Research on factors associated with lung cancer

Yueyang Jin[1, 4,†]**, Yudi Qiao**[2, †]**, Haiqing Yang**[3, †]

[1]Mathematics at Case Western Reserve University, OH, US;
[2]The High School Affiliated to Renmin University of China, Beijing, China;
[3]Shijiazhuang Foreign Language School, Hebei, China.
[†]These authors contributed equally.

[4]yxj606@case.edu

Abstract. Lung cancer has become one of the most widely infected cancers around the world and is getting more attention from scientists and the public. Thus, based on the dataset Lung Cancer uploaded on Kaggle in 2022, this paper reviews previous research on risk factors about their effects and mechanisms on lung cancer. Besides, by transforming categorical variables into numerical ones and using chi-square tests of independence, this paper examines whether independent variables in the dataset are associated with lung cancer development. This paper finds that variables 'gender', 'smoking', 'chronic disease' and 'shortness of breath' are not tested associated with lung cancer development; variable 'yellow finger' and the other nine independent variables are tested as factors contributing to lung cancer. Identifying these risk factors and analyzing their mechanisms can effectively help people prevent lung cancer and support the development of lung cancer treatment.

Keywords: lung cancer, risk factors, chi-square tests.

1. Introduction

Lung cancer is the leading cause of cancer death worldwide, with an average of 1.7 million deaths a year due to tobacco use worldwide [1]. In 2020, the World Health Organization (WHO) published the incidence and death rates of lung cancer worldwide. About 2.2 million new cases of lung cancer were reported in 2020, accounting for 11.4 percent of all malignancies, ranking second. By 2020, lung cancer will cause 1.79 million deaths, accounting for 18 percent of all malignancies, ranking first [2].

Lung cancer is cancer that begins in the lungs. Cancer begins when cells in the body begin to grow out of control. Lung cancer occurs when cells in the lungs grow abnormally and form tumors [3]. This can be seen as a nodule or mass on a chest X-ray or CT(" CAT ") scan. There are two main types of cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), and about 80 to 85 percent of lung cancers are NSCLC. Lung cancer is now the leading cause of cancer death worldwide, killing 1.59 million people in 2018. Most cases of lung cancer are caused by smoking, but exposure to air pollution is also a risk factor. Air pollution may be linked to an increased risk of lung cancer, even among non-smokers, a new study has found [4]. Based on the data set, gender, air pollution, smoking, stress... These factors can affect people's health and lead to lung cancer. In many high-income countries, women ages 30 to 49 are diagnosed with lung cancer at a higher rate than men of the same age, according to a new study published in the International Journal of Cancer. Outdoor air pollution is a major

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

contributor to the global lung cancer burden, as the majority of the world's population now lives in places with high levels of air pollution. Lung cancer is more common in people who smoke. Smoking is the main cause of lung cancer; Fifty-five percent of lung cancer deaths in women and more than 70 percent in men are caused by smoking. In summary, these factors promote the occurrence of lung cancer to varying degrees [5].

Therefore, it is particularly necessary to analyze the factors that cause lung cancer. By plotting line charts and using the Chi-square test of independence, the relationship between 15 variables and lung cancer development is tested. After analyzing the data, the risk factors for lung cancer are identified.

2. Methodology

2.1. Data

2.1.1. Data source. The dataset is downloaded from Kaggle and is updated a year ago. It includes 15 independent variables that might be associated with lung cancer development and 1 dependent variable which shows whether a subject has developed lung cancer. 309 subjects are included in the dataset.

2.1.2. Variable information. Basic information of the variables is shown in Table 1.

Table 1. Variable specification.

Category	Variable	Variable details
Independent variable	GENDER	M(male), F(female)
	AGE	Age of the subject
	FATIGUE	YES:2, NO:1
	YELLOW_FINGERS	
	CHRONIC_DISEASE	
	PEER_PRESSURE	
	ANXIETY	
	SMOKING	
	ALLERGY	
	ALCOHOL CONSUMING	
	WHEEZING	
	COUGHING	
	CHEST PAIN	
	SWALLOWING DIFFICULTY	
	SHORTNESS OF BREATH	
Dependent variable	LUNG_CANCER	YES, NO

2.1.3. Data processing and transformation. Among the 16 variables, 'GENDER' and 'LUNG_CANCER' are categorical and the others are numerical. The characters in columns 'GENDER' and 'LUNG_CANCER' are replaced by numbers for the convenience of the follow-up study.

