COPD predictors such as age and gender.

Risk factors for chronic obstructive pulmonary disease (COPD) include tobacco smoking, environmental factors (such as air pollution and a deterioration in air quality due to the burning of biomass fuels), and a predisposition to the disease due to genetics [11] The aging of the population and continued exposure to environmental COPD risk factors will, over the course of time, lead to an increase in the prevalence of COPD [12, 13].

The best current medicines aim to lessen the likelihood of a COPD exacerbation and improve functionality and improve quality of life and lower symptoms. The most common non-pharmacological management strategies for stable COPD include quitting smoking, increasing one's level of physical activity (including participation in pulmonary rehabilitation), along with making several other changes to one's lifestyle, including taking medication [14].

Exacerbations of chronic obstructive pulmonary disease (COPD) are associated with less-than-desirable outcomes in terms of health, a larger possibility of additional exacerbations, a reduction in lung function, a deterioration in quality of life, and an increased risk of mortality are some of the potential negative outcomes [15][16]. Most COPD-associated medical expenses are also due to exacerbations; the cost is significantly raised by those who require hospitalization [17]. As a result, preventing exacerbations is one of the key goals of managing COPD [18].

Exacerbations of COPD are known to be more likely in people with certain disease characteristics, such as a history of prior exacerbations, higher airway obstruction, more severe symptoms, and coexisting diseases such diabetes, cancer, heart failure, and acid reflux [19, 20]. Additionally, the eosinophil count in blood is a moderator of the therapeutic response to inhaled corticosteroids (ICSs) and a predictor of the risk of exacerbation, with higher eosinophil counts leading to greater exacerbation rate reductions [21-23]. Therefore, it's crucial to customize therapies based on the unique patient variables that raise the likelihood of aggravation.

While different techniques are used for prediction in healthcare studies, machine learning models have been developed and used to improve the ability to predict disposition of patients with COPD [24].

III. RESEARCH METHOD

Regression models are used as the basis for predictions through a causal relationship between multiple predictors and a predicted outcome. Three regression models are commonly used in healthcare research: linear, log-linear, and nonlinear regression. The main aim of using such models underlies its capability to help the analyst determine the connection between the response and dependent variables Y and other explanatory/covariate variables xi, for i= 1, . . . , n, where i is the unit of observation, calculate the value of the depended variable Y, estimates likelihood parameter values, and evaluate the performance of the model in addition to what-if analysis. Statistical models are required to perform regression, where a model fits a set of observed data.

In normal linear regression models, we assume that observations are independent, and the variance is homoscedasticity i.e. unaffected by the covariate values, linear regression models indicate that

$$Y_i = \beta_0 + \grave{\beta} x_i + \varepsilon_i \tag{1}$$

the intercept is β0 and Yi is the dependent variable (an estimated parameter), $\grave{\beta}$ is the slope of the line, x_i is the independent variable, and ε_i is the distance between the predicted value and the observed value. The sum of ε_i^2 is minimized for better fit. $\grave{\beta}$ is calculated as:

$$\grave{\beta} \ = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2}$$

It is crucial to ensure that the mean is non-negative by using transformations of the mean function so that

$$g(\mu_i) = \beta_0 + \beta x_i \tag{3}$$

A link function $g(\cdot)$ is sometimes used to ensure a non-negative mean e.g. $g(x) = log\ x$ and that the mean is scaled to a probability value between 0 and 1 e.g. $(x) = log\ (x/(1 - x)$.

On the other hand, a multiple linear regression model indicates that

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_{p-1} x_{ip-1} + \epsilon_i \tag{4}$$

Matrix notation can also be used to represent the model as follows:

$$Y = X\beta + \epsilon \tag{5}$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \tag{6}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \tag{7}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (8)$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{bmatrix} \quad (9)$$

Likewise, a link function can be used to limit the covariate values to a scale of probability between 0 and 1. Because of the binary nature of the response variable, disease status, the logit link function, is considered as the canonical link function for a binomial generalized linear model (GLM). It is represented by the following equation:

$$logit(\theta_i) = ln(\frac{\theta_i}{1-\theta_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \quad (10)$$

Where θ_i is the probability of interest for the ith observation and $x_{i1}, x_{i2}, \dots, x_{in}$ are the values for the n covariates of interest measured at the observation. β_0 is the intercept value. While $\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients that determine the size of the effect of the respective covariate.

