FACTORS MEDIATING LUNG CANCER IN NEVER SMOKERS IDENTIFIED FROM ADVANCED IMAGING AND EPIDEMIOLOGIC ANALYSIS

PRESENTER: KEVIN GUO, CLASS OF 2024
FACULTY MENTOR: DR. FAROUK DAKO, ASSISTANT PROFESSOR, DIAGNOSTIC RADIOLOGY, PERELMAN SCHOOL OF MEDICINE

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 containing more than 155,000 participants and more than 36,000 never-smokers with LCINS was analyzed using R and Excel software to determine risk factors and imaging features of LCINS. The factors analyzed for predictive power in LCINS incidence were age, height, weight, BMI, race, income, family history, and secondary smoke exposure. Multiple statistical methods, including t-tests, ANOVA tests, and logistical regression, were implemented to assess each factor. Through comparison and corroboration of results from the statistical methods, age and race were the key factors that had statistically significant evidence as potential influences in LCINS incidence. 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. These results could be used in future studies to explore deep learning techniques that enable cross-sectional imaging analysis for predictive factors of LCINS or other lung cancers.

BACKGROUND

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. Research literature indicates that LCINS is distinct from smoking-related lung cancer with differences in molecular triggers and treatment responses. 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. 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 commonly considered to be associated with lung cancer are age, ethnicity, genetics, and gender. 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 damages over time. In addition, some researchers hypothesize that the East Asian subtype of lung cancer is partially explained by genetic differences. 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. 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" (Thun, 2008). 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 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. It contains almost 36,000 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. The statistical methods used in this study can be applied to other datasets where the outcome is similarly binary.

RESEARCH QUESTIONS

- Which risk factors would best predict LCINS in the general population?
- What is the ranking of these factors in affecting lung cancer incidence?
- How can these factors be used in LCINS diagnosis?
- After processing the data, what statistical methods should be utilized to determine statistical significance of the identified risk factors?
- Would there be significant differences in demographics between populations with lung cancer and those without lung cancer LCINS?

INDEPENDENT AND DEPENDENT VARIABLES

- Independent variable: age, height, weight, BMI, race, income, family history, secondary smoke exposure
- Dependent variables: lung cancer incidence

HYPOTHESIS

Null ($\alpha = 0.05$)
- The difference between the mean ages of the populations with and without lung cancer is zero
- The incidence of lung cancer across all major races and income brackets is the same
- The proportion of patients with and without a family history of lung cancer are the same
- The patient groups with heavy exposure, medium exposure, and no exposure to secondary smoke have equal incidences of lung cancer

Alternative ($\alpha = 0.05$)
- The difference between the mean ages of the populations with and without lung cancer is greater than zero
- The incidence of lung cancer across all major races and income brackets is not the same
- The proportion of patients with and without a family history of lung cancer are the same
- The patient groups with greater exposure to secondary smoke will have higher incidences of lung cancer

METHODS

Software:

R

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. Furthermore, it can create easy-to-customize plots that visualize statistical results effectively. Furthermore, it can manage large amounts of data in forms of arrays, matrices, and other types of dataframes. Users can also download packages to perform multiple types of data analyses according to their needs and specifications. In this study, the data compiled from the PLCO dataset was filtered for nonsmokers by selecting all participants who have not smoked any cigarette packs at the time of the study. This resulted in a total of 36,335 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. Using the readxl package, the data was imported into R and analyzed with logistic regression and ANOVA testing.

Statistical methods and comparison:

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. In this study, t-tests were applied for analyzing all parameters except for secondary smoke exposure due to there being more than two groups for the risk factor.

*If two different sample groups' proportion means need to be tested, a two-sample Z-test is used instead. However, for the sake of consistency, t-test will be used.

ANOVA – ANOVA (Analysis Of Variance) assesses whether the population means of more than two sample groups are all equal to each other. In the case of family history, ANOVA test would not be appropriate since there are only two groups: patients with family history of lung cancer and patients with no history of LCINS.

Logistic Regression – Logistic regression is a statistical model that is primarily used in datasets with only two outcomes for the dependent variable. 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.