2.2. Chi-square test of independence

The chi-square (X2) test of independence is a nonparametric hypothesis test [6]. It is used to test for a relationship between two categorical variables. It is known that if two variables that are categorical are independent, then it can be written as P(A)=P(A|B). The χ^2 test utilizes this information to calculate expected values for cells in a two-way contingency table assuming that the two categorical variables are independent, which is called the null hypothesis.

The χ^2 test is used in this study to examine whether the variables are associated with lung cancer development. The null hypothesis here is that A (the tested variable) is independent of lung cancer development. All tests are at the confidence level(α) of 0.05.

3. Statistical analysis

3.1. Data visualization

3.1.1. The relationship between age and lung-cancer development. Figure 1 is plotted to show the age distribution of subjects who have developed lung cancer.

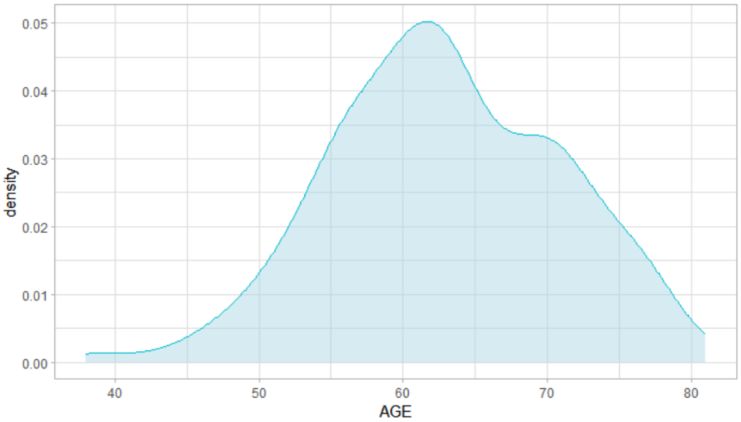


Figure 1. Line chart of age distribution (Photo credit: Original).

Figure 1. shows that the data approximates a normal distribution, with a mean of about 62 years old.

3.1.2. Gender. Figure 2 displays the percentage of males and females who got lung cancer based on the dataset from Kaggle. From this chart, we know that more males got lung cancer than females in this dataset, which is reasonable as typical males are more likely going to smoke and drink in their lives compared to females according to the article published by City of Hope last year in February [7].

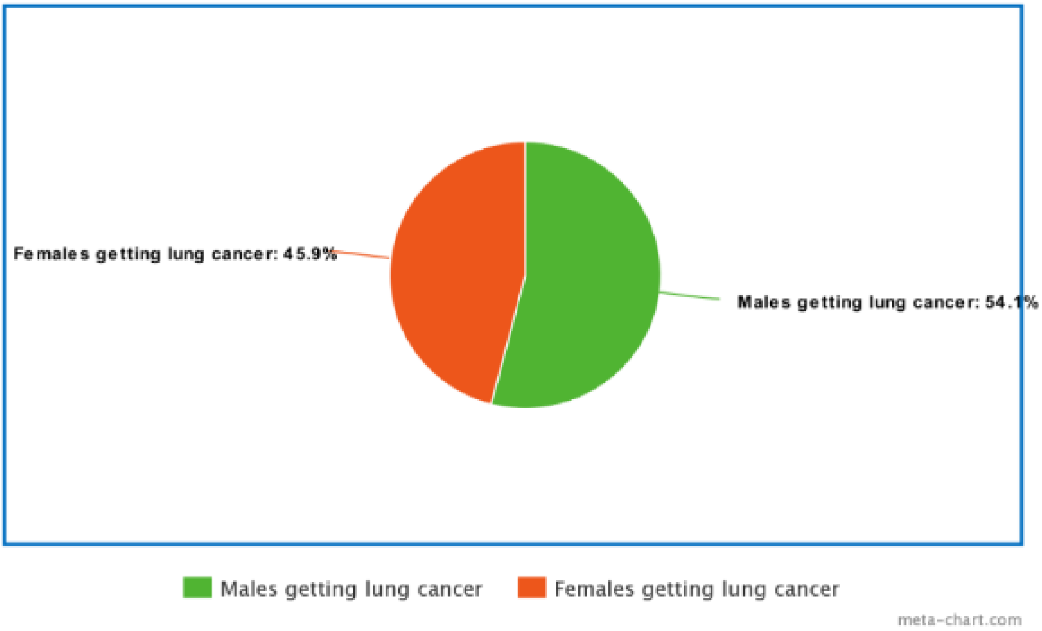


Figure 2. Distribution of males and females who got lung cancer (Photo credit: Original).

