Journal of General Medicine

Research Article

Factors Mediating Lung Cancer in Never-Smokers Identified from Epidemiologic Analysis

Kevin Guo[1*], William C Cho[2] and Farouk Dako[3]

[1]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

[2]Department of Clinical Oncology, Queen Elizabeth Hospital, Kowloon, Hong Kong, China

[3]Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Lung Cancer in Never-Smokers (LCINS) is one of the leading causes of cancer patient deaths in the United States. Unlike lung cancer onset by cigarette smoking, LCINS is not as readily understood and research on the subject has been conflicting. Thus, early diagnosis and prevention are key in reducing mortality among LCINS patients. In this study, the Prostate, Lung, Colorectal, and Ovarian (PLCO) dataset contains more than 155,000 participants and over 36,000 never-smokers. Data from more than 5,000 patients with LCINS were analyzed using R and Excel software to determine risk factors of LCINS. The factors analyzed for predictive power in LCINS incidence were age, height, weight, body mass index, race, income, family history, and secondary smoke exposure. Multiple statistical methods, including t-tests, ANOVA tests, and logistic regression, were implemented to assess each factor. Through comparison and corroboration of results from these statistical methods, age and race were identified as the key factors that had statistically significant evidence as potential influences in LCINS incidence. This was indicated by strong positive correlation between age/race and the predicted probability of LCINS development with a statistically significant p value. Other physical characteristics did not appear to have a significant impact on the likelihood of developing LCINS. In addition, the statistical method that provided the most information regarding a factor's power was logistical regression due to the binomial outcome of whether or not a patient has LCINS. It enables a deeper elucidation of how individual factors influence the probability of LCINS development. These results provide valuable insights into demographic risk factors for LCINS and lay the groundwork for future investigations, potentially including the application of advanced analytical techniques such as deep learning algorithms to explore additional predictive factors for LCINS and other lung cancer subtypes.

Keywords: Lung Cancer in Never-Smokers (LCINS); PLCO (Prostate, Lung, Colorectal, and Ovarian) data set; Statistical analysis; Risk factors R

INTRODUCTION

Lung Cancer in Never-Smokers (LCINS) is one of the leading causes of cancer mortality, leading to approximately 20,000 deaths in the U.S. and contributing to 10-20% of lung cancer deaths with increasing incidence [1]. Research literature indicates that LCINS is distinct from smoking-related lung cancer with differences in molecular triggers and treatment responses [2]. Furthermore, several epidemiologic studies suggest that a unique genetic subtype of lung adenocarcinoma from East Asian never-smokers is distinct from other geographical subtypes of cancer driven primarily by targetable oncogenic drivers [3]. As a result of these differences, lung cancer incidence and mortality have been cited to be slightly lower in never-smokers when compared to smokers.

Some risk factors are commonly considered to be associated with lung cancer are age, ethnicity, genetics, and gender [4,5] Age is often an implicit modifier, meaning that never-smokers could be exposed to lung carcinogens (i.e., secondary smoke from cigarettes, radon gas) and accumulate DNA and cellular damages over time. In addition, some researchers hypothesize that the East Asian subtype of lung cancer is partially explained by genetic differences [6]. Further investigation into the risk factors of LCINS often yields conflicting and inconclusive results. For example, Schwartz et al. hypothesized from a case control study in Michigan that African American never-smokers do not have higher incidence of lung cancer than white people [7]. However, Thun et al. concluded from their review of the literature that "African American women never smokers had significantly higher incidence rates from lung cancer than women of European descent who had never smoked" [8,9]. A more extensive query into lung cancer screening datasets may improve early diagnosis and prevent LCINS in susceptible populations.

In this study, the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer dataset [10] was migrated from Excel to R and analyzed via t-test, ANOVA test, and logistical regression. The dataset is derived from a large population-based randomized trial of approximately 155,000 participants in the United States enrolled between 1993 and 2001 [11]. It contains more than 5,000 never-smokers with LCINS, serving as an epidemiologic and imaging resource to identify risk factors and imaging features of LCINS. The parameters under investigation were age, height, weight, race, family history of lung cancer, and exposure to smoking during one's lifetime. The combination of logistic regression, t-tests, and ANOVA (Analysis Of Variance) tests indicates that age and race are significant contributing factors to LCINS. These results could be used in future studies involving machine learning that could improve prevention and diagnosis of LCINS and other lung cancers [12]. The statistical methods used in this study can be applied to other datasets where the

Submitted: 13 August 2024 | **Accepted:** 25 August, 2024 | **Published:** 27 August, 2024

***Corresponding author:** Kevin Guo, Department of Biology, University of Pennsylvania, 3451 Walnut Street, Philadelphia, PA 19104, USA.

Citation: Guo K, Cho WC, Dako F (2024) Factors Mediating Lung Cancer in Never Smokers Identified from Epidemiologic Analysis . J Gen Med 5(1): 1022.

outcome is similarly binary.

MATERIALS AND METHODS

Data Compilation and Analysis

R is a versatile open-source platform that can help run more advanced statistical models and techniques such as partial least-squares regression and logistic regression [13]. It can create easy-to-customize plots that visualize statistical results effectively. Additionally, R can manage large amounts of data in the forms of arrays, matrices, and other types of data frames. Users can also download packages to perform multiple types of data analyses according to their needs and specifications.

The PLCO dataset contains anonymized data from the participants of the PLCO Cancer Screening Trial, which was a large, randomized controlled trial conducted between 1993 and 2009 to evaluate the efficacy of screening for prostate, lung, colorectal, and ovarian cancers

In this study, the data compiled from the PLCO dataset were filtered for nonsmokers by selecting all participants who had not smoked any cigarette packs at the time of the study. This resulted in a total of 5,124 LCINS participants being analyzed. Each nonsmoker was then designated either 0 for no LCINS or 1 for LCINS. The proportions of sample groups for each risk factor were calculated by taking the average of each data column containing the designated numbers. The data was imported into R and analyzed with logistic regression, t-test, and ANOVA testing using the readxl package.

