Risk and Demographic Factors Examination Models for Lung Cancer Incidences in the United States

Jiayu Chen

Abstract. Lung cancer has been the leading cause of cancer deaths worldwide since the spread of smoking and the decline in air quality. This paper analyzes the risk factors and demographic variables that contribute to lung cancer incidences in the United States. Risk factors include smoking and environmental pollution, while demographic variables include education, income level, age group, and ethnicity. Multiple linear regression model and random forest model are used for variable interpretation and lung cancer prediction. In the model evaluation, the multiple linear regression model is found with better prediction accuracy than the random forest model by using cross-validation with resampling. The random forest model has 6.82, 0.70, and 5.21 of RMSE, R squared, and MAE, while the multiple linear regression model has 5.70, 0.74, and 4.70. Smoking is the most contributing risk factor based on the result, and the most significant demographic variables are Hispanic and high school education proportions. The discussion section includes the interpretation of the groups of variables, potential limitations in the models, and the future direction of finding deeper correlations between lung cancer incidences and the selected risk and demographic variables. This analysis illustrates the importance of demographic variables in cancer and shows more directions for research on demographic variables.

Keywords: lung cancer, random forest model, linear regression, smoking, socioeconomic status (SES).

INTRODUCTION

Lung cancer has been one of leading cancer in the United States of America, occupying about 25% of total cancer deaths. It is estimated to have 236740 new incidences and 130180 deaths of lung cancer in the United States in 2022. The main age group of lung cancer is around 70 years old, and the ratio of male to female incidence is approximately 2:1 [1]. Therefore, lung cancer treatment has always been a significant medical problem that needs to be overcome. Lung cancer morbidity and mortality remain high in the United States due to smoking, air pollution, radiation exposure, and fungal infections in the lungs. Studying and reducing these risk factors are necessary to reduce lung cancer death and incidences.

Smoking is always the major causing factor of Lung cancer. A few centuries ago, lung cancer was an extremely rare disease, and the large increase in modern lung cancer is closely related to the increase of carcinogens in human life. Among them smoking is the most important carcinogenic substance provider. When cigarettes burn, they release 38 toxic chemicals, including tar, carbon monoxide, and so on [2]. Tar, for instance, can damage the trachea, and lungs . It is easy for tar to deposit on the villi of the lungs, resulting in these villi losing their original functions. Long term accumulation will make lung organs unable to function normally, and there is a greater chance of lung cancer [3]. Smoking is also the fastest and most prolific way to produce free radicals, at least 100,000 of which are produced with each puff of a cigarette, which can cause cancer and many chronic diseases.

Until now, most of Lung cancer incidences and deaths are directly correlated with smoking histories. The linkage between Lung cancer and smoking was strengthened in the mid 20th century, and after decades of research, this relationship has only become clearer. The trend of cigarettes had a direct correlation with the epidemic of Lung cancer around the world historically. According to Fig. 1 from Our World Data, the historical trend of Lung cancer deaths per 100,000 men has an almost identical trend with the cigarette consumption trend with a lapse [4]. From a longitudinal view, the reduction of smoking cigarettes has a significant contribution to lung cancer deaths.