3.1.3. Smoking. Among subjects who have developed lung cancer, the proportion of those who smoke and those who do not smoke can be seen in Figure 3.

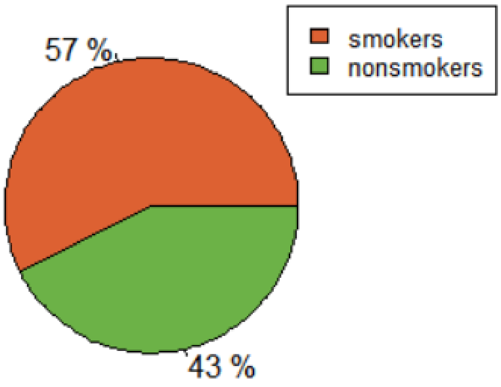


Figure 3. The proportion of smokers and nonsmokers among lung cancer patients (Photo credit: Original)

Figure 3 shows that smokers make up a larger proportion of lung cancer patients than non-smokers. This is probably because chemicals in cigarettes can affect smokers by making large changes to their genes. These genetic mutations are contributors to lung cancer development [8].

3.1.4. Yellow fingers. Among subjects who have developed lung cancer, the proportion of those with yellow fingers and those without yellow fingers can be seen in Figure 4.

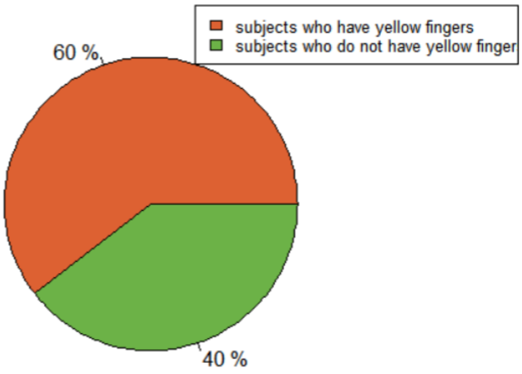


Figure 4. The proportion of patients with yellow fingers (Photo credit: Original).

Figure 4 shows that people with yellow fingers make up a larger proportion of lung cancer patients than people without yellow fingers. Yellow fingers are one of the characteristics of people who smoke for a long time. So the relationship between yellow fingers and lung cancer development is similar to that between smoking and lung cancer development.

3.1.5. Anxiety. Among subjects who have developed lung cancer, the proportion of those with anxiety and those without anxiety can be seen in Figure 5.

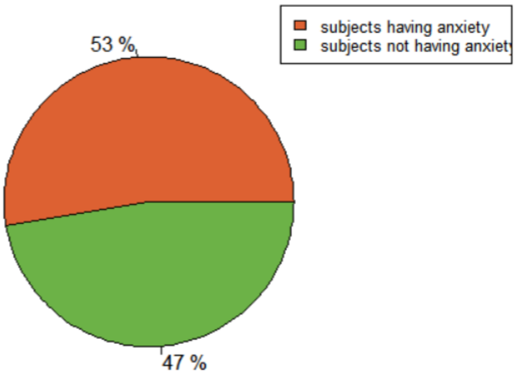


Figure 5. The proportion of patients with anxiety (Photo credit: Original)

Figure 5. shows that there are more lung cancer patients with anxiety than those without anxiety. Research has shown that psychological anxiety will bring disorders to the vegetative nerves of the human body, thus affecting the endocrine system and immune systems, resulting in a period of a sharp decline in human immunity [9]. This would lead to increased lung cancer risk [10].

3.1.6. Peer pressure Among subjects who have developed lung cancer, the proportion of those having peer pressure and those not having peer pressure can be seen in Figure 6.

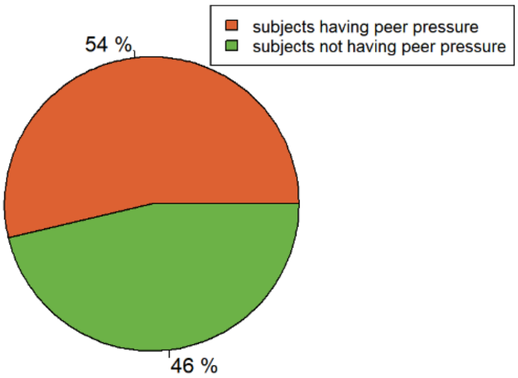


Figure 6. The proportion of lung cancer patients having and not having peer pressure (Photo credit: Original).

Figure 6 shows that there are more lung cancer patients having peer pressure than not having peer pressure. This is because pressure can cause a decrease in immunity [11], thereby increasing the risk of lung cancer development [10].

