Imperial College Business School

The Investigation of the Impact of Smoking Behaviour on Lung Cancer Diagnosis

BUSI70501 Introduction to Health Analytics

Module leader: Samantha Burn

Submitted by Group S

Word count:2498

Submission date:10/02/2025

Content

1 Introduction

Lung cancer is one of the most prevalent and deadly malignancies worldwide, primarily caused by cigarette smoking, and smoking behaviors are still increasing in many countries (Parkin et al., 1994). In the United States, an estimated 226,650 new cases of lung cancer and 124,730 deaths from lung cancer are expected in 2025 while smoking remains the significant cause (American Cancer Society, 2023). While the link between smoking behaviors and lung cancer is well established, the extent to which different aspects of smoking behaviour influence lung cancer risk remains unsolved (B. Rachet et al., 2004). Previous studies suggest that factors such as the age of smoking initiation, smoking duration, and cumulative smoking exposure may contribute differently to the likelihood of developing lung cancer (Matos et al., 1998; Engeland et al., 1996; Khuder, 2001). While some studies have reported the effect of age of first smoking and smoking duration, others have found no significant correlation to lung cancer (Benhamou, Benhamou & R Flamant, 1987; Hegmann et al., 1993; Brown & Chu, 1987). However, most of these studies have failed to account for confounding by smoking duration, which influences lung cancer risk. Investigating these components in detail can provide valuable insights into how smoking behaviour affects cancer risk and inform more targeted public health interventions and behavioral strategies for smoking cessation.

By analysing how smoking duration interacts to shape lung cancer risk, this study aims to investigate the impact of smoking duration on lung cancer risk using quantitative methods. Quantitative conclusions were drawn by performing logit regression analysis on data from the IPUMS NHIS database to reveal the correlation between smoking behaviour and lung cancer.

2 Data

2.1 Dataset

This study utilizes the IPUMS NHIS (National Health Interview Survey) dataset, which provides comprehensive health data representative of the U.S. population (Adeyemi et al., 2021). It is well-suited for analyzing the relationship between smoking behavior and lung cancer risk, containing detailed self-reported smoking histories, demographic information, and health statuses (American Cancer Society, 2023).

The study focuses on current smokers, excluding former smokers to eliminate the impact of smoking cessation on smoking duration and lung cancer risk (Benhamou, Benhamou & Flamant, 1987). Data from three time periods (2000, 2005, and 2010) were extracted, as the only years with available data on all key variables in IPUMS NHIS, guaranteeing an adequate

sample size and temporal variance for analysis (Engeland et al., 1996).

2.2 Main variables

2.2.1 Dependent variable

The dependent variable, lung cancer diagnosis (CNLUNG), is a binary indicator of whether an individual has been diagnosed with lung cancer (1 = Yes, 0 = No) (Gutiérrez-Torres et al., 2024). The dataset records individuals who reported a lung cancer diagnosis, and all invalid responses were removed to maintain data quality. This variable serves as the primary outcome measure, enabling assessment of the impact of smoking behavior on lung cancer risk (Khuder, 2001). The descriptive statistic of dependent variable is presented in Appendix 1.

2.2.2 Explanatory variables

The key independent variable is smoking duration, which measures the number of years between the initiation of smoking and diagnosis of lung cancer, as long-term smoking is hypothesized to increase lung cancer risk(Matos et al., 1998). This variable is calculated as the difference between the respondent's current age (AGE) and the age at which they first smoked regularly (SMOKAGEREG) (Hegmann et al., 1993). Consistency checks ensured that smoking initiation age did not exceed the respondent's age at diagnosis. As a continuous variable, it reflects the dose-response relationship observed in epidemiological studies. The descriptive statistic of dependent variable is in appendix 2.

2.2.3 Control variables

To reduce the impact of confounding factors on the results, we added the following control variables to the model based on previous studies. Three control variables were introduced into the model of this study, including gender, income, and family cancer history.

Gender (SEX) is a binary variable indicating male (1) or female (2), based on self-reported data. Responses marked as uncertain or refused were excluded. Gender is included due to established differences in smoking patterns and susceptibility to lung cancer (Radzikowska, Głaz & Roszkowski, 2002). A detailed description of variable is shown in Appendix 3.