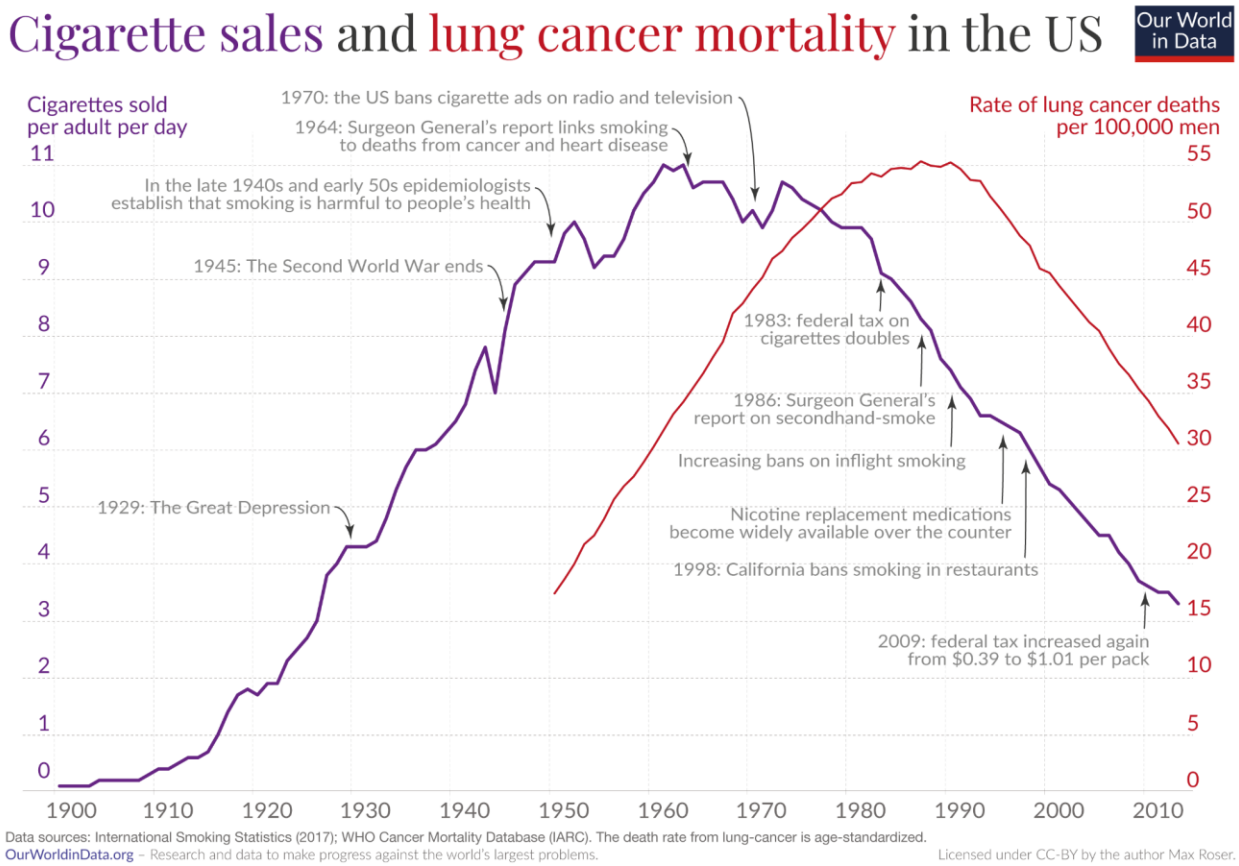

FIGURE 1. International Smoking Statistics (2017); WHO Cancer Mortality Database (IARC) [4].

Other than smoking, the other major way humans are exposed to carcinogens is through air pollution. As a result of industrial production, motor vehicle emissions and a variety of daily exhaust emissions, ambient air contains many pollutants, and the resulting health losses are enormous worldwide. In 2004, there were 62,000 lung cancer deaths due to ambient air pollution [5]. In addition to carcinogens, there is another very important component of air pollution, which is called particle pollution. This is especially true of PM 2.5 (particles with a diameter of 2.5 μm or less), which are one of the most dangerous carcinogens since they are easy to appear, and they stay in the air for a long time, many of them are inhaled into the bronchi and alveoli. They can deposit in the lungs and damage lung tissue, making them more vulnerable to damage.

However, while comparing the two contributing factors of Lung cancer to the Lung cancer incidence rate, smoking and air pollution in population level are not as significant as expected in individual level [6]. While carcinogens are the cancer-causing molecules, other demographical or social factors can be the secondary contributing or reducing factors that affect Lung cancer prediction.

Socioeconomic status is one of the leading demographical factors contributing to Lung cancer. Socioeconomic status is an aggregate measure of an individual's economic and social status relative to that of others through income, education, and occupational factors. Studies have found a strong link between socioeconomic status and smoking: lower levels of education, job type evaluation, and economic level were associated with more frequent smoking, leading to increased lung cancer rates in similar SES categories [7].

Race is also an important determinant of lung cancer. This does not mean that there are genetic differences among different races, but it contains complex social factors, including medical services, economic differences, living environment and so on. People of color get less care in America [8]. Compared with whites, blacks have fewer opportunities for early diagnosis and surgery, which makes their lung cancer more difficult to treat [9]. Like socioeconomic status, ethnic factors have a profound impact on the relationship between biological carcinogens and lung cancer.