3.1.7. Chronic disease. Among subjects who have developed lung cancer, the proportion of those having chronic diseases and those not having a chronic disease can be seen in Figure 7.

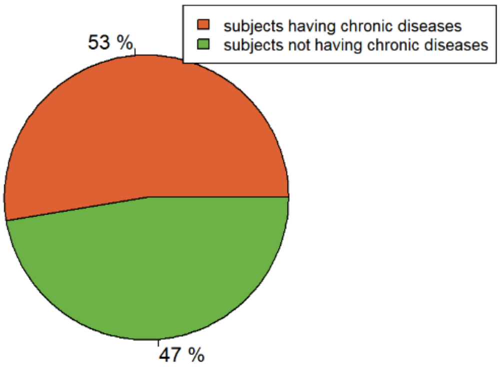


Figure 7. The proportion of patients with chronic diseases (Photo credit: Original).

Figure 7 shows that there are more lung cancer patients having chronic diseases than those not have chronic diseases. Multiple studies have shown that chronic lung diseases can increase the risk of lung cancer: non-smokers with COPD had a 167% increased risk of lung cancer compared with non-smokers without COPD [12]; TB damages lung tissue and causes fibrosis, scarring, and genetic changes, increasing the risk of lung cancer by 1.7 times [13]; in a study of 90, 000 people who had been hospitalized for asthma, 713 developed lung cancer, a 58% higher incidence rate than in the general population [14].

3.1.8. Fatigue. Among subjects who have developed lung cancer, the proportion of those having fatigue and those not having fatigue can be seen in Figure 8.

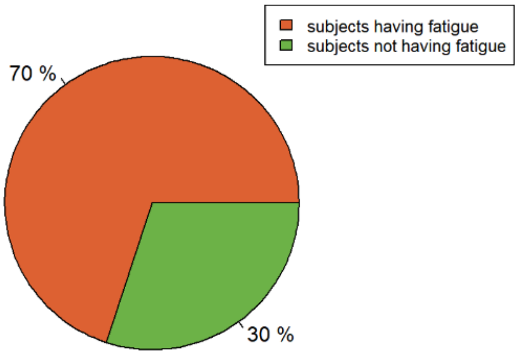


Figure 8. the proportion of lung cancer patients having and not having fatigue (Photo credit: Original).

Figure 8 shows that people having fatigue make up a larger proportion of lung cancer patients than people not having fatigue. This is because fatigue can cause a decrease in immunity [9] just like peer pressure, which can contribute to increased lung cancer risk [10].

3.1.9. Allergy. Among subjects who have developed lung cancer, the proportion of those having allergies and those not having an allergy can be seen in Figure 9.

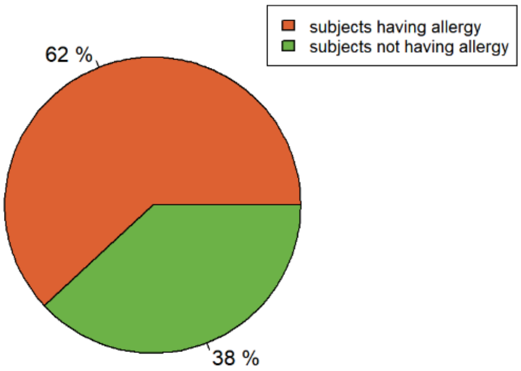


Figure 9. The proportion of lung cancer patients having and not having allergies (Photo credit: Original).

Figure 9 shows that there are more lung cancer patients having allergies than not having an allergy. However, allergy can decrease the risk of lung cancer development, which contradicts the results shown in Figure 9. This is probably due to the limitations of this dataset(for detailed analysis, see 5. Discussion).

3.1.10. Wheezing. When we first look at Figure 10 generated by the dataset we have for wheezing, we noticed that among the lung cancer patients, there are definitely more people who also got wheezing, as wheezing is one of the "results of lung tumor", or lung cancer in more general terms according to the article "Signs and Symptoms of Lung Cancer" published by HealthLine in September 2021 [15].

