Proceedings of the 4th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/84/20240773

Data analysis of lung cancer incidence

Minhao Li[1,5], Qiyue Liu[2], Baoyi Zhang[3], Jiachen Sun[4]

[1]School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Changsha, 215028, China
[2]Global Education, Guangzhou, 510630, China
[3]The Affiliated International School of Shenzhen University, Shenzhen, 518057, China
[4]Capital Normal University High School, Beijing, 100048, China

[5]Minhao.li20@student.xjtlu.edu.cn

Abstract. Lung cancer has been widely considered as a global high incidence-disease, and its incidence is related to many factors, like smoking, genetics, and other diseases. This paper aims to model and analyse a database to further study the relationship between the incidence of lung cancer and other factors by utilizing a clinical database containing the patient information such as disease diagnoses, age, gender, and smoking. The model finds that patients with allergy and swallowing difficulty have significantly high incidence of lung cancer, while smoking does not show a highly positive association with lung cancer in the model which is widely confirmed should be highly related with lung cancer. The problem could be related to the sample in the database, since most of people in the dataset are getting respiratory diseases who are more likely to quit smoking. Finally, the paper discusses the associated causes of lung cancer and potential risks that other diseases may have based on the model result.

Keywords: Lung cancer, disease, data analysis, risk factors, screening

1. Introduction

Lung carcinoma, also known as lung cancer, is a kind of malignancy that occurs in the lung. To be more specific, lung cancer is the genetic damage of the lung cell. For example, when people inhaling inhale oxygen and release carbon dioxide by exhaling, some harmful chemicals and particles enter the lung and damage the DNA. There are several symptoms of lung cancer. Firstly, coughing that lasts for a long time. Next, chest pain and shortness of breath. Also, wheezing, coughing up blood and dramatically weight loss can be the symptoms of lung cancer. People can do some precautionary measures and diagnose it earlier. Despite the causes of lung cancer being almost the respiratory carcinogens in the environment, the substantial individual variation in susceptibility to them is one of the key factors. In this case, the risk of lung cancer can be understood as the combined consequences of the interrelationship between exposure to causative agents and the individual susceptibility to these agents [1]. Besides, among all cancers, lung cancer is not only the most common cancer, but also has a high incidence. In the year 2000, lung cancer accounted for 12.3% of all cancer cases. It's estimated that approximately 1.2 million new cases of lung cancer occurred that year. Smoking is the most important cause of lung

DOI: 10.54254/2755-2721/84/20240773

cancer. More than 80% tobacco smokers are suffering from lung cancer [2]. However, as our research, factors below may have influenced the morbidity of lung cancer.

1.1. Gender

Cigarette smoking is actually a substantial determinant to affect the sex differences of lung cancer. Generally, men are more likely to be addicted to tobacco smoking than women. The data in 2015 illustrates that there was 16.7 percent of adult males have a history of smoking, and for adult females, there was only 13.6 percent. Such results are related to a combination of some factors (lifestyle, hormones, and cultural factors). For instance, study indicates that there is higher testosterone, androstenedione, estradiol, estrone, 17-OHP, and SHBG levels in the body of postmenopausal women who smoke [3].

1.2. Age Group

Study shows that lung cancer can own by both 0.03% of men and women up to the age of 39. This rate dramatically increases around 45-49 years old. The percentage peaked in the 85-89 age group. Lung cancer frequently occurs in the elders, most of the cases of lung cancer occurring patients who over the age of 65 years old. people of any age can suffer from lung cancer, but it is most common at the range of 54-75 years old. We should aware the symptoms of the disease to diagnose it early, but the diagnosis of lung cancer can vary depending different age group. The symptoms of lung cancer of older people are usually persistent cough and high blood pressure, whereas young people are more likely to have the symptoms such as severe fatigue [4].