The purpose of this paper is to consider both demographical and the common risk factors in order to create a comprehensive view on potential lung cancer risk factors. Therefore, various risk factors and demographical variables are selected to analyze their correlations with lung cancer. The result of the study shows the most contributing factors when prediction lung cancer incidences and interpretations to different demographical variables in according to lung cancer incidences.

METHODOLOGY

The dataset used for this paper is based on online, public statistics. Most of the data used are selected from National Cancer Institute (NCI), State Cancer Profile, which provides data of Lung cancer incidences from 2014 to 2018, demographic variables, and risk factors variables at the state level on average. After filtering unnecessary demographic and risk factors like diet and vaccines, a total of 12 variables are selected. These include Lung cancer age-adjusted incidences per 100,000, current 18+ smokers percentage, Percent of State Population with Smokefree laws, people with high school and college degree, poverty, elders age groups, and each ethnicity populations percentage in state level. Other variables including air quality index, radon level and average personal income are based on other credential public sources. More details are provided in Table 1 below:

TABLE 1. Variable Descriptions

State	States in the US	Source
Lung_Cancer_cases	Lung & Bronchus, Age-Adjusted Incidence Rates by Cancer Site, All Stages (2014-2018)	NCI State Cancer Profile [10]
current_smoker_percent	Current Smoker, Ages 18+, 2020	NCI State Cancer Profile
smoke_law	Percent of State Population with 100% Smokefree Laws, 2021	NCI State Cancer Profile
Personal_income	Personal per capita income, 2020 average	Bureau of Economic Analysis [11]
Air quality	Air quality index, 2010-2014 average	U.S. Air Quality Index State Rank [12]
radon_level	Radon Levels by State 2022	Radon.com [13]
high_school	Citizens with at least high school degrees by state, 2015-2019, Ages 25+	NCI State Cancer Profile
college	Citizens with at least Bachelor's degree by state, 2015-2019, Ages 25+	NCI State Cancer Profile
poverty	Persons below poverty by the federal government's official poverty definition, 2015-2019	NCI State Cancer Profile
elders_65	Age 50 and over proportion by state, 2015-2019	NCI State Cancer Profile
white	White proportion by state, 2015-2019 average	NCI State Cancer Profile
black	Black proportion by state, 2015-2019 average	NCI State Cancer Profile
asian_pacific	Asian and Pacific Islander proportion by state, 2015-2019 average	NCI State Cancer Profile
hispanic	Hispanic proportion by state, 2015-2019 average	NCI State Cancer Profile

In data cleaning, every observation is checked for its completeness. This means that NA values and any unusual values (0s, negative numbers) in the data is checked to prevent errors. These missing or incorrect values are checked with the data description of sources and by adding up the results in a lower geographic level (by county level) and add up to state level to analyze why those missing value exist and calculate the missing values if possible. During this process, two areas, Puerto Rico, and District of Columbia, are found with missing values for multiple variables, including radon level, green score, public smoking law, and average personal income. These variables also do not have county-level data, so these missing values cannot be calculated. In addition, after comparing Puerto Rico and District of Columbia to other 50 states, considering their incomplete data and extreme data, these two areas may not be representative compared to other states and may affect the accuracy and completeness of the whole dataset. For this reason, these two areas are not considered in this study. The result dataset contains 15 variables and 50 observations, corresponding to the 50 states in the US.

Two models are used in this study: The multiple Linear Regression Model and Random Forest. They are built to find the most significant factors contribute to state-level lung cancer incidences among all variables. After finding the significant variables, Five-Fold Cross-Validation is used for prediction. In this step, the whole dataset is randomly

divided into five subsets, where four become the training group and the remaining one becomes the testing group. Accuracies of the predictions are also calculated through this process.

RESULTS

Exploratory Data Analysis

All the variables are examined first with exploratory data analysis before building models to ensure their correlation with Lung cancer Incidences. In this process, public smoking law variable is filtered because of its unequal distribution of values: more than half of the states have 100% smokefree law. This is not suitable for building model since it can skew the result prediction with the unbalanced data.

