Insurance Coverage and Lung Cancer Diagnosis in the United States of America — A Causal Inference Analysis

Afraa Noureen
Daniela Jiménez Lara
Revanth Chowdary Ganga
Luopeiwen (Tina) Yi

1. Executive Summary

Lung cancer ranks as the second leading cause of death in the United States, necessitating effective strategies for prevention and early detection. Access to low-dose CT screening is crucial, particularly among high-risk groups like smokers aged 50 to 80. However, disparities persist across demographics, with women, racial minorities, low-income individuals, and certain occupations facing elevated risks.

The Affordable Care Act (ACA) has significantly expanded healthcare coverage since 2010, leading to a decline in uninsured rates and an increase in cancer diagnoses among insured individuals. Despite progress, healthcare disparities remain, highlighting the need for targeted interventions to improve access to care and reduce inequities.

We aim to deliver insights on how healthcare coverage influences lung cancer outcomes, guiding the development of targeted healthcare interventions. To determine the impact of insurance coverage on lung cancer diagnosis rates, our study utilized a rigorous methodology that includes data segmentation of pre- and post-2010 to reflect changes from the Affordable Care Act, as well as a combined data from 2000 to 2018. We implemented matching techniques to pair insured individuals with similar uninsured counterparts, ensuring comparability. Our analysis employed linear probability regression with demographic controls, allowing us to isolate the specific influence of insurance on lung cancer rates and establish a causal relationship. This method provides a robust basis for informed healthcare policy decisions.

Our findings reveal a consistent positive impact of health insurance coverage on the likelihood of receiving lung cancer diagnosis across all time periods. Transitioning from no insurance to coverage is associated with a significant increase in diagnosis rates, highlighting the role of insurance in facilitating cancer detection and management. Furthermore, our investigation into the reasons for the lack of insurance coverage reveals cost as the primary barrier, with unemployment also contributing. Addressing these barriers through policy changes could expand coverage, enabling timely diagnosis of lung cancer. Our analysis offers guidance for policymakers to develop effective interventions aimed at enhancing screening efforts and reducing the burden of lung cancer nationwide.

2. Context

Lung cancer is the second leading cause of death in the United States of America and the most salient of the three lung diseases that exist within the top 10 causes of death in 2019[1]. This type of cancer has also experienced a 3.2% increase in death rates from 2009 to 2019[2].This persistence of high mortality rates in lung cancer emphasizes the need for comprehensive strategies in prevention, treatment, and early detection. For the latter, the CDC guideline for screening this disease is a low-dose computed tomography test (CT scan), done yearly for people with a history of smoking and between the ages of 50 and 80 years old[3].As detection requires this particular exam, having the appropriate healthcare coverage that guarantees this service is essential to combat the disease.

Lung cancer persists in a differentiated manner across demographics in the United States population. Using survey information from the National Health Service Survey (NHIS) we can analyze how lung cancer behaves differently among them and conclude that cancer incidence varies across race, gender, and occupations in the country. For instance, from 2000 to 2018, the proportion of the female population with lung cancer has been higher than that of men for most years (refer to Appendix D Figure. 8).

As we continue analyzing through the demographics, differences in race can also be found; where the proportion of the Asian population that has lung cancer has been lower than the proportions for White, Black and American Indians groups each (refer to Appendix D Figure 9). These three groups also show an increasing trend of lung cancer throughout the period but with different proportions for their populations. However, number wise the White population group has the highest number of lung cancer cases each year, as expected by the size of this racial group.

In addition to race differences, income level and occupation are associated with different rates of lung cancer in the population. From 2010 to 2019, although the military comprises one of the smallest occupational groups, it exhibits the highest rate of lung cancer diagnoses among all occupations (refer to Appendix D Figure 5 and 6). As we further analyze demographic differences, and focus on income level, it's notable how for the same period of time, the lowest income ($0 - $34,999) share of the population steadily maintains the highest proportion of lung cancer diagnosis (refer to Appendix D Figure 7).

Finally, as one of the main causes of lung cancer is smoking habits we can see the clear association of lung cancer and smoking habits. For the period of 2000 to 2018, consistently less

---

[1] Institute for Health Metrics, “United States of America Profile”
[2] Ibid.
[3] CDC. “Who should be screened for lung cancer”

than 10% of the people diagnosed with lung cancer indicated that they had never smoked (refer to Appendix D Figure 10).

Early detection of lung cancer, via CT screening, significantly improves its survival rate, as a recent study by Mount Sinai researchers demonstrates. Their results show a 10-year lung cancer–specific survival rate of 81% and a 20-year lung cancer–specific survival rate of 81% in patients who were diagnosed with lung cancer through early CT screening. Among patients who were diagnosed with stage I, the survival rate was 95%[4]. Given this information, a needed strategy to combat this cancer is making early screening available for people at risk. Although universal healthcare is not a reality in the country, currently Medicare and other insurance plans help pay for recommended CT scans[5].

The biggest effort to make healthcare available for all U.S. citizens was the Affordable Care Act (ACA) of 2010. This comprehensive reform law has three primary goals: 1). Make affordable health insurance available to more people by providing consumers with subsidies that lower costs for households with incomes between 100% and 400% of the federal poverty level (FPL). 2). Expand the Medicaid program to cover all adults with income below 138% of the FPL. 3). Support innovative medical care delivery methods designed to lower the costs of health care generally[6].

The ACA has resulted in a rising healthcare coverage noticeable in representative health surveys. Results show that healthcare coverage has increased yearly since 2010 reaching a rate of 90% coverage by 2018. Previously, from 2000 to 2009, the percentage of uninsured people maintained a steady rate of 17%, as shown in Figure 1.

---

[4] Lokaj, “Early CT Lung Cancer Screening Significantly Improves Survival in Lung Cancer”
[5] CDC, ‘Who should be screened for Lung Cancer’.
[6] HSS. “About the Affordable Care Act”.

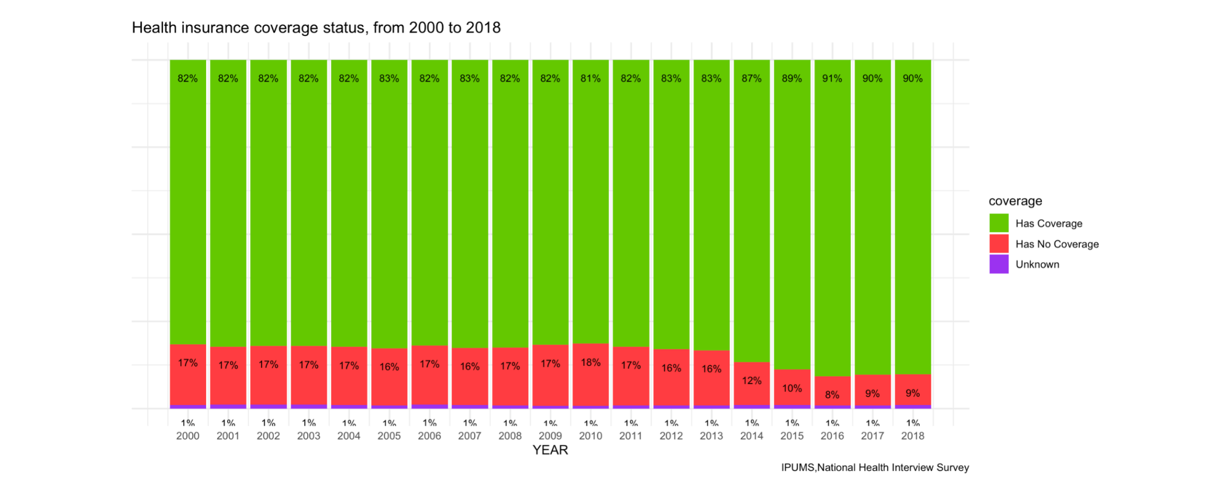

Figure 1. Proportion of insured and uninsured population per year from 2000 to 2010, where a yearly increase is notable after the implementation of ACA in 2010.

Parallel to the increase in healthcare coverage, cancer diagnosis has increased steadily for people with health insurance coverage. Cancer diagnosis for people without insurance has also increased but at a slower rate and smaller percentage change, as seen in Figure 2.

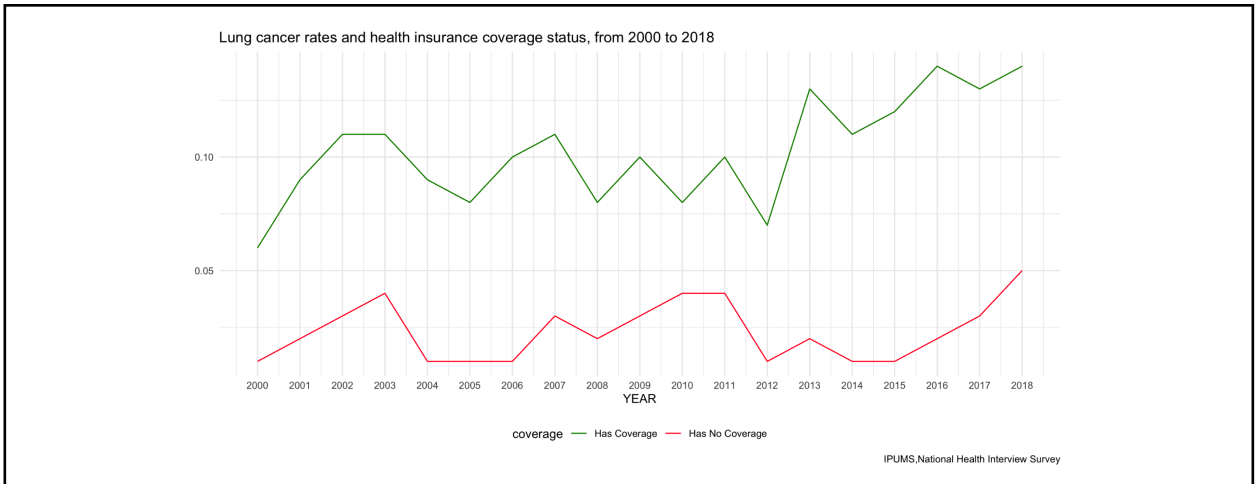

Figure 2. Share of lung cancer diagnosis from 2000 to 2010 for population with and without health insurance coverage

Understanding the association between insurance coverage and lung cancer diagnosis rates underscores the importance of healthcare access in detecting and managing cancer, highlighting the potential benefits of policies aimed at improving insurance coverage and reducing healthcare disparities.

3. Question

Does increasing the access to medical care, indicated by having insurance coverage, lead to higher rates of lung cancer diagnosis?

This is the central question driving our investigation. We aim to determine if individuals **with health insurance coverage** are **more likely** to receive a lung cancer diagnosis **compared to those without coverage**. This inquiry is essential for understanding the role of healthcare accessibility, particularly through insurance, in influencing lung cancer incidence rates.

4. Data

We utilized data sourced from the National Health Interview Survey (NHIS), a comprehensive repository of health-related information spanning over 50 years (1963-present). The NHIS serves as a critical resource for understanding various health indicators, including general health status, illness distribution, healthcare access, insurance coverage, and health behaviors like exercise, diet, and substance use. On average, the NHIS surveys approximately 100,000 individuals across 45,000 households annually[7].

The NHIS dataset used for our analysis provides a detailed overview of demographic factors and lung cancer status among the U.S. population from 2000 to 2018. We chose to analyze data up to 2018 due to the emergence of the COVID-19 pandemic, which significantly impacted the economy and healthcare systems. This decision helps avoid potential uncertainties in our findings' validity, further supported by the decline in data availability for insurance coverage status post-2019. It contains essential data such as age, gender, occupation, race, smoking status, family income, insurance coverage, and lung cancer diagnosis. Each entry in the dataset corresponds to a unique individual surveyed, enabling us to aggregate data across all years without duplications.

To account for the impact of the ACA enacted in 2010, we segmented the NHIS data into pre- and post-2010 periods. By considering the implications of the ACA, we aim to discern how shifts in healthcare policy may have influenced lung cancer diagnosis patterns.

We examined insurance coverage status and lung cancer diagnosis as key variables. We also took into account potential related factors such as occupation, race, gender, smoking status, and income level. By considering these variables, we aimed to isolate the specific effects of insurance coverage and lung cancer diagnosis. Our initial comparison of individuals with and without insurance coverage revealed significant disparities between these groups, as illustrated in Figure 3. below.