STATISTICAL METHODS

t-test – This is an inferential statistical test to determine if the unknown population means of two sample groups are equal. t-tests are usually used when the population variances are not known (14). In this study, t-tests were applied for analyzing the parameters including age, height, weight, and BMI.

The t-value is compared to a critical t-value to assess whether the difference between the sample means is statistically significant. If the absolute t-value exceeds the t-critical value at a chosen significance level (α) which is typically 0.05, the null hypothesis that the means are equal is rejected. Otherwise, there is not enough evidence to suggest a significant difference between the groups.

The t-value is calculated by dividing the difference between the two sample means by the standard error of the difference. The formula for the t-value is:

$\bar{X}_1$ and $\bar{X}_2$ are the sample means.

s_1^2 and s_2^2 are the variances of the two groups.

n_1 and n_2 are the sample sizes of the two groups.

Table 1: Comparison of the advantages and disadvantages of statistical tests and models.

Statistical Test/Model	Pros	Cons
T-test	• Can be used to compare the means of two different groups and determine whether the difference of means is statistically significant • Allow for both populations variances to not be equal (assumption in statistics that is usually needed to compare two different groups) • Can be used when the sample sizes of the two groups are small or the population variances are not known • Easy to understand and interpret the results	• Not applicable for three or more independent variable groups. • Requires assumptions that both populations are normally distributed • Lower degree of freedom requires higher t-values to reach t-test significance
Logistical Regression	• Easy to implement and interpret • No assumptions are made about the distributions of the variables involved • Provides coeffcient size and direction of association (positive or negative) • Provides high accuracy results when dataset is binomially distributed • Less prone to overfitting than other models	• Cannot be used if number of observations is less than number of variables • Assumes linearity between dependent and independent variables • Dependent variables must be discrete (i.e.counting numbers as opposed to in-between numbers like height)
ANOVA Test	• Can be used to assess whether multiple sample means are equal • Limits the type I error (false positive rate) • Overall a more powerful statistical test than t-test	• Can only determine that one group mean is different, not which one • Assumes that each case is independent, all distributions are normal, and variance of data in groups are homogeneous

Z-test – If two different sample groups' proportion means need to be tested, a two-sample Z-test is preferably used to compute Z-value in determining whether the two proportion means are different. In this study, Z-tests were used for analyzing the parameters including race and family history.

"Ppool", a parameter called pooled proportion, is typically used for a two-proportion Z-test. The pooled proportion combines the data from two groups to create an overall proportion estimate, which is then used in the calculation of the Z-value.

The Z-value (or Z-score) is a statistical measurement that describes the position of a data point relative to the mean of a group of data, expressed in terms of standard deviations. It shows how many standard deviations a particular data point is away from the mean.

The Z-value needs to be compared to a critical Z-value to decide whether the difference between two proportions is statistically significant. The critical Z-value is a threshold in hypothesis testing that marks the boundary between the rejection and non-rejection regions for the null hypothesis. It is used to determine whether a test statistic (Z-value) is extreme enough to reject the null hypothesis.

For a two-tailed test, the critical Z-value is approximately 1.96 at the significant level (α) of 0.05. If the calculated Z-value is greater than 1.96, the null hypothesis is rejected because the Z-value falls in the rejection region.

Analysis of Variance (ANOVA) – ANOVA assesses whether the population means of more than two sample groups are all equal to each other. In this study, ANOVA was utilized to analyze the parameters including race, income, and secondary smoke exposure. In the cases of age, height, weight, BMI, and family history, ANOVA test would not be appropriate since there are only two groups: patients with family history of LCINS and patients with no history of LCINS.

To perform an ANOVA test, the F-value, which is the ratio of expected variation to unexpected variation, must be calculated and compared to a critical F-value. This is done by first calculating the grand mean, also referred to as G/N. G represents the sum of all the data points, and N is the total sample size.

The F-value is represented by comparing the mean squares (between) to the mean squares (within). This is achieved through the following formulas:

Let k = number of groups

x_i = individual data value

$\bar{x}$ = group average

n = total sample size across all groups

Sum of squares within (SSW) =

Degrees of freedom within (df_within) = n – k

Mean squares within (MSW) = SSW/df_within

Sum of squares between (SSB) =

Degrees of freedom between (df_between) = k - 1

Mean squares between (MSB) = SSB/df_between

F-value = MSB/MSW

This F-value is compared to a critical F-value dictated by the appropriate degrees of freedom and number of groups. If the F-value is greater than the critical F-value, then the differences between the groups are statistically significant.

Logistic Regression – Logistic regression is a statistical model that is primarily used in datasets with only two outcomes for the dependent variable (15). Logistic regression may not be applicable for secondary smoke exposure and family history since the number of patient groups is smaller than most other risk factors where logistic regression is appropriate.

The comparison of these statistical test and models is provided in table 1.

RESULTS

Impact of Age on Incidence of LCINS

Previous studies indicated that age significantly impacts the incidence of lung cancer [16,17]. This investigation further demonstrates that age is also an important determinant in occurrence of lung cancer in never-smokers. While lung cancer can affect both smokers and never-smokers, the risk factors and patterns differ. As individuals age, several factors come into play. First, the risk of developing lung cancer increases with age, with a notable rise after 50. In women, hormonal changes associated with menopause can influence the risk [18]. Environmental and occupational exposures, like prolonged exposure to pollutants, asbestos, and secondhand smoke, accumulate over time, raising the risk. Genetic factors may become more pronounced with age, and a weakening immune system can impact the body's ability to combat cancer cells [19]. Cumulative exposures and age-related changes in the respiratory system and lung tissue make the lungs more susceptible to damage and carcinogenesis. While the risk of lung cancer in never-smokers is generally lower than in smokers, it is not negligible, and age is a significant factor in its development [7]. Early detection and prevention measures are crucial to reduce the risk as never-smokers age [20].