The top 5 states with the most lung cancer cases are Kentucky, West Virginia, Arkansas, Mississippi, and Tennessee. The top 5 states with the least lung cancer cases are Utah, New Mexico, California, and Wyoming.

Figure 2 shows the correlation between the lung cancer cases and the contributing variables, and most of the variables display correlations with Lung cancer Incidence.

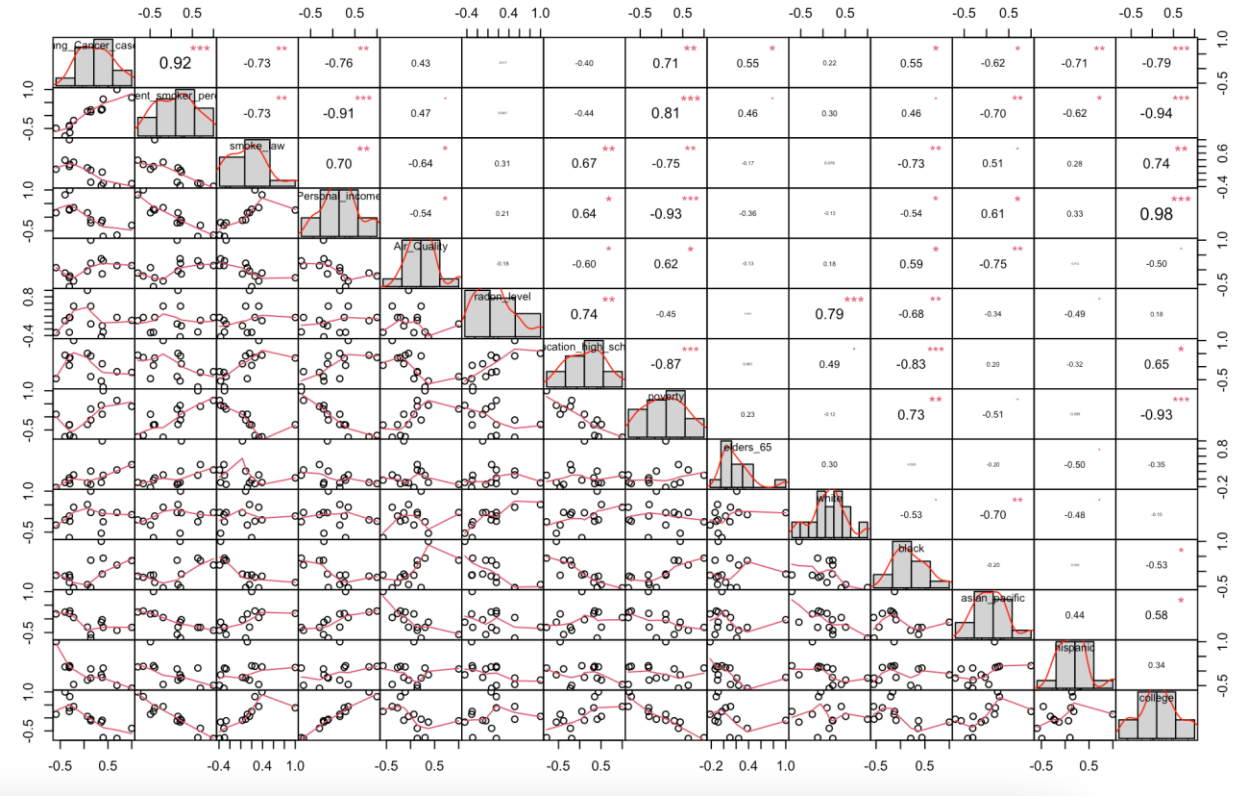

FIGURE 2. Lung Cancer Matrix plot; Correlation codes: 1.0 '***' 0.8 '**' 0.65 '*' 0.50 '. ' 0.45. (Photo credit: Original).