[7] IPUMS Health Surveys. “About Ipums”

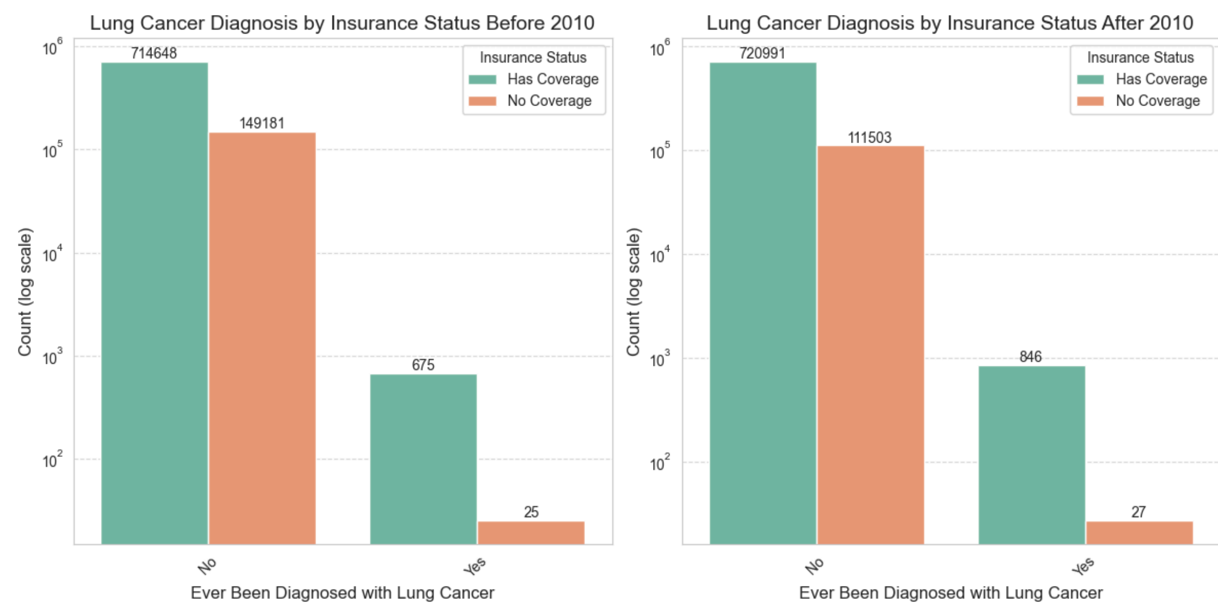

Figure 3. Distribution of lung cancer diagnosis cases by insurance status before and after 2010

There was a marked imbalance in the number of individuals with insurance coverage compared to those without. This imbalance can be likened to having an unequal representation of data, which may distort our analysis. Upon examining various demographic factors such as income levels, racial backgrounds, genders, occupations, and smoking habits, we observed wide variations in the proportion of individuals with insurance coverage across these categories (refer to Appendix F Table 8). These discrepancies, termed "imbalance," have the potential to skew our analysis and lead to erroneous conclusions. There's a risk of extrapolating relationships estimated from data of individuals without insurance coverage to estimate the characteristics of individuals with insurance coverage in areas where we lack actual data. Essentially, by attempting to fill in the gaps for the group of individuals with insurance coverage that we lack data on, we may arrive at misleading conclusions. To develop our analysis considering these caveats, we will be using a matching approach (refer to Methodology).

During data preprocessing, we refined the dataset to enhance its quality and relevance for modeling analyses. We chose to prioritize family income over individual earnings to better represent the financial aspect of healthcare access, considering the strong correlation between individual and family income within households (refer to Appendix C Table 5). Including both variables could introduce multicollinearity issues, affecting regression results. Prioritizing family income allowed us to capture a broader socioeconomic indicator while avoiding redundant information and model instability.

To ensure the reliability of our subsequent regression analysis, we focused on data purity by excluding variables with significant missing values, such as age at first cancer diagnosis (refer to Appendix C Table 4), instead of resorting to imputation. Imputation for variables with high missing values would have introduced artificial data, potentially compromising the representativeness of the dataset.

A significant portion, approximately 75%, of our dataset fell under the category of out-of-scope data for the individuals surveyed (Refer to Appendix E Table 6). Analysis revealed that this unknown data was not randomly distributed (Refer to Appendix E Table 7), challenging the dataset's representativeness. Instead of discarding this data, we opted to retain it. Omitting these values based on our analysis would result in a considerable loss of data and could potentially skew our results, given their substantial presence. By preserving this data and including it in our analysis, we aimed to ensure a thorough examination of the dataset, considering all available information.

5. Design

5.1 Methodology

Our methodology for examining the relationship between insurance coverage and lung cancer diagnosis rates was carefully selected to ensure meaningful results. Initially, we divided our dataset into two distinct time periods: pre-2010 and post-2010, which was essential for us to analyze any shifts or patterns in lung cancer diagnosis rates following the implementation of the ACA in 2010. We also conducted an analysis spanning from 2000 to 2018 to evaluate the long-term impact of health insurance coverage. This broader scope provided insights into long-term patterns of insurance impact on lung cancer diagnoses.

To address the inherent imbalance in our data (refer to appendix F, Figure 11) for each time period in our analyses, we use matching, a common technique in observational studies. Matching serves to create a more balanced comparison between individuals with and without insurance coverage. Our goal is to pair individuals from the insurance group with similar individuals from the non-insurance group based on observable characteristics such as income, race, gender, occupation, and smoking habits. This ensures that each individual with insurance coverage has a counterpart without insurance who shares similar traits. Through this approach, we establish a more balanced dataset, enabling us to accurately assess the impact of insurance coverage on lung cancer diagnosis rates.

Following this, we conducted linear probability regression analyses with controls - incorporating factors such as gender, occupation, race, smoking status, and income level - to explore the relationship between insurance coverage and lung cancer diagnosis rates across our different time-based analyses. Using this regression with controls allowed us to evaluate the likelihood of a lung cancer diagnosis while accounting for potential confounding variables. We aimed to isolate the specific impact of insurance coverage on lung cancer diagnosis rates, free from the influence of other factors.

We also explored the factors contributing to individuals' lack of insurance coverage, such as financial constraints, unemployment, and changes in coverage status. This investigation was crucial as it provided insight into the fundamental barriers to healthcare access. Understanding these factors could guide targeted policy interventions aimed at enhancing overall healthcare accessibility for vulnerable populations along with providing visibility on the potential impact the change in policy may have on insurance coverage.

5.2 Causal Inference

Causal inference is essential to our analysis, ensuring that our findings are grounded in evidence-based research rather than superficial correlations. It's essential to recognize that correlation does not always imply causation.

Consider a scenario where we may have observed a correlation between regions with higher insurance coverage and increased rates of lung cancer diagnoses. While this correlation may suggest a potential relationship between insurance coverage and diagnosis rates, it doesn't provide evidence of causation, as there may be other factors at play. Regions with higher insurance coverage may also have better access to healthcare facilities, leading to more frequent cancer screenings and diagnoses. Conversely, regions with lower insurance coverage may face barriers to accessing healthcare services, resulting in underdiagnosis or delayed diagnosis of lung cancer. In this case, the correlation between insurance coverage and diagnosis rates may be influenced by factors such as healthcare infrastructure, socioeconomic status, and healthcare-seeking behavior, rather than a direct causal relationship between insurance coverage and lung cancer diagnosis.

In our study, we go beyond simple correlations to explore causation. We carefully account for confounding variables such as demographics and socioeconomic status. This detailed approach helps us isolate the influence of insurance coverage on lung cancer diagnosis rates, allowing us to draw meaningful conclusions about causality rather than solely relying on observed associations.

Understanding the disparity between correlation and causation, we provide policymakers with practical insights derived from rigorous causal inference techniques. Our goal is to support informed decision-making in healthcare policy and practice, ultimately driving enhancements in healthcare accessibility and outcomes.

5.3 Modeling

5.3.1 Matching

In our analysis, we employ a matching technique to refine our dataset, addressing differences between individuals with and without insurance coverage (as mentioned in our methodology section). This step is crucial to ensure that our analysis accurately captures the relationship between insurance coverage and lung cancer diagnosis rates.

The matching technique is implemented across three key periods: 2000-2018, before 2010, and after 2010. The detailed matching process can be referred to in the Appendix G. Once the matching process is complete, the resultant data is balanced with respect to the variables considered. This matched data is then used for subsequent regression analysis, aiming to estimate the causal impact of variable health insurance coverage on lung cancer diagnosis rates. This methodical approach ensures that our dataset is optimally prepared for accurate causal inference, minimizing biases inherent in observational studies and enhancing the reliability of our conclusions.

5.3.2 Linear Probability Model

A regression is necessary in this context due to several confounding features that likely obscure the true relationship between insurance coverage and lung cancer diagnosis rates. For instance, individuals in certain occupations often receive health insurance coverage as part of their employment benefits. This coverage may encourage them to engage in preventive health behaviors and undergo regular screenings, thereby increasing their chances of early diagnosis. Conversely, individuals in different occupations may not have access to such benefits, potentially reducing their engagement in preventive healthcare measures. Therefore, insurance coverage might not only reflect accessible healthcare but also working benefits from different occupations. To accurately extract this correlation and assess the impact of insurance coverage independently from other confounding variables such as income, race, gender, smoking status, and occupation, we have adopted a linear probability model.

Following the meticulous matching process, we proceed with a weighted least squares (WLS) regression model, a type of Linear Probability Model (LPM), using the matched dataset to explore the causal relationship between insurance coverage and lung cancer diagnosis. LPM is ideally suited for analyzing binary outcome variables, such as lung cancer diagnosis status, because it models the linear relationship between predictors, such as insurance coverage, and the probability of the outcome. To ensure our results are dependable, especially when there's uneven variability in our data, we use a technique called robust standard errors. These adjustments help us get more accurate estimates, which are important for making reliable decisions based on the data.

This model is particularly effective for capturing the potential impact of health insurance coverage on lung cancer diagnosis, controlling for a set of confounding characteristics. By controlling the impact of those variables on lung cancer diagnosis, the model allows us to more accurately isolate and identify the independent effect of insurance coverage on the likelihood of lung cancer diagnoses. This rigorous approach ensures that any observed effect can be attributed more confidently to changes in insurance status rather than other demographic factors, aiding in the formulation of evidence-based healthcare policies.

6. Results

	2000-2009	**2010-2018**	**2000-2018**
Intercept	-0.0005	-0.0007	-0.0006
Insurance Coverage	0.0008***	0.0010***	0.0009***
	($5.6e^{-5}$)	($6.2e^{-5}$)	($4.0e^{-5}$)
R^2	0.005	0.006	0.005
Adjusted R^2	0.005	0.006	0.005
No. of observations	856283	827558	1690664

Robust Standard errors given in parentheses.

***p < 0.01; **p < 0.05; *p < 0.1.

Table 1. Regression Results: Impact of Insurance Coverage on Lung Cancer Diagnosis Rates Across Three Time Periods

In the pre-2010 regression analysis, we observe a positive coefficient of 0.0008 for insurance coverage with controls. This implies that transitioning from no insurance coverage to having insurance coverage is associated with a small but significant increase of 0.08 percentage points, considering the incidence rate of lung cancer in our dataset is ~0.001% (refer to Appendix H). in the probability of being diagnosed with lung cancer. The p-value of less than 0.01 indicates that this relationship is statistically significant. The positive correlation between insurance coverage and lung cancer diagnoses may stem from detection bias, where individuals with insurance are more likely to pursue screening tests, leading to increased diagnoses.

Upon examining the post-2010 regression analysis, we find a slight increase in the coefficient for insurance coverage with controls, rising to 0.0010. This suggests that the relationship between insurance coverage and lung cancer diagnosis rates may have strengthened following the implementation of the Affordable Care Act. The statistically significant p-value of less than 0.01 reinforces the robustness of this association.

The ACA played a substantial role in reducing the number of uninsured Americans by expanding Medicaid coverage and establishing health insurance marketplaces, leading to increased access to healthcare services. The slight but significant increase in the likelihood of being diagnosed with lung cancer among insured individuals post-2010 suggests the ACA mandated coverage of preventive services, including cancer screenings, likely contributed to the observed increase in diagnoses[8] [9].

The consistent positive coefficients observed across all periods, including the combined data from 2000 to 2018, further emphasizes the relationship between insurance coverage and increased likelihood of lung cancer diagnosis. The low p-value of less than 0.01 indicates the continued statistical significance of this association.

After conducting regression analysis, which revealed a significant association between insurance coverage and the diagnosis of lung cancer, we proceeded to investigate the underlying reasons for the lack of insurance coverage among individuals. It is important to determine if factors beyond financial constraints contribute to the absence of insurance coverage, as even policy implementation may not alter lung cancer diagnosis rates if uptake remains low.

---

[8] Sabik, et al., ‘The ACA and Cancer Screening and Diagnosis’.