The analysis of PLCO data revealed a strong positive association between age and lung cancer incidence, as indicated by a high linear correlation with an R-squared value of 0.92 (Figure 1A). Additionally, the logistic regression model demonstrated a clear upward trend between age and the predicted probability of developing lung cancer, supported by a statistically significant p-value ($p < 0.05$) (Figure 1B). The results of a t-test further confirmed the significance of age in relation to LCINS incidence, with a t-value of approximately 2.57, surpassing the critical t value of 1.65 (Table 2). Hence, these findings collectively establish age as a substantial and noteworthy factor contributing to the incidence of lung cancer.

Evaluation of the Effects of Height, Weight, and BMI on the Incidence of LCINS

There is evidence suggesting that height, weight, and BMI may impact the incidence of LCINS. The exact relationship between these factors and lung cancer risk is complex and not fully understood, requiring further investigation [21].

Several studies have found a positive association between height and the risk of LCINS. For example, a study published in 2017 found that taller height was associated with an increased risk of lung cancer among never-smoking women in the United States [22]. One possible explanation is that taller people have larger lungs, which may lead to increased exposure to environmental toxins and pollutants that can cause lung cancer. However, other studies have found no association between height and lung cancer risk in never-smokers. Therefore, more research is needed to confirm the relationship between height and the risk of LCINS.

Body weight and BMI may also play a role in lung cancer risk, although the evidence is not consistent. Obesity has been linked to an increased risk of several types of cancer, including LCINS [23]. One possible explanation is that excess body fat can cause inflammation and increase levels of insulin and other growth factors that promote the development of cancer [24]. However, the relationship between BMI and lung cancer risk is not entirely clear. Some studies have suggested that higher BMI may be associated with a decreased risk of LCINS [25]. The reasons for these conflicting findings are not fully understood and may be the result of differences in study design, population characteristics, and other factors. More investigation is needed to determine the impact of body weight and BMI on LCINS.

The results of our analysis indicate that weight, height, and BMI are not significant factors in the incidence of LCINS. All three t-tests, which assessed the impact of these variables on LCINS incidence, yielded t-values lower than the t-critical value of 1.65 (Table 3A, Table 3B, Table 3C). This suggests that these factors do not play a significant role in influencing the risk of LCINS. In fact, only t-values higher than the t-critical value would suggest a significant effect on LCINS incidence. Moreover, this finding is further reinforced by the logistic regression analysis, where high p-values exceeding 0.05 (though specific values are not provided in this summary) indicate that weight, height, and BMI are not statistically significant predictors of LCINS incidence. Therefore, it can be concluded that these physical characteristics do not appear to have a substantial impact on the likelihood of developing LCINS.

Race

There is some evidence indicating that race can play a role in the incidence of LCINS. White people overwhelmingly represent LCINS cases, followed by Asian, black, Hispanic, and Native American people. Interestingly, logistic regression reveals a lower incidence of Asians with LCINS compared to other races. Nevertheless, several studies have suggested that Asian populations may be at higher risk of developing lung cancer in never-smokers compared to other racial groups [26]. The higher prevalence of driver mutations in genes such as EGFR and KRAS [27] may account for the higher incidence of lung cancer in never-smokers among Asian/Pacific Islanders. In addition to genetic factors, environmental factors, such as exposure to secondhand smoke, radon, and other air pollutants, may also increase the risk of LCINS. These exposures may vary depending on race.

The results of our investigation, involving a combination of statistical tests and a comprehensive evaluation of race as a potential factor in LCINS, provide compelling evidence (Figure 2A). Specifically, the Z-tests conducted for various population groups suggest that the Caucasian population exhibits a statistically significant proportion of LCINS individuals with a Z-value greater than the critical Z-value of 1.96 (Table 4A), indicating that race may play a crucial role. Moreover, our logistic regression model underscores this by revealing that Caucasian individuals generally face a higher risk of lung cancer when compared to other racial groups included in this study (Figure 2B). This conclusion is reinforced by a high p-value obtained from the logistic regression, further emphasizing the significance of race as a contributing factor. Additionally, the ANOVA test results, with an F-value surpassing the critical F-value of 2.21, decisively reject the null hypothesis (Table 4B), providing strong evidence that race indeed plays a significant and noteworthy role in LCINS incidence.

Income

There is evidence to suggest that income may play a role in the incidence of lung cancer among never-smokers [28]. People with lower incomes, especially those within the $20,000 - $49,000 and $50,000 - $99,000 income brackets are significantly affected. However, this effect was not seen in the logistic regression analysis. Even though income may not directly impact LCINS, it could potentially influence other factors, primarily social determinants of health that affect patients' lifestyles. This includes living in areas with higher levels of air pollution and have less access to healthcare, including cancer screening and treatment [29]. Studies have also found that people with lower incomes are more likely to work in jobs with exposure to harmful substances, such as asbestos and radon, which can increase the risk of lung cancer [30].