Linear Regression Results

Table 2 shows the multiple linear regression results done by R. The linear regression model gives an R squared value of 0.8586 and the adjusted R squared value is 0.8127. Current_smoker_percent, elders_65, radon_level, college, white, and black are in positive correlations; Personal_income, poverty, high_school, and Air_Quality, asian_pacific, and hispanic are in negative correlations with Lung cancer incidences. Among all the variables, only Personal income, current smoker percent, and avg_co2_percapita are statistically significant. Among them, the current smoker percent, hispanic and high_school has the most effect to the regression model. This follows the common risk factor of Lung cancer incidences which is smoking, the common age group of Lung cancer incidences, and the ethnicity with least

lung cancer incidences. The most statistically significant variables give that every one percent rise of current smoker percent increases the lung cancer cases by 2.048, every one percent rise of Hispanic population decreases Lung cancer cases by 7.623, and every one percent rise of people with more than high school degree decreases the lung cancer cases by 5.179.

TABLE 2 Linear Regression Model Results

Residuals					
Min	**1Q**	**Median**	**3Q**	**Max**	
-9.8057	-2.2762	0.3781	2.5762	10.7415	
Coefficients	**Estimate**	**Std. Error**	**t Value**	**Pr(>|t|)**	
(Intercept)	4.75E+02	9.86E+01	4.821	2.45E-05	***
current_smoker_percent	1.97E+00	5.45E-01	3.609	0.000906	***
Personal_income	-4.00E-05	1.81E-04	-0.222	0.825897	
poverty	-2.12E+00	8.23E-01	-2.573	0.014231	*
high_school	-5.05E+00	8.49E-01	-5.953	7.28E-07	***
elders_65	1.61E+00	4.93E-01	3.27	0.002329	**
radon_level	6.62E-01	4.82E-01	1.373	0.177932	
college	7.30E-01	3.72E-01	1.965	0.056981	
Air_Quality	-4.38E-01	2.06E-01	-2.127	0.040119	*
white	1.52E-01	2.61E-01	0.583	0.56358	
black	1.28E-02	2.81E-01	0.046	0.963784	
asian_pacific	-1.23E-01	4.15E-01	-0.296	0.769261	
hispanic	-7.57E-01	1.57E-01	-4.818	2.47E-05	***
Residual standard error: 4.932 on 37 degrees of freedom					
Multiple R-squared: 0.8586, Adjusted R-squared: 0.8127					
F-statistic: 18.71 on 12 and 37 DF, p-value: 3.563e-12					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' .

The model shows that poverty has a negative coefficient, which means more poverty leads to less lung cancer cases. While previous papers states that lower economic status of people is linked with higher chances of smoking and getting lung cancer, the negative coefficient is inconsistent with the linkage.

Table 3 shows the evaluation of the Multiple Linear Regression model. Five-fold cross validation is used for the evaluation. This method divides the total dataset into 5 subsets, while one of them being the testing group and the other four of them being the training group. The final result is showing the accuracy of predicting the testing group based on the training group. This method is resampled for 100 times and the result is an average of the 100 resampling cross validation result. The RSME is 5.70, R squared is 0.74, and MAE is 4.70.

TABLE 3 Linear Regression Accuracy

Linear Regression (No pre-processing)			
Samples	**Predictors**	**Resampling**	**Summary of Sample Sizes**
50	12	Cross-Validated (10-fold, repeated 20 times)	46, 45, 45, 44, 44, 46, ...
Resampling Results			
RMSE	**R-Squared**	**MAE**	
5.582672	0.7545769	4.571417	

Random Forest Results

The random forest model gives that the smoker percent, hispanic, black, college, and Personal income are the top 5 significant factors. The second common risk factors, radon level, is in the lower significance, which is unusual. Figure 3 shows the random forest result done by R.