[9] Kominski, el al., ‘The Affordable Care Act’s Impacts on Access to Insurance and Health Care for Low-Income Populations’

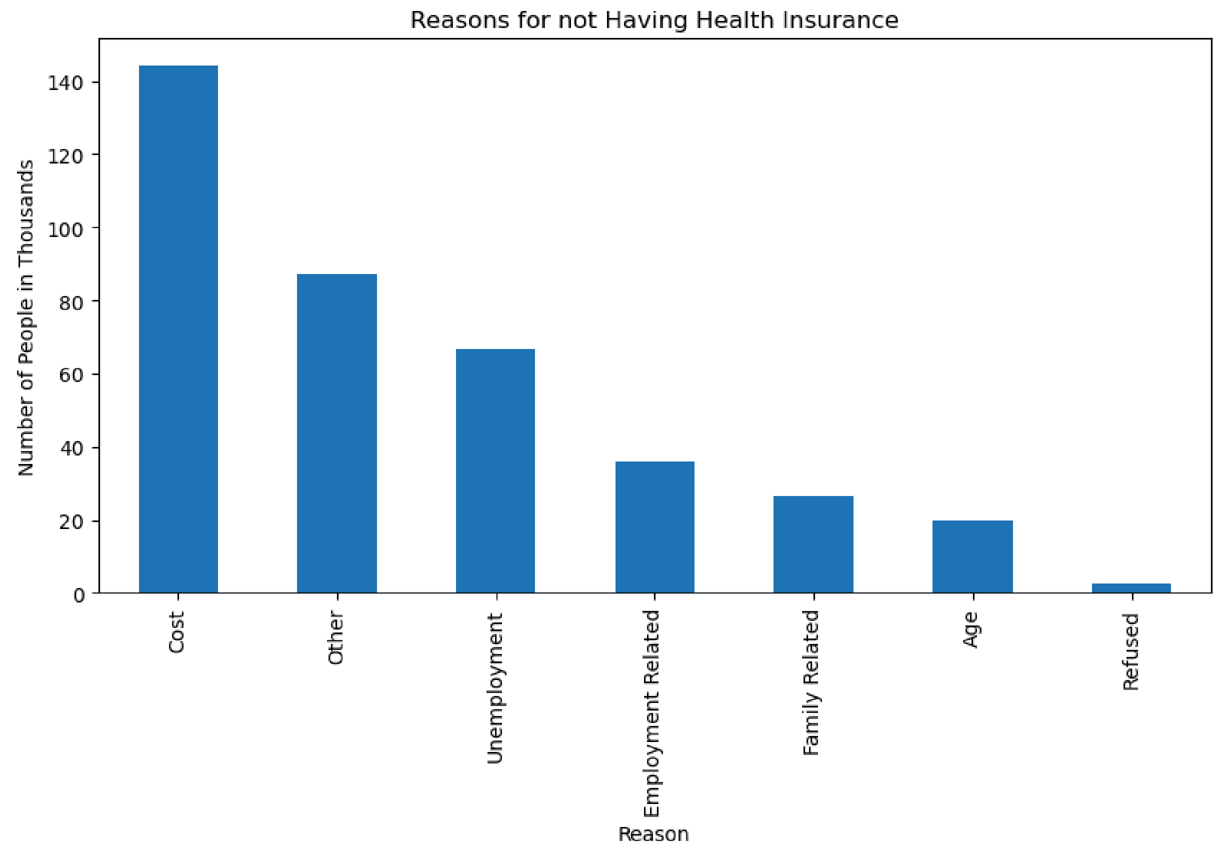

Figure 4 : Reasons for not having insurance coverage, the most predominant reason is the cost of insurance, the second most explicitly mentioned reason is "unemployment" which can also be closely related to cost and affordability of Insurance policies.

Our analysis focused on individuals who explicitly stated their lack of insurance coverage is illustrated in Figure 4. As depicted above, the predominant reason cited for this absence was related to the high cost of insurance. Additionally, the mention of 'unemployment' as a contributing factor can also be linked to financial constraints, given that unemployment often correlates with the inability to afford insurance premiums.

Based on our findings, implementing policy changes could potentially expand insurance coverage among the populace. This, in turn, may facilitate early diagnosis of lung cancer and potentially mitigate other health conditions.

7. Conclusion

Our analysis suggests that sustained efforts to improve and expand health insurance coverage could be key to lung cancer detection and management of health conditions nationwide, offering a significant payoff in terms of public health and well-being. The findings indicate a consistent positive impact of insurance coverage on the diagnosis of lung cancer.

Given the linkage between insurance coverage and increased lung cancer diagnosis rates, there might be a need for targeted initiatives that focus on enhancing access to cancer screening and detection services, particularly among populations that were previously uninsured or more prone to lung cancer. As early diagnosis of lung cancer leads to higher survival rates, these policies could also have an effect on the mortality rates associated with this disease.

Our dataset primarily explores demographic differences in access to health services broadly, rather than focusing exclusively on lung cancer. Consequently, it contains relatively fewer observations of individuals with lung cancer. However, since our survey is representative of the U.S. population, the proportions observed accurately reflect those in the general population. We are therefore confident that, despite the small number of cases, our results genuinely represent the impact of increased health insurance coverage on lung cancer diagnoses. Thus, an observed increase in diagnoses likely indicates improved access to healthcare rather than a deterioration in lung health.

To build upon our current work, in the future, further analysis may be performed on targeted survey data to study the impact of insurance coverage on the diagnosis of lung and other cancers in general to address the data imbalance and low observation constraints. Additional areas of focus could be on the current healthcare infrastructure to see if it can handle the additional influx of patients and if they are well equipped for diagnosing the various diseases.

8. Appendix

Appendix A. Data Cleaning Process

The data cleaning process aimed to enhance the dataset's quality and usability for subsequent analyses. Various steps were undertaken to address missing values, standardize variable representations, and ensure data consistency.

1. Initial Data Cleaning:
- Columns with over 80% missing values were dropped and not imputed to prevent artificial or unreliable data.
- Unnecessary columns for analysis were removed.
- Entries from years beyond the scope of our analysis, 2000-2018 were excluded, resulting in a dataset size of 1,762,659 observations.

2. Standardization and Recategorization:
- Smoking habits were standardized into binary categories of "Yes" and "No," simplifying interpretation and analysis.
- Occupation categories outside the intended scope or flagged as unknown were consolidated into a single category, aiding interpretation while retaining valuable data.

3. Detailed Data Cleaning Steps:
1. Insurance Coverage: Entries with '9.0' indicating "unknown" insurance coverage were removed. Categories '1.0' and '2.0' denoted insured and uninsured individuals, respectively.
2. Lung Cancer Status: Unknown statuses ('9.0' and '7.0') and irrelevant categories ('1.0') were dropped. '0.0' was recoded to '1.0' to represent non-diagnosis of lung cancer.
3. Lung Cancer Age: Categories '96,' '97,' and '99' were merged into '99' to represent non-applicable/unknown ages.
4. Family Income: Values '97,' '98,' and '99' were recoded to '0' for unspecified/unknown income, simplifying income brackets.
5. Smoking Status: Recategorized into numerical codes: 'Yes' to '1,' 'No' to '2,' and 'Other' to '0.'
6. Occupation: Remapped occupation codes based on the mapping dictionary, which varied across different years.
7. Race, Individual Income, and Gender: Preserved original categorizations for detailed demographic analysis.
8. Reason for no insurance variables: For variables indicating individuals who lacked insurance coverage, we specifically focused on responses indicating the reason for the lack of coverage. We excluded categories with a significant number of missing values, grouping them under the label "Other." Explicit "yes" responses were tallied under relevant categories, while all other

responses were treated as "no." Instances where individuals provided multiple "yes" responses were left unprocessed to maintain the integrity of the data.

4. Further Processing Decisions:

- Dropped variable related to age first being diagnosed with lung cancer due to high number of non-applicable/unknown values.(refer Appendix Section C, Table 4. for further details on age first being diagnosed with lung cancer)
- Evaluated correlation between family income and individual income, opting to retain family income due to fewer unspecified/unknown values (refer Appendix Section C, Table 5. for further details on the test for correlation)

Appendix B. Dataset Documentation

This section contains detailed dataset documentation for all the variables used in our regression and subsequent analyses.

Variable	Description	Values
SEX	Gender of the individual.	1: Male, 2: Female
RACENEW	Race/Ethnicity of the individual.	100: White Only, 200: Black/African American only, 300: American Indian/Alaska Native only, 400: Asian only, 520: Other Race, 530: Race Group not Releasable, 541: Multiple Race (1999-2018: Including American Indian/Alaska Native)
INCFAM97ON2	Family Income Bracket.	10.0: $0 - $34,999, 20.0: $35,000-$74,999, 30.0: $75,000 and over, 31.0: $75,000-$99,999, 32.0: $100,000 and over, 96.0: $20,000 or more (no detail), 0.0: NIU/Unknown
HINOTCOVE	Insurance Coverage Status.	1.0: has coverage, 2.0: has no coverage (2.0 is later recoded to be 0.0 for matching using de_flame)
CNLUNG	Lung Cancer Diagnosis Status.	1.0: Not diagnosed (mentioned) lung cancer, 2.0: Diagnosed (mentioned) lung cancer
SMK	Smoking Status.	1: Yes, 2: No, 0: Others
Occupation_Code	Occupation Category.	0: NIU/Unknown, 1: Office and Administrative Support Occupations, 2: Management Occupations, 3: Sales and Related Occupations, 4: Production Occupations, 5: Education, Training, and Library Occupations, 6: Transportation and Material Moving Occupations, 7: Healthcare Practitioners and Technical Occupations, 8: Business and Financial Operations Occupations, 9: Construction and Extraction Occupations, 10: Food Preparation and Serving Related Occupations, 11: Building and Grounds Cleaning and Maintenance Occupations, 12: Installation, Maintenance, and Repair Occupations, 13: Personal Care and Service Occupations, 14: Healthcare Support Occupations, 15: Architecture and Engineering Occupations, 16: Community and Social Service Occupations, 17: Protective Service Occupations, 18: Arts, Design, Entertainment, Sports, and Media Occupations, 19: Professional Specialty Occupations, 20: Administrative Support Occupations, Including Clerical, 21: Computer and Mathematical Occupations, 22: Executive, Administrative, and Managerial Occupations, 23: Legal Occupations, 24: Military, 25: Service Occupations, 26: Life, Physical, and Social Science Occupations, 27: Sales Occupations, 28: Farming, Fishing, and Forestry Occupations, 29: Precision Production, Craft, and Repair Occupations, 30: Operators, Fabricators, and Laborers, 31: Technicians and Related Support Occupations, 32: Farming, Forestry, and Fishing Occupations, 33: Handlers, Equipment Cleaners, Helpers, and Laborers

Table 2. Documentation for demographics variables, insurance coverage status, and lung cancer diagnosis status

Variable	Description	Values
HINOUNEMPR	Reason for No Insurance Coverage: Unemployment.	Binary: 1 for present, 0 for absent
HINOEMPR	Reason for No Insurance Coverage: Employment-related reason.	Binary: 1 for present, 0 for absent
HINOFAMR	Reason for No Insurance Coverage: Family-related.	Binary: 1 for present, 0 for absent
HINOAGER	Reason for No Insurance Coverage: Aged out of family plan.	Binary: 1 for present, 0 for absent
HINOCOSTR	Reason for No Insurance Coverage: Too expensive.	Binary: 1 for present, 0 for absent
HINOREFUSER	Reason for No Insurance Coverage: Poor health/refused coverage.	Binary: 1 for present, 0 for absent
OTHER	Reason for No Insurance Coverage: None of the above reasons/other.	Binary: 1 for present, 0 for absent

Table 3. Documentation for variables related to reasons for no insurance coverage

Appendix C. Dropped Variables

Age	Percentage
NIU/Unknown	99.908
65.0	0.004
70.0	0.004
66.0	0.004
60.0	0.004
64.0	0.003
62.0	0.003
69.0	0.003
67.0	0.003
73.0	0.003
68.0	0.003
72.0	0.003
74.0	0.003
55.0	0.003
56.0	0.003
59.0	0.002
75.0	0.002
50.0	0.002
71.0	0.002
58.0	0.002
63.0	0.002
77.0	0.002
85.0	0.002
57.0	0.002
79.0	0.002
54.0	0.002
78.0	0.002
51.0	0.002
61.0	0.002
76.0	0.001
52.0	0.001
80.0	0.001
53.0	0.001
81.0	0.001
48.0	0.001
49.0	0.001
84.0	0.001
83.0	0.001
44.0	0.001
47.0	0.001
43.0	0.001
45.0	0.001
40.0	0.001
35.0	0.001
82.0	0.0
42.0	0.0
18.0	0.0
39.0	0.0
2.0	0.0
31.0	0.0
46.0	0.0
41.0	0.0
3.0	0.0
30.0	0.0
38.0	0
37.0	0
28.0	0
32.0	0
27.0	0
4.0	0
1.0	0
36.0	0
34.0	0
23.0	0
25.0	0
22.0	0
17.0	0
10.0	0
29.0	0
14.0	0
15.0	0
26.0	0
7.0	0
6.0	0
12.0	0
16.0	0
24.0	0
13.0	0
20.0	0
33.0	0
21.0	0
9.0	0
5.0	0

Table 4. Percentage distribution of age at lung cancer diagnosis

This table displays the percentage distribution of age at which individuals were first diagnosed with lung cancer. Due to the high proportion of NIU/Unknown data, this field was not considered in our analysis.

Statistical Measure	Value	Interpretation
Cramer's V	0.30882381772887607	Indicates a moderate association between the variables.
Theil's U	0.9999999999999999	Suggests that one variable almost fully explains the other.

Table 5. Correlation check between family income and individual income

We conducted a correlation check between family income and individual income using Cramer's V and Theil's U statistics. Both indicated a moderate to strong redundancy, with Theil's U nearing 1, suggesting that one variable nearly fully explains the other. Consequently, given the fewer NIU/Unknown values in the family income, we retained the family income variable and dropped the individual income variable for a cleaner and more focused dataset.