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 • Allows 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 assumption that both populations are normally distributed • Lower degrees of freedom require higher t-values to reach t-test significance
Logistical regression	• Easy to implement and interpret • No assumptions are made about the distribution of the variables involved • Provides coefficient 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 variable 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 variances of data in groups are homogeneous

RESULTS

Age

Figure 1A: Effect of age on LCINS

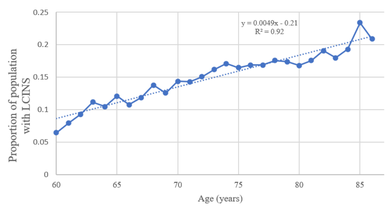

Figure 1B: Logistic regression model of predicted probability LCINS by age

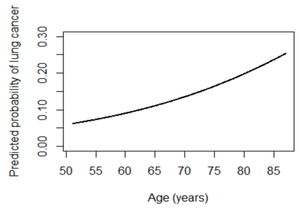

Table 1: T-test of age between those with no LCINS and those with LCINS

Age	No history	History
Average	71.00	72.34
Sample Variance	35.03	34.82
Sample size	31124	5210
t-value	2.57	
t-critical value	1.65	

- A high linear correlation between age and lung cancer incidence is observed, with an R^2 value of 0.92.
- The logistic regression model shows an increasing trend between age and predicted probability of lung cancer. This result is further corroborated by a p-value less than 0.05 (data not shown).
- The t-test has a t value of approximately 2.57, which is greater than the t critical value of 1.65. Thus, age is a significant factor for LCINS incidence.

Height, Weight, and BMI

Table 2A: T-test of weight between those with no LCINS and those with LCINS

Weight	No history	History
Average	168.92	170.45
Sample Variance	1299.79	1327.25
Sample size	31124	5210
t-value	0.08	
t-critical value	1.65	

Table 2C: T-test of BMI between those with no LCINS and those with history of LCINS

BMI	No history	History
Average	27.17	27.14
Sample Variance	25.02	24.46
Sample size	31124	5210
t-value	-0.10	
t-critical value	1.65	

Table 2B: T-test of height between those with no LCINS and those with history of LCINS

Height	No history	History
Average	66.02	66.35
Sample Variance	15.63	16.29
Sample size	31124	5210
t-value	1.37	
t-critical value	1.65	

- All three t-tests involving weight, height, and BMI yielded t-values lower than their t-critical value of 1.65, suggesting that they are not significant factors in LCINS incidence.
- Only t-values higher than the t-critical value would indicate that the risk factor plays a significant role in LCINS incidence.
- This result is further corroborated by high p-values greater than 0.05 from logistic regression (data not shown).

Race

Figure 3A: Bar graph of proportion of each group with LCINS and those without LCINS

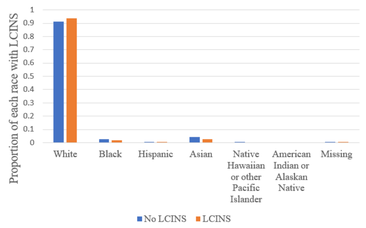

Figure 3B: Logistic regression model of predicted probability of LCINS by race

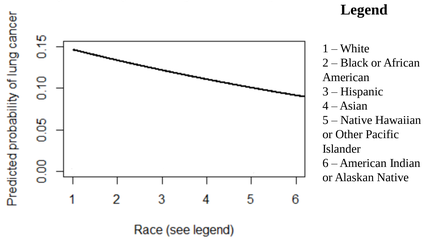

Table 3A: Two-sample t-test of proportions of white population white population between those with no LCINS and those with history of LCINS

Race (Caucasian)	No history	History
Average	0.91	0.94
Population size	31124	5210
p pooled	0.92	
Z value	6.01	
Z-critical (abs)	1.96	

Table 3B: ANOVA test results accounting for the difference of means between the six races

Race					
	Average	Sum of data values	Number of data values	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	0.10	28.00	271.00	4430.56	4.03
Asian	0.10	146.00	1448.00	Mean squared (within)	Mean squared (between)
Native Hawaiian or other Pacific Islande	0.12	18.00	152.00	0.12	0.81
American Indian or Alaskan Native	0.09	3.00	32.00	f value	f critical value
				2.21	6.57

- Through all t-tests* conducted (see Methods section), the one for the Caucasian population indicates a statistically significant proportion of Caucasian people with LCINS.
- The logistic regression model further indicates that Caucasian people generally have a higher risk of lung cancer compared to other races. This result is further corroborated by a high p-value greater than 0.05 from logistic regression (data not shown).
- The ANOVA test yields a f value higher than the f-critical value, rejecting the null hypothesis. This indicates that race is a significant factor for LCINS incidence.

RESULTS cont'd.