The analysis of the data regarding income and its potential association with LCINS reveals a complex picture. When examining a bar graph, it seems to suggest that individuals with no LCINS tend to have higher incomes (Figure 3A). However, when subjected to more rigorous statistical tests, such as the logistic regression model and the ANOVA the results indicate that the difference in LCINS incidence among income groups is not statistically significant. The data from analysis using logistic regression model showed that the incidence of LCINS across income groups is comparable (Fig. 3B). Consistently, the ANOVA test revealed that the population means of income groups are not statistically different with

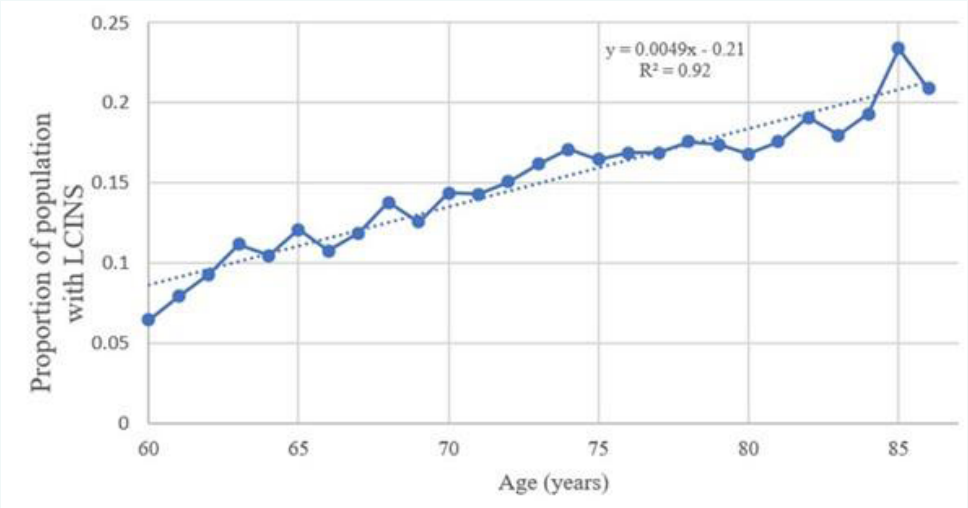


Figure 1A: Prediction of association between age and Lung Cancer in Never Smokers (LCINS) incidence.
1A). Linear association of age of the patient and proportion of LCINS.

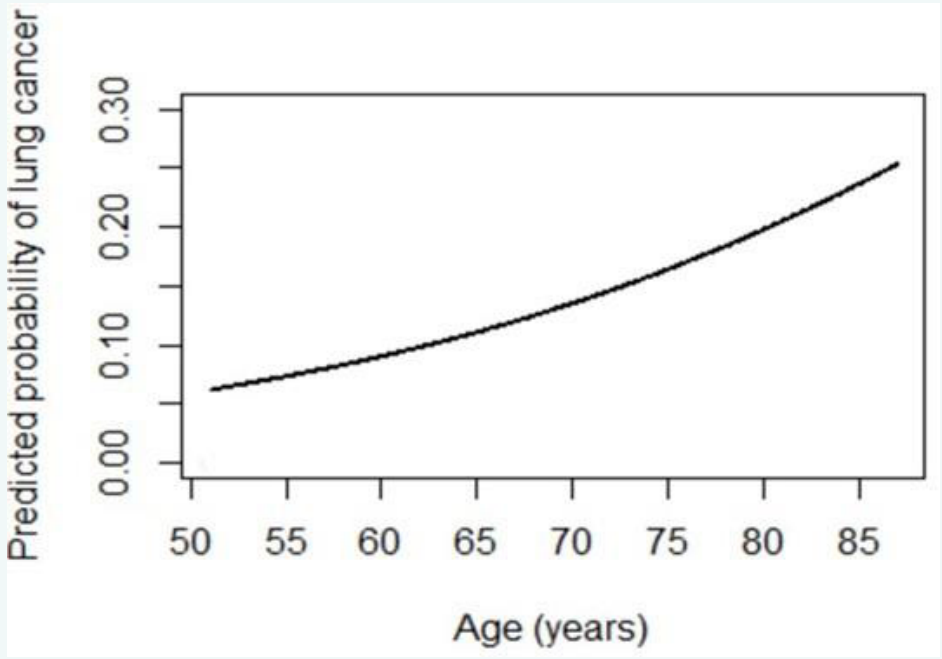


Figure 1B: Logistic regression of the probability of LCINS due to age.

Table 2: T-test of age between those with no LCINS and those with LCINS.

Age	No History	History
Average	71	72.34
Sample Variance	35.03	34.82
Sample Size	31124	5210
t-value	2.57	N/A
Critical t-value	1.65	N/A

Table 3A: Student's t-tests of patients with or without a history of lung cancer in never smokers due to A) height, B) body weight, and C) Body Mass Index (BMI).

A

Height	No History	History
Average	66.02	66.35
Sample Variance	15.63	16.29
Sample Size	31124	5210
t-value	1.37	N/A
Critical t-value	1.65	

B

Weight	No History	History
Average	168.92	170.45
Sample Variance	1299.79	1327.25
Sample Size	31124	5210
t-value	0.08	N/A
Critical t-value	1.65	

C

BMI	No History	History
Average	27.17	27.14
Sample Variance	25.02	24.46
Sample Size	31124	5210
t-value	0.10	N/A
Critical t-value	1.65	N/A

F-value lower that the critical F-value of 2.76 (Table 5).

As a result, there is insufficient statistical evidence to support income as a significant factor contributing to LCINS incidence. This underscores the importance of relying on robust statistical methods to draw meaningful conclusions from data, even when visual trends may initially suggest a relationship.

Family History and Secondary Exposure

While smoking is the primary risk factor for lung cancer, accounting for the majority of cases [9], there are cases of lung cancer that occur in

individuals who have never smoked. The incidence of Lung Cancer in Never-Smokers (LCINS) has been documented in various studies [31]. Family history can play a significant role in the incidence of lung cancer in never-smokers. Furthermore, exposure to secondhand smoke especially enclosed spaces, e.g., homes and workplaces, on a regular basis may increase the risk of developing lung cancer in never-smokers.