Appendix D. Lung Cancer Diagnosis based on Demographics

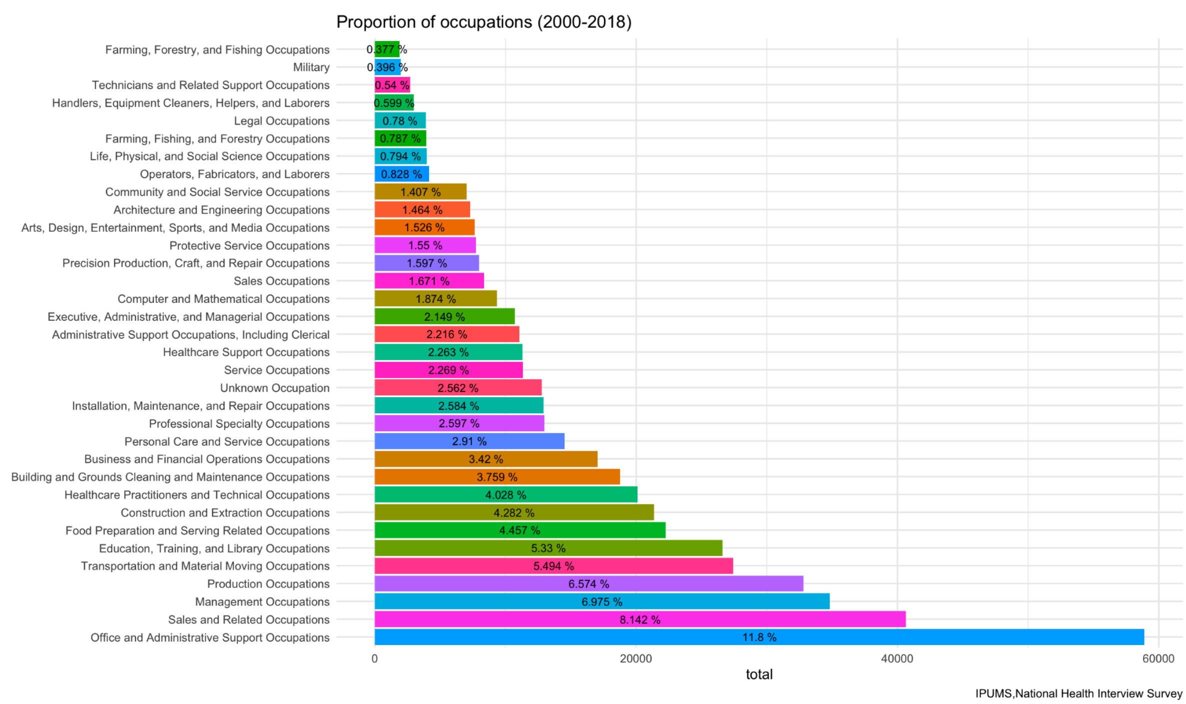

Figure 5. Percentage of occupations in the U.S. in the period from 2000 to 2018, where office and administrative support occupations maintain the highest value

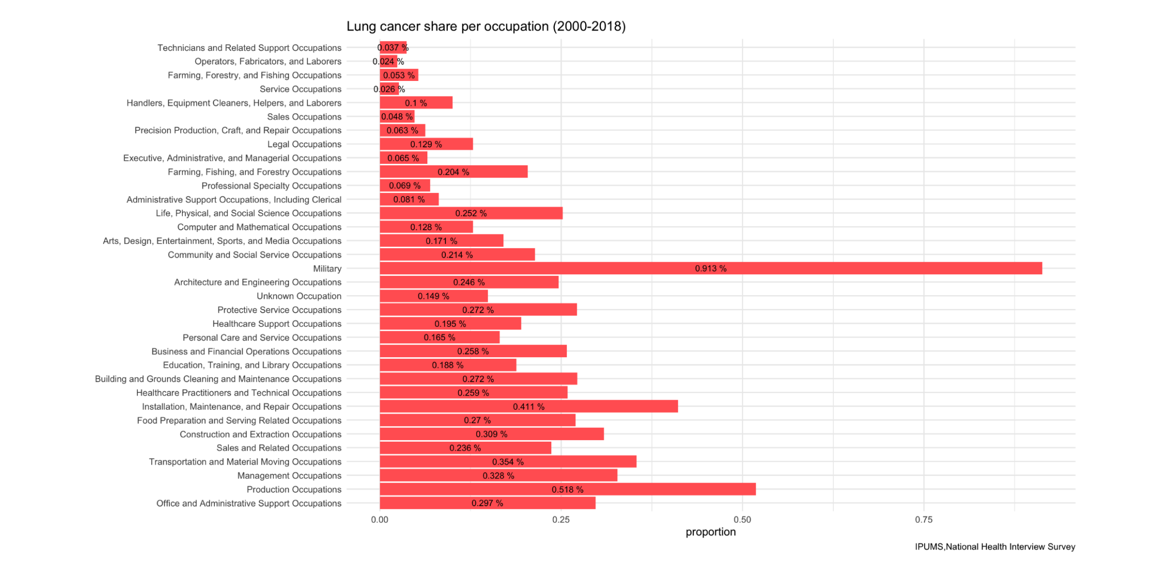

Figure 6. Lung cancer percentage per occupation in the period from 2000 to 2018, where the military holds the highest proportion

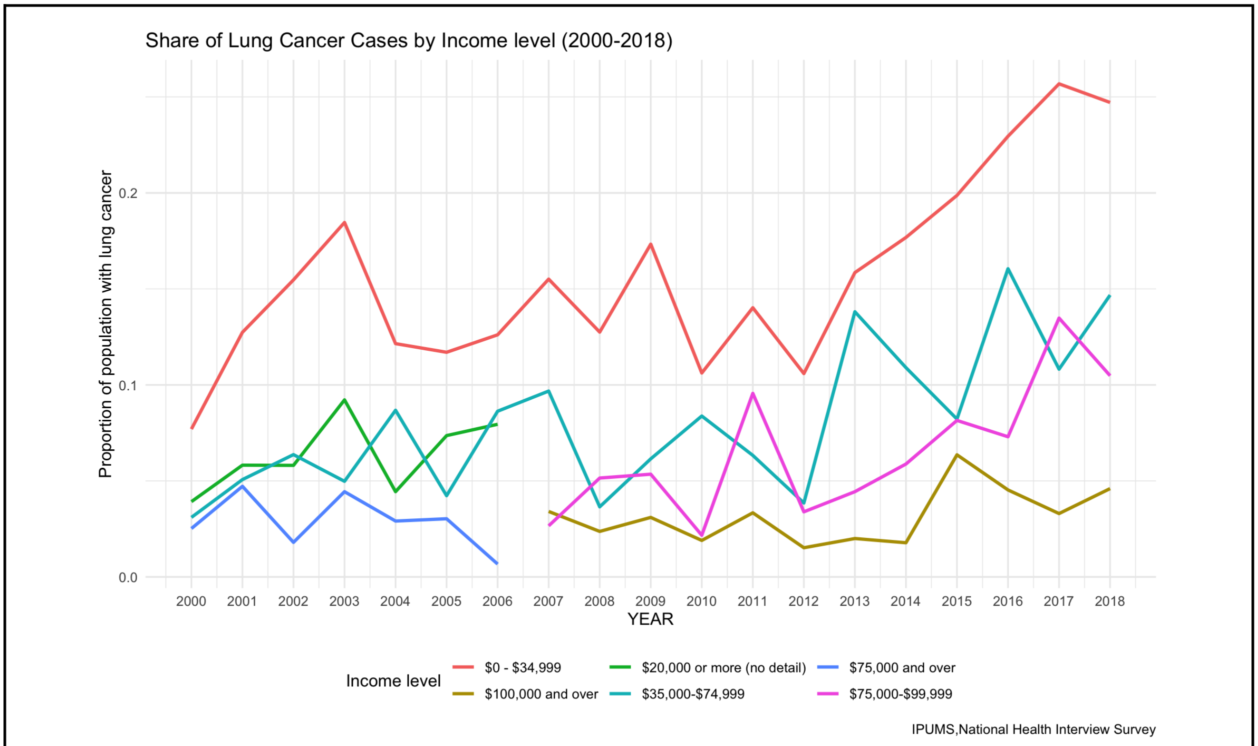

Figure 7. Lung cancer rate per income bracket population from 2000 to 2018. The lowest income level shows the highest proportion of lung cancer diagnosis

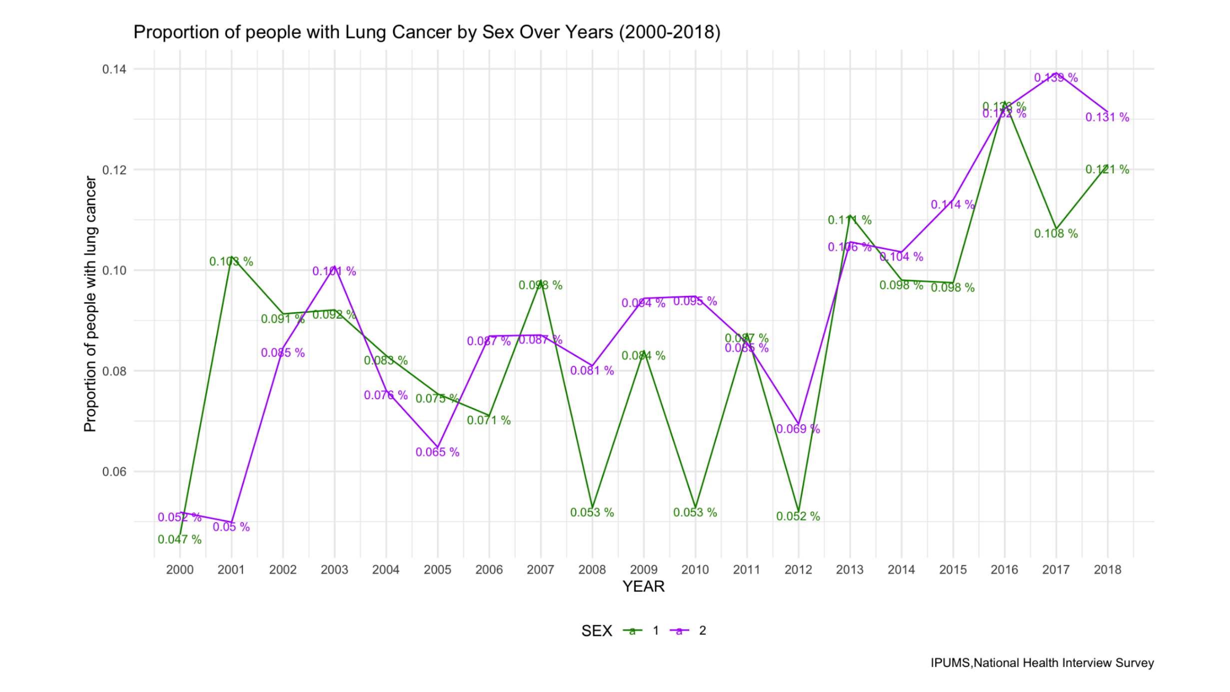

Figure 8. Share of people with lung cancer diagnosed by sex over years (2000-2018), where women hold a higher proportion of lung cancer over the years

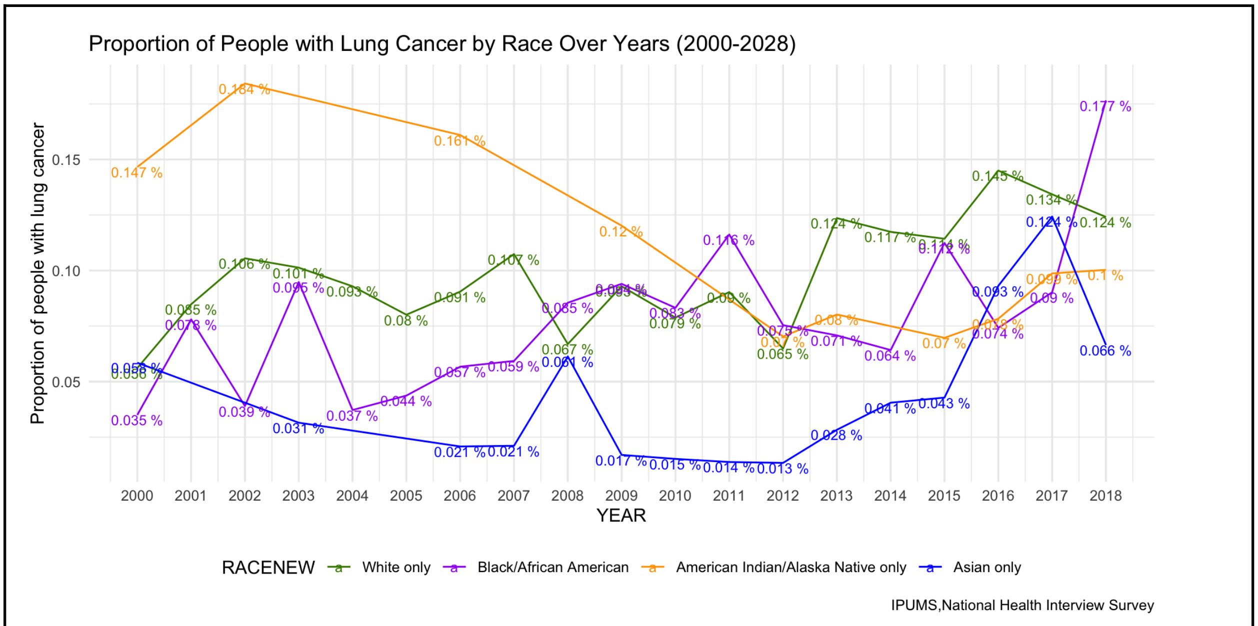

Figure 9. Percentage of people with lung cancer by race over years (2000-2018). The Asian racial group consistently presents the lower rates of lung cancer.