Given the previous representation of the logit function of θ_i , θ_i can then be represented as:

$$\theta_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_U x_{iU})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_U x_{iU})} \quad (11)$$

The residual sum of squares for model kth, the residual sum of squares is $SSE_k = \| \boldsymbol{Y} - \hat{\mu}_k \|^2$, considering $\hat{\mu}_k = \boldsymbol{X}_k \hat{\beta}_k$ is the calculated mean vector, while $\hat{\beta}_k$ is the estimated value using least squares of β_k,

For a nested class of models ,the residual sum of squares $SSE_k = \| \boldsymbol{Y} - \boldsymbol{X}_k \hat{\beta}_k \|^2$ declines, and the greatest number of parameters reach the minimum of $\{SSE_k\}$. Therefore, choosing the optimal model does not always depend on minimizing SSEk across k models. However, $D_k = \mathrm{E}[\| \boldsymbol{\mu} - \hat{\mu}_k \|^2]$,the anticipated squared distance between μ and $\hat{\boldsymbol{\mu}}_k$,provides a decent indicator of how μ is given by accurately $\hat{\boldsymbol{\mu}}_k$ estimates.

It is ideal to select the model for which D_k is the least, however this is not possible because D_k depends on the unknown population parameter, So, after getting a decent estimate $\widehat{D}_k$ of D_k, one must minimize $\widehat{D}_k$ over the range of k = 1, …, K.it is thought that $\{\boldsymbol{X}_{\hat{k}} \hat{\beta}_{\hat{k}}\}$the most accurate estimate of the mean vector μ if the minimum is reached at $k = \hat{k}$ .In order to determine an impartial or nearly unbiased estimate of D_k. Akaike's FPE and Mallows' criteria are used.

Model selection is the process of fitting various models to a specific dataset, determining how well they perform and how difficult they are, and finally selecting the best model. This can be applied to both supervised e.g. predictive model for regression or classification or unsupervised machine learning e.g. clustering model. A considerable amount of literature has been published, pointing out different approaches for model selection. They were all developed from various academic disciplines. The Bayesian Information Criterion (BIC) and frequentist probability are the sources of the Akaike Information Criterion (AIC) (BIC). The probability that is derived from Bayesian theory. The Minimum Description Length (MDL), a second strategy, was derived from information theory. AIC and BIC are utilized to evaluate a model's log-likelihood, number of parameters, minimum description, and complexity.$BIC(k) = nlog\hat{\sigma}_k^2 + log(n)p_k$ (12)

$$AIC(k) = nlog\hat{\sigma}_k^2 + 2p_k \quad (13)$$

To determine the prediction error of the kth fitted model, cross-validation (CV) and generalized cross-validation (GCV) are used. Whereas log-likelihood comes from Maximum Likelihood Estimation i.e., the process of increasing the conditional likelihood of witnessing data in a particular probability distribution. This method is used to discover or improve a model's parameters in response to training data.

IV. Explorartory Data Analysis

A. Introduction to Exploratory Data Analysis(EDA)

Exploratory data analysis (EDA) examines a dataset to find patterns and anomalies, test hypotheses, and use statistical measures. A proper EDA process reveals ground truths about the dataset while avoiding assumptions. There are various steps to perform EDA, including problem definition, data preparation, data analysis, and development and representation of results.

B. Exploratory Data Analysis Results

In this cross-sectional study, a cohort dataset that consists of 1262 records was analyzed. The dataset includes the following variables: disease status (categorical), age (continuous), gender (categorical), smoking status (categorical), and ethnicity (categorical).

Figures 1 to 4 demonstrate the first quartile, median, third quartile, and maximum values for Age as a continuous variable, as well as a five-number summary for the categorical variable's disease state, smoking status, sex, and ethnicity.