Income level(EARNIMP1) is a categorical variable representing an individual's total personal earnings in the previous calendar year. It includes both reported and imputed values to address missing data, ensuring completeness in income-related analyses. The imputed income variables in the IPUMS NHIS dataset have already been merged with other survey data from each survey year, allowing for consistent use in analysis. In the dataset, EARNIMP1 is recorded as a numeric categorical variable, with predefined income brackets from $1 to $115,000 and over in 31 categories. Since income level influences both smoking behavior

and health outcomes, controlling for income helps account for socioeconomic factors that may confound the relationship between smoking duration and lung cancer risk. The distribution of responses among different income categories is illustrated in Appendix 4.

Family history of lung cancer (family_lung_cancer) variable is constructed as a binary indicator to account for genetic predisposition to lung cancer. It is derived from six related variables in the IPUMS NHIS dataset, capturing whether a respondent's biological relatives, including mother, father, full siblings, and children, have been diagnosed with lung cancer. The family_lung_cancer variable is coded as 1 if any of these six variables are recorded as "Mentioned" (Value = 2), indicating that at least one immediate family member was diagnosed with lung cancer. Conversely, it is coded as 0 if all six variables are recorded as "Not mentioned" (Value = 1), indicating no reported family history of lung cancer. Observations with invalid responses(Values = 7, 8, 9) are excluded from the analysis to maintain data reliability. Including family_lung_cancer as a control variable helps adjust for genetic factors. The descriptive statistics of family history are displayed in Appendix 5.

3 Empirical Strategy

3.1 Model Construction

This report estimates a logistic regression model of the risk variation of an individual being diagnosed with lung cancer on smoking duration, controlling for sex, income, and family history of lung cancer.

To assess the impact of smoking duration on lung cancer risk the model is specified as follows:

$$y_i = \beta_0 + \beta_1 \bullet smoking\ duration_i + \gamma \bullet Z_i + \varepsilon_i$$

where:

- y_i is a binary variable indicating whether an individual ever has been diagnosed with lung cancer (y_i =1 if diagnosed, y_i =0 otherwise).

- $smoking\ duration_i$ is a continuous variable representing the number of years an individual has been smoking, reflecting cumulative exposure effect of smoking on

lung cancer risk.

- Z_i represents a vector of control variables, including:
    - sex_i: a binary variable indicating the individual's sex (1= male, 2= female).
    - $income_i$: a category variable representing the individual's income level.
    - $family\ lung\ cancer_i$: a binary variable indicating whether the individual has a family member with diagnosed lung cancer (1= Yes, 0= No).
- ε_i is the error term.

3.2 Hypothesis Test

To evaluate the statistical significance of the variables in the model, we conduct hypothesis testing based on the estimated coefficients:

$$H_0: \beta_i = 0 \text{ vs } H_1: \beta_i \neq 0$$

- Null Hypothesis (H_0): The variable has no effect on lung cancer diagnosis.
- Alternative Hypothesis (H_1): The variable has a significant effect on lung cancer diagnosis.

Statistical significance will be determined based on the p-values associated with each coefficient in the regression output. If $p < 0.05$, we reject the null hypothesis and conclude that the variable significantly influences lung cancer risk.

3.3 Assumption

This logistic regression model analysis relies on several key assumptions to ensure valid results. It assumes that all observations come from the same underlying process, with no extreme outliers or influential points distorting the estimates. The logit link function is appropriate for use in modeling the relationship between smoking duration and binary lung cancer diagnosis, assumed linearity on the log-odds scale. Additionally, the variance of the

response variable follows the expected form for a binomial distribution: $\mathrm{Var}(y_i) = \mu_i(1 - \mu_i)$. A constant dispersion parameter (ϕ) is considered across all observations.

Independence of observations is also assumed, meaning that one individual's diagnosis does not influence another's.

3.4 Standard Error Adjustment

To account for the possibility of non-constant variance in the error term, heteroskedasticity-robust standard errors will be used. This ensures that the estimated standard errors, t-statistics, and p-values remain valid even if the assumption of homoscedasticity is violated.