The results of the t-test analysis indicate that there is no statistically significant difference in means between patients with a family history of LCINS and those without such a history. In other words, the Z-value is lower than the critical Z-value of 1.96, suggesting that the presence or absence of a family history of LCINS does not lead to a significant variation in in the LCINS incidence, as per the statistical analysis (Table 6). These findings underscore the importance of exploring other potential factors that may be more relevant in understanding the incidence of lung cancer within the studied population, as family history alone does not appear to be a statistically significant determinant in this context.

For secondary exposure, the results of the ANOVA test have provided valuable insights into the relationship between secondary smoke exposure and LCINS (Table 7). The test did not produce a sufficiently high F-value that surpasses the critical F-value of 3.00 and to reject the null hypothesis.. This outcome suggests that, based on the data and analysis performed, secondary smoke exposure is not a statistically significant factor contributing to LCINS incidence. In other words, there is no strong evidence to support the notion that exposure to secondary smoke is a significant driver of LCINS incidence within the studied population.

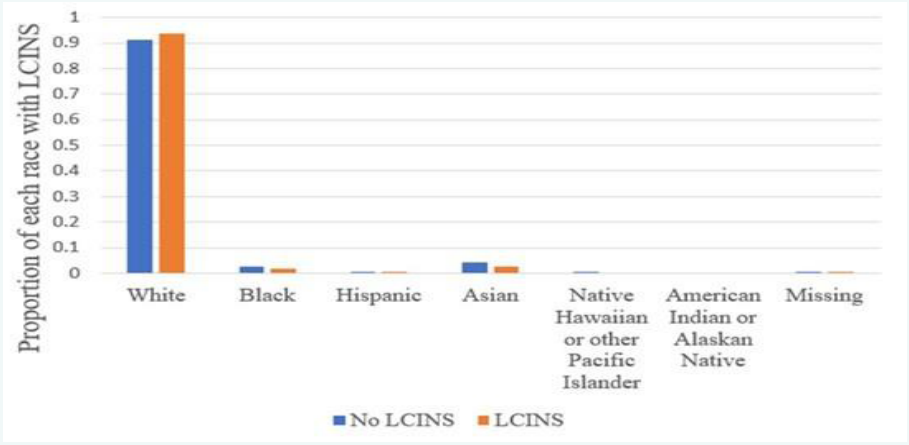


Figure 2A: Lung Cancer in Never Smokers (LCINS) disproportionately affects patients by race. 2A). Bar chart of the proportion of each patient with and without LCINS.

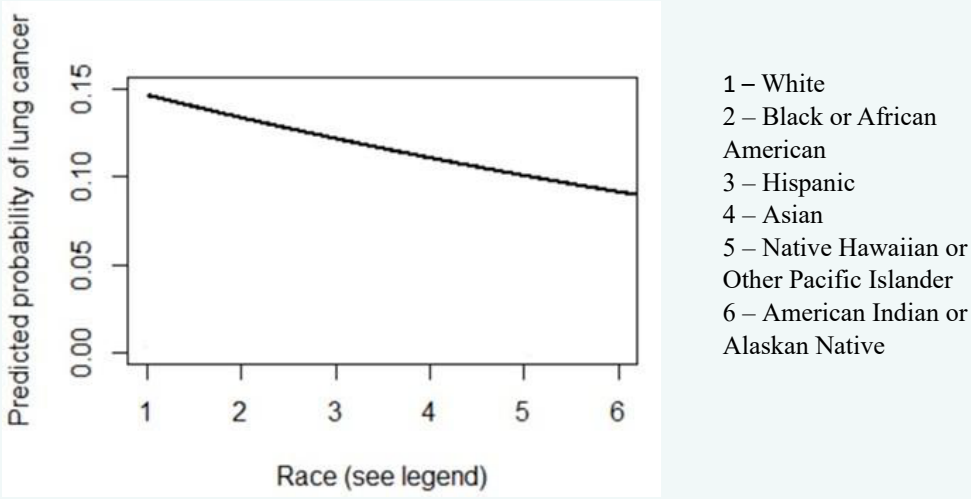


Figure: Logistic regression model of the predicted probability of LCINS by race.

Table 4: Two-sample Z-test of proportions of white population between those with no LCINS and those with history of LCINS (A), and ANOVA test results indicates statistically significant difference in means among the six races (B)*.

A

Race (Caucasian)	No History	History
Proportion	0.91	0.94
Population size	31124	5210
p pooled	0.92	N/A
Z-value	6.01	
Critical Z-value	1.96	

B

Race	Average	Sum of Data Values	Sample Size	G/N	Sum of Squares (total)
White	0.15	4884.00	33276.00	0.14	4434.59
Black	0.11	99.00	887.00	Sum of squares (within)	Sum of squares (between)
Hispanic	0.1	28.00	271.00	4430.56	4.03
Asian	0.1	146.00	1448.00	Mean squares (within)	Mean squares (between)
Native Hawaiian or other Pacific Islandres	0.12	18.00	152.00	0.12	0.81
American Indian or Alaskan Native	0.09	3.00	32.00	F-value	Critical F-value
				6.57	2.21

DISCUSSION

The comprehensive analysis of the dataset focused on determining the factors influencing LCINS patient incidence has yielded crucial insights, with age and race emerging as the primary determinants. Across various statistical methods employed, age and race consistently demonstrated the highest levels of statistical significance. The association between increasing age and elevated LCINS incidence indicates a progressive risk with advancing age [17]. Additionally, the observation that whites exhibit higher rates of LCINS compared to other racial groups underscores the importance of considering demographic factors in understanding and addressing this health issue.