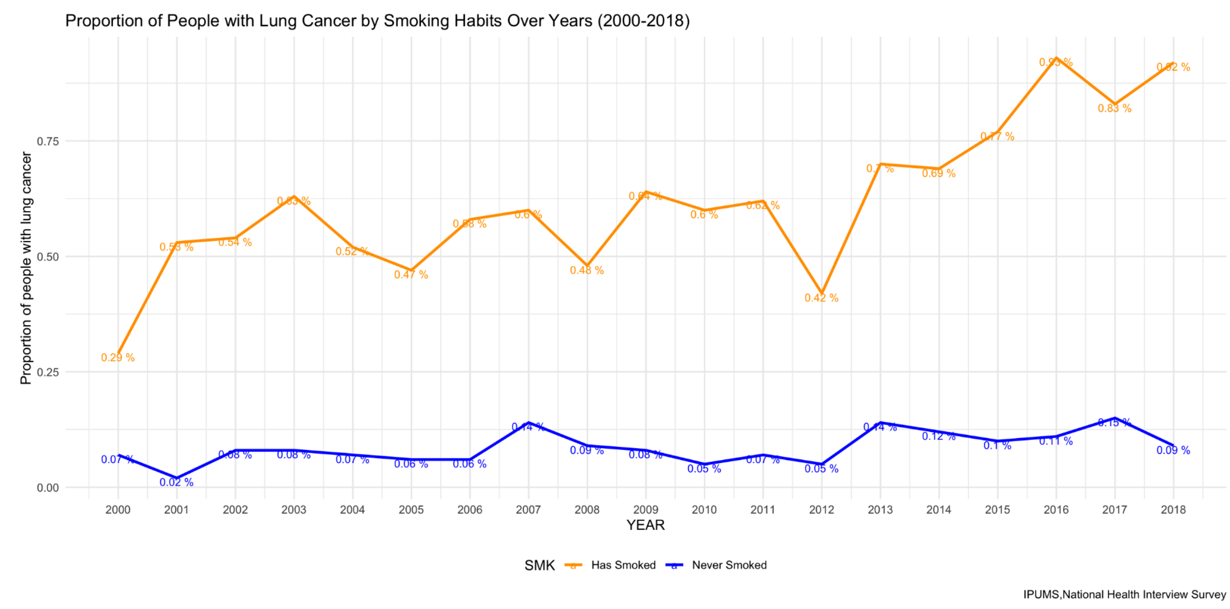

Figure 10. Share of people with lung cancer diagnosis by smoking habits over years (2000-2018)

Appendix E. Not-In-Universe/Unknown Data

Variable	Proportion of NIU/Unknown Values
Race	0.023984
Income	0.103091
Smoking	0.694351
Occupation	0.737989
Any Null	0.758738
Any Null (Excluding Occupation)	0.719782
Any Null (Excluding Occupation and Income)	0.699823

Table 6. Proportion of missing values in different variables

This figure displays the proportion of missing values for variables we consider in our dataset. Variables include race, income, smoking status, occupation, and the presence of any null values. The last two columns represent the proportion of missing values excluding the occupation variable, and excluding both occupation and income variables, respectively

Insurance Coverage Status	Lung Cancer Diagnosis Status	Proportion of records dropped after omitting NIU/Unknown Values
Insured	No Lung Cancer	0.761727
Uninsured	No Lung Cancer	0.745082
Insured	Lung Cancer	0.297173
Uninsured	Lung Cancer	0.192308

Table 7. Proportion of records dropped after excluding individuals with not-in-universe/unknown values for each group

This table illustrates the proportion of records dropped from the dataset following the exclusion of individuals with any NIU/Unknown values for any of the demographic factors within each group, categorized by lung cancer diagnosis status and insurance coverage status. The depicted proportion reflects the data loss incurred after eliminating individuals with NIU/Unknown insurance coverage status from each respective group.

Appendix F. Insurance Coverage Related Analysis

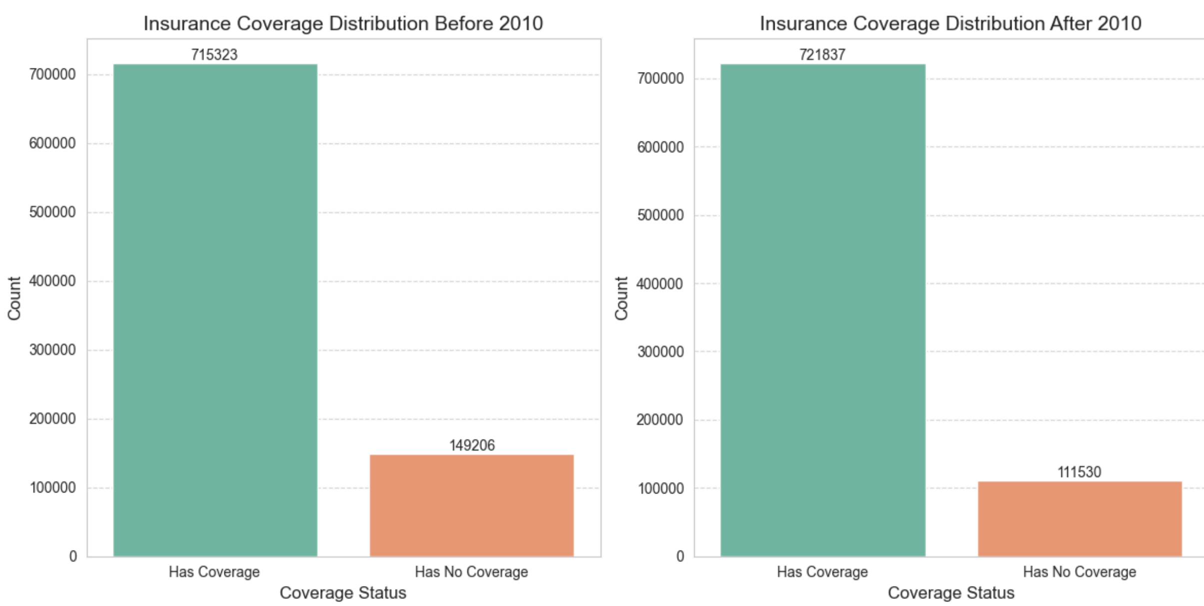

Figure 11. Insurance coverage distributions for the periods before and after ACA implementation. For both periods the number of observations with health coverage exceed no health coverage.

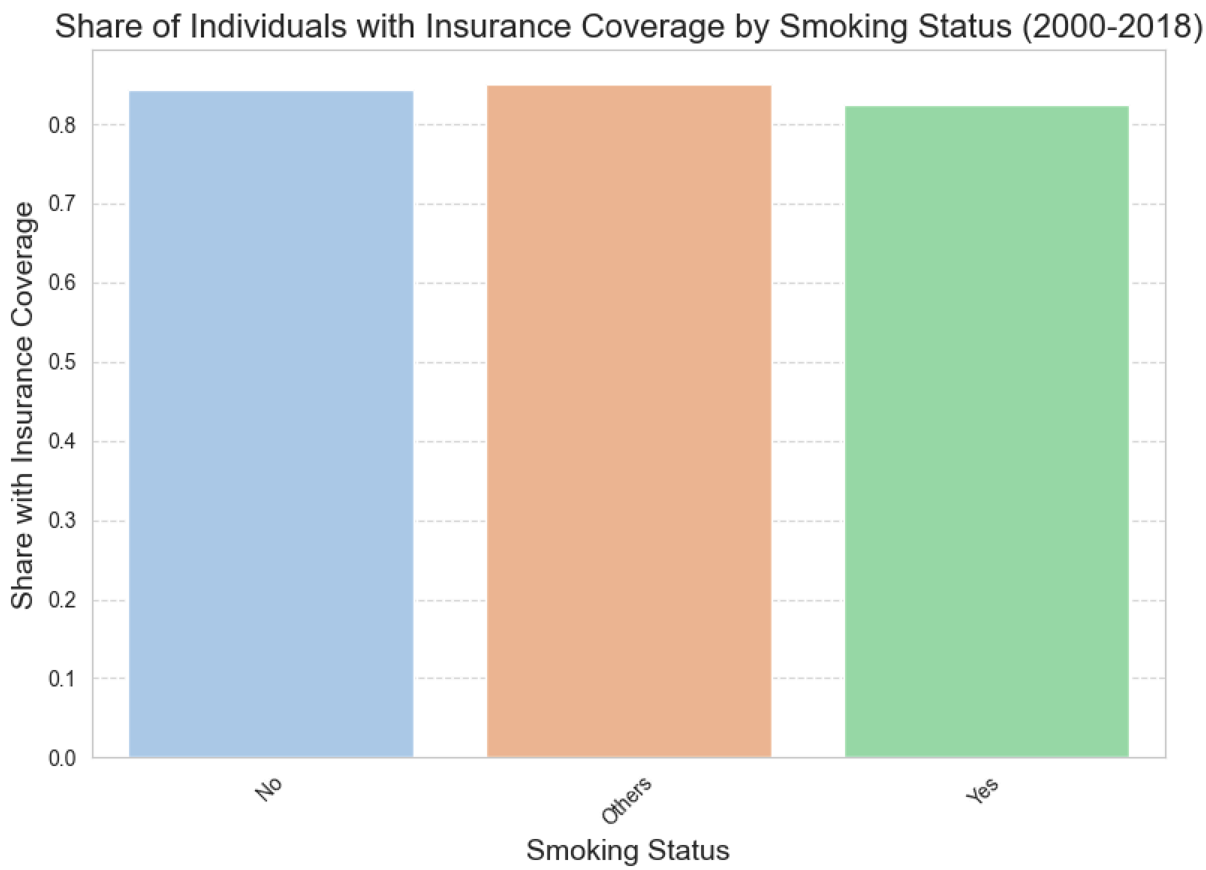

Figure 12. Share of individuals with health insurance coverage by smoking status

The figure shows the distribution of individuals based on their health insurance status, categorized by smoking status: 'Yes', 'No', and 'Other'. The graph presents the proportion of individuals within each smoking category who have health insurance, offering insights into the relationship between smoking habits and access to health coverage.

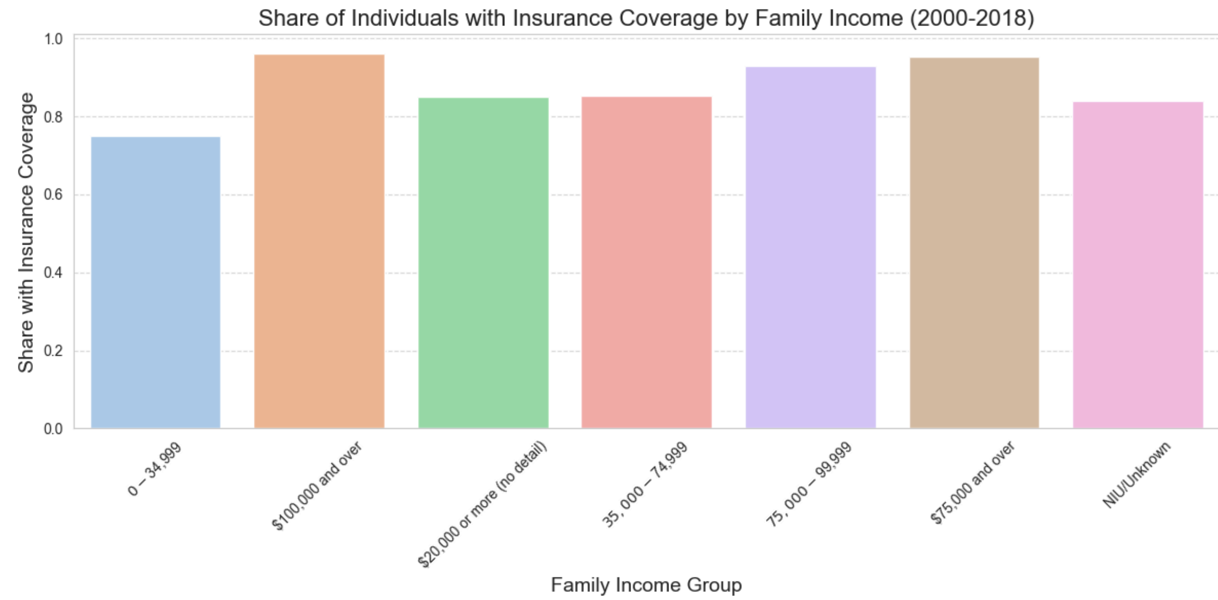

Figure 13. Proportion of people with insurance coverage based on family income groups, as observed, the relatively poor people with family income less than $35,000 have the least proportion of insurance coverage as compared to other groups.