4 Result

4.1 Results

The results of the fitted logistic regression model are expressed in terms of log-odds, following the logit link function:

$$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \alpha + \beta x_i$$

where β represents the log odds ratio for one unit increase in x_i. To facilitate analysis, we convert the coefficients into odds ratios:

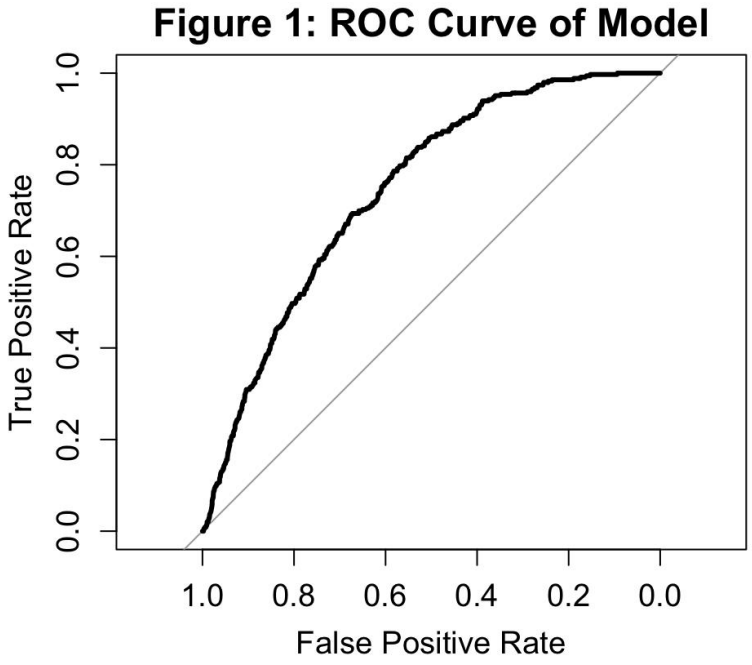
$$\mathrm{OR} = e^{\beta}$$

The model estimates are presented in Table 1, includes key statistics to assess the overall fit of the model. The results show that smoking duration is a significant predictor of lung cancer diagnosis. However, income does not demonstrate a statistically significant effect (p = 0.549) on lung cancer risk in this model. The reduction in residual deviance (from 8257.3 to 7570.8) illustrates that the model explains a meaningful portion of the variation in lung cancer diagnosis compared to the null hypothesis. As shown in Figure 1, the performance was assessed using the Receiver Operating Characteristic Curve(ROC), with an Area Under the Curve(AUC) value of 0.7449, indicating moderate predictive ability, and effective distinction between positive and negative class samples.

Variable	Estimate	Standard Error	Z-value	p-value[a]	OR
Intercept	-5.364014	0.148391	-36.148	< 2e-16	0.004682
Smoking Duration	0.041226	0.002130	19.357	< 2e-16	1.042
Sex (Female = 2)	0.737046	0.068823	10.709	< 2e-16	2.09
Income	-0.001424	0.002376	-0.600	0.549	Not significant
Family Lung Cancer	0.562769	0.065079	8.647	< 2e-16	1.76
Statistics			**Value**		
Number of Observations			18,480		
Null Deviance			8257.3 (on 18,480 degrees of freedom)		
Residual Deviance			7570.8 (on 18,475 degrees of freedom)		
AIC (Akaike Information Criterion)			7580.8		

Table 1: Results of fitted Logistic Regression Model

[a] Null hypothesis: this variable has no effect in this model



To examine potential multicollinearity among the predictor variables, we calculated the Variance Inflation Factor (VIF), which is shown in last column of Table 2. All VIF values are near 1, indicating that multicollinearity is not a significant issue in this model. Additionally, we used robust standard errors to account for heteroskedasticity, ensuring more reliable coefficient estimates. The robust standard errors, also presented in Table 2, provide adjusted measures of variability that are less sensitive to violations of model assumptions. The results suggest that the logistic regression model estimates are mostly stable, with small differences in the coefficients.