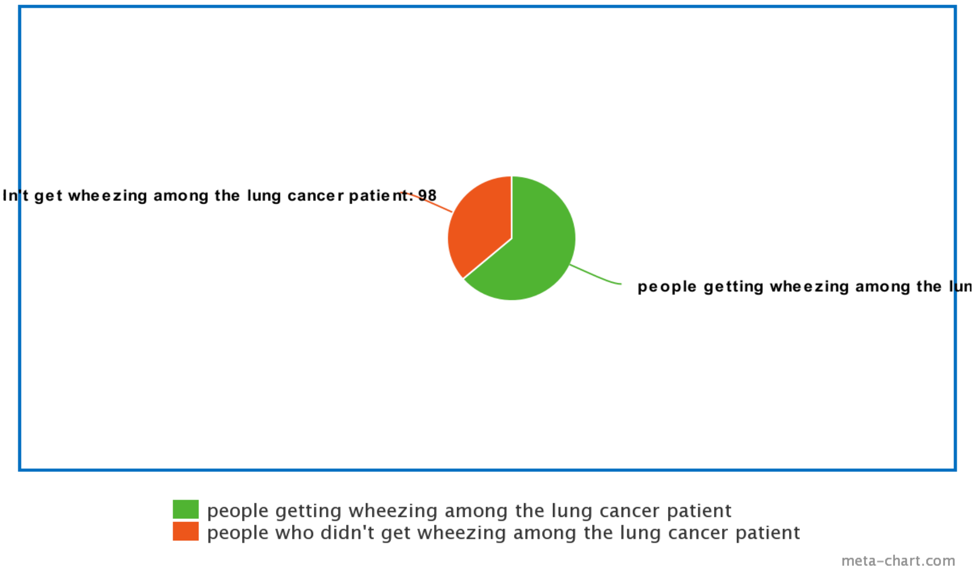


Figure 10. Wheezing as a lung cancer factor (Photo credit: Original).

3.1.11. Alcohol consuming. From Figure 11, people who consumed alcohol are more likely going to get lung cancer than those who did not, which is fair because “heavy drinkers are more likely to develop the disease” [16] and it’s very easy to become heavy drinkers as people like to drink at large dinners and parties with friends.

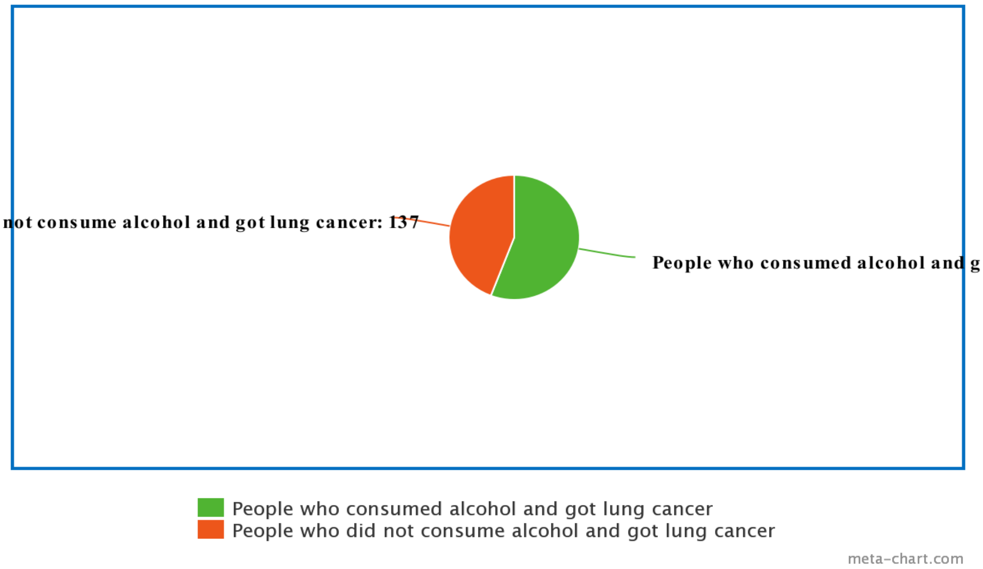


Figure 11. Alcohol consuming as a factor of lung cancer (Photo credit: Original).

3.1.12. Coughing. As shown in Figure 12, about 66.67% of patients in this dataset coughed when they had lung cancer, which showed us that coughing is an important factor that causes lung cancer and a majority of illnesses.