1.3. Smoking

Smoking is one of the leading risk factors for cancer in the world. Carcinogens in cigarette can weaken people's immune system and damage the lung tissues. Carcinogens can also affect cellular DNA and can lead to the development of cancer. When people smoke, the carcinogens they inhale cause mutations in our DNA. According to a 2016 study, when consume a pack of tobacco cigarettes a day on average has approximately 150 mutations Trusted Source in the DNA of any given lung cell each year [5].

We collect clinical data about lung cancer and analyze the dataset using R code. The purpose of this essay is to predict the morbidity of lung cancer by researching the relationship between lung cancer and the possible factors that may affect lung cancer.

2. EDA

2.1. Data Pre-processing

Since the dataset got is complete enough on its own and had no outliers after checking, we just did some simple pre-processing. Because most of the data only represents whether a person has the disease or not, the "yes" and "no" in the dataset had been replaced by numbers "1"and "0", and convert them to factor form to make R easier to run.

At the beginning of the data analysis, since we had many groups of factors associated with lung cancer, we need to know which ones we may focus on in the following research. We choose to using grafts to briefly compare the relationship between different factors and lung cancers, and here are some important factors in Figure 1 we might need to pay more attention to.

DOI: 10.54254/2755-2721/84/20240773

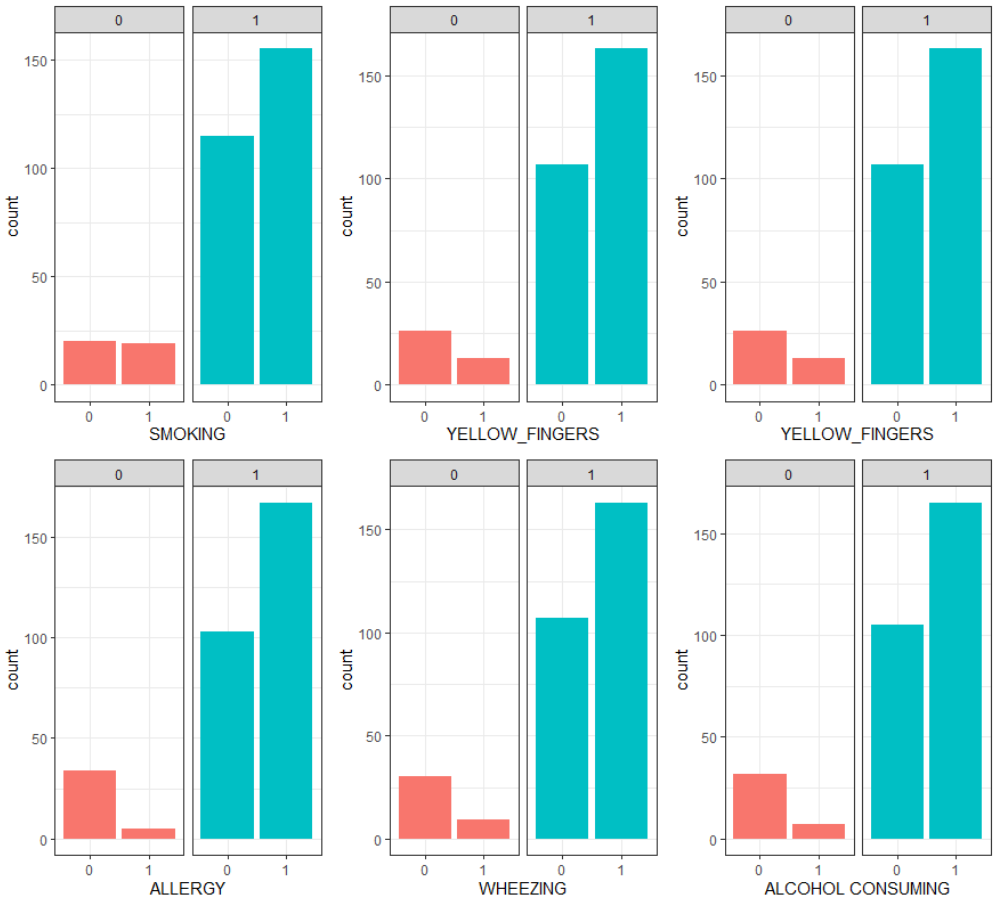

Figure 1. Some important factors' comparing graft (1 represent having disease, 0 represent not having disease).