Variable	Estimate	Std. Error	Z value	p-value	VIF value
(Intercept)	-5.3640145	0.1315728	-40.7684	<2e-16	
Smoking Duration	0.0412256	0.0020837	19.7848	<2e-16	1.059860
Sex2	0.7370465	0.0672561	10.9588	<2e-16	1.043029
Income	-0.0014243	0.0025344	-0.5620	0.5741	1.059820
Family Lung Cancer	0.5627688	0.0664811	8.4651	<2e-16	1.030699

Table 2: Robust Standard Errors and VIF Values for Logistic Regression Models

4.2 Robustness Check

To ensure the stability and reliability of logistic regression results, we conducted a robustness check by modifying the fitted model (Model A) in two ways: removing the income variable (Model B) and adding age as a control variable (Model C). The results are summarized in Table 3.

Variables	Coefficient	p-value[a]	AIC
Model A			**7580.8**
Model B			**7579.1**
Intercept	-5.394949	<2e-16	
Smoking duration	0.041446	<2e-16	
Sex2	0.744645	<2e-16	
Family Lung Cancer	0.561737	<2e-16	
Model C			**7570.1**
Intercept	-5.742847	<2e-16	
Smoking duration	0.021418	0.000189	
Sex2	0.706637	<2e-16	
income	-0.001379	0.560900	
Age	0.021350	0.000234	
Family Lung Cancer	0.552217	<2e-16	

Table 3: Robustness Check of Logistic Regression Models

[a] Null hypothesis: this variable has no effect in this model

Removing income (Model B) slightly improves model fit (AIC: 7579.1 to 7580.8) while keeping explanatory variables significant, confirming that income has no meaningful impact on lung cancer diagnosis. Adding age (Model C) further improves model fit (AIC: 7570.1) and is statistically significant ($p = 0.000234$), suggesting older individuals face higher lung cancer risk. Including age reduces the smoking duration coefficient (0.0414 to 0.0214), indicating age absorbs part of its effect. Coefficients remain similar across models, validating the robustness of our initial model and supporting the decision to exclude income while considering age as a key determinant in lung cancer diagnosis probability.

4.3 Heterogeneity Analysis

To explore whether the relationship between smoking duration and lung cancer varies by gender, we estimated separate logistic regression models for males (Model D) and females (Model E). The results indicate that smoking duration has a stronger effect on lung cancer risk for males compared to females. This difference may be due to biological factors, lifestyle, or variations in smoking intensity. Family history of lung cancer is a significant predictor for both genders, though its effect appears to be slightly stronger for females. The model fit is notably better for males, indicating that the model more effectively explains lung cancer risk in men than in women.

An interaction model (Model F) was also established to assess whether the effect of smoking duration on lung cancer risk differs between males and females, introducing an interaction term between smoking duration and sex. While females exhibit a higher baseline risk of lung cancer, the increase in risk per additional unit of smoking duration is more pronounced in males(the interaction term is -0.0332). Furthermore, the strong effect of family history underscores the importance of genetic predisposition in lung cancer risk, regardless of gender.

The results from all these models provide strong evidence that gender moderates the relationship between smoking duration and lung cancer risk, are presented in Table 4.

Variable	**Model D** (Estimate, p-value)	**Model E** (Estimate, p-value)	**Model F** (Estimate, p-value)
Intercept	-5.5725 (**p** < 2e-16)	-3.4900 (**p** < 2e-16)	-5.4597 (**p** < 2e-16)
Smoking Duration	-3.4900 (**p** < 2e-16)	0.0293 (**p** < 2e-16)	0.0632 (**p** < 2e-16)
Income	0.0054 (**p** = 0.0899)	-0.0063 (**p** = 0.0818)	-0.0002 (p=0.936)
Family Lung Cancer	0.5171 (**p** < 4.3e-06)	0.5867 (**p** < 1.75e-13)	0.5653 (**p** < 2e-16)
Sex (Female=2)			1.9203 (**p** < 2e-16)
Smoking duration:sex2			-0.0332 (**p** <1.55e-13)
AIC	2659.8	4864.5	7526.4
Residual Deviance	2651.8	4856.5	7514.4

Table 4: Logistic Regression Results for Lung Cancer Diagnosis by Sex and Interaction Effects

5 Discussion and interpretation

This study provides strong evidence that smoking duration significantly influences lung cancer risk, reinforcing prior research on smoking behavior and lung cancer incidence. The logistic regression analysis reveals that each additional year of smoking increases the risk of lung cancer by 4.2% (OR = 1.042), confirming findings from previous studies (Moolgavkar, Dewanji & Luebeck, 1989; Rachet et al., 2004; Gutiérrez-Torres et al., 2024). By incorporating sex and family history of lung cancer as control variables, this study refines the understanding of this relationship.