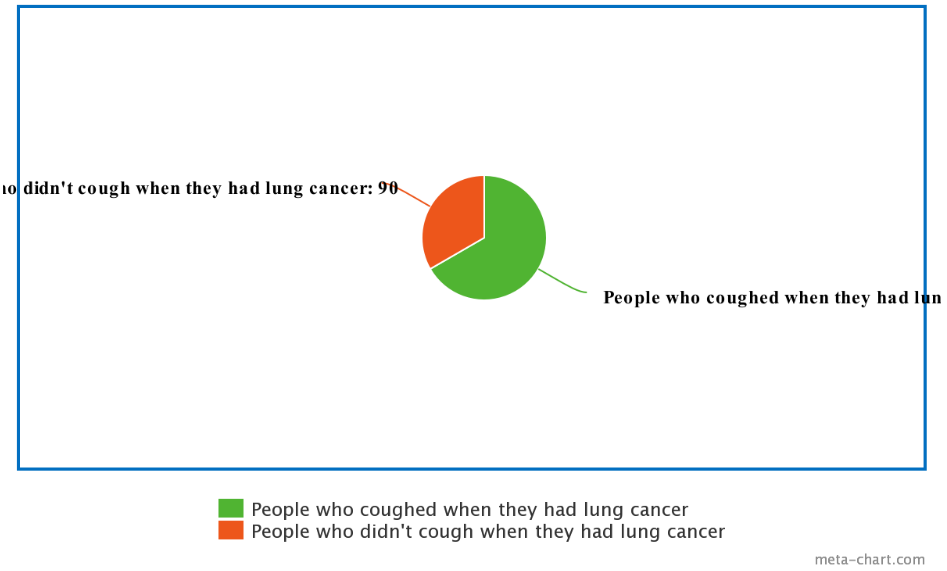


Figure 12. Coughing as a factor of lung cancer (Photo credit: Original).

3.1.13. Breathing problems. As shown in Figure 13, about 73.7% of patients in this dataset had breathing problems when they had lung cancer, which is very logical as lungs "take in oxygen when you inhale and release carbon dioxide when you exhale" according to Mayo Clinic and we usually breathe by inhaling oxygen and exhaling carbon dioxide [17].

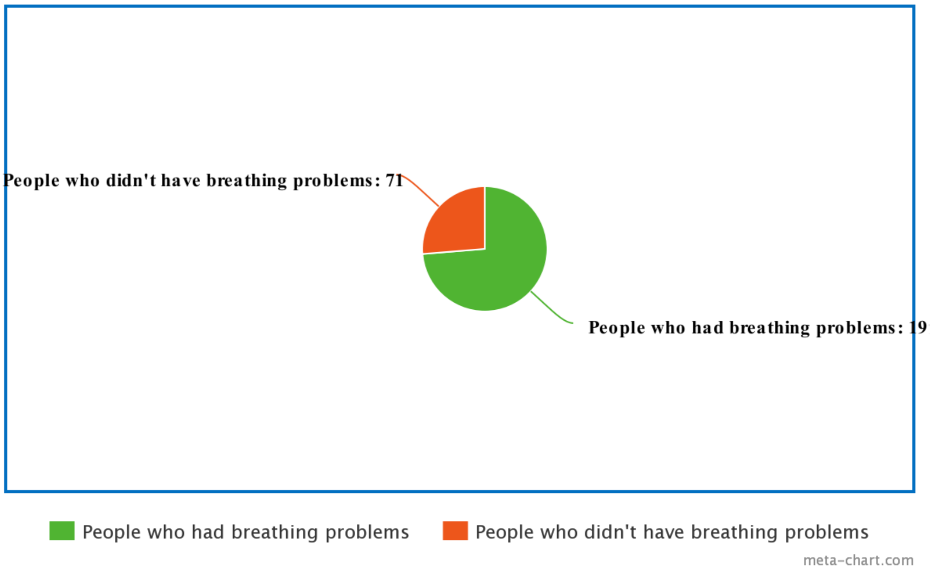


Figure 13. Distribution of lung cancer patients who have breathing problems (Photo credit: Original).

3.1.14. Swallowing difficulty. From Figure 14 shown above, about 53.7% of lung cancer patients had swallowing problems, which matches with the result we get from doing the chi-square test on the dataset(for detailed analysis, see 3.2).