We can find some elements that have higher correlation with lung cancer, for example, the rates of people who get fatigue and shortness of breath are significantly higher among people with lung cancer, and the rates of people who have alcohol consuming and swallowing difficulty are significantly lower among people who not have lung cancer. And also, we can find from the graft that age is the only factor contain multiple numbers and most people in the dataset are over 40 years old, which means the prediction made from the dataset may not be correct for under 40. Those can be important reference for our later data analysis. Last, we divide age into four groups: under50, 50-59, 60-69, over 70, in order to facilitate the analysis of age.

2.2. Accuracy

Then we use two different classification methods to test if the data can give us a satisfying prediction with lung cancer. LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) import the dataset to fit the Gaussian distribution and estimate the mean and covariance matrix to build a model. Then the model classification results are compared with the formal data of the test set to determine the accuracy of the model. QDA generally performs better when the decision boundary is non-linear. However, the QDA gives dataset the 96.12% accuracy with 1.129e-07 p-value and LDA shows 93.85% accuracy with 0.0001427 p-value, which means the dataset is trustable and have high possibility to deduce a good prediction model. Furthermore, since QDA gives a better score in accuracy, we can infer that the dataset in non-linear.

DOI: 10.54254/2755-2721/84/20240773

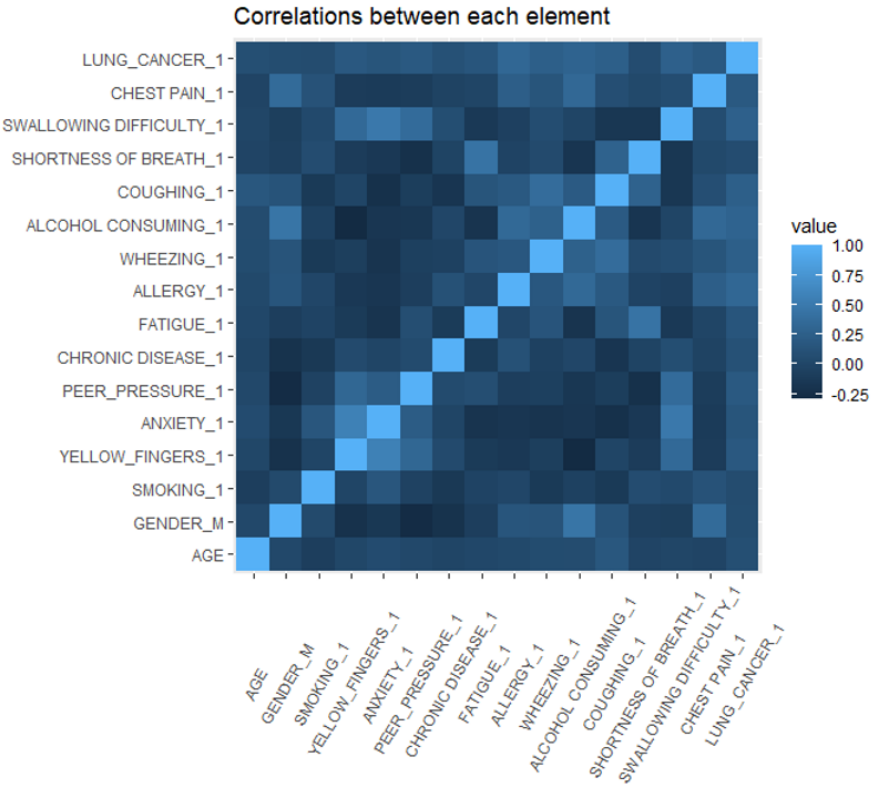

Figure 2. The correlation graft.