A key finding is that gender variable moderates the effect of smoking duration on lung cancer risk. The heterogeneity analysis indicates larger smoking duration has stronger effect in males on increasing the probability of diagnosis of lung cancer. This could be due to

differences in lung physiology, hormonal influences, or smoking intensity. The interaction model confirms that sex significantly moderates smoking duration's impact on lung cancer ($p < 1.55e-13$). While men have traditionally been viewed as more at risk (Xing et al., 2011), newer studies suggest that women may be more biologically sensitive to tobacco carcinogens, increasing their lung cancer susceptibility even at lower exposure levels (Radzikowska, Głaz, & Roszkowski, 2002; International Early Lung Cancer Action Program Investigators, 2006). Since smoking intensity was not included in this study, future research should explore sex-specific differences in smoking patterns and lung cancer risk.

Another key finding is that the family history of lung cancer is a significant predictor of lung cancer risk. Individuals with a family history were 76% more likely to develop lung cancer (OR = 1.76), highlighting the role of genetic susceptibility and shared environmental exposures (Matakidou, Eisen & Houlston, 2005; Yin et al., 2021). Even after adjusted, family history remains a strong risk factor, suggesting that genetic predisposition amplifies the carcinogenic effects of smoking. This finding supports prioritizing high-risk populations for targeted smoking cessation programs and early screening strategies.

Income level did not significantly influence lung cancer risk ($p = 0.549$). Even after removing income from the model, the other predictors remained stable, suggesting that socioeconomic status does not directly impact lung cancer risk in this study (Redondo-Sánchez et al., 2022; Roque et al., 2023). While lower-income populations generally have higher smoking rates, once smoking duration is controlled for, income appears to have limited predictive power.

Limitation

Firstly, the lack of smoking intensity as a variable is a major limitation. While smoking duration is important, cumulative exposure, typically measured by pack-years, is a stronger predictor of lung cancer risk (Khuder, 2001). The omission of smoking intensity may lead to underestimation of smoking's impact. Second, excluding former smokers limits generalizability. Research shows that former smokers remain at an elevated risk compared to never-smokers, though their risk declines over time (Engeland et al., 1996). Future studies should analyze smoking cessation effects. In addition, the cross-sectional design prevents causal inference. Longitudinal studies are needed to assess how changes in smoking behavior affect lung cancer incidence over time. Confounders such as genetic predisposition, occupational exposures, and comorbidities could bias the results (Matakidou, Eisen & Houlston, 2005). Future research should incorporate additional confounders to improve validity.

Reference List

Adeyemi, O.J., Gill, T.L., Paul, R. and Huber, L.B. (2021). Evaluating the association of self-reported psychological distress and self-rated health on survival times among women with breast cancer in the U.S. PLOS ONE, 16(12), p.e0260481. doi:https://doi.org/10.1371/journal.pone.0260481.

American Cancer Society (2023). Lung Cancer Statistics | How Common is Lung Cancer? [online] www.cancer.org. Available at: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html.

B. Rachet, J. Siemiatycki, M. Abrahamowicz and K. Leffondré (2004). A flexible modeling approach to estimating the component effects of smoking behavior on lung cancer. Journal of Clinical Epidemiology, 57(10), pp.1076–1085. doi:https://doi.org/10.1016/j.jclinepi.2004.02.014.

Benhamou, E., Benhamou, S. and R Flamant (1987). Lung cancer and women: results of a French case-control study. British Journal of Cancer, 55(1), pp.91–95. doi:https://doi.org/10.1038/bjc.1987.19.

Brown, C.C. and Chu, K.C. (1987). Use of multistage models to infer stage affected by carcinogenic exposure: Example of lung cancer and cigarette smoking. Journal of Chronic Diseases, 40, pp.171S179S. doi:https://doi.org/10.1016/s0021-9681(87)80020-6.