While age and race surfaced as influential factors, the study also examined a spectrum of variables including height, weight, BMI, income, family history, and exposure to secondary smoke. However, the evidence generated failed to reject the null hypotheses for these factors, suggesting that, within the scope of this study, they do not play a significant role in LCINS incidence. This nuanced understanding contributes to the delineation of factors that demand more attention in future research and clinical practice. Meanwhile, other factors, e.g., gender and environmental factors including smog, PM2.5, radon that were not included in this study would warrant further investigation [32].

The most appropriate statistical model for our data analysis is logistic regression, which aligns with the binary nature of the dependent variable predicting LCINS incidence in patients. This modeling choice enables a deeper elucidation of how independent variables relate to the likelihood of LCINS occurrence. Furthermore, the incorporation of t-tests and ANOVA tests alongside logistic regression corroborate evidence, providing additional layers of validation through t-values and p-values.

The statistical analysis identified age as an important risk factor for occurrence of lung cancer in never-smokers. This aligns with the results from previous studies revealing that age significantly impacts the incidence of lung cancer [16]. This further indicates the validity of current model of statistical analysis. Moreover, the statistical methods employed in this study can be applied to other datasets with binary outcomes. This expands the potential impact of the research and underscores the broader utility of our analytical approach in addressing other health-related issues and their outcomes.

The implications stemming from this study hold promising avenues for further research and practical applications. It would be a significant outcome if a Machine Learning (ML) model can be created for LCINS diagnosis, which considers key risk factors and epidemiological elements with differing weights. This model could serve as a valuable tool for healthcare professionals, enabling them to identify key target populations susceptible to LCINS. For example, a preliminary ML model using Electronic Health Records (EHR) data also yielded age, race, and ethnicity as top predictors of lung cancer, with an additional factor being diagnosis of chronic obstructive pulmonary disease [12]. This, in turn, facilitates early lung cancer prevention strategies and enhances awareness for symptom recognition among high-risk individuals.

However, there are opportunities for further refinement and expansion of this research. The inclusion of additional parameters, such as gender, medication history, and environmental factors, could unveil additional influential factors affecting LCINS incidence. Assessing correlations between independent variables is crucial for uncovering potential confounding variables that might impact the observed associations.

In future studies, exploring alternative statistical models, like Partial Least Squares Regression (PLSR), can offer further detailed insights, especially when dealing with myriad factors [33,34]. PLSR has been shown to effectively handle large numbers of predictors and multicollinearity issues [35], making it a promising approach for complex data analysis. The consideration of different platforms, such as SAS [36] and S [37], for data management and analysis, may enhance efficiency and facilitate the handling of large datasets.

In addition to comparing LCINS to lung cancer in smokers within the PLCO dataset, it may be valuable to compare LCINS to other cancers in the PLCO dataset, such as colorectal and ovarian cancer. This comparison is particularly relevant because these cancers share similar driver mutations in the KRAS and p53 genes [27]. By examining the similarities and differences in the incidence, risk factors, and outcomes of these cancers, researchers can gain a more comprehensive understanding of the underlying molecular mechanisms and potential shared risk factors.

However, the PLCO dataset is not the only relevant database to explore when investigating LCINS. Other databases, such as the National Cancer Institute's SEER (Surveillance, Epidemiology, and End Results) Program [38] and The Cancer Genome Atlas (TCGA) [39], also collect and publish data detailing cancer incidence and survival rates from nationwide health registries.

The SEER Program, established in 1973, collects data on cancer cases from various locations and sources throughout the United States, covering approximately 34.6% of the population [34]. This database provides valuable information on cancer incidence, prevalence, survival,

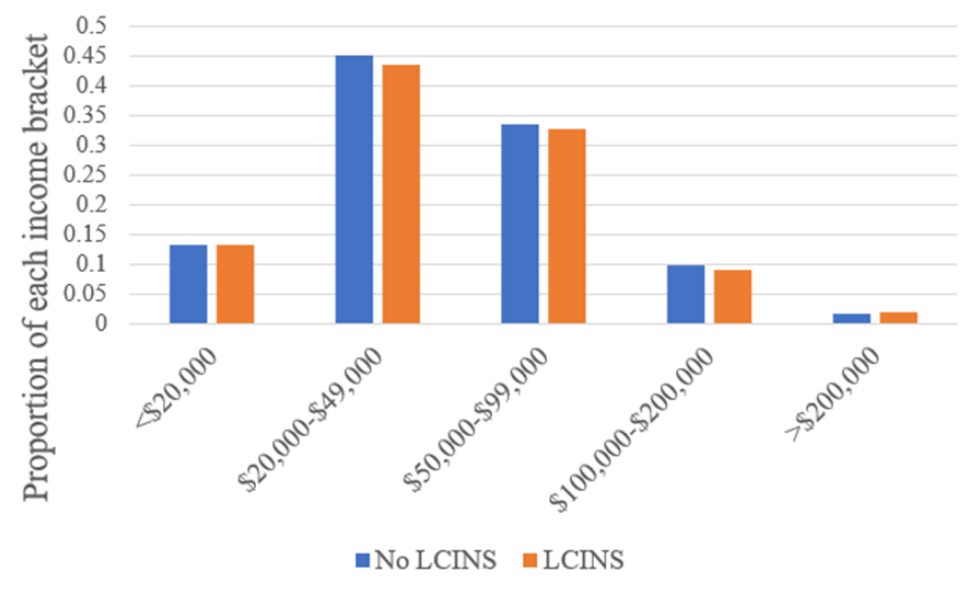
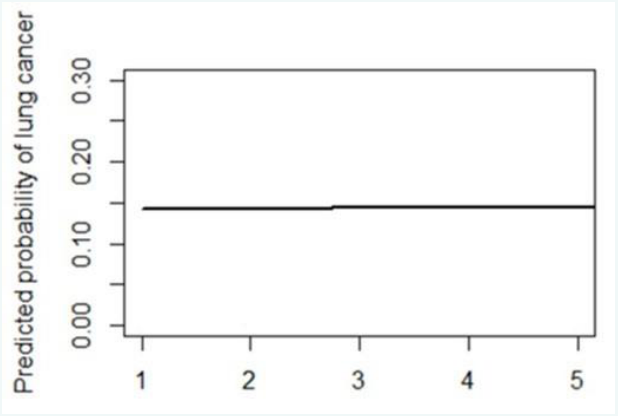


Figure 3: There is no conclusive evidence to suggest that income impacts Lung Cancer in Never Smokers (LCINS). 3A). Bar chart comparing proportions of patient populations in each income bracket.