2.3. Correlation Between Each Element

To find the relationship between each element, we create a graft to visualize it. As shown in Figure 2. And by selecting the element with more than 0.4 cor-value, we find four pairs of elements that have strong correlation: gender & alcohol consuming, anxiety & yellow fingers, swallowing difficulty & anxiety, fatigue & shortness of breath. And to find which factor could be more important in later lung cancer prediction, we calculate the correlation of lung cancer with others and list the few factors with relatively stronger correlation: Allergy, wheezing, alcohol consuming, coughing and swallowing difficulty.

3. Methods to Analysis

As for the method, we decide to using a Decision tree as the method to analyze. Decision tree is a tree-like model of decisions and their possible consequences. In the proposed work, we predict the chance of one having lung cancer through their pre-existing condition, and the model perfectly fits our needs. First, we split the data into two separate data frames, 80 percent of the data is used to train the model, and the rest 20% will be used to make predictions. Since the original data is not listed in a particular way, we randomly select those data and check the lung cancer proportion of each group to verify if the randomization process is correct. After that, we import the training group data into the decision tree model, and by comparing the xerror (cross-validation relative error) came from a different model, we get the model with four different variables which have the lowest xerror and an appropriate number of nodes. The result has shown in Figure 3.

DOI: 10.54254/2755-2721/84/20240773

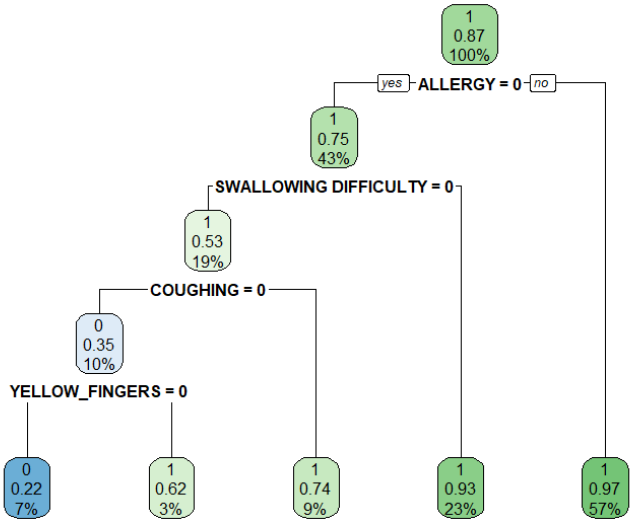

Figure 3. The first model result.

To test if the model is accurate enough, we import the prepared test group to do the prediction. To be exact, we first use the model to predict if one has lung cancer and then compare if it is correct. By using confusion matrix, we can get the accuracy rating, the function of accuracy has been shown below:

$$accuracy = (TP + TN)/(TP + TN + FP + FN)$$

There we got a score of 87 percent for the test set, which shows the model did a good prediction.

3.1. Model Diagnostics

The problem is, even though the score of accuracy seems good, our dataset is too small that the p-value of the prediction is only 0.59 which is too high to guarantee the score is correct. The problem also influences the result of our randomization process, using different randomly selected data can lead the model different from others. In this case, we try to use different selected data to build the model and even one with importing the entire dataset. The result is, the different groups give the different probability of getting lung cancer, even though there are gaps between them, the conclusion always shows that people who have allergy, swallowing difficulty have significantly more probability of getting lung cancer. Besides, patients who get chest pain, alcohol consuming, yellow figure and peer pressure have higher incidence of lung cancers compared to other factors.