Dosemeci, M., Alavanja, M., Vetter, R., Eaton, B. and Blair, A. (1992). Mortality among Laboratory Workers Employed at the U.S. Department of Agriculture. Epidemiology, 3(3), pp.258–262. doi:https://doi.org/10.1097/00001648-199205000-00012.

Engeland, A., Haldorsen, T., Andersen, A. and Tretli, S. (1996). The impact of smoking habits on lung cancer risk: 28 years' observation of 26,000 Norwegian men and women. Cancer Causes and Control, 7(3), pp.366–376. doi:https://doi.org/10.1007/bf00052943.

Gutiérrez-Torres, D.S., Kim, S., Albanes, D., Weinstein, S.J., Inoue-Choi, M., Albert, P.S. and Freedman, N.D. (2024). Changes in smoking use and subsequent lung cancer risk in the ATBC study. Journal of the National Cancer Institute, [online] p.djae012. doi:https://doi.org/10.1093/jnci/djae012.

Haldorsen, T. (1999). Cohort analysis of cigarette smoking and lung cancer incidence among Norwegian women. International Journal of Epidemiology, 28(6), pp.1032–1036. doi:https://doi.org/10.1093/ije/28.6.1032.

Hegmann, K.T., Fraser, A.M., Keaney, R.P., Moser, S.E., Nilasena, D.S., Sedlars, M., Higham-Gren, L. and Lyon, J.L. (1993). The effect of age at smoking initiation on lung cancer risk. Epidemiology (Cambridge, Mass.), [online] 4(5), pp.444–448. doi:https://doi.org/10.1097/00001648-199309000-00010.

International Early Lung Cancer Action Program Investigators (2006). Women's susceptibility to tobacco carcinogens and survival after diagnosis of lung cancer. JAMA, [online] 296, pp.180–184. doi:https://doi.org/10.1001/jama.296.2.180.

Khuder, S.A. (2001). Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. Lung cancer (Amsterdam, Netherlands), [online] 31(2-3), pp.139–48. doi:https://doi.org/10.1016/s0169-5002(00)00181-1.

Kreuzer, M., Kreienbrock, L., Gerken, M., Heinrich, J., Bruske-Hohlfeld, I., Muller, K.-M. . and Wichmann, H.E. (1998). Risk Factors for Lung Cancer in Young Adults. American Journal of Epidemiology, 147(11), pp.1028–1037. doi:https://doi.org/10.1093/oxfordjournals.aje.a009396.

Matakidou, A., Eisen, T. and Houlston, R.S. (2005). Systematic review of the relationship between family history and lung cancer risk. British journal of cancer, [online] 93(7), pp.825–33. doi:https://doi.org/10.1038/sj.bjc.6602769.

Matos, E., Vilensky, M., P Boffetta and M Kogevinas (1998). Lung cancer and smoking: A case-control study in Buenos Aires, Argentina. Lung Cancer, 21(3), pp.155–163. doi:https://doi.org/10.1016/s0169-5002(98)00055-5.

Moolgavkar, S.H., Dewanji, A. and Luebeck, G. (1989). Cigarette Smoking and Lung Cancer: Reanalysis of the British Doctor's Data. JNCI Journal of the National Cancer Institute, 81(6), pp.415–420. doi:https://doi.org/10.1093/jnci/81.6.415.

Parkin, D.M., Pisani, P., Lopez, A.D. and Masuyer, E. (1994). At least one in seven cases of cancer is caused by smoking. Global estimates for 1985. International Journal of Cancer, 59(4), pp.494–504. doi:https://doi.org/10.1002/ijc.2910590411.

Radzikowska, E., Głaz, P. and Roszkowski, K. (2002). Lung cancer in women: age, smoking, histology, performance status, stage, initial treatment and survival. Population-based study of 20 561 cases. Annals of Oncology, 13(7), pp.1087–1093. doi:https://doi.org/10.1093/annonc/mdf187.