1 – $<20,000
2 – $20,000 - $49,000
3 – $50,000 - $99,000
4 – $100,000 - $200,000
5 – >$200,000

Figure: Logistic regression model of the predicted probability of LCINS by income.

Table 5: ANOVA test results indicates no difference in means among the five income brackets*.

Income	Average	Sum of Data Values	Sample Size	G/N	Sum of Squares (total)
< $20,000	0.15	586.00	4009.00	0.14	3843.63
$20,000 - $49,000	0.14	1947.00	13714.00	**Sum of squares (within)**	**Sum of squares (between)**
$50,000 - 99,000	0.14	1468.00	10216.00	3843.07	0.56
$100,000 - $200,000	0.13	396.00	2940.00	**Mean squares (within)**	**Mean squares (between)**
> $200,000	0.17	87.00	521.00	0.12	0.14
N/A				**F-value**	**Critical F-value**
				1.14	2.76

*Degrees of freedom (between) = 4; degrees of freedom (within) = 31395

Table 6: Two sample Z-test of family history between those with and those without a history of lung cancer in never smokers

Family History	No History	History
Proportion	0.13	0.13
Population size	3423	586
p pooled	0.13	N/A
Z-value	0	
Critical Z-value	1.96	

Table 7: ANOVA test results indicate no difference in means among the three groups of heavy, medium, and no exposure to secondary smoke*.

Secondary Smoke Exposure	Average	Sum of Data Values	Sample Size	G/N	Sum of Squares (total)
1	0.14	2576	18045	0.14	5596.72
2	0.15	1026	6756	**Sum of squares (within)**	**Sum of squares (between)**
3	0.14	2936	20613	5596.22	0.50
N/A				**Mean squared (within)**	**Mean squared (between)**
				0.12	0.25
				F-value	**Critical F-value**
				2.01	3.00

*Degrees of freedom (between) = 2; degrees of freedom (within) = 45412

and mortality rates, as well as patient demographics and tumor characteristics. By leveraging the SEER database, researchers can conduct large-scale epidemiological studies to identify trends and risk factors associated with LCINS and compare them to other cancers.

The Cancer Genome Atlas (TCGA) is another comprehensive database that can provide insights into LCINS. TCGA is a landmark cancer genomics program that has molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types [39]. This database generates a vast resource of genomic, epigenomic, transcriptomic, and proteomic data, allowing researchers to explore the molecular underpinnings of various cancers, including LCINS. By comparing the genomic profiles of LCINS to other cancers in the TCGA database, researchers can identify common and unique molecular alterations that may contribute to the development and progression of these diseases.

In summary, while the PLCO dataset provides a valuable resource for comparing LCINS to other cancers with shared driver mutations, such as colorectal and ovarian cancer, it is essential to explore other relevant databases like SEER and TCGA. These databases offer additional information on cancer incidence, survival rates, and molecular characteristics, enabling researchers to gain a more comprehensive understanding of LCINS and its relationship to other cancers.

CONCLUSIONS

In summary, this study not only sheds light on the specific factors influencing LCINS incidence but also lays the groundwork for future research endeavors. The results contribute to a more precise understanding of risk factors, opening avenues for targeted interventions, and the broader applicability of the statistical methods ensures the potential for advancements in predictive modeling beyond the scope of this study. The integration of additional parameters such as genetic background and a patient's measured socioeconomic status and exploration of alternative statistical models will undoubtedly contribute to a more comprehensive understanding of the complex factors influencing LCINS.

ACKNOWLEDGEMENT

Not applicable.

FUNDING

This research was generously supported by the Penn Undergraduate Research Mentor (PURM) Program. We extend our sincere gratitude for their funding and assistance.

AVAILABILITY OF DATA AND MATERIALS

The PLCO (Prostate, Lung, Colorectal and Ovarian) Cancer Screening Trial dataset is a publicly available resource that can be accessed through the National Cancer Institute's (NCI) Biometric Research Branch (BRB) Data Archive.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Since the PLCO dataset contains anonymized and de-identified data from study participants, the requirement for individual consent to participate in our analysis is waived. The PLCO Cancer Screening Trial has already obtained the necessary ethics approvals and participant consent for the original data collection and consent for publication.