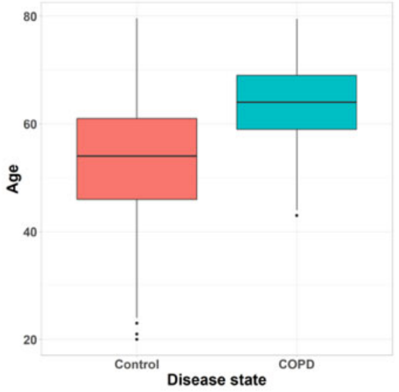

Fig. 1 Disease state versus Age

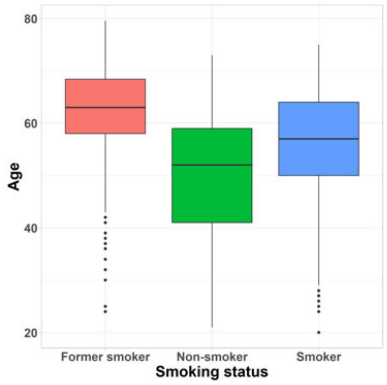

Fig. 2 Smoking status versus Age

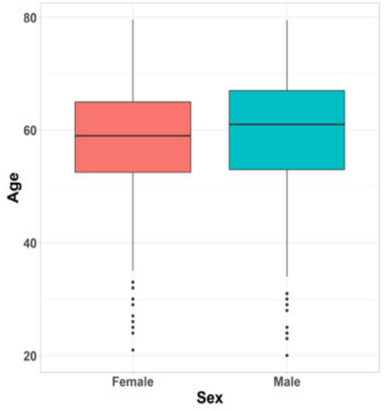

Fig. 3 Sex versus Age

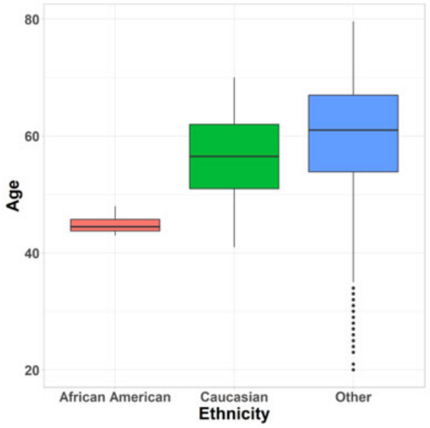

Fig. 4 Ethnicity versus Age

This EDA concluded that patients with COPD were of higher age group. Other descriptive statistics of the studied dataset are demonstrated in tables 1-3.

TABLE 1 FREQUENCY DISTRIBUTION OF DISEASE VERSUS SMOKING STATE

	Former Smoker	Nonsmoker	Smoker	Total
Control	140	182	252	**574**
COPD	521	1	166	**688**
Total	**661**	**183**	**418**	**1262**

TABLE 2 FREQUENCY DISTRIBUTION OF GENDER VERSUS SMOKING STATE

	Former Smoker	Nonsmoker	Smoker	Total
Male	251	65	154	**470**
Female	410	118	264	**792**
Total	**661**	**183**	**418**	**1262**

TABLE 3 FREQUENCY DISTRIBUTION OF GENDER VERSUS SMOKING STATE

	Nonsmoker	Smoker	Total
Male	211	363	**574**
Female	259	429	**688**
Total	**470**	**792**	**1262**

The statistical test Chi-Square is used to calculate the significance level of association between the predicted variables and the predictors by looking at the value of p. A value of p that is less than 0.05 indicates a confidence level of 95%. The results show that smoking (chi-squared = 356.69, df = 2, p < 2.2e-16) and sex (chi-squared=0.13603, df = 1, p-value = 0.7123 are Significantly connected to the disease condition, specifically COPD. With a value of p>0.05, the disease status is not substantially correlated with gender or age.

V. Results and Analysis

A binomial logistic regression model was developed to predict the COPD status accounting for multiple covariables including smoking status, sex, age, and ethnicity. Model performance and validation were carefully measured to ensure accurate results.

Both backward stepwise regression (also known as backward elimination) and forward selection approaches were used during model selection. the process started with a full saturated model. Following backward elimination approach, the number of predictors was reduced to achieve highest accuracy. Similarly, the forward selection was also used and began with only an intercept. Variables were added gradually while observing its performance.

The selected model is based on BIC and the regression model is described as follows:

$$\eta_i = \textit{-5.3970} \; \text{-19.0724}\times \textit{nonsmoker} \;\; \textit{-1.3668}\times \textit{Smoker} + 0.1095\times \text{Age} \quad (14)$$

Where the intercept value β_0=-5.3970, β_1=19.0724, β_2 = -1.3668, and β_3 0.1095.

The model indicates the key covariates are age and smoking status with corresponding *p*-values of the coefficients ≤ 0.05, indicating the significance of these parameters. The model was used to run predictions on a randomly selected test set and the resulting accuracy is 73%.