Income

Figure 4A: Bar graph comparing proportions of patient populations in each income bracket

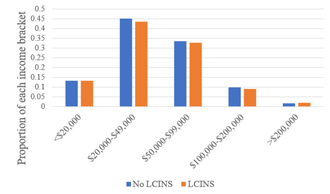

Figure 4B: Logistic regression model of predicted probability of lung cancer in never-smokers by income

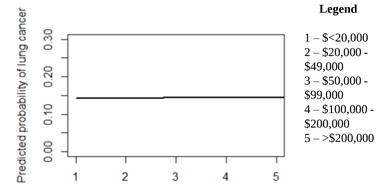

Table 4: ANOVA test results accounting for the difference of means between the five income brackets

Income					
	Average	Sum of data values	Number of data values	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 squared (within)	Mean squared (between)
>$200,000	0.17	87.00	521.00	0.12	0.14
				f value	f critical value
				1.14	2.76

- Bar graph suggests that patients with no LCINS are associated with higher incomes.
- However, the logistic regression model and ANOVA test indicate the population means are equal. There is not enough statistical evidence to suggest income as a factor in LCINS incidence.

Family History

Table 5: Two sample t-test of family history between those with no LCINS and those with history of LCINS

Family history	No history	History
Average	0.13	0.13
Population size	3423	586
p pooled	0.13	
t value	0.00	
t-critical (abs)	1.96	

- The t-test yielded no statistical evidence of a difference in means between patients with a family history of LCINS and those without LCINS.

Secondary Smoke Exposure

Table 6: ANOVA test results accounting for the difference of means between the three groups of heavy, medium, and no exposure to secondary smoke

Secondary Smoke Exposure	Average	Sum of data values	Number of data values	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
				Mean squared (within)	Mean squared (between)
				0.12	0.25
				f value	f critical value
				2.01	3.00

- The ANOVA test did not yield a high enough f critical value to reject the null hypothesis, indicating that secondary smoke exposure is not a significant factor for LCINS incidence.

CONCLUSION

- The results indicate that age and race are the most important factors in LCINS patient incidence.
- Age and race showed the most statistical significance across all statistical methods. As age increases, LCINS incidence increases. Caucasian people tend to have higher rates of LCINS when compared to other races
- The other factors (height, weight, BMI, income, family history, and secondary smoke) did not generate enough evidence to reject the null hypotheses and therefore do not play a significant role in LCINS incidence.
- The most appropriate statistical model for current dataset is logistical regression due to the binomial nature (whether or not a patient had LCINS) of the dependent variable.
  - However, t-tests and ANOVA tests can be used in conjunction with logistical regression to corroborate evidence with t-values or p-values.
- The implications with the results of this study:
  - A machine learning model for LCINS diagnosis could be developed by accounting for major risk and epidemiology factors with differing weights.
  - It allows doctors to identify key target populations who are susceptible to LCINS and educate them for early prevention or symptom recognition.
  - The statistical methods used to analyze the PLCO dataset can be applied to other datasets where the outcome is similarly binary, especially in populations outside of the U.S.

FUTURE DIRECTIONS

- Include more parameters such as gender and medication history to determine whether any more important factors might influence LCINS incidence.
- Assessing the correlation between the independent variables would help determine if there are any confounding variables at play.
- Using other statistical models such as partial least squares regression (PLSR) could be considered in future studies involving many factors.
- These models could be explored on different platforms such as SAS and S, a direct successor to R, to better manage data.
- The approach to the data analysis models used in this study may be extended to cancer-related and noncancer-related datasets and could help create a machine learning model that can predict lung cancer based on weighted factors.

REFERENCES

- Lemjabbar-Alaoui, H., Hassan, O. U., Yang, Y.-W., & Buchanan, P. (2015, December). Lung cancer: Biology and treatment options. Biochimica et biophysica acta. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4663145/.
- McCarthy, W. J., Meza, R., Jeon, J., & Moolgavkar, S. H. (2012, July). Chapter 6: Lung cancer in Never SMOKERS: Epidemiology and risk prediction models. Risk analysis : an official publication of the Society for Risk Analysis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3485693/.
- Okazaki, I., Ishikawa, S., Ando, W., & Sohara, Y. (2016, December 1). Lung adenocarcinoma in Never Smokers: Problems of primary prevention from aspects of SUSCEPTIBLE genes and carcinogens. Anticancer Research. https://ar.iiarjournals.org/content/36/12/6207.
- Samet, J. M., Avila-Tang, E., Boffetta, P., Hannan, L. M., Olivo-Marston, S., Thun, M. J., & Rudin, C. M. (2009, September 15). Lung cancer in never smokers: Clinical epidemiology and environmental risk factors. Clinical Cancer Research. https://clincancerres.aacrjournals.org/content/15/18/5626.
- Subramanian, J., & Govindan, R. (2016, September 21). Article tools. Journal of Clinical Oncology. https://ascopubs.org/doi/10.1200/JCO.2006.06.8015.
- Zhou, F., & Zhou, C. (2018, August). Lung cancer in never smokers-the East Asian experience. Translational lung cancer research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131183/.