Redondo-Sánchez, D., Petrova, D., Rodríguez-Barranco, M., Fernández-Navarro, P., Jiménez-Moleón, J.J. and Sánchez, M.-J. (2022). Socio-Economic Inequalities in Lung Cancer Outcomes: An Overview of Systematic Reviews. Cancers, [online] 14(2), p.398. doi:https://doi.org/10.3390/cancers14020398.

Roque, K., Nanto Caparachin, Castro-Mollo, M., Galvez-Nino, M., Ruiz, R., Coanqui, O., Valdivieso, N., Hurtado, M., Amorin, E., Ebert Poquioma, Cruzado, J. and Mas, L. (2023). Abstract C076: Income and incidence of lung cancer in the capital of an upper-middle country. Cancer Epidemiology Biomarkers & Prevention, 32(1_Supplement), pp.C076–C076. doi:https://doi.org/10.1158/1538-7755.disp22-c076.

Xing, X., Liao, Y., Tang, H., Chen, G., Ju, S. and You, L. (2011). [Gender-associated differences of

lung cancer and mechanism]. PubMed, 14(7), pp.625–30. doi:https://doi.org/10.3779/j.issn.1009-3419.2011.07.12.

Yin, X., Chan, C.P.Y., Seow, A., Yau, W.-P. and Seow, W.J. (2021). Association between family history and lung cancer risk among Chinese women in Singapore. Scientific Reports, 11(1). doi:https://doi.org/10.1038/s41598-021-00929-9.

Appendices

Appendix1:

Variable	Count	Proportion
Diagnosed with lung cancer	1085	0.05870894
Not diagnosed with lung cancer	17396	0.94129106

Appendix2:

Variable	Mean	Min.	Max.	Standard Deviation
Smoking duration	24.93583	0	76	14.99111

Appendix3:

Variable	Count	Proportion
Male	9238	0.4998647
Female	9243	0.5001353

Appendix4:

Code	Label	Count	Proportion
0	NIU	4954.0	0.2680590877
1	$1 to $4,999	1493.0	0.0807856718
2	$5,000 to $9,999	1371.0	0.0741842974
3	$10,000 to $14,999	1644.0	0.0889562253
4	$15,000 to $19,999	1430.0	0.0773767653
5	$20,000 to $24,999	1472.0	0.0796493696
10	$25,000 to $34,999	1754.0	0.0949082842
11	$25,000 to $29,999	286.0	0.0154753531
12	$30,000 to $34,999	251.0	0.0135815162
20	$35,000 to $44,999	1094.0	0.059195931
21	$35,000 to $39,999	227.0	0.0128288851
22	$40,000 to $44,999	179.0	0.0098658213
30	$45,000 to $54,999	661.0	0.0357664629
31	$45,000 to $49,999	121.0	0.0065472468
32	$50,000 to $54,999	109.0	0.0058979492
40	$55,000 to $64,999	356.0	0.0192630269
41	$55,000 to $59,999	64.0	0.003463011
42	$60,000 to $64,999	72.0	0.0038958931
50	$65,000 to $74,999	219.0	0.0118500081
51	$65,000 to $69,999	58.0	0.0031383583
52	$70,000 to $74,999	54.0	0.0029219198
60	$75,000 and over	402.0	0.0217520697
61	$75,000 to $79,999	38.0	0.0020561658
62	$80,000 to $84,999	32.0	0.001731508

63	$85,000 to $89,999	28.0	0.00151150695
64	$90,000 to $94,999	21.0	0.0011363021
65	$95,000 to $99,999	11.0	0.0005952059
66	$100,000 and over	N/A	N/A
67	$100,000-$104,999	19.0	0.0003246578
68	$105,000-$109,999	6.0	0.0029760294
69	$110,000-$114,999	55.0	0.0010280829
70	$115,000 and over	N/A	N/A

Appendix5:

Variable	**Count**	**Proportion**
Family member had lung cancer	5676	0.3071262
Family member did not have lung cancer	12805	0.6928738

Appendix6: Contribution

- Conceptualization & Research: Zengming An, Yuetian Ma, Pengxv Pan, Qianqian Cai, Moran
- Tian, Zihan Ma
- Data analysis: Yuetian Ma
- Interpretation of results: Zengming An
- Drafting report: Zengming An, Yuetian Ma, Pengxv Pan