4. Discussion

No one expects such a result that those factors like age and smoking which have been proved to be associated with lung cancer do not show on the list, but it's reasonable to get the conclusion. The database was collected in an unfair method, since about 90% of samples in the dataset are getting lung cancer which means. Besides, the database contains patients only and most get respiratory diseases, which means we basically did not use the healthy people for the comparison. The conclusion could be completely changed if we add few young and healthy people in the dataset, since the weight coefficient of age could be significantly increasing. And people who get respiratory diseases are more likely to quit smoking for their recovery, which could be the reason why smoking is not on the list. Even though many researches provide overwhelming evidence that smoking causes lung cancer, the lung cancer patients are most likely to abandon smoking at the same time. There is some further research we made about some factors associated with lung cancer incidence.

Nowadays, the incidence of cancer is increasing at an alarming rate. Lung cancer, regardless of gender, is one of the most prevalent types in many countries. Various factors contribute to its

DOI: 10.54254/2755-2721/84/20240773

development, including genetic susceptibility, unhealthy diet, exposure hazards, and air pollution. These factors can act independently or in conjunction with tobacco smoking to shape the descriptive epidemiology of lung cancer [6]. Our research that contributes to lung cancer. The association between smoking and lung cancer has been extensively studied and firmly established. Smokers face up to a 30-fold higher risk of developing this disease compared to nonsmokers. In the United States alone, lung cancer accounts for 31% and 26% of all male and female cancer-related deaths respectively [7,8]. Smoking stands out as the primary cause behind approximately 85% of all diagnosed cases. When tobacco is smoked, it releases numerous harmful chemicals into the lungs including carcinogens that damage DNA in lung cells leading to tumor formation.

The development of lung cancer is driven by multiple genetic mutations extensively investigated by various researchers [9], yet our understanding regarding its molecular pathogenesis remains incomplete. In order to comprehensively comprehend how lung cancer develops, it is crucial to consider not only the tumor microenvironment (TME), but also the inflammatory pathways involved in carcinogenesis [10]. Carcinoma in the lungs occurs alongside neighboring inflamed cells as well as structural and stromal cells. Lung diseases associated with high risks for developing inflammation include COPD [11,12]. Profound abnormalities can be observed within inflammatory pathways related to COPD [13,14]. Notably among various cytokines, growth factors, and mediators present in the developing TME, interleukin (IL)-1β, prostaglandin E2 (PGE2), and transforming growth factor (TGF)-β have been found to simultaneously pave the way for both epithelial–mesenchymal transition (EMT) and destruction of specific host cell–mediated immune responses against tumor antigens [15,16].

Even though the model not perform well in some aspects, the result can be used to research the potential risk that lung cancer patients may have. As the model shows, people who get allergy or swallowing difficult are predicted to have more than 90% probability of getting lung cancer. Indeed, more than 60% of the sample who get lung cancer also get one of the two diseases in the dataset, which means allergy and swallowing difficulty somehow have a strong correlation with lung cancer. There could several reasons lead to the association between allergy and lung cancer, however, many essays mentioned about the inverse association between allergies and cancer. For example, a study examined the lung cancer risk in woman with history of asthma or hay fever and found the probability of developing lung cancer is significantly decreased [17]. The reason could be highly relevant to immunosurveillance that allergy may enhance, but the association between allergy and lung cancer is site specific and need further research to verity. As for swallowing difficultly, lung cancer may directly cause the patients dysphagia due to direct tumor invasion or nerve compression, and the lung cancer treatment and concurrent diseases may also exacerbate swallowing difficulty [18]. Swallowing difficultly could impact patients' quality of life, while there is limited literature mention about it. In this case, people should be aware of this as the potential risks that lung cancer may causes. The other factors that highly associated with lung cancer, like coughing, wheezing and fatigue, are the common symptom in lung cancer patients. The causes could be similar to swallowing difficultly which is related to the abnormal respiratory status.

5. Conclusion

Overall, this study provides a prediction model based on a clinical database, using data mining to research about the correlation between lung cancer and other health factors, which can offer support for diagnosis and treatment. Furthermore, we point out the data issue which leads some factors less associated and discuss about the associated causes and potential risk that shows in the model result. However, lung cancer prediction is a complex challenge, this study still requires further research to validate and optimize the model performance.