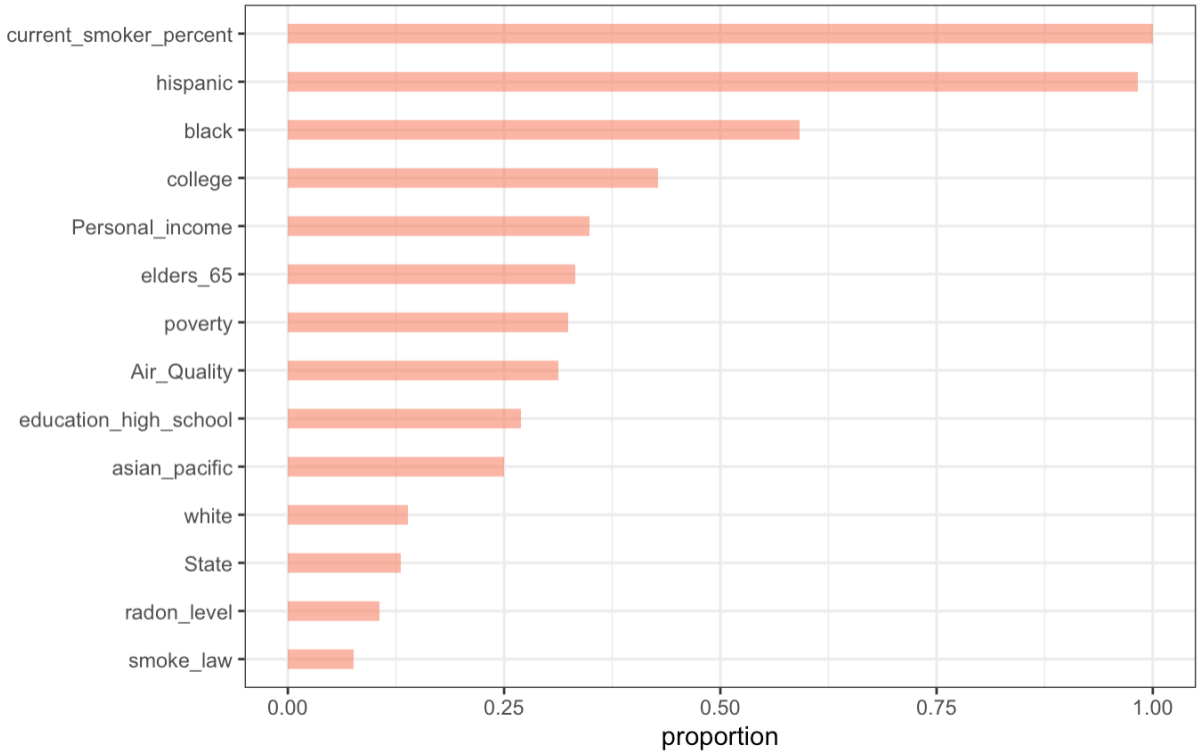

FIGURE 3. Random Forest Model Results Rank (Photo credit: Original).

Table 4 shows the RMSE, R squared, and MAE values of the random forest model and mtry 12 is used for the final value. The final RMSE, R squared, and MAE are 6.82, 0.70, and 5.21. The ultimate mode, mtry 12, is selected by the smallest RMSE score.

TABLE 4. Random Forest Model Accuracy

Random Forest (No pre-processing)			
Samples	**Predictors**	**Resampling**	**Summary of Sample Sizes**
50	12	Cross-Validated (10-fold, repeated 20 times)	46, 45, 45, 44, 44, 46, ...
Resampling Results across Tuning Parameters			
MTRY	**RMSE**	**R-Squared**	**MAE**
2	7.704423	0.6464862	5.781695
7	6.996594	0.6862587	5.289622
12	6.822520	0.7036213	5.214590

Findings Based on the Results

The most contributing factor is smoking according to the two models. Hispanic ethnicity is displaying a strong significance and followed by education and economic category where Linear Regression model suggest high school degree and poverty, and Random Forest model suggest college degree and personal income level. Overall, the multiple

linear regression has a better prediction performance than the random forest model due to its lower RMSE, higher R squared, and lower MAE values when compared to the random forest model.

DISCUSSION

Aligning to current studies on significance of smoking to lung cancer, smoking is also shown to be the most contributing factor to lung cancer incidences in state level: Current smoking adults percentage is significant among both approaches of building models. While smoking to individual health is severe, its damage to public health is hard to quantify since there are too many uncontrollable factors, including different environmental conditions and social factors. For ongoing smokers, the effect includes both individuals and the surrounding because of secondhand smoke and the trigger of smoking temptation. It is a plausible explanation to why current smokers are significant since either increasing intake of carcinogens from cigarettes to people nearby or triggering more smoking behaviors can contribute to lung cancer risks in community health.