Assuming that the data are independently distributed, the response (dependent) variable assumes a binomial distribution (part of the exponential family), The GLM model does not assume a linear relationship between the Disease state and any of the independent variables, and it uses the maximum likelihood estimate (MLE) to estimate the parameters (coefficients).

The calibration graph generated for the studied dataset indicated good calibration, and the test for the Hosmer Lemeshow c-statistic demonstrated no lack of fit as shown in figure 5.

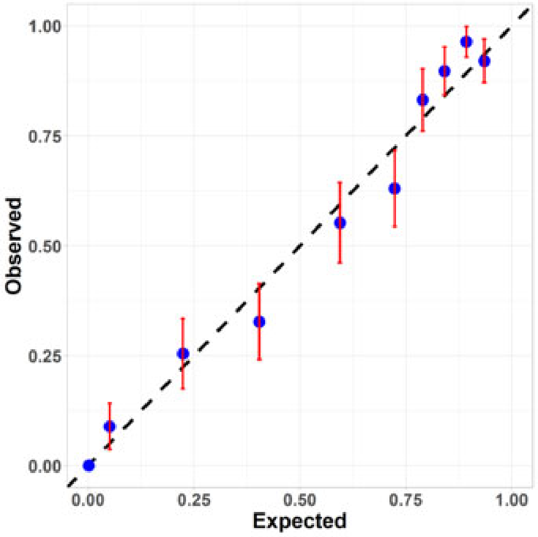

Figure 5 Hosmer Lemshow graph for the dataset

VI. Conclusions

COPD is a progressive lung condition that is known third-leading cause of mortality globally. A statistical model was developed to predict the probability of a patient contracting COPD. The model developed is a binomial regression model with a logit link function. Following rigorous model selection process, the results obtained indicated the significance of the smoking status and age on COPD disease state. The model achieved a 73% accuracy on a randomly selected independent test set. By identifying those who are at a high risk for COPD, the model could be used as a tool for early detection programs.

The challenge of using regression is recognised in epidemiology since the models rely on observational data. Hence, proper measures should be considered to ensure covariates are correctly measured and adjusted.

References

[1] WHO. Chronic obstructive pulmonary disease (COPD) 2020. Available from: https://www.who.int/ news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[2] Alqahtani JS, Oyelade T, Aldhahir AM, Mendes RG, Alghamdi SM, Miravitlles M, et al. (2021) Reduction in hospitalised COPD exacerbations during COVID-19: A systematic review and meta-analysis. PLoS ONE 16(8): e0255659. https://doi.org/10.1371/journal.pone.0255659

[3] Li, T., Zhou, H. P., Zhou, Z. J., Guo, L. Q., & Zhou, L. (2021). Computed tomography-identified phenotypes of small airway obstructions in chronic obstructive pulmonary disease. Chinese Medical Journal, 134(17), 2025-2036.

[4] Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. 2021 [cited 2021 01/03/2021]. Available from: https://goldcopd.org/wp-content/uploads/2019/11/GOLD-2020-POCKET-GUIDE-FINALpgsized-wms.pdf.

[5] Hurst JR, Vestbo J, Anzueto A, Locantore N, Mu¨llerova H, Tal-Singer R, et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. The New England journal of medicine. 2010; 363 (12):1128–38. Epub 2010/09/17. https://doi.org/10.1056/NEJMoa0909883 PMID: 20843247.

[6] Moreira, A. T. A. D., Pinto, C. R., Lemos, A. C. M., Assunção-Costa, L., Souza, G. S., & Martins Netto, E. (2021). Evidence of the association between adherence to treatment and mortality among patients with COPD monitored at a public disease management program in Brazil. Jornal Brasileiro de Pneumologia, 48.

[7] Chang JT, Meza R, Levy DT, Arenberg D, Jeon J (2021) Prediction of COPD risk accounting for time-varying smoking exposures. PLoS ONE 16(3): e0248535. https://doi.org/10.1371/journal.pone.0248535

[8] Athlin, Å., Giezeman, M., Hasselgren, M., Montgomery, S., Lisspers, K., Ställberg, B., ... & Sundh, J. (2021). Prediction of Mortality Using Different COPD Risk Assessments–A 12-Year Follow-Up. International Journal of Chronic Obstructive Pulmonary Disease, 16, 665.

[9] M. Cazzola, C. P. Page, L. Calzetta, and M. G. Matera, Pharmacology and Therapeutics of Asthma and COPD. Springer, Cham, 2012, pp. 2-8.