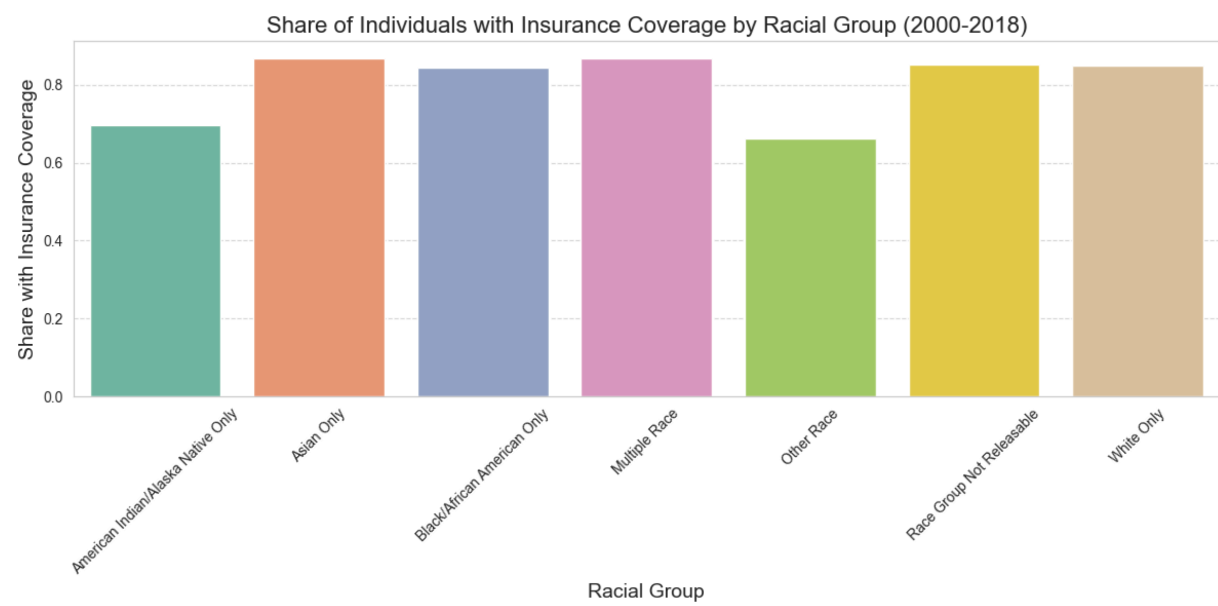

Figure 14. Proportion of people with insurance coverage based on racial groups. Out of the groups which are explicitly specified, American Indians have the lowest levels of insurance coverage as compared to other groups.

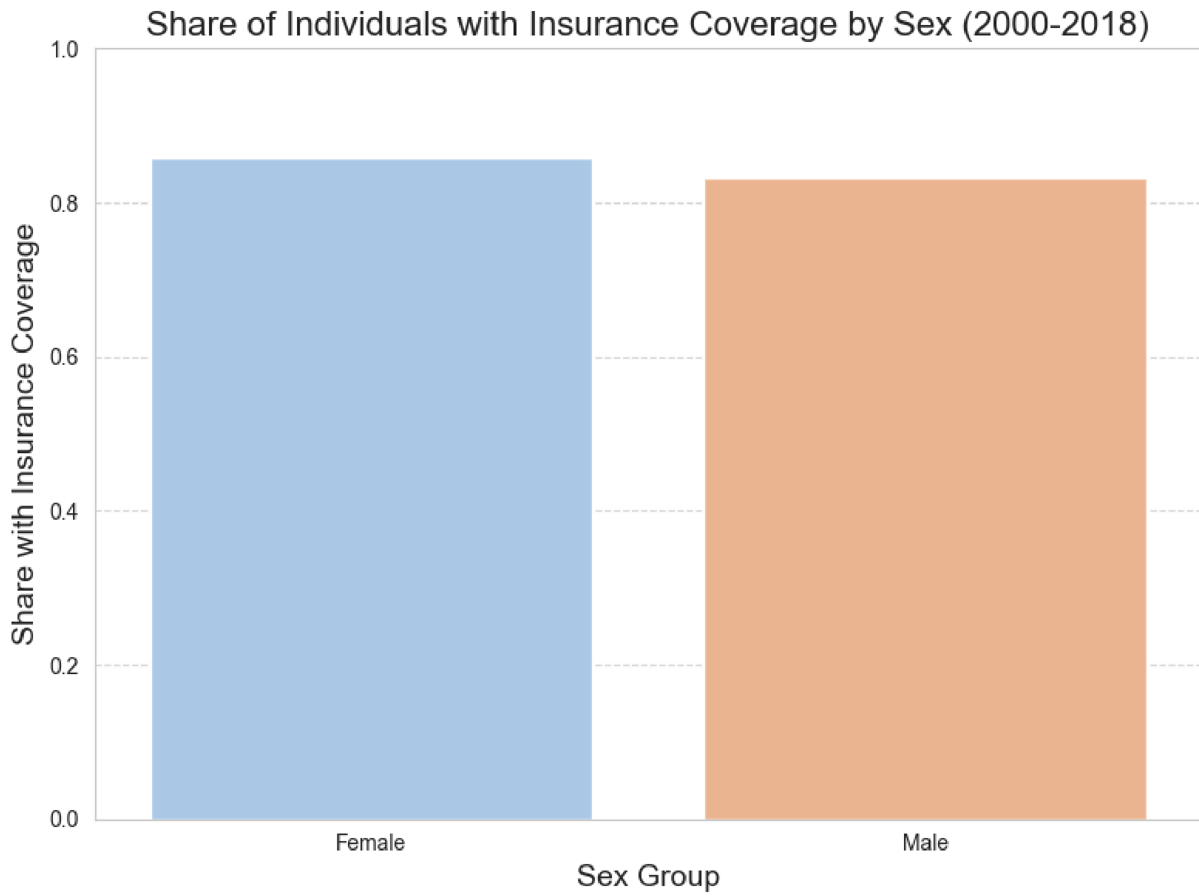

Figure 15. Proportion of people with insurance coverage based on gender, both male and female individuals have approximately the same proportion of insurance coverage

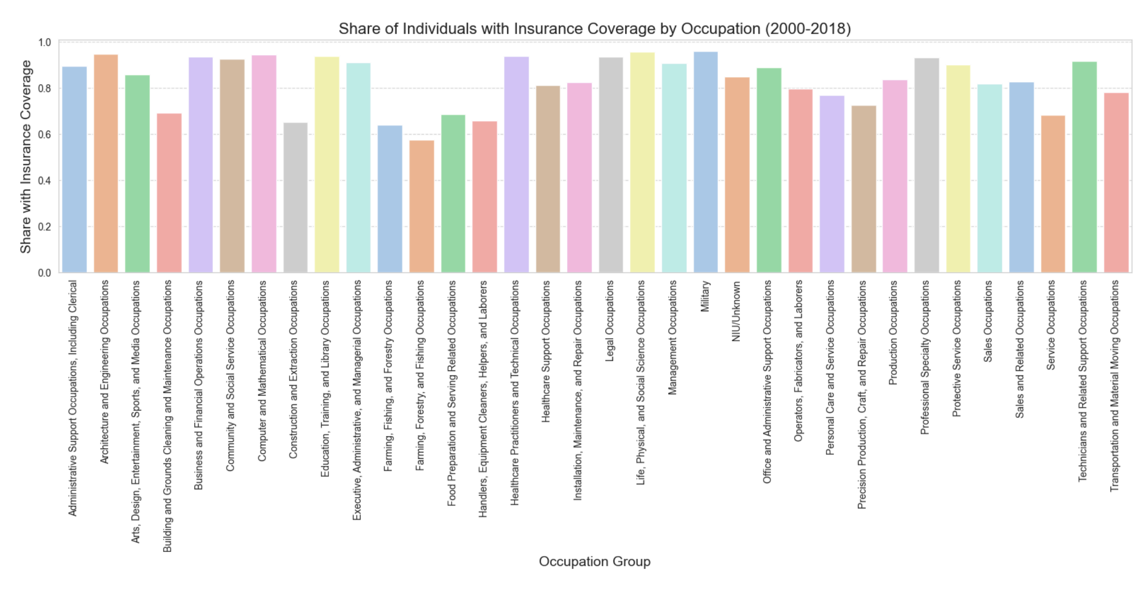

Figure 16. Share of individuals with insurance coverage based on occupation, the differences in insurance coverages suggest a potential link between occupation/income and insurance coverage.

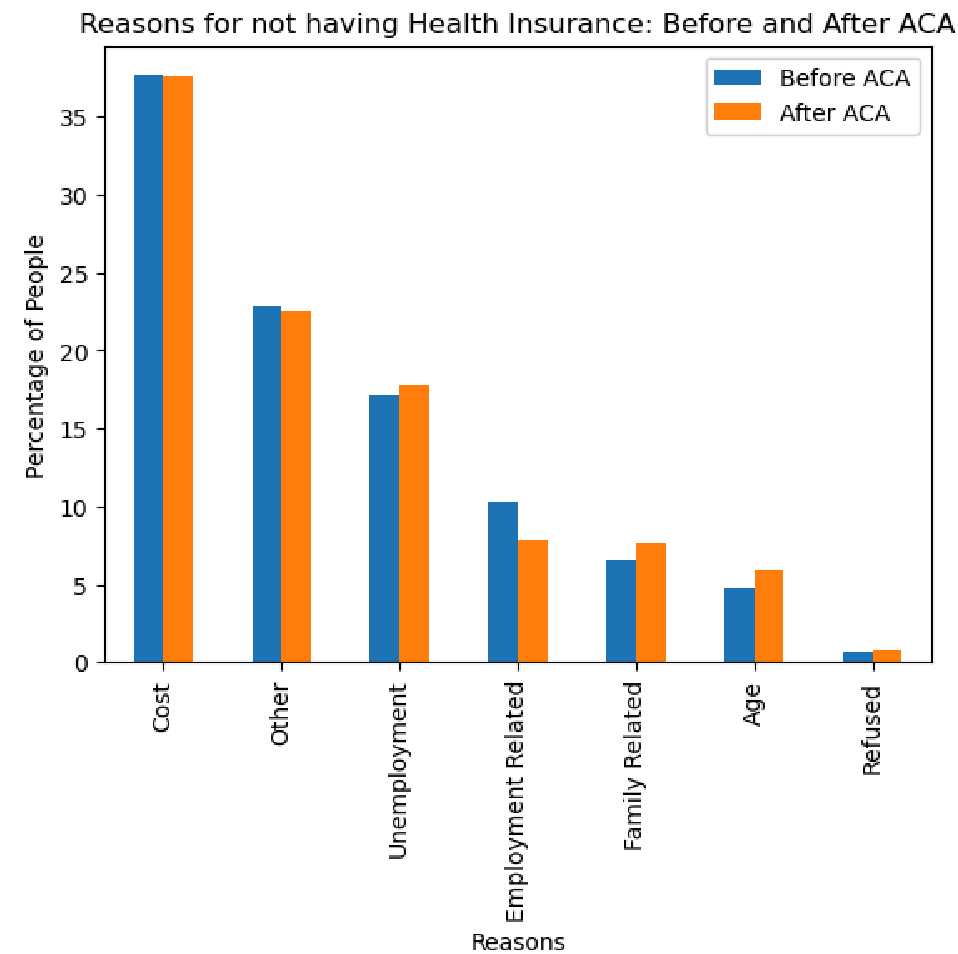

Figure 17. Reasons for not having insurance before and after the Affordable Care Act (ACA), while there is minor variation in the percentages, the overall trend and order is similar before and after the ACA.

Variable	Chi-squared Test Statistic	P-value
Racial Group	7829.59	0.000
Sex	2289.91	0.000
Family Income	76105.13	0.000
Smoking Status	1033.07	0.000
Occupation	29789.14	0.000

Table 8. Summary of Chi-squared Test Statistics and P-values for The Differences In Insurance Coverage Across Different Variable Groups

The above figure presents the results of chi-squared tests, which assess the association between insurance coverage and various demographic variables. Chi-squared test statistics quantify the strength of these associations, with higher values indicating a more pronounced relationship. The accompanying p-values indicate the significance of these associations; values below a chosen threshold (often 0.05) signify statistical significance. Across demographic factors such as racial group, sex, family income, smoking status, and occupation, all variables exhibit low p-values (0.000), indicating that the differences in insurance coverage across each demographic group are statistically significant.

Appendix G. Modeling

For each dataset (2000-2018, pre-2010, post-2010), we perform matching. The matching process is structured as follows:

1. Initial Matching: Identify all observations that exactly match across all specified variables (race, gender, smoking status, family income, and occupation). This comprehensive matching helps ensure that comparisons are made between comparable groups.
2. Variable Reduction: Analyze the utility of each matching variable in predicting the outcome variable, lung cancer diagnosis. The least predictive variable is dropped, narrowing the set of variables used for further matching. This step is iterated to optimize the matching process.
3. Iterative Refinement: Continue the process of dropping the least useful variables and re-matching the data. This iterative refinement proceeds until all observations are matched or no further variables can be reasonably dropped. We then observe the prediction errors of each

iteration, and select an iteration number before the match quality falls off substantially as a our early stopping threshold to obtain a good match.
4. Rerun matching with the early stopping threshold and obtain the matched dataset.