References

[1] Alberg, A. J., & Samet, J. M. (2003). Epidemiology of lung cancer. Chest, 123(1), 21S-49S.

[2] Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). Focus on Lung Cancer. Cancer Cell, 1(1), 49–52. https://doi.org/10.1016/s1535-6108(02)00027-2

[3] Brand, J. S., Chan, M.-F., Dowsett, M., Folkerd, E., Wareham, N. J., Luben, R. N., van der Schouw, Y. T., & Khaw, K.-T. (2011). Cigarette smoking and endogenous sex hormones in postmenopausal women. The Journal of Clinical Endocrinology & Metabolism, 96(10), 3184–3192. https://doi.org/10.1210/jc.2011-1165
[4] Lynne Eldridge, M. (2022, March 23). How lung cancer affects different age groups. Verywell Health. https://www.verywellhealth.com/lung-cancer-age-5216079
[5] Sato M, Shames DS, Gazdar AF, Minna JD. A translational view of the molecular pathogenesis of lung cancer. J Thorac Oncol 2007;2:327–343.
[6] Jyoti Malhotra, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, Paolo Boffetta European Respiratory Journal 2016 48: 889-902; DOI: 10.1183/13993003.00359-2016
[7] Youlden DR, Cramb SM, Baade PD. The international epidemiology of lung cancer: geographical distribution and secular trends. J Thorac Oncol 2008;3:819–831.
[8] Proctor RN. Tobacco and the global lung cancer epidemic. Nat Rev Cancer 2001;1:82–86.
[9] Sato M, Shames DS, Gazdar AF, Minna JD. A translational view of the molecular pathogenesis of lung cancer. J Thorac Oncol 2007;2:327–343.
[10] Prendergast GC. Inflammatory mediators in cancer etiology and targets for therapy and prevention. Cancer Reviews Online 2008;9:17–18.
[11] Taraseviciene-Stewart L, Voelkel NF. Molecular pathogenesis of emphysema. J Clin Invest 2008;118:394–402.
[12] Sevenoaks MJ, Stockley RA. Chronic obstructive pulmonary disease, inflammation and co-morbidity: a common inflammatory phenotype? Respir Res 2006;7:70.
[13] Reynolds PR, Cosio MG, Hoidal JR. Cigarette smoke-induced Egr-1 upregulates proinflammatory cytokines in pulmonary epithelial cells. Am J Respir Cell Mol Biol 2006;35:314–319.
[14] Kim V, Rogers TJ, Criner GJ. New concepts in the pathobiology of chronic obstructive pulmonary disease. Proc Am Thorac Soc 2008;5:478–485.
[15] Dohadwala M, Yang SC, Luo J, Sharma S, Batra RK, Huang M, Lin Y, Goodglick L, Krysan K, Fishbein MC, et al. Cyclooxygenase-2-dependent regulation of E-cadherin: prostaglandin E(2) induces transcriptional repressors ZEB1 and snail in non-small cell lung cancer. Cancer Res 2006;66:5338–5345.
[16] Leng Q, Bentwich Z, Borkow G. Increased TGF-β, Cbl-b and CTLA-4 levels and immunosuppression in association with chronic immune activation. Int Immunol 2006;18:637–644.
[17] Merrill, R. M., Isakson, R. T., & Beck, R. E. (2007). The association between allergies and cancer: what is currently known?. *Annals of Allergy, Asthma & Immunology*, *99*(2), 102-117.
[18] Brady, G. C., Roe, J. W., O'Brien, M., Boaz, A., & Shaw, C. (2018). An investigation of the prevalence of swallowing difficulties and impact on quality of life in patients with advanced lung cancer. *Supportive Care in Cancer*, *26*, 515-519.