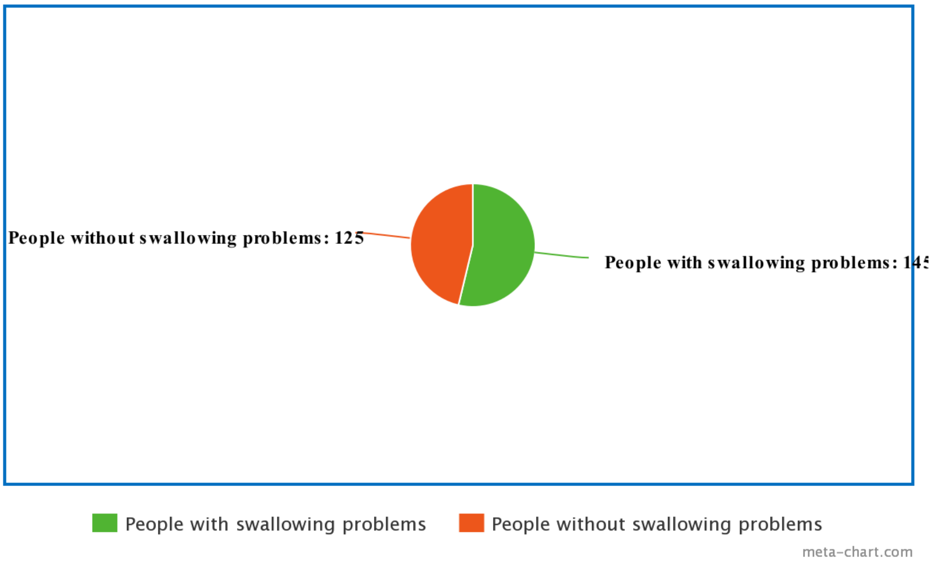


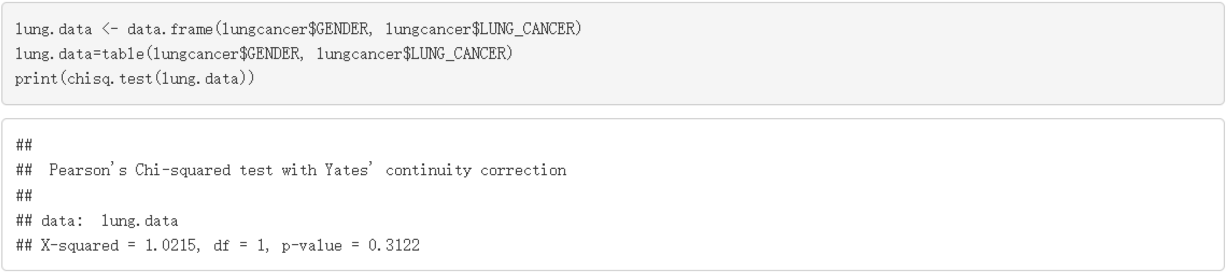
Figure 14. Distribution of lung cancer patients who have swallowing difficulties (Photo credit: Original).

3.2. Chi-square test of independence

Take the first categorical variable ‘Gender’ for example:

Null hypothesis: H0=GENDER is independent of LUNG_CANCER.

The code used can be seen in Figure 15.



```
lung.data <- data.frame(lungcancer$GENDER, lungcancer$LUNG_CANCER)
lung.data=table(lungcancer$GENDER, lungcancer$LUNG_CANCER)
print(chisq.test(lung.data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lung.data
## X-squared = 1.0215, df = 1, p-value = 0.3122
```

Figure 15. The χ^2 test (Photo credit: Original).

At the confidence level of 0.05, there is enough evidence to support the original hypothesis, so the gender of an object is independent of whether the object has developed lung cancer, the two variables are not associated with each other.

Other categorical variables that are independent of lung cancer development are listed in Table 2.

Table 2. Other categorical variables unrelated to lung cancer development.

H0=A is independent of LUNG_CANCER	relation	variable(A)	p-value(α=0.05)

valid	independent	smoking
		chronic disease
		shortness of breath
			0.3953(>0.05)
			0.07541(>0.05)
			0.3739(>0.05)

Categorical variables that are dependent of lung cancer development are listed in Table 3.

Table 3. Other categorical variables related to lung cancer development.

H0=A is independent of LUNG_CANCER	relation	variable(A)	p-value(α=0.05)
invalid	dependent	yellow finger	0.002573(<0.05)
		anxiety	0.01747(<0.05)
		peer pressure	0.001902(<0.05)
		fatigue	0.01366(<0.05)
		allergy	0.0000000281(<0.05)
		wheezing	0.00002555(<0.05)
		alcohol consuming	9.607e-07(<0.05)
		coughing	2.717e-05(<0.05)
		swallowing difficulty	1.113e-05(<0.05)
		chest pain	0.001496(<0.05)

Among the 15 variables, the p-values of 'gender' 'smoking', 'chronic disease' and 'shortness of breath' are bigger than 0.05, which means that they are not associated with lung cancer development; However, the p-values of variable 'yellow finger' and the other nine independent variables are smaller than 0.05, so there is no enough evidence to support the null hypotheses contraposing these 10 variables, which means that they are tested as factors contributing to lung cancer.