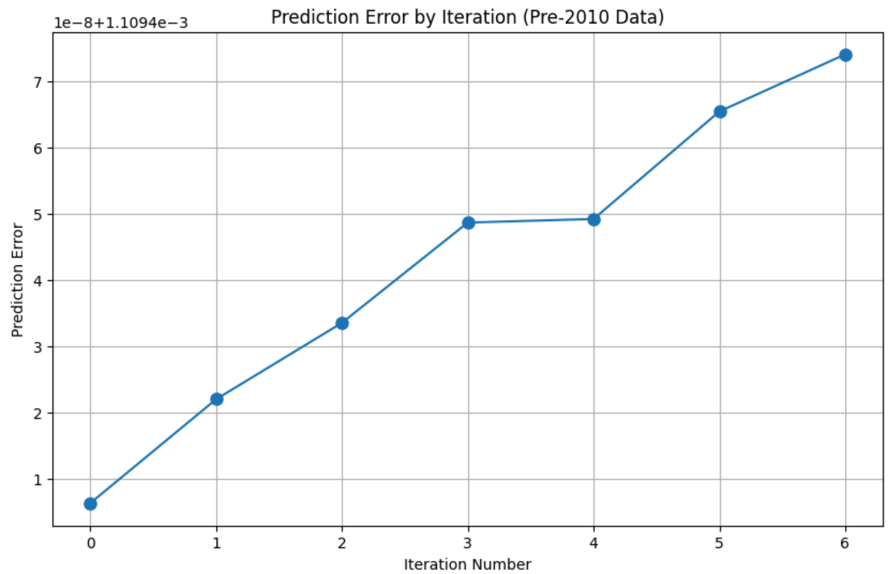

Figure 18. Prediction Error by Iteration (Pre-2010 Data)

By analyzing the graph, there's a sharp increase in prediction error after iteration 0, The increase in prediction error suggests the match quality seems to fall off dramatically. At iteration 0, the sample size of matched data is 856283, which ensures enough number of observations for the later regression analysis without a big increase in prediction error. Thus, the stopping point for matching is chosen at iteration number 0.

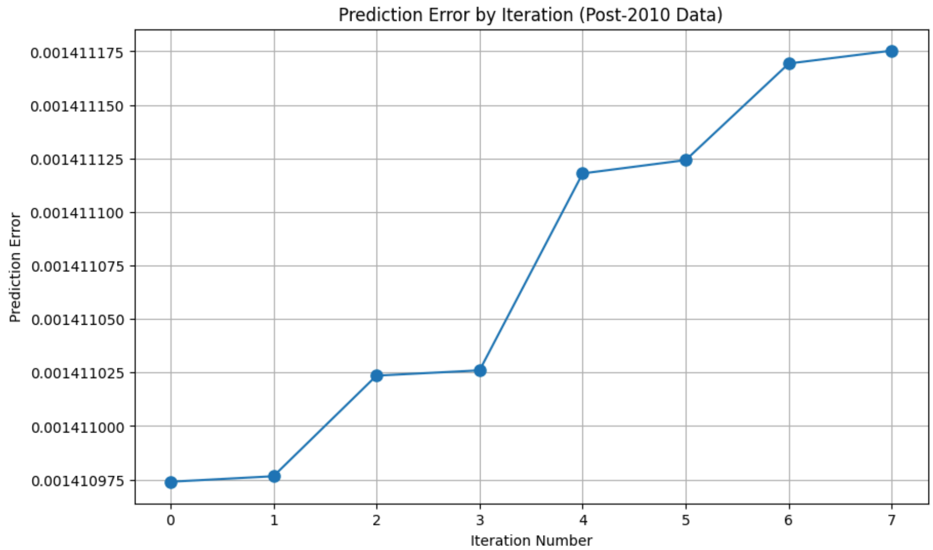

Figure 19. Prediction Error by Iteration (Post-2010 Data)

By analyzing the graph, there's a sharp increase in prediction error between iteration 1 and iteration 2, the increase in prediction error spikes between iteration 3 and iteration 4, and there's another sharp increase in prediction error between iteration 5 and iteration 6. The increase in prediction error suggests the match quality seems to fall off dramatically. At iteration 1, the sample size of matched data is 827558, which ensures enough number of observations for the later regression analysis without a big increase in prediction error. Thus, the stopping point for matching is chosen at iteration number 1.

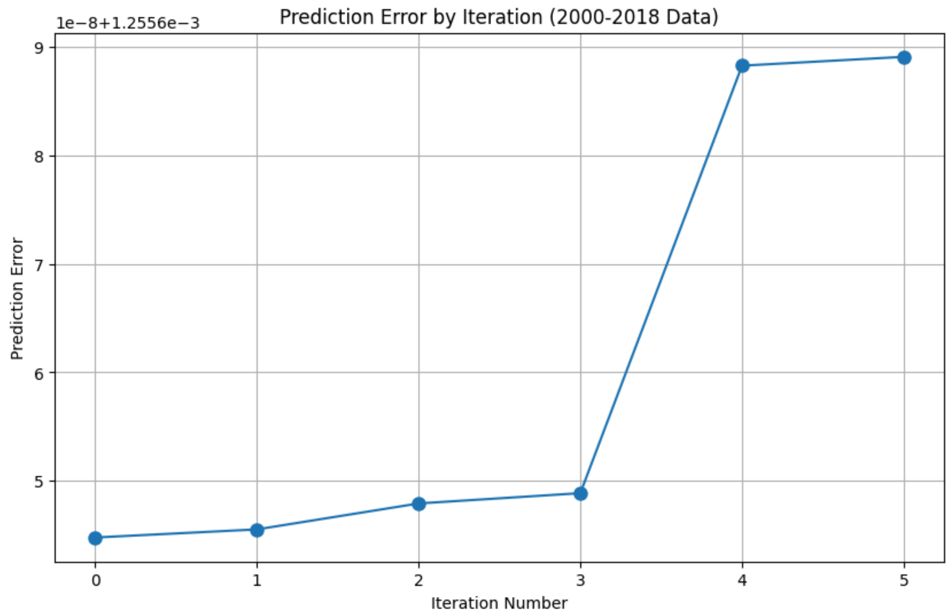

Figure 20. Prediction Error by Iteration (2000-2018 Data)

By analyzing the graph, there's a sharp increase in prediction error between iteration 3 and iteration 4, The increase in prediction error suggests the match quality seems to fall off dramatically. At iteration 3, the sample size of matched data is 1690664, which ensures enough number of observations for the later regression analysis without a big increase in prediction error. Thus, the stopping point for matching is chosen at iteration number 3.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0005	9.73E-05	-4.736	0	-0.001	0
Insurace Coverage	0.0008	5.66E-05	13.598	0	0.001	0.001
Sex - Female	-0.0001	5.99E-05	-2.42	0.016	0	-2.76E-05
Race - Black/African	-0.0002	8.07E-05	-2.966	0.003	0	-8.12E-05
Race - American Indian	-0.0001	0	-0.326	0.744	-0.001	0.001
Race - Asian	-0.0002	0	-1.519	0.129	0	5.91E-05
Race - Other	-0.0004	0	-3.453	0.001	-0.001	0
Race - Not Releaseable	-0.0002	4.17E-05	-5.536	0	0	0
Race - Mutiple	-0.0003	0	-1.909	0.056	-0.001	8.02E-06
Family Income - $0 - $34,999	0.0003	0	3.094	0.002	0	0.001
Family Income - $35,000-$74,999	-0.0002	0	-1.526	0.127	0	4.44E-05
Family Income - 75000+	-0.0003	0	-2.682	0.007	0	-7.25E-05
Family Income - $75,000-$99,999	-0.0004	0	-2.673	0.008	-0.001	-9.71E-05
Family Income - 100000+	-0.0003	0	-2.606	0.009	-0.001	-7.54E-05
Family Income - 20000+	-1.09E-05	0	-0.093	0.926	0	0
Smoke - Yes	0.0074	0	19.149	0	0.007	0.008
Smoke - No	0.0029	0	11.042	0	0.002	0.003
Occupation - Office and Administrative Support	-0.0018	0	-3.758	0	-0.003	-0.001
Occupation - Management	-0.0021	0.001	-3.463	0.001	-0.003	-0.001
Occupation - Sales and Related	-0.0025	0.001	-4.982	0	-0.004	-0.002
Occupation - Production	-0.0004	0.001	-0.456	0.648	-0.002	0.001
Occupation - Education, Training, and Library	-0.0023	0.001	-4.515	0	-0.003	-0.001
Occupation - Transportation and Material Moving	-0.0022	0.001	-3.432	0.001	-0.003	-0.001
Occupation - Healthcare Practitioners and Technical	-0.0023	0.001	-3.638	0	-0.004	-0.001
Occupation - Business and Financial Operations	-0.0014	0.001	-1.656	0.098	-0.003	0
Occupation - Construction and Extraction	-0.0021	0.001	-2.446	0.014	-0.004	0
Occupation - Food Preparation and Serving Related	-0.0003	0.001	-0.272	0.786	-0.002	0.002
Occupation - Building and Grounds Cleaning and Maintenance	-0.0031	0.001	-4.44	0	-0.004	-0.002
Occupation - Installation, Maintenance, and Repair	-0.0024	0.001	-2.59	0.01	-0.004	-0.001
Occupation - Personal Care and Service	-0.0026	0.001	-3.378	0.001	-0.004	-0.001
Occupation - Healthcare Support	-0.0036	0.001	-5.417	0	-0.005	-0.002
Occupation - Architecture and Engineering	-0.0028	0.001	-2.795	0.005	-0.005	-0.001
Occupation - Community and Social Service	-0.0032	0.001	-4.031	0	-0.005	-0.002
Occupation - Protective Service	-0.0024	0.001	-2.041	0.041	-0.005	-9.38E-05
Occupation - Arts, Design, Entertainment, Sports, and Media	-0.0027	0.001	-2.677	0.007	-0.005	-0.001
Occupation - Professional Specialty	-0.0036	0	-9.928	0	-0.004	-0.003
Occupation - Administrative Support , Including Clerical	-0.0038	0	-9.783	0	-0.005	-0.003
Occupation - Computer and Mathematical	-0.0035	0.001	-6.119	0	-0.005	-0.002
Occupation - Executive, Administrative, and Managerial	-0.0041	0	-10.683	0	-0.005	-0.003
Occupation - Legal	-0.0041	0	-13.713	0	-0.005	-0.004
Occupation - Military	-0.0007	0.006	-0.118	0.906	-0.012	0.01
Occupation - Service	-0.0047	0	-13.764	0	-0.005	-0.004
Occupation - Life, Physical, and Social Science	-0.002	0.002	-1.113	0.266	-0.005	0.001
Occupation - Sales	-0.0043	0	-10.912	0	-0.005	-0.004
Occupation - Farming, Fishing, and Forestry	-0.0039	0.001	-3.274	0.001	-0.006	-0.002
Occupation - Precision Production, Craft, and Repair	-0.0048	0	-10.942	0	-0.006	-0.004
Occupation - Operators, Fabricators, and Laborers	-0.0049	0	-11.925	0	-0.006	-0.004
Occupation - Technicians and Related Support	-0.0044	0	-8.985	0	-0.005	-0.003
Occupation - Farming, Forestry, and Fishing	-0.0043	0.001	-4.899	0	-0.006	-0.003
Occupation - Handlers, Equipment Cleaners, Helpers, and Laborers	-0.0037	0.001	-4.442	0	-0.005	-0.002

Table 9. Full regression table for data before 2010, i.e. before the implementation of the Affordable Care Act (ACA) in USA

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0007	0	-5.975	0	-0.001	0
Insurace Coverage	0.001	6.21E-05	15.39	0	0.001	0.001
Sex - Female	0.0002	6.45E-05	3.544	0	0	0
Race - Black/African	-0.0002	0	-1.996	0.046	0	-3.61E-06
Race - American Indian	-0.0006	0	-2.296	0.022	-0.001	-8.23E-05
Race - Asian	-0.0002	8.83E-05	-2.343	0.019	0	-3.38E-05
Race - Other	-0.0003	3.92E-05	-8.796	0	0	0
Race - Multiple	-0.0001	0	-0.637	0.524	0	0
Income - 0 - 34,999	0.0003	0	2.257	0.024	3.92E-05	0.001
Income - 35,000 - 74,999	-0.0002	0	-2.081	0.037	0	-1.44E-05
Income - 75,000+	-0.0004	0	-3.18	0.001	-0.001	0
Income - 100,000+	-0.0006	0	-5.339	0	-0.001	0
Smoking - Yes	0.0071	0.001	14.08	0	0.006	0.008
Smoking - No	0.001	0	2.332	0.02	0	0.002
Occupation - Office and Administrative Support	8.11E-07	0.001	0.002	0.999	-0.001	0.001
Occupation - Management	0.0006	0.001	0.95	0.342	-0.001	0.002
Occupation - Sales and Related	-0.0007	0.001	-1.227	0.22	-0.002	0
Occupation - Production	0.0024	0.001	3.245	0.001	0.001	0.004
Occupation - Education, Training, and Library	-0.0004	0.001	-0.814	0.416	-0.002	0.001
Occupation - Transportation and Material Moving	0.0004	0.001	0.516	0.606	-0.001	0.002
Occupation - Healthcare Practitioners and Technical	4.66E-05	0.001	0.074	0.941	-0.001	0.001
Occupation - Business and Financial Operations	-0.0003	0.001	-0.443	0.658	-0.002	0.001
Occupation - Construction and Extraction	-0.0002	0.001	-0.215	0.83	-0.002	0.001
Occupation - Food Preparation and Serving Related	-0.001	0.001	-1.473	0.141	-0.002	0
Occupation - Building and Grounds Cleaning and Maintenance	0.0004	0.001	0.455	0.649	-0.001	0.002
Occupation - Installation, Maintenance, and Repair	0.0012	0.001	1.237	0.216	-0.001	0.003
Occupation - Personal Care and Service	-0.0014	0.001	-2.213	0.027	-0.003	0
Occupation - Healthcare Support	-0.0007	0.001	-0.937	0.349	-0.002	0.001
Occupation - Architecture and Engineering	-9.31E-05	0.001	-0.105	0.916	-0.002	0.002
Occupation - Community and Social Service	-0.0005	0.001	-0.625	0.532	-0.002	0.001
Occupation - Protective Service	-0.0007	0.001	-0.812	0.417	-0.002	0.001
Occupation - Arts, Design, Entertainment, Sports, and Media	-0.001	0.001	-1.312	0.19	-0.003	0.001
Occupation - Computer and Mathematical	-0.0011	0.001	-1.803	0.071	-0.002	9.93E-05
Occupation - Legal	-0.0015	0.001	-1.633	0.102	-0.003	0
Occupation - Military	0.0103	0.005	2.243	0.025	0.001	0.019
Occupation - Life, Physical, and Social Science	6.92E-05	0.001	0.059	0.953	-0.002	0.002
Occupation - Farming, Fishing, and Forestry	-0.0001	0.001	-0.097	0.922	-0.003	0.003