[10] Esteban, C., Moraza, J., Sancho, F., Aburto, M., Aramburu, A., Goiria, B., ... & Capelastegui, A. (2015). Machine learning for COPD exacerbation prediction.

[11] Molfino, N. A. (2007). Genetic predisposition to accelerated decline of lung function in COPD. International Journal of Chronic Obstructive Pulmonary Disease, 2(2), 117.

[12] Lopez, A. D., Shibuya, K., Rao, C., Mathers, C. D., Hansell, A. L., Held, L. S., ... & Buist, S. (2006). Chronic obstructive pulmonary disease: current burden and future projections. European Respiratory Journal, 27(2), 397-412.

[13] Hurst, J. R., Siddiqui, M. K., Singh, B., Varghese, P., Holmgren, U., & de Nigris, E. (2021). A systematic literature review of the humanistic burden of COPD. International Journal of Chronic Obstructive Pulmonary Disease, 16, 1303.

[14] Halpin, D. M., Criner, G. J., Papi, A., Singh, D., Anzueto, A., Martinez, F. J., ... & Vogelmeier, C. F. (2021). Global initiative for the diagnosis, management, and prevention of chronic obstructive lung disease. The 2020 GOLD science committee report on COVID-19 and chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine, 203(1), 24-36.

[15] Halpin, D. M., Decramer, M., Celli, B. R., Mueller, A., Metzdorf, N., & Tashkin, D. P. (2017). Effect of a single exacerbation on decline in lung function in COPD. Respiratory medicine, 128, 85-91.

[16] Seemungal, T. A., Donaldson, G. C., Paul, E. A., Bestall, J. C., Jeffries, D. J., & Wedzicha, J. A. (1998). Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine, 157(5), 1418-1422.

[17] Blasi, F., Cesana, G., Conti, S., Chiodini, V., Aliberti, S., Fornari, C., & Mantovani, L. G. (2014). The clinical and economic impact of exacerbations of chronic obstructive pulmonary disease: a cohort of hospitalized patients. PloS one, 9(6), e101228.

[18] Asia Pacific COPD Roundtable Group. (2005). Global Initiative for Chronic Obstructive Lung Disease strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease: An Asia–Pacific perspective. Respirology, 10(1), 9-17.

[19]Müllerová, H., Shukla, A., Hawkins, A., & Quint, J. (2014). Risk factors for acute exacerbations of COPD in a primary care population: a retrospective observational cohort study. BMJ open, 4(12), e006171.

[20] Hurst, J. R., Vestbo, J., Anzueto, A., Locantore, N., Müllerova, H., Tal-Singer, R., ... & Wedzicha, J. A. (2010). Susceptibility to exacerbation in chronic obstructive pulmonary disease. New England Journal of Medicine, 363(12), 1128-1138.

[21] Bafadhel, M., Peterson, S., De Blas, M. A., Calverley, P. M., Rennard, S. I., Richter, K., & Fagerås, M. (2018). Predictors of exacerbation risk

and response to budesonide in patients with chronic obstructive pulmonary disease: a post-hoc analysis of three randomised trials. The Lancet Respiratory Medicine, 6(2), 117-126.

[22] Pascoe, S., Barnes, N., Brusselle, G., Compton, C., Criner, G. J., Dransfield, M. T., ... & Singh, D. (2019). Blood eosinophils and treatment response with triple and dual combination therapy in chronic obstructive pulmonary disease: analysis of the IMPACT trial. The Lancet Respiratory Medicine, 7(9), 745-756.

[23] Ferguson, G. T., Rabe, K. F., Martinez, F. J., Fabbri, L. M., Wang, C., Ichinose, M., ... & Reisner, C. (2018). Triple therapy with budesonide/glycopyrrolate/formoterol fumarate with co-suspension delivery technology versus dual therapies in chronic obstructive pulmonary disease (KRONOS): a double-blind, parallel-group, multicentre, phase 3 randomised controlled trial. The Lancet Respiratory Medicine, 6(10), 747-758.

[24] Rabe, K. F., Martinez, F. J., Ferguson, G. T., Wang, C., Singh, D., Wedzicha, J. A., ... & Dorinsky, P. (2020). Triple inhaled therapy at two glucocorticoid doses in moderate-to-very-severe COPD. New England Journal of Medicine, 383(1), 35-48.

[25] Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. CNS neuroscience & therapeutics, 24(11), 1037-1052.