Air pollution has also been one of the most rising risks of lung cancer. This study analyzes the air pollution from air quality index and radon level. Air quality index is the most straightforward way to measurement of air quality in an area. It only shows significance in the multiple linear regression model, and its relatively low significance in the models may be caused by the low fluctuation and that air quality of most states stay in a relatively healthy level (lower or around 50). Comparisons with higher air quality index countries or cities can be a further approach for this risk factor.

Other than common risk factor, demographical values are also selected to evaluate its significance and adjust the models. Among these variables elders aged above 65 years old displayed significance in both model approaches. This result is consistent to the current research, as most of them suggests the age group with the most number of diagnosis with lung cancer is greater than 65 years old [14]. The reason is that the older age related weaker immune system and more carcinogens accumulation. Other than biological reasons, Lung cancer diagnosis are usually tested in later ages. Suggested by recent studies that only 16% of lung cancer diagnoses are in the early stage and most diagnosis are in third stage, which lies in the range of 65 years old or older [14].

Socioeconomic status is considered a new leading risk of lung cancer. Education level and income level are included in this study to evaluate their importance to lung cancer. By the random forest models, personal income level is the fifth leading causing factors to lung cancer. This correlation is mainly due to smoking, as low education and low-income level are associated with more smoking frequencies [7]. It can also be shown by the negative coefficient of both variables in the multiple linear regression models, which means that age-adjusted lung cancer cases per 100,000 can decrease by 2 cases when either of these two variables rise up one percent.

Ethnicity is an important group to consider when predicting lung cancer, since it is one of the most revealing variables of social difference. Among the 4 groups of races, only Hispanic group displays a great significance. This result shows that there are fewer cases of lung cancer in the states with more Hispanics, which proves statistically that Hispanics have a low incidence of lung cancer. However, these two models are not enough to further explain the underlying reasons for the low incidence of Hispanics. More detailed demographic variables and more data are needed to find the factors that reduce the incidence of lung cancer by comparing Hispanics with other races.

Limitations and Improvements

Indicating as second most significant cause of lung cancer, radon does not display significance in the models as expected. This insignificance can be due to large geographical classifications in collecting data (by state level), which is one of the limitations of the study. State level is too broad for environmental measurement, as some cities with unhealthy AQI number can be neutralized by collecting with other healthy AQI cites in their state. City or county level of air pollution can be a reasonable approach for future analysis. In addition, more models can be used in the future.

Another limitation is that measurement of second handed smoke has not been taken into consideration in this dataset. The public smoking law variable cannot represent the second handed smoke comprehensively, and more specific measurement is needed on particular public places. The importance of secondhand smoke on lung cancer can be better concluded after the comparison and analysis of secondhand smoke content in different types of public areas. It is also a feasible measure to collect people's impression of smoking in public places by means of questionnaire.

While socioeconomic status has a correlation with lung cancer shown by previous study, the study does not include the job evaluations in state level. The reason is lack of data while searching for public datasets on this variable: most of the values are not representative in state level and are case specific.

Future Research

As smoking is determined to be the most significant factor to state level lung cancer cases, future approaches can focus on the other variables, like environmental pollution and other demographic variable. More specific carcinogens level in the air may be used in the models. While the air quality of states is mostly on healthy level, finding areas with the most carcinogens level in air and examine the health condition of surrounding neighborhoods is one possible method.

Besides variable selection, future research can also broaden the focus from the United States to other countries in the world, as some countries have larger differences in some variables which may help conclude correlations.

CONCLUSION

Both linear regression model and random forest model have more than 70% variance. Although the accuracy of linear model is higher than that of random forest, it is similar: RSME of linear model is 5.70, MAE is 4.70; The RMSE of random forest is 6.82, and MAE is 5.21.

In terms of demographic variables, both models point out that Hispanics are statistically strongly related to lung cancer and are as important as smoking. The linear model indicates that the correlation is negative. Unlike other demographic variables, there is not much research on the association between Hispanics and lung cancer, and this strong negative correlation indicates that Hispanics may have other characteristics, which makes them stand out among different races in the United States. Another strongly related demographic variable is education. The linear model of high school diploma and the university diploma have relatively high influence in the random forest model respectively. Its importance may not be direct carcinogenesis, but higher education is related to less exposure to carcinogens and more cancer prevention.