After reviewing the PLCO study protocols and the associated ethics approvals, we have determined that our analysis, which used the anonymized PLCO dataset, does not require additional ethics approval.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

1. de Alencar VTL, Figueiredo AB, Corassa M, Gollob KJ, Cordeiro de Lima VC. Lung cancer in never smokers: Tumor immunology and challenges for immunotherapy. Front Immunol. 2022; 13: 984349.
2. Kerrigan K, Wang X, Haaland B, Adamson B, Patel S, Puri S, et al. Real world characterization of advanced non-small cell lung cancer in never smokers by actionable mutation status. Clin Lung Cancer. 2021; 22: 260-267.e2.
3. Kim HK, Lee B, Sohn I, Choi YL, Shin SW, Shim JJ, et al. Distinct genomic profile and mutational signature of lung adenocarcinoma in never-smokers. Journal of Clinical Medicine 2021; 10: 1489.
4. Couraud S, Zalcman G, Milleron B, Morin F, Souquet PJ. Lung cancer in never smokers--a review. Eur J Cancer. 2012; 48: 1299-1311.
5. Subramanian J, Govindan R. Lung cancer in never smokers: A review. J Clin Oncol. 2007; 25: 561-570.
6. Yang Y, Yin W, He W, Jiang C, Zhou X, Song X, et al. Phenotype-genotype correlation in multiple primary lung cancer patients in China. Sci Rep. 2016; 6: 36177.
7. Samet JM, Avila-Tang E, Boffetta P, Hannan LM, Olivo-Marston S, Thun MJ, et al. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clin Cancer Res. 2009; 15: 5626-5645.
8. Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, et al. Lung cancer incidence and risk factors in a cohort study of US women: Findings from the Women's Health Initiative. American Journal of Epidemiology. 2008; 168: 192-202.
9. Thun MJ, Carter BD, Feskanich D, Freedman ND, Prentice R, Lopez AD, et al. 50-year trends in smoking-related mortality in the United States. N Engl J Med. 2013; 368: 351-364.
10. Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, Crawford ED, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. Control Clin Trials. 2000; 21: 273S-309S.
11. Ries LA, Melbert D, Krapcho M, Stinchcomb DG, Howlader N, Horner MJ, et al.. SEER cancer statistics review, 1975-2005, National Cancer Institute. 2008.
12. Chandran U, Reps J, Yang R, Vachani A, Maldonado F, Kalsekar I. Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiol Biomarkers Prev. 2023; 32: 337-343.
13. Spector P. Data manipulation with R. New York. Springer Science & Business Media. 2008.
14. Glen S. Student's t-Test: Definition, Examples. StatisticsHowTo.com: Elementary statistics for the rest of us!. 2016.
15. Peng CYJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. The Journal of Educational Research. 2002; 96: 3-14.
16. Aunan JR, Cho WC, Søreide K. The biology of aging and cancer: A brief overview of shared and divergent molecular hallmarks. Aging Dis. 2017; 8: 628-642.
17. Ma Z, Zhu C, Wang H, Ji M, Huang Y, Wei X, et al. Association between biological aging and lung cancer risk: Cohort study and Mendelian randomization analysis. iScience. 2023; 26: 106018.
18. 18. Jeon KH, Shin DW, Han K, Kim D, Yoo JE, Jeong SM, et al. Female reproductive factors and the risk of lung cancer in postmenopausal women: A nationwide cohort study. Br J Cancer. 2020; 122: 1417-1424.
19. Li Y, Wang C, Peng M. Aging immune system and its correlation with liability to severe lung complications. Front Public Health. 2021; 9: 735151.
20. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine. 2011; 365: 395-409.
21. Sanikini H, Yuan JM, Butler LM, Koh WP, Gao YT, Steffen A, et al. Body mass index and lung cancer risk: A pooled analysis based on nested case-control studies from four cohort studies. BMC Cancer. 2018; 18: 220.
22. Wang F, Xu X, Yang J, Min L, Liang S, Chen Y. Height and lung cancer risk: A meta-analysis of observational studies. PLoS One. 2017; 12: e0185316.
23. Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: A systematic review and meta-analysis of prospective observational studies. Lancet. 2008; 371: 569-578.
24. Kolb R, Sutterwala FS, Zhang W. Obesity and cancer: Inflammation bridges the two. Curr Opin Pharmacol. 2016; 29: 77-89.
25. Zhu H, Zhang S. Body mass index and lung cancer risk in never smokers: A meta-analysis. BMC Cancer. 2018; 18: 635.
26. Zhou F, Zhou C. Lung cancer in never smokers-the East Asian experience. Transl Lung Cancer Res. 2018; 7: 450-463.
27. Ha SY, Choi SJ, Cho JH, Choi HJ, Lee J, Jung K, et al. Lung cancer in never-smoker Asian females is driven by oncogenic mutations, most often involving EGFR. Oncotarget. 2015; 6: 5465-5474.
28. Sidorchuk A, Agardh EE, Aremu O, Hallqvist J, Allebeck P, Moradi T. Socioeconomic differences in lung cancer incidence: A systematic review and meta-analysis. Cancer Causes Control. 2009; 20: 459-471.
29. Brock BA, Mir H, Flenaugh EL, Oprea-Ilies G, Singh R, Singh S. Social and biological determinants in lung cancer disparity. Cancers (Basel). 2024; 16: 612.
30. Shankar A, Dubey A, Saini D, Singh M, Prasad CP, Roy S, et al. Environmental and occupational determinants of lung cancer. Transl Lung Cancer Res. 2019; 8: 31-49.
31. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer. 2007; 7: 778-790.
32. Cheng TY, Cramb SM, Baade PD, Youlden DR, Nwogu C, Reid ME. The international epidemiology of lung cancer: Latest trends, disparities, and tumor characteristics. J Thorac Oncol. 2016; 11: 1653-1671.

33. Wold S, Sjöström M, Eriksson L.PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. 2001; 58: 109-130.

34. Geladi P, Kowalski BR. Partial least-squares regression: A tutorial. Analytica Chimica Acta 1986; 185: 1-17.

35. Abdi, H., Partial least squares regression and Projection on Latent Structure Regression (PLS Regression). Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2: 97-106.

36. SAS Institute Inc. SAS/STAT 15.2 User's Guide. Cary NC. SAS Institute Inc. 2021.

37. Becker RA, Chambers JM, Wilks AR. The New S Language: A programming environment for data analysis and graphics. 1988; 702.

38. National Cancer Institute, 2021. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission.

39. Cancer Genome Atlas Research Network; Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013; 45: 1113-1120.