Table 10. Full regression table for data post 2010, i.e. after the implementation of the Affordable Care Act (ACA) in USA

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0006	7.52E-05	-7.546	0	-0.001	0
Insurace Coverage	0.0009	4.00E-05	22.053	0	0.001	0.001
Sex - Female	4.16E-05	4.41E-05	0.943	0.346	-4.48E-05	0
Race - Black/African	-0.0002	6.40E-05	-3.854	0	0	0
Race - American Indian	-0.0004	0	-1.771	0.076	-0.001	4.19E-05
Race - Asian	-0.0002	6.52E-05	-3.812	0	0	0
Race - Other	-0.0004	0	-3.508	0	-0.001	0
Race - Not Releaseable	-0.0006	3.85E-05	-15.846	0	-0.001	-0.001
Race - Mutiple	-0.0002	0	-1.663	0.096	0	3.72E-05
Family Income - $0 - $34,999	0.0003	8.74E-05	3.643	0	0	0
Family Income - $35,000-$74,999	-0.0002	7.90E-05	-2.844	0.004	0	-6.98E-05
Family Income - 75000+	-0.0004	8.16E-05	-4.856	0	-0.001	0
Family Income - $75,000-$99,999	-0.0003	9.89E-05	-3.194	0.001	-0.001	0
Family Income - 100000+	-0.0005	7.58E-05	-6.248	0	-0.001	0
Family Income - 20000+	-0.0001	0	-0.999	0.318	0	0
Smoke - Yes	0.0075	0	24.242	0	0.007	0.008
Smoke - No	0.0022	0	9.445	0	0.002	0.003
Occupation - Office and Administrative Support	-0.0011	0	-3.195	0.001	-0.002	0
Occupation - Management	-0.0008	0	-1.87	0.061	-0.002	3.72E-05
Occupation - Sales and Related	-0.0018	0	-4.766	0	-0.003	-0.001
Occupation - Production	0.0009	0.001	1.672	0.095	0	0.002
Occupation - Education, Training, and Library	-0.0016	0	-4.43	0	-0.002	-0.001
Occupation - Transportation and Material Moving	-0.0011	0	-2.257	0.024	-0.002	0
Occupation - Healthcare Practitioners and Technical	-0.0012	0	-2.729	0.006	-0.002	0
Occupation - Business and Financial Operations	-0.0013	0	-2.78	0.005	-0.002	0
Occupation - Construction and Extraction	-0.0013	0.001	-2.316	0.021	-0.002	0
Occupation - Food Preparation and Serving Related	-0.0012	0.001	-2.35	0.019	-0.002	0
Occupation - Building and Grounds Cleaning and Maintenance	-0.0013	0.001	-2.428	0.015	-0.002	0
Occupation - Installation, Maintenance, and Repair	-0.0006	0.001	-0.8	0.423	-0.002	0.001
Occupation - Personal Care and Service	-0.0023	0	-4.985	0	-0.003	-0.001
Occupation - Healthcare Support	-0.0022	0.001	-4.112	0	-0.003	-0.001
Occupation - Architecture and Engineering	-0.0015	0.001	-2.232	0.026	-0.003	0
Occupation - Community and Social Service	-0.0017	0.001	-2.62	0.009	-0.003	0
Occupation - Protective Service	-0.0016	0.001	-2.408	0.016	-0.003	0
Occupation - Arts, Design, Entertainment, Sports, and Media	-0.0022	0.001	-3.723	0	-0.003	-0.001
Occupation - Professional Specialty	-0.0032	0	-9.539	0	-0.004	-0.003
Occupation - Administrative Support , Including Clerical	-0.0035	0	-9.748	0	-0.004	-0.003
Occupation - Computer and Mathematical	-0.0024	0	-5.377	0	-0.003	-0.002
Occupation - Executive, Administrative, and Managerial	-0.0037	0	-10.653	0	-0.004	-0.003
Occupation - Legal	-0.0026	0.001	-4.02	0	-0.004	-0.001
Occupation - Military	0.0067	0.003	2.077	0.038	0	0.013
Occupation - Service	-0.0044	0	-14.547	0	-0.005	-0.004
Occupation - Life, Physical, and Social Science	-0.0013	0.001	-1.558	0.119	-0.003	0
Occupation - Sales	-0.004	0	-10.902	0	-0.005	-0.003
Occupation - Farming, Fishing, and Forestry	-0.002	0.001	-2.024	0.043	-0.004	-6.33E-05
Occupation - Precision Production, Craft, and Repair	-0.0045	0	-11.083	0	-0.005	-0.004
Occupation - Operators, Fabricators, and Laborers	-0.0046	0	-12.174	0	-0.005	-0.004
Occupation - Technicians and Related Support	-0.004	0	-8.708	0	-0.005	-0.003
Occupation - Farming, Forestry, and Fishing	-0.0039	0.001	-4.486	0	-0.006	-0.002
Occupation - Handlers, Equipment Cleaners, Helpers, and Laborers	-0.0034	0.001	-4.057	0	-0.005	-0.002

Table 11. Full regression table - based on data from 2000 to 2018 in USA

Appendix H: Lung Cancer Diagnosis Distribution

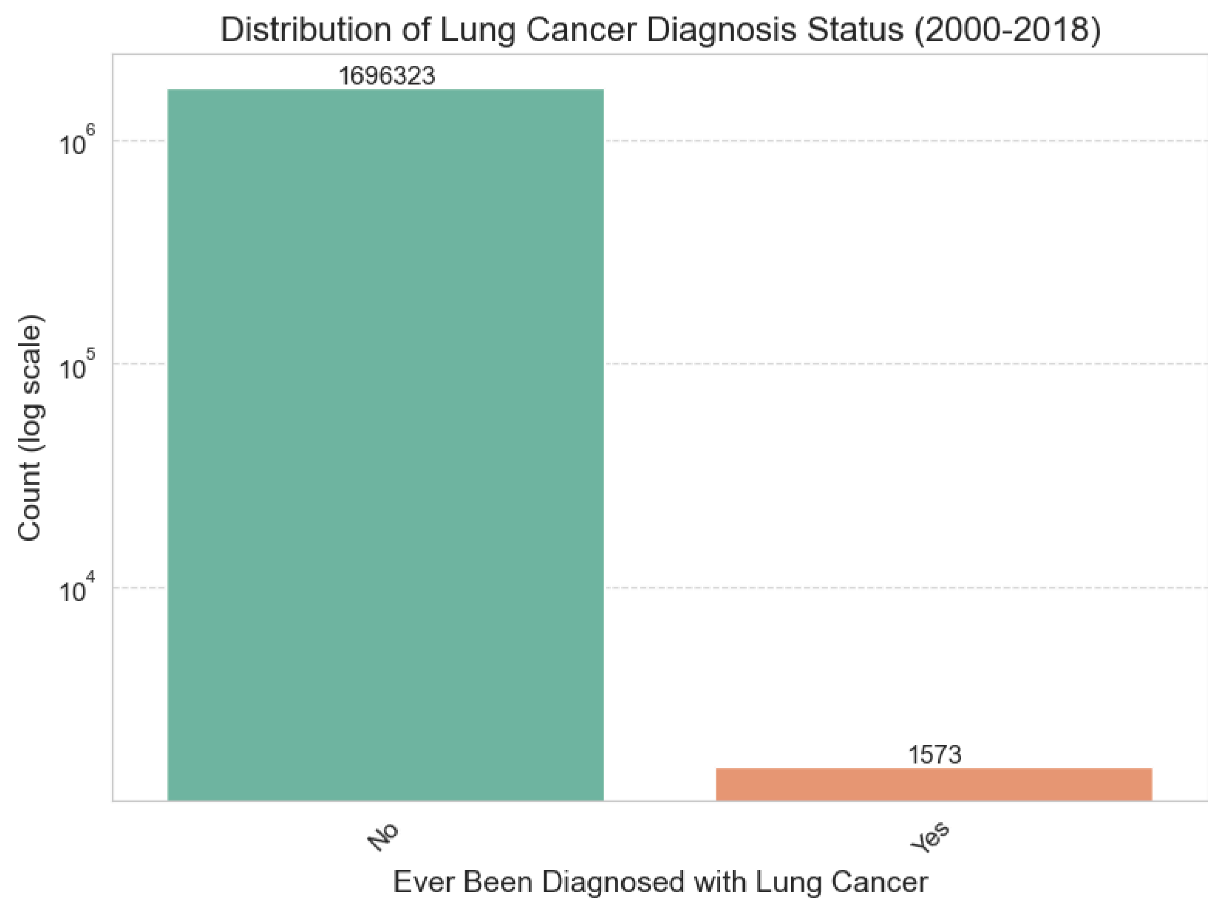

Figure 21. Distribution of Lung Cancer Diagnosis Status (2000 - 2018)

9. Bibliography

Center for Disease Control and Prevention. Who Should be Screened for Lung Cancer. Accessed April 27, 2024. https://www.cdc.gov/cancer/lung/basic_info/screening.htm#:~:text=may%20work%20better.-,The%20only%20recommended%20screening%20test%20for%20lung%20cancer%20is%20low,minutes%20and%20is%20not%20painful.

Center for Disease Control and Prevention. "*Smoking and Tobacco Use*. Retrieved from Centers for Disease Control and Prevention". Accessed April 28, 2023. https://www.cdc.gov/tobacco/basic_information/health_effects/respiratory/index.htm#:~:text=The%20best%20way%20to%20prevent,and%20if%20you%20smoke%2C%20quit.&text=Talk%20with%20your%20doctor%20about,cigarettes%2C%20cigars%2C%20or%20pipes .

Center for Disease Control and Prevention. "500 Cities: Chronic Obstructive Pulmonary Disease among Adults Aged >=18 Years | Data | Centers for Disease Control and Prevention." n.d. Data.cdc.gov. Accessed March 12, 2024. https://data.cdc.gov/500-Cities-Places/500-Cities-Chronic-obstructive-pulmonary-disease-a/xzgm-d4ya/data_preview

Center for Disease Control and Prevention."Download USCS Data Tables | CDC." 2021.. https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm.

U.S. Department of Health & Human Services. 2022. "About the Affordable Care Act." U.S. Department of Health & Human Services. March 17, 2022. https://www.hhs.gov/healthcare/about-the-aca/index.html.

Institute for Health Metrics and Evaluation. "United States of America". Accessed February 9, 2023. https://www.healthdata.org/research-analysis/health-by-location/profiles/united-states .

Kominski, Gerald F., Narissa J. Nonzee, and Andrea Sorensen. 2019. "The Affordable Care Act's Impacts on Access to Insurance and Healthcare for Low-Income Populations." Annual Review of Public Health 38 (1): 489–505. https://doi.org/10.1146/annurev-publhealth-031816-044555

Lokaj, Rina. Early CT Lung Cancer Screening Significantly Improves Survival in Lung Cancer. Accessed March 9, 2023.
https://www.cancernetwork.com/view/early-ct-lung-cancer-screening-significantly-improves-survival-in-lung-cancer

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Annie Chen, Stephanie Richards, and Michael Westberry. IPUMS Health Surveys: National Health Interview Survey, Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2023.
https://doi.org/10.18128/D070.V7.3

Laura Dwyer-Lindgren, M., Amelia Bertozzi-Villa, M., Rebecca W. Stubbs, M., & al, e. (2017). Trends and Patterns of Differences in Chronic Respiratroy Disease Mortality Among Us Counties, 1980-2014. *National Library of Medicine*, 1136–1149.

World Health Organization. (2020). WHO Package of essential noncommicable (PEN) disease interventions for primary healthcare. 28-38.

World Health Organization . (2024, Februrary 12). *World Health Organization Fact Sheets*. Retrieved from World Health Organization:
https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health

PDQ Screening and Prevention Editorial Board. PDQ Lung Cancer Prevention. Bethesda, MD: National Cancer Institute. Updated 10/23/2024 Available at:
https://www.cancer.gov/types/lung/patient/lung-prevention-pdq . Accessed 02-10-2024 [PMID: 26389497]

Sabik, Lindsay M., and Georges Adunlin. 2017. "The ACA and Cancer Screening and Diagnosis." *Cancer Journal (Sudbury, Mass.)* 23 (3): 151–62.
https://doi.org/10.1097/PPO.0000000000000261.