4. Conclusion

Based on the analyses of the dataset given by Kaggle regarding Lung Cancer, we realized that a majority of factors listed are directly related to the probability of whether or not people will get lung cancer based on their unique lifestyles. This dataset is very reliable, as the p-values of most factors are under 0.05 after we ran the chi-square test. More specifically, it provides us with information on the leading factors and obvious signs of lung cancer so that we can prevent from doing it and thus decrease the chance of getting lung cancer. For example, from the pie chart we created using the given dataset, we can see that more than half of the people in the survey consumed alcohol and were diagnosed with lung cancer, so we should try to consume less alcohol each week in order to decrease the chance of getting it. Another information we can obtain from the analyses is that most people who were diagnosed with lung cancer have breathing problems, as fluid in the lungs can gather around them, making it more difficult for people to breath if their lungs are affected. From the information above, we can conclude that people who have major breathing problems have a large chance of getting lung cancer, as it is a major signal that you have been diagnosed with it.

Although the dataset is reliable at this time, it has some limitations. Firsts, this dataset did not include the family history of people getting lung cancer, as there's a higher chance of getting lung cancer when one of the family members carries a mutated gene. Second, the factors leading to lung cancer listed in the dataset are not enough, as other factors such as exposure to carcinogens and radon gas can also lead to lung cancer. Overall, this dataset provides us with more information regarding lung cancer, which helps us gain more knowledge on cancer as it is one of the main causes of death at a young age.

References

[1] American thoracic society. Lung Cancer - (n.d.). Retrieved January 30, 2023, from https://www.thoracic.org/patients/patient-resources/resources/lung-cancer-intro.pdf

[2] Outdoor Air Pollution and cancer: An overview of the current evidence. Retrieved January 30, 2023, from https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21632

[3] Smoking as a risk factor for lung cancer in women and men. Retrieved January 30, 2023, from https://bmjopen.bmj.com/content/bmjopen/8/10/e021611.full.pdf

[4] American Cancer Society. Study: Young women now have higher rates for lung cancer than men

worldwide. (2020, February 14). Retrieved January 31, 2023. from https://www.cancer.org/latest-news/study-young-women-now-have-higher-rate-for-lung-cancer-than-men-worldwide.html

[5] What is lung cancer? Types of Lung Cancer. Retrieved January 31, 2023, from https://www.cancer.org/cancer/lung-cancer/about/what-is.html

[6] Juan Sentana. Tests for independence between categorical variables. Economics Letters,Volume 220,2022,110850,ISSN 0165-1765.

[7] Why Men get cancer more than Women and How they can manage their risk - City of Hope(2022 February 9). Retrieved February 21, 2023, from https://www.cancercenter.com/community/blog/2022/02/men-and-cancer-risk

[8] Lawrence A. Loeb, Virginia L. Emster, Kenneth E. Warner, John Abbotts, John Laszlo. Smoking and Lung Cancer: An Overview. Cancer Research 44,5940-5958, December 1984.

[9] Zhenghong Yu. Factors contributing to decrease in human immunity. Open books for good (Seeking medical advice),2011(12):40.

[10] MedicineNet. Can a Weak Immune System Cause Cancer? 11/29/2021. From: https://www.medicinenet.com/can_a_weak_immune_system_cause_cancer/article.html

[11] Poller WC, Downey J, Mooslechner AA; et al. Brain motor and fear circuits regulate leukocytes during acute stress [published online ahead of print, 2022 May 30]. Nature. 2022;10.1038/s41586-022-04890-z. doi:10.1038/s41586-022-04890-z

[12] Park HY, Kang D, Shin SH, Yoo KH, Rhee CK, Suh GY, Kim H, Shim YM, Guallar E, Cho J, Kwon OJ. Chronic obstructive pulmonary disease and lung cancer incidence in never smokers: a cohort study. Thorax. 2020 Jun;75(6):506-509.

[13] Engels EA, Shen M, Chapman RS; et al. Tuberculosis and subsequent risk of lung cancer in Xuanwei, China. Int J Cancer 2009;124: 1183-11837.

[14] Laney AS, Weissman DN. The classic pneumoconioses: new epidemiological and laboratory observations. Clin Chest Med 2012;33:745-58

[15] Lung Cancer Symptoms: Coughing, Wheezing, and More - healthline(2021 September 6). Retrieved February 21, 2023, from https://www.healthline.com/health/lung-cancer-symptoms

[16] How Alcohol Affects Lung Cancer Risk and Outcomes -- verywellhealth(2022 November 28). Retrieved February 21, 2023 from https://www.verywellhealth.com/alcohol-and-lung-cancer-risk-2248986

[17] Lung Cancer -- Symptoms and causes -- Mayo Clinic(n.d). Retrieved February 21, 2023 from https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620.