Smoking and air pollution have been generally recognized as the key pathogenic factors of lung cancer, and the reasons are carcinogens in cigarettes and air. Smoking and air quality are also proved to be two of the important pathogenic factors of lung cancer in the model of this analysis. Among all elements, air quality is not one of the highest statistical significances and ranks low in the random forest model. Moreover, most demographic variables are related to smoking while air quality is not. The reason may be the lack of data on air particulate levels, like PM2.5.

This paper finds the significance of Hispanic race and education variable in considering lung cancer in populational level, and demonstrates that other than carcinogens, lung cancer can also be reduced by considering demographic risk factors. Based on the results of this analysis, the future research direction can be placed on the research of Hispanics and education. And can refine the regional area, for example, in cities level. In the choice of variables, more examinations can be focused about the relationship between these two demographic variables and carcinogens.

REFERENCES

1. Xiong, Y., Cheng, K., & Gu, Y. (n.d.). Lung cancer pathological distribution study. Medical Science and Technology. 26 (25), 69-71 (2022).
2. Lemjabbar-Alaoui, H., Hassan, O. U. I., Yang, Y.-W., & Buchanan, P. Lung cancer: Biology and treatment options. Biochimica Et Biophysica Acta (BBA) - Reviews on Cancer, 1856(2), 189–210 (2015).
3. Harris, J. E.. Cigarette tar yields in relation to mortality from lung cancer in the cancer prevention study II prospective cohort, 1982-8. BMJ, 328(7431), 72 (2004).
4. Roser, M. Smoking: How large of a global problem is it? and how can we make progress against it? Our World in Data. Retrieved October 2, 2022, from https://ourworldindata.org/smoking-big-problem-in-brief
5. Cohen, A. J. Air pollution and lung cancer: What more do we need to know? Thorax, 58(12), 1010–1012 (2003).
6. Kamis, A., Cao, R., He, Y., Tian, Y., & Wu, C. Predicting lung cancer in the United States: A multiple model examination of public health factors. International Journal of Environmental Research and Public Health, 18(11), 6127 (2021).
7. Hovanec, J., Siemiatycki, J., Conway, D. I., Olsson, A., Stücker, I., Guida, F., Jöckel, K.-H., Pohlabeln, H., Ahrens, W., Brüske, I., Wichmann, H.-E., Gustavsson, P., Consonni, D., Merletti, F., Richiardi, L., Simonato, L., Fortes, C., Parent, M.-E., McLaughlin, J., … Behrens, T. Lung cancer and socioeconomic status in a pooled analysis of case-control studies. PLOS ONE, 13(2), e0192999 (2018).
8. Centers for Disease Control and Prevention. (n.d.). Morbidity and mortality weekly report (MMWR). Centers for Disease Control and Prevention. Retrieved October 2, 2022, from

https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5944a2.htm#:~:text=Racial%20and%20ethnic%20disparities%20in,Hispanics%20have%20had%20lower%20rates.
9. Tong, L., Yang, D., & Bai, C. (n.d.). The Enlightenment of American Lung Cancer Prevention and Control to China. Medical and Health Technology.
10. State Cancer Profiles. (n.d.). Retrieved October 3, 2022, from https://statecancerprofiles.cancer.gov/
11. Bureau of Economic Analysis. U.S. Bureau of Economic Analysis (BEA). (n.d.). Retrieved October 2, 2022, from https://www.bea.gov/
12. U.S. Air Quality Index State Rank. USA.com. (n.d.). Retrieved October 2, 2022, from http://www.usa.com/rank/us--air-quality-index--staterank.htm?hl=&hlst=&wist=&yr=&dis=&sb=DESC&plow=&phigh=&ps=
13. Detailed radon information. Radon.com. (n.d.). Retrieved October 2, 2022, from https://www.radon.com/info_by_state/
14. Chen, S., & Wu, S. Identifying lung cancer risk factors in the elderly using deep neural networks: Quantitative Analysis of Web-based survey data. Journal of Medical Internet Research, 22(3), e